User-oriented medical image retrieval

MARKONIS, Dimitrios

Abstract
This thesis studies the user-oriented design and development lifecycle of a medical image retrieval system. It investigates the user information needs of a target user group in Radiology. These needs are translated into requirements and specifications of a system that could assist radiologists to gain easy and quick access to medical information found in the open access literature and the internal hospital databases. The development of Parallel Distributed Image Search Engine is described and its architecture and applications are presented. The system is evaluated both empirically on the widely used ImageCLEF medical test challenge and on user tests with real users to assess its usability.

Reference

URN : urn:nbn:ch:unige-433251
DOI : 10.13097/archive-ouverte/unige:43325

Available at:
http://archive-ouverte.unige.ch/unige:43325

Disclaimer: layout of this document may differ from the published version.
User–oriented medical image retrieval

THÈSE

présentée à la Faculté des sciences de l’Université de Genève
pour obtenir le grade de Docteur ès sciences, mention informatique

par

Dimitrios Markonis

de

Grèce (GR)

Thèse N° 4739

GENÈVE
2014
La Faculté des sciences, sur le préavis de Messieurs H. MUELLER, professeur et directeur de thèse (Faculté de médecine, Département de radiologie et informatique et Haute Ecole Spécialisée de Suisse Occidentale Valais, Sion, Suisse), S. MARCHAND-MAILLET, professeur et codirecteur de thèse (Département d’informatique) et S. ANTANI, docteur (Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland, USA), autorise l’impression de la présente thèse, sans exprimer d’opinion sur les propositions qui y sont énoncées.

Genève, le 2 décembre, 2014
Contents

Résumé v
Abstract vii
Acknowledgements ix

1 Introduction and thesis overview 1
1.1 Medical informatics in the early 2000s 1
1.2 Modern medical informatics 3
1.2.1 User-centered design 4
1.3 Medical informatics in Radiology 4
1.3.1 Content-based image retrieval 5
1.4 Thesis overview and contributions 6
1.4.1 Thesis overview 6
1.4.2 Scientific contributions 6

2 User requirements and system specifications 9
2.1 User requirement investigation methods 10
2.1.1 Observation 10
2.1.2 Interviews 10
2.1.3 Survey 11
2.2 User requirements investigation results 12
2.2.1 Observation results 12
2.2.2 Interviews’ results 12
2.2.3 Survey results 13
2.3 Discussion 19

3 System design and development 21
3.1 The Parallel Distributed Image Search Engine 21
3.1.1 Specifications and System Design 21
3.1.2 Architecture 22
3.1.3 Applications 34
3.2 Additional components 43
3.2.1 Density-based visual vocabulary creation 44
3.2.2 Modality classification 45
3.2.3 Compound figure separation 46
3.2.4 Multi-modal relevance feedback 48
3.2.5 Parallel image indexing 49

4 Empirical evaluation 53
4.1 Datasets 53
4.2 Local Features 53
4.3 Global image descriptors 56
# CONTENTS

4.4 Relevance feedback ............................................. 56
4.5 Modality Classification .......................................... 59
4.6 Compound figure separation .................................... 60
4.7 Indexing and retrieval efficiency  .............................. 62
  4.7.1 Indexing .................................................. 62
  4.7.2 Retrieval ................................................ 64

5 User–oriented evaluation .......................................... 67
  5.1 First evaluation round .......................................... 67
    5.1.1 Methods and Materials .................................. 68
    5.1.2 Results ............................................... 72
    5.1.3 Discussion ............................................ 81
  5.2 Second evaluation round ....................................... 84
    5.2.1 Methods and Materials .................................. 85
    5.2.2 Results ............................................... 86
    5.2.3 Discussion ............................................ 96

6 Conclusions and Perspectives .................................... 99
  6.1 Achievements and limitations ................................ 99
    6.1.1 Image use and search behavior survey .................. 99
    6.1.2 System design and empirical evaluation ................. 100
    6.1.3 User–centered evaluation ................................ 101
  6.2 Perspectives .................................................. 102
    6.2.1 Radiology information needs ............................ 102
    6.2.2 Future of medical document image retrieval ............ 102

A Survey questionnaires ........................................... 105

B User Tests protocol ............................................... 119
  B.1 CONSENT FORM ............................................. 123

C User Tests tasks .................................................. 125

D User Tests questionnaires ........................................ 129
  D.1 DEMOGRAPHIC SURVEY ..................................... 129
  D.2 USABILITY SURVEY ........................................ 130

Notation ............................................................... 137

Glossary ............................................................... 139

List of Figures ....................................................... 140

List of Tables ........................................................ 143

Bibliography .......................................................... 145

Index ................................................................. 159
Résumé

L'informatique médicale est présente dans le domaine de la santé depuis plus de 20 ans. Un de ses principaux objectifs est d’aider les médecins à accéder et gérer des informations médicales. Avec une évolution de la mentalité de la Médecine fondée sur les faits vers un modèle plus axé sur le patient, la nécessité de trouver de cas semblables est plus que jamais d’actualité.

En radiologie, un domaine basé sur l’imagerie médicale informatique, une énorme quantité de données est produite chaque année. Le développement de nouvelles techniques d’imagerie médicale, ainsi que l’augmentation du nombre de détails de l’image (tranches fines, séries temporelles, plus haute résolution, acquisitions multi-modales), une quantité constamment croissante d’informations visuelles s’accumule. Les radiologues sont souvent surchargés par la grande quantité de données visuelles et un rapport européen estime que les images médicales occupent 30 % de tout l’espace de stockage numérique dans le monde entier. L’accès aux dossiers électroniques à l’hôpital est souvent rigide et uniquement possible par l’identifiant du patient, tandis que la recherche d’images sur Internet est souvent effectuée à l’aide de moteurs de recherche d’usage général, souvent avec des résultats de qualité douteuse. Les systèmes d’accès et de gestion aux images radiologiques, doivent donc permettre un accès facile et rapide à l’information visuelle disponible. Bien que la recherche d’images médicales est un domaine de recherche très actif, seules quelques applications atteignent l’environnement clinique. Afin de répondre à des situations réelles et à des besoins d’information réels des utilisateurs, les systèmes de recherche d’images médicales doivent suivre une approche de conception centrée sur l’utilisateur.

Dans cette thèse, le cycle de vie de développement axé sur l’utilisateur d’un système innovateur de recherche d’images médicales est présenté. Le comportement d’utilisation et de recherche d’images de radiologues est étudié à l’aide d’un sondage et les besoins des utilisateurs sont identifiés. Les spécifications du système Parallel Distributed Image Search Engine (ParaDISE) sont générées et il est implémenté en utilisant des fonctionnalités telles que la recherche d’images par le contenu, le retour de pertinence et la fusion multi-modale d’informations. De complexes pipelines d’indexation et de recherche sont conçus afin de répondre aux exigences du système qui ont été définies par les besoins des utilisateurs. Les méthodes qui constituent les composantes de ces pipelines sont évaluées de manière empirique et l’usabilité d’un système intégré est évaluée au travers d’une série de tests utilisateurs conduits avec les radiologues.

Les résultats montrent que les radiologues ont des besoins d’information non satisfaits. La possibilité de rechercher par l’information visuelle peut aider lorsque des pathologies inconnues sont rencontrées, ou pour le diagnostic différentiel. Elle peut également compléter la recherche textuelle afin d’accélérer le processus itératif de recherche à l’aide du retour de pertinence. Lier des images à des cas et une navigation rapide entre les deux ont également été jugés importants. La recherche d’images par le contenu doit encore être améliorée pour être réellement digne de confiance dans la pratique clinique. ParaDISE constitue une plate-forme où les nouvelles caractéristiques visuelles, les représentations d’image et les structures d’index peuvent être testées et évaluées à grande échelle.
Abstract

Medical informatics have been present in healthcare for more than 20 years. One of their main goals is to assist the physicians to access and manage medical information. With the mentality shifting from evidence-based medicine towards a more patient-based model, the need for finding closely similar cases is more relevant than ever.

In Radiology, a domain based on imaging informatics, a huge amount of data is produced yearly. The development of new medical imaging techniques and the increasing number of image details (thin slices, temporal series, higher resolution, multi-modal acquisitions) accumulate to an ever-growing amount of visual information. Radiologists are often overloaded with the large amount of image data, while a European report estimates medical images to occupy 30% of the total digital storage worldwide. Accessing the electronic records at the hospital is often rigid and solely patient-based, while image search on the Internet is often performed in general-purpose search engines, often with results of questionable quality.

Radiology image access and management systems, thus, need to provide easy and quick access to visual information available. While medical image retrieval is a very active research field, only few applications reach the clinical environment. In order to address real life scenarios and meet real information needs of users, medical image retrieval systems need to follow a user-centered design approach.

In this thesis, the user-centered development lifecycle of a novel medical image retrieval system is presented. The image use and search behavior of radiologists is investigated through a survey and user information needs are identified. Specifications of the Parallel Distributed Image search Engine (ParaDISE) system are generated and it is implemented using features such as content-based image retrieval, relevance feedback and multi-modal information fusion. Complex indexing and retrieval pipelines are designed to meet the system requirements that have been specified by the user needs. The methods constituting the components of these pipelines are evaluated empirically and the usability of an integrated system is assessed in a series of user studies with radiologists.

The results show that radiologists have unmet information needs. The ability to search by visual information can assist when unknown pathologies are met, or for differential diagnosis. It can also complement textual search to speed up the search iteration process with the use of relevance feedback. Linking of images and cases and quick navigation between the two was also found to be important. Content-based retrieval quality needs further improvement to be really trustworthy in the clinical practice. ParaDISE constitutes a platform where new visual features, image representations and index structures can be tested and evaluated in large scale.
Acknowledgements

I am grateful to my thesis supervisor, professor Henning Müller. Throughout the four years of this thesis he has always been present when I needed advice or guidance. He gave me directions for concrete research methodology and ethics. At the same time, he taught me to be independent, by showing trust to my judgement and my skills. I was always amazed by his wealth of knowledge in information retrieval, image processing and evaluation and I learnt a lot from him. I will be following his example as closely as I can in my future research career.

I also want to warmly thank Stephane Marchand-Maillet for being available and willing to help when I needed him, giving advice or facilitating administrative procedures. His comments were accurate and relevant and helped me improve my research and presentation skills.

I give my thanks to my thesis expert Sameer Antani from the National Library of Medicine of the United States of America for accepting to become a member of the board of examiners of this thesis. I have been following his work and it has been an inspiration to me during this thesis.

I address my thanks to Dr. Lianyi Wang for welcoming me to China during my research exchange trip and Chunyan Wang for facilitating my stay there.

My gratitude to the clinical partners who made the usability evaluation possible by helping recruiting participants but also participating in the user tests themselves: Dr. Alexandra Tsikrika, Dr. Pierre-Alexandre Poletti and Dr. Elmar Kotter.

Many thanks to my colleagues at HES-SO. Roger Schaar co-founder of ParaDISE, brilliant programmer and a friend, without whom this study would not be possible. Antonio Foncubierta, for our discussions on image representation on our first research steps and teaching me Spanish. My office-mates Alba Garcia Seco de Herrera and Oscar Jimenez for putting up with me and making this thesis an enjoyable experience. Theodora Tsikrika, who even though she came abruptly to HES-SO for a small amount of time, taught me a lot on information retrieval, research and academia. Adrien Depeursinge, for his guidance, his ideas and willingness to help. Also for the unforgettable “ski and cheese” weekends in his chalet on the Alps. Antoine Widmer, for bringing ParaDISE to the future, with the innovative applications he developed. Ajad Chatkuli for his work on the compound figure separation. Ivan Eggel for his help with parallel infrastructure, technical problem solving and cigarette supply. Stephano Bromury for our fruitful discussions on dimensionality reduction and modality classification during our table tennis sessions.

Many thanks also to my colleagues in the KHRESMOI project. Alan Harbury for his scientific coordination, his critical comments and his honest advice. Sascha Kriewel for putting up with my requests for changes of the user-interface and always be able to deliver. Markus Holzer and Frederic Baroz for our collaboration in the user tests planning, setting up and conduction. Rene Donner and Georg Langs for our collaboration in connecting the internal and external search prototypes. Konstantin Pentechev for our collaboration in combining semantics with image search and classification. Every person whose name is missing here should know that (s)he equally have her/his place here.

My deepest gratitude to my family for all the support these years. My parents, Nicolas and Paraskevi, for the values they instilled me, combining critical thought imagination and creativity. My brothers, Yannis and Stavros, who I know I can always trust and rely on. Finally, I would like to thank my lovely Dimitra, who put the "wow" factor in my life.
Chapter 1

Introduction and thesis overview

The study of Medicine has always been of great importance to human civilization. The will to extend life expectancy has been present throughout the history of mankind. With the triumph of science over superstition and obscurantism after the middle ages, medicine has not been excluded by the new mindset. Healthcare moved away from nostrums and quackery towards observation and experiments. In the end of the 20th century a ground-breaking shift in medical practice mentality took place, when Evidence-based Medicine (EBM) started to gain ground among the medical doctors. The core concepts of EBM included applying epidemiological principles, such as randomized controlled trials (RCTs) to the practice of patient care. A definition of EBM is given in [49]:

“...evidence based medicine is rooted in five linked ideas: firstly, clinical decisions should be based on the best available scientific evidence; secondly, the clinical problem – rather than habits or protocols – should determine the type of evidence to be sought; thirdly, identifying the best evidence means using epidemiological and biostatistical ways of thinking; fourthly, conclusions derived from identifying and critically appraising evidence are useful only if put into action in managing patients or making health care decisions; and, finally, performance should be constantly evaluated.”

The emergence of electronics and informatics during the second half of the 20th century would not leave Medicine unaffected, either. Electronic devices were able to monitor and provide information about the patient condition, while information systems were introduced to medical institutions. The paper-based organization of hospitals of the western countries quickly became digital. As healthcare had become an information-based field, information systems were considered as a powerful tool, assisting with the access to and management of information. The human knowledge was gradually transferred into digital form and became easily accessible via the World Wide Web.

<table>
<thead>
<tr>
<th>Medical informatics category</th>
<th>Focus of study and applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioinformatics</td>
<td>molecular and cellular processes, such as gene sequences and maps</td>
</tr>
<tr>
<td>Imaging informatics</td>
<td>tissues and organs, such as radiology imaging systems</td>
</tr>
<tr>
<td>Clinical informatics</td>
<td>clinicians and patients, including applications of clinical specialities</td>
</tr>
<tr>
<td>Public Health informatics</td>
<td>population, such as disease surveillance systems</td>
</tr>
</tbody>
</table>

Table 1.1: The categorization of medical informatics, as it is proposed in [104].

1.1 Medical informatics in the early 2000s

By the start of the 21st century, health informatics had been applied in several fields of healthcare. A categorization of medical informatics was given in [104] and presented in Table 1.1. A summarization of the first applications of one of the four categories, clinical informatics, was made
in [84]. The same study proposes the discrimination between patient–specific applications and knowledge–based ones. The three categories of clinical informatics proposed in [84] are:

- Electronic Medical Records (EMR): systems keeping the patient records in health institutions in digital format, allowing this way for easy access and management;
- Information Retrieval: with the tremendous growth of the Internet knowledge–based information became available in online journals, medical textbooks, clinical practice guidelines etc.;
- Decision support systems: knowledge–based applications emerging from artificial intelligence and expert system research fields found use in assisting clinicians with patient care by detecting critical situations and errors.

EMR were massively adopted by health institutions. However, the adoption was slow because of clinicians concerns that privacy, patient safety, quality of care and efficiency will decline [118]. Moreover, in the beginning, informatics use such as computerized physician order point (CPOE) added extra time to the clinicians [84]. A survey on 86 health information systems (HIS) implementations in 7 countries reports that HIS themselves were not the source of any of the above concerns, but rather the processes around the systems facilitated errors [118].

With the wide use of HIS in health institutions, a big discussion about their impact and the evaluation of it on the clinicians and institutional performance started. In [9] evaluation of the impact of HIS was presented as well as three different models of how systems affect organizations, including:

1. the computer system as an external force;
2. the system design determined by user information needs;
3. complex social interactions as determinants of the system use.

The same categorization was proposed in [97]. In this study the third model was adopted and it was claimed that the organizational and social interactions affect the impact and use of an HIS and at the same time they are affected by the system. These interactions have to be taken into account when designing an HIS and evaluating its impact [23, 24].
1.2 Modern medical informatics

Many medical applications in the early 2000s were technology–driven, thus following the development process presumed in category 1) of [9]. In [98] several research studies are referred where it is shown that the rate of adoption of HIS in medical institutions is low. In the same study it is attempted to identify the reasons behind this low rate adoption. One of the main factors identified is that the application development process was detached from the clinicians. Developers tended to put the blame on clinicians for system misuse instead of taking feedback with a critical view to the system usability. Secondly, systems were often developed in an “one–fits–all” manner without taking into account the different types of users and their needs. Finally, the evaluation of the system impact was done using measures that were not corresponding to the truth, such as the amount of use.

At the same time criticism about EBM started to emerge. The main critical points are described in [44]:

- reliance on empiricism: the main criticism is that EBM is based on experiments that try to minimize bias instead of being based on physiology theory;
- narrow definition of evidence: EBM grades evidence. It considers RCT and statistical methods as more trustworthy sources of evidence; However, RCTs and meta–analysis have not been shown to be more reliable than other research methods, the questions that EBM can answer are limited and EBM has failed to provide a means to integrate other, non–statistical, forms of medical information, such as professional experience and patient specific factors;
- lack of evidence of efficacy: EBM assumes that it will improve the quality of health care. This assumption is ”not self–evident”. While the definition states that performance should be constantly evaluated, it is claimed in [44] that “More than ten years after the inception of the practice of EBM, there is no evidence of its effectiveness in providing higher quality healthcare.”
- limited usefulness for the individual patient: epidemiology is based on the field of statistics. Consequently, the information that EBM provides the clinician is statistically–based; clinical studies and randomized trials uncover trends and average behaviors of a group of patients. Individual patients do not present average behavior, the treatment is either beneficial for a person or not;
- threats to the autonomy of the doctor/patient: this criticism claims that EBM reduces the decisions made by the patient with the doctors assistance through these enforced guidelines.

A more patient–oriented approach in medical practice has been starting to gain ground the last years. Taking into account the patient characteristics to apply personalised diagnosis and treatment became feasible with the recent advances in genomics. Medical informatics followed this trend with the growth of the Internet providing easy access to health–related information and the use of smart phones and personal monitoring devices [103]. Patient–oriented medical practice has applications to be used by and patients and applications to be used by physicians in a patient–based mentality [17, 47].
CHAPTER 1. INTRODUCTION AND THESIS OVERVIEW

1.2.1 User-centered design

In order to respond to the shortcomings of the HIS in the early years of their implementation to clinical environments, medical informatics developers started to adopt a more user-oriented approach. User-centered design (UCD) [186] has been used for many years in industry [86, 95]. In this approach the design of a product is driven by the user requirements and feedback to improve the product’s usability and the user experience. In [162] a model is proposed for the user-oriented medical system development lifecycle that unifies some of the most important lifecycle models [80, 120, 133, 144]. The model is presented in Figure 1.1.

It consists of five basic steps that are followed in an iterative manner. First there is the investigation of user needs. The user needs will then dictate the requirements of a system. The requirements are then translated into technical specifications. Then comes the implementation of these technical specifications and the release of the system. Each of these steps requires a verification using the previous step and a hazard analysis (HA). The implementation step also needs to validate that it covers the system requirements and which is usually performed by empirical tests with users, called user-centered evaluation. In the end of each iteration a post-deployment corrective and preventative actions (CAPA) step is taken.

Medical applications that have used UCD in their development lifecycle include an antibiotic decision support system that was tested on 21 patients [179] and a personal monitoring device [47]. In [66] an iterative UCD approach is described for the development of an application for collecting and managing data by emergency personnel at mass casualty disaster scenarios. Even a framework for redesigning user interfaces of systems that have not been designed in a user-oriented manner have been proposed [94].

1.3 Medical informatics in Radiology

The use of informatics in Radiology is so extensive that it is considered as one of the main categories of medical informatics in [104] as seen in Table 1.1. The introduction of filmless imaging and Picture Archive and Communication Systems (PACS) has improved the performance and efficiency in radiology [135, 153, 154]. Many computer-aided diagnosis (CAD) systems were developed for several pathologies and imaging modalities [4, 25, 27, 54, 59].
1.3. MEDICAL INFORMATICS IN RADIOLOGY

Figure 1.2: The block diagram of a state-of-the-art CBIR pipeline, Bag-of-Visual-Words [170]. Local features from a training group are clustered and produce a “visual vocabulary”. The query image and the database images are then expressed as distributions of “visual words”.

Few UCD approaches in imaging informatics exist, such as a usability test study on integrating a CAD on bone age assessment [74] and a usability study on an image retrieval system for cervical cancer images [12]. Often, approaches that claim to follow a UCD omit several steps and do not follow in general the full development lifecycle depicted in Figure 1.1.

Following the trend of the era it has been proposed that imaging informatics should follow a more patient-oriented approach [158]. Right imaging procedures for each patient need to be chosen. A more personalised imaging record of the patient needs to be created, including examinations taken from different institutions. Similar cases where diagnosis and treatment is reported need to be found when rare pathologies are met. This information while being present in the hospital databases or on the Internet is either difficult to access or not trustworthy. Using the visual content of the images in order to find similar cases has been proposed in the past [140]. The description of this approach is presented in the following section. Recent studies on human perception of medical images [62] and the role of the experience level in the domain in image description [187] provide more information on how the visual content should be used for such a purpose.

1.3.1 Content-based image retrieval

A technique that gained much attention in the information retrieval field during the last 20 years is Content-Based Image Retrieval (CBIR). This technique uses an example image instead of query keywords in order to find relevant images in an image collection (Figure 1.2). A CBIR system uses a mathematical model to represent the visual content of the images and applies a distance metric to measure the similarity between the query image and each image in the collection. The images of the collection are then returned to the user, sorted by their calculated similarity to the query image.

Research on CBIR has been carried out in several fields, such as object and scene retrieval [170] and remote sensing [48]. In the early years, mathematical models were used to represent the visual content of the image in a holistic manner [99, 148]. Later, local descriptors [117] modelling the information around specific points or a region of interest (ROI) were shown to outperform global descriptors in several tasks [57, 134]. While local descriptors allowed for partial matching of images and showed scale and rotation invariance, they were inefficient for search within large-scale image collections. For this reason, more compact representations inspired from text-based information retrieval such as Bags-of-Visual-Words [170] have been developed. Efficient indexing structures such as the Inverted Index have also been employed to allow for fast real-time search [142].
CHAPTER 1. INTRODUCTION AND THESIS OVERVIEW

CBIR is an active research field and has been proposed to be promising for medical information retrieval [140] since there are scenarios in which keywords may not be the most optimal choice for initiating a search. Recent studies investigating the image search behavior and requirements in radiology provided an overview of the daily clinical work flow of radiologists [129]. The need for additional information often occurs when abnormalities of an unknown pathology for the radiologist are encountered in a new case. This may occur to trainees and residents but also to experienced clinicians in rare pathologies. In such situations, choosing the correct keywords can be a difficult task and may require several search refinement iterations. Cases where relevant information and evidence is needed for differential diagnosis can also be very time-consuming using text-based search. Mobile devices, such as tablets and smartphones are increasingly integrated into clinical routines and extensive interaction using typing of keywords can be impractical.

In the radiological use cases of CBIR the image example can be a diagnostic image from a newly introduced medical case. ROIs can be marked to narrow down the search. Retrieved images and their relevant ROIs are returned together with their associated cases allowing for differential diagnosis.

Several attempts of applying CBIR in the medical field can be found. Retrieval from mammography image databases has been proposed in [63]. CBIR systems on x-ray databases have also been developed [151]. In [169] a retrieval system using visual content for epilepsy is described, accessing a database of Magnetic Resonance (MR) and Single-Photon Emission Computed Tomography (SPECT) images. High resolution computed tomography (HRCT) image collections have also served as databases for CBIR [56, 168].

While solely visual content of the images may not always be enough for effective retrieval, it provides extra information that complements the textual search [70]. Thus, CBIR can replace text-based search in cases where correct keywords are difficult to find, or complement a text query to reduce the number of keywords needed as well as the number of search iterations. Additional navigational techniques that often accompany CBIR, such as relevance feedback (RF) [157], allow for even easier and faster search refinement.

1.4 Thesis overview and contributions

This section provides an overview of this thesis and its organization in Section 1.4.1 and summarizes the scientific contributions and their associated papers in Section 1.4.2.

1.4.1 Thesis overview

This thesis describes a user-oriented development cycle for a novel information system for radiologists for retrieving images and cases from the medical literature. It begins in Chapter 2 where a survey on the image use and search behavior of radiologists is presented. The system design and implementation is detailed in Chapter 3 along with the applications that is integrated into. The components of the complex indexing and retrieval pipelines are presented. The results of the empirical evaluation of the system and the methods that it implements are presented and discussed in Chapter 4. The usability of the system is assessed on an iterative user-study with radiologists in Chapter 5. Finally the study is concluded in Chapter 6 with a critical view on its limitations and weaknesses.

1.4.2 Scientific contributions

This thesis was carried out in the context of the Knowledge Helper for Medical and Other Information Users (KHRESMOI) project. KHRESMOI\(^1\) is a four-year project that aims at creating a multilingual, multi-modal search and access system. This system can assist general practitioners, the general public and radiologists in accessing trustable biomedical information. These three target groups have different search behavior, goals and information needs. Thus, the system has

\(^1\)http://www.khresmoi.eu/
been divided into subsystems, designed to correspond to the requirements of the target groups. One of the main functionalities is to allow efficient access to the visual information available in electronic records and the open access medical literature on the Internet. The system applies several novel information extraction and retrieval techniques, such as CBIR, relevance feedback and the use of semantics in 2D and 3D medical image search.

The main scientific contributions of this thesis are in the fields of medical image retrieval and classification, user–centered system design and evaluation:

- One of the contributions in user–centered design lies on the identification of user needs for medical image retrieval with the conduction of the survey published in [129]. Thus, in–depth insights are provided for the image use and search behavior of image user groups, such as radiologists.

- System requirements and specifications were derived of the results of the survey and a medical image retrieval system was designed and developed [124, 125, 163]. This constitutes the first user–oriented approach for medical image retrieval systems, using features such as CBIR, relevance feedback and semantics.

- Complex indexing and retrieval pipelines were developed [123] to include features such as modality filtering [67] and compound figure separation [38] in retrieval. These features allow for better usability of a medical image search system and can contribute in higher retrieval performance.

- Implementation and evaluation of parallel visual indexing using the MapReduce framework was published in [130]. Results include the use of an easy to setup hardware cluster for parallel computations that provide significant speedup of the indexing of the visual features of large datasets.

- A contribution in medical image retrieval concerns the use of density–based clustering for the creation of visual vocabularies [126]. The algorithm can detect irregular clusters of high-dimensional large-scale datasets in a time efficient way.

- Empirical evaluation of state–of–the–art local features and image representations is conducted for medical image datasets.

- Semi-supervised machine learning is applied for improving k–Nearest Neighbour classifiers (kNN) for medical image categorization [67]. The results show an improvement over simple kNN classifiers and the method is applicable to more complex classifiers.

- Relevance feedback techniques using visual, textual and multi–modal information are evaluated in [131]. It is shown that all evaluated relevance feedback techniques really improve the initial query results. The performance over the various techniques is evaluated for several result list sizes and proposed techniques using multi–modal information are shown to have competitive performance.

- The first –to our knowledge– extensive user–oriented evaluation of a medical image retrieval engine is conducted in two rounds of user tests, providing a user study protocol for assessing medical image retrieval systems, insights about the usability of novel techniques in image retrieval and discussions on what could be common pitfalls of sophisticated medical information systems [122, 127, 128].
Chapter 2

User requirements and system specifications

The development of applications and products based on real user needs has been a standard procedure in many other fields [135] including medical information retrieval [58, 155]. As is stressed in [79], “...by asking from the perspective of the user, ‘what should a successful system do’, relevant variables can be identified”. For CBIR systems such user studies were initially only very rarely performed. Approaches were rather technology–driven in terms of applications than based on real user requirements. On a wider scale, many studies exist on the information needs and the use of information retrieval (IR) systems. In [85], a systematic framework for evaluating the use of medical IR systems is proposed. While the full framework is out of the scope of our study, important concepts, such as user satisfaction and search failure, were taken into account.

A theoretical analysis on information needs is attempted in [193]. Moreover, a model for information seeking behavior is proposed using four categories. Categories a and b include search strategies by the user, that may or may not depend on a mediator/IR system. Categories c and d, include the search strategies employed by the mediators/IR systems to satisfy the user’s demands for information. Our study focuses on the first two categories in the field of radiology, while the two latter will be the subject of user testing and system evaluation.

One of the first user analyses for image retrieval was [121], on the behaviour of journalists when searching for an image. For this purpose, observation of the journalists in their work, interviews with them and analysis of a sample of their queries were used. Analyzing queries is a common approach for investigating visual information search needs and behavior in specialized fields, such as history [41] and art [114], or general large scale studies [78]. However, as it is well explained in [114], “the use of image retrieval systems varies in different fields because users have their own specific information–seeking behavior and need unique features designed for their tasks”.

Similar approaches have been followed in the medical domain. Some studies rely on log files from either media search engines [136] or U.S National Library of Medicine’s MEDLINE system [139] to find out more information on how clinicians search for images. Others [83, 137] perform interviews and surveys among clinicians, but these were done on a small scale and were not focused on radiologists specifically. The study described in this paper aims to shed light onto the search for images in the field of radiology and identify the requirements for a specialized image search engine.

In general, a variety of methods exist for obtaining information on system use:

- observation of the behavior of users (which can include direct observation [105, 121] or the analysis of log files that record the behavior of users [41, 78, 114, 121, 136, 199]);
- interviews with stakeholders [43, 83, 105, 121];
- surveys [135, 137, 199].

While observation can provide useful insights on the behavior of the subjects, it often lacks to provide clear and in–depth information about important concepts, such as user satisfaction,
unmet needs and desired functionalities. On the other hand, interviews assist in obtaining such information but mainly on a qualitative level as such interviews are usually time-consuming. Finally, structured surveys with stakeholders can give quantitative results, but need careful design, as questions should be easy to understand and on target and need to be performed on a relatively large scale in order to have statistical significance. As the role of these methods is complementary, all of them were applied in this study.

In addition to direct observation, another way to monitor image use behavior is observation using eye-tracking equipment. Eye gaze tracking has been widely used in user interface design and evaluation [15, 22, 46, 105] but also in the radiology field, eye-tracking analysis was used as input for the design of workstations [15, 19]. Moreover, studies of eye movement tracking were used to analyze image interpretation and decision making of radiologists [19, 106, 107, 181]. However, most of these studies either concentrated on specific anatomy locations or on certain modalities.

2.1 User requirement investigation methods

This section describes the techniques that were used for obtaining information on the image use and search behavior of radiologists. All tasks were performed in the radiology departments of the Vienna University hospitals, Austria and the University hospitals of Geneva, Switzerland, two large teaching hospitals. The number of responses is low as only known radiologists were targeted in the institutions to guarantee a high quality of responses.

2.1.1 Observation

In order to learn more about information behavior of radiologists, watching them perform standard tasks and then analyzing information needs at specific moments was the first step. These observations were used to help constructing the survey and the questions. To obtain information on the workflow three steps were taken:

1. Listen to experienced radiologists describing the main steps of the radiology image analysis process; This includes a description of the workflow through diagnosing cases and preparing the radiology report;

2. Case discussions were followed where interesting cases are explained including the reasoning process, imaging data required, and evidence provided by several exams; This process also described the workflow and the use of external knowledge to support the workflow;

3. Eye tracking experiments were performed, where radiologists diagnose cases while being eye-tracked; to obtain information on areas of the images that are of particular interest and about the way a clinician browses through an image series.

The setup of the eye-tracking consisted of a stand-alone workstation at the University hospital of Vienna, which was not connected to the PACS network. There was one workstation PC connected to one 23” LCD-Monitor. The eye-tracking software and hardware were installed and patient studies were imported to the work station via a CD.

2.1.2 Interviews

After a first survey form was constructed using the observation results and the literature review, several structured interviews with the draft survey form were performed. The interviews where used to learn whether the questions were understandable and whether responses correspond to the study’s target of interest. The detailed analysis had the goal to find problematic parts and analyze whether the forms were understandable for radiologists.

Three detailed interviews were performed in Geneva with successive versions of the survey form, where a clinician filled in the form explaining aloud how the questions were understood and why a particular answer was given. Each time the form was adapted based on the comments of the
previous interview. In Vienna, two rounds of structured interviews with the survey forms were performed and the form adapted accordingly.

2.1.3 Survey

Starting point for the survey questions was a user study previously performed in Portland, Oregon, USA and then later in Geneva, Switzerland [137]. Based on the questions in this survey a form was adapted to comments from local radiologists to correspond to the specific group of radiologists instead of clinicians in general, as the first surveys did.

Three main tasks were identified to evaluate the specific needs:

1. clinical work on patients;
2. work regarding teaching, as in the preparation of lectures;
3. research work that can include a variety of tasks.

Besides the search requirements, basic demographic data on the radiologists was acquired to better interpret results. The final version of the survey consisted of 4 sections: general data, clinical work, teaching, and research. In the general section there were questions regarding:

- age;
- gender;
- specialization within radiology;
- country of radiology specialization;
- type of hospital;
- years of experience in radiology;
- activity distribution between teaching / research / clinical work.

A common set of questions was used for the three activity domains. The first part is focused on the current image search behaviour of radiologists:

- the tasks where images other than those of the patient treated, need to be found;
- which sources are searched for images;
- how the search is performed;
- how relevance of an image can be determined;
- how often search for images fails;
- why the search fails;
- how much time is taken before stopping;
- how much time is taken when relevant images are found.

In the second part the participants were asked to propose services and tools useful for their search, and imagine a perfect image search system for their needs. Questions were:

- What are useful additions for search systems?
- What would a perfect search system be like?
- How can visual information of images be exploited?
- Which tools for an automatic annotation of images would be useful?
- Are medical terminologies or ontologies being used?

The questionnaire form used in this survey can be found in Appendix A.
2.2 User requirements investigation results

This section describes the results obtained in the study.

2.2.1 Observation results

The radiology workflow starts usually with opening a case for which an imaging exam was requested. The images are then transferred from the PACS server to the local viewing workstation, and then the viewing process can start. The viewing options are set depending on requirements such as size and number of views per screen. The setup depends on the imaging modality and on the radiologist’s preferences and can be changed during the analysis process. Before starting to analyze the images the patient’s medical history and anamnesis are reviewed. The radiologist then analyzes the images by adjusting the brightness/contrast and scrolling through the slices. The sets of images can be changed using thumbnail previews. Tools such as for measuring sizes are available for specific organs and pathologies. Once a pathology or abnormality is found there are two possibilities:

The abnormality is known: potential diagnosis and differential diagnosis are given and the medical finding is described.

The abnormality is unknown: search for additional information is needed.

A common way to handle unknown abnormalities is to ask an experienced colleague for help. This sometimes ends up in group discussion about possible pathologies, corresponding to “information exchange” as in [193]. Often, the radiologist has to search through the literature (Internet, books, scientific articles). For this, the pathology needs to be described as well as possible. With the potential diagnosis and differential diagnosis the medical finding is completed. If, from the scientific or teaching point of view, the study is interesting for the radiologists it is being marked for future reference.

In case discussions, the workflow is similar to the description above starting with the anamnesis and history of the patient. One of the important aspects when presenting cases is sharing the experience between radiologists particularly for cases occurring rarely. Important steps in the diagnosis process are the comparison of findings with the state of the art in the literature and request of additional exams (imaging, laboratory, etc.) to assure that the probability of a correct interpretation is high. This means that access to and knowledge of the literature is important in practicing evidence-based medicine also in radiology. Justifying decisions is important and links to related cases are essential.

Other important aspects mentioned in the seminars are the temporal nature of images, for example comparing images of the patient over time, measuring for example the growth of a tumour. Computer-aided detection, such as highlighting particular abnormal regions or visualizing results are also important.

For the eye tracking, the system was calibrated for each participant. When calibration was finished successfully, the participants were performing their image viewing and analysis tasks described above. The study included 3 sessions, each with a different radiologist. Two persons were working on the same studies (head CTs, mammography, chest x-rays) and the third one on a knee MRI. As the system was on a separate workstation these were chosen cases and not cases they would have worked on anyway. The task was to perform the usual analysis for diagnosis. The radiologists were explaining their tasks while they were performing the actions and the results of the experiments were visualized among others as heat map images (Figure 2.1).

2.2.2 Interviews’ results

The main outcomes of the structured interviews to adapt the forms were:

- radiologists are not familiar with visual retrieval (search using visual characteristics of an image) and providing examples in the questionnaires could be better;
2.2. USER REQUIREMENTS INVESTIGATION RESULTS

Figure 2.1: Heat map images for knee MRI (left) and a chest x-ray (right), showing that often a particular region attracts the main attention as in the knee MRI or several distinct regions as with the chest x-ray.

- many persons store images locally on their computers for future use and this has to be taken into account (although this is not a desired practice in most hospitals where data acquisition needs to be validated by an ethics committee);

- radiologists found it hard to separate between the three proposed tasks of teaching, research and clinical work and mixed things when filling the survey, sometimes mentioning the overlap between the tasks;

- many formulations were modified as clinicians referred to it as computer science jargon that could be hard to understand.

The results of the structured interviews were used for modifying the forms.

2.2.3 Survey results

This section describes the outcomes of the survey to which 34 radiologists responded. Ten persons filled in the paper form after a seminar at the University of Geneva, whereas all other participants filled in the electronic form. The ten paper forms were transcribed into the electronic form for a homogeneous treatment.

Demographics:

As many questionnaires were filled in seminars, about half of the population is under 30 years. The other half is evenly distributed between 30–55 years. Two thirds are male and one third is female, highlighting that radiology is one of the few domains with a majority of men in medicine. As could be expected from performing the survey mainly in Vienna and Geneva, most persons have had their radiology education in either Austria or Switzerland. All other countries are from Western Europe or the US, meaning that education is comparable.

In terms of radiology specializations there are no surprises. 23 persons specialize in general radiology, 1–2 persons each in musculoskeletal radiology, thoracic radiology, radiology informatics, neuroradiology, orthopaedic radiology and body imaging. One person mentioned to specialise in CT and another one was still a student. 28 persons work in public hospitals and two in private clinics, with four persons mentioning to mainly work in research at the University. The rather junior sample, is seen in the years of experience, with eleven persons having less than two years of radiology experience. Otherwise, the distribution is relatively even.

For the work time distribution it was possible to weight the time spend on clinical work, teaching and research on a scale from 0–5. This allows estimating the percentage spent on each of the activities. Most persons perform all three activities, few have no teaching or no research and all clinical work.
Clinics — Teaching — Research:

All 34 participants mention tasks where they search for images other than of the person being diagnosed. There are several reasons for searching images apart from the ones they are currently assessing. The main reasons to search for images are finding material for presentations (mentioned by 8), differential diagnosis during a medical finding for difficult cases or in case of an unclear pathology (mentioned by 13) or performing clinical research (mentioned by 3). Specific examples listed are lung fibrosis, brain- or bone tumours or lesions in brain, liver or other structures. A task mentioned, where images can be useful is also the grading of a disease.

For teaching, the main focus of clinicians is to find similar cases. Depending on the class they are teaching they look for easy, advanced or tricky cases. The image type depends on the current topic and ranges from plain x-rays, CT scans to typical pathologies such as primary brain tumours or lesions. It also includes differential diagnosis. Links with images of the scientific literature were also mentioned as useful.

While performing the search the most frequently used source is the Internet using keywords (Figure 2.2a) mentioned by 14. Google (5) is used as well as public medical databases (PubMed, Goldminer, e-anatomy, Eurorad were mentioned each by 1–3 ) and Wikipedia (mentioned by 2). Most of the clinicians also have personal files stored on desktop PCs to search images (sometimes with keywords), as mentioned by 12. The local patient record is queried using the patient name or ID, which are sometimes stored on PDAs or local PCs regarding interesting or typical cases (mentioned by 9). Other options for finding information is looking at books (8) and asking colleagues (4). There is no significant difference between clinical and teaching activities. clinicians focusing on teaching seem to have more organized and larger personal databases. Generally keyword search is mentioned and no CBIR systems.

When an image is found, it needs to be decided whether or not it is useful based on experience and comparison with a reference case. This corresponds to the notion of relevance in information retrieval.

The correct image properties (e.g. modality, contrast, patient age/gender, record date, mentioned by 7), as well as the quality of the images and the reliability of the sources, define suitability of the found images. The availability of a detailed description or of comments on the image also has an influence. Asking colleagues for their opinion is another option, mentioned by two persons. Figure 2.2b depicts the responses for clinical work and teaching. Personal experience is the most important criterion followed by image properties that can include for example the modality.

A notion mentioned by clinicians was the trust that they have in the image or the diagnosis attached to it. If a diagnosis was, for example, only taken based on the images there is less trust than when a biopsy confirmed the diagnosis. For images found on the Internet this information is not always documented.

Computing the average of the clinicians’ responses , a 75% success rate searching for images

![Figure 2.2: (a) Information sources used for finding images. (b) Defining relevance or suitability of images](image-url)
2.2. USER REQUIREMENTS INVESTIGATION RESULTS

(based on self-assessment) is found. This can be an overestimation as people might not be aware of all available data and potentially more relevant items not found. In Figure 2.3a the percentage of persons with a success rate below 40% is low. When comparing teaching and clinical work it becomes clear that clinical work has a higher risk of failed search as all success rates below 40% are in this category (as seen in Figure 2.3a). This may highlight that clinical work is less well defined and has harder search tasks than for example teaching. Based on the question on the success rate of searches, in clinical work the time taken for searches is found to be lower than for teaching and research, which also explains the higher failure rate.

Figure 2.3: Self-assessment of the success rate of image searches, (a) the overall percentages and (b) a comparison between clinical work and teaching.

The clinicians think that most of the time the desired images are available but cannot be found due to various reasons. The main reason for not finding a relevant image is that the topic or pathology is rare, too new and sometimes also too general. It needs to be noted that not all storage systems are fully searchable (e.g. scanned reports). Time pressure has a negative impact on finding relevant images as well. Figure 2.4 compares the responses for clinical work and teaching but both categories lead to similar results.

Figure 2.4: Reasons given for unsuccessful image search.

When comparing search times of successful image search, it becomes clear that around 65% of successful searches finish after ten minutes or less (Figure 2.5a). Only a few persons search longer to successfully find images. For unsuccessful image search a few persons already stop after 5 minutes or less, but most often 10 minutes or even over 15 minutes are mentioned before stopping (Figure 2.5b). This highlights the importance of image search in the workflow. It also highlights the room for improvement with optimized search tools to find relevant information quickly.

Figure 2.6 compares the search time for clinical work, teaching and research. The length of successful search for clinical work is rather short and considerably below 10 minutes. For research only few responses were obtained but for teaching the time to successfully find images is much longer and many search for 10 minutes or more and still find relevant images.
CHAPTER 2. USER REQUIREMENTS AND SYSTEM SPECIFICATIONS

Figure 2.5: Time taken to find relevant images (a) and times before an unsuccessful search is abandoned (b)

Figure 2.6: Time taken for successful (a) and failed (b) image search compared for clinical work, teaching and research.
2.2. USER REQUIREMENTS INVESTIGATION RESULTS

For unsuccessful search, times for research are often more than 15 minutes. During clinical activities, the average time before quitting the search is significantly shorter, between 5 and 10 minutes, probably due to time pressure. This difference balances the overall failed search time distribution. Since about one fourth of all searches are unsuccessful it is clear that having tools to more easily find out whether or not relevant images even exist can be very helpful and reduce the amount of time lost.

23 of 34 persons responded to the question on potentially useful additions for image search, sometimes in several categories such as teaching, research and clinical work with slight differences between the categories. The functionalities most often suggested are search by pathology (23) and modality (19), followed by search for similar images (17) and patient demography (8 times). Apart from the predefined options the radiologists mentioned the need for multilingual retrieval and proposed other additions, such as pathology or even symptom classification (using for example an ontology), query by text and image and semantic retrieval based on image characteristics. The search for reconstructed 3D images was also mentioned as was the need to connect radiology images with histopathology or other criteria allowing to judge the confidence of a diagnosis.

19 responses were obtained for desired input possibilities and 22 for desired result formats. As there were no major differences between clinical work/teaching/research, the three are combined. The perfect image search system should use images (with possible regions of interest) as well as keywords as input (Figure 2.7a). Keywords could vary strongly, from describing the anatomical structure, the pathology and histology, up to more demographic information like patient age.

![Figure 2.7: Desired (a) input and (b) output data for a medical image search.](image)

The output should include image examples and a corresponding description (Figure 2.7b). If available, differential diagnosis could be provided by the search engine. More detailed information including references would be helpful. A few people mentioned that information supporting the diagnosis would be useful such as a biopsy, to raise the level of trust in the information supplied.

There was a relatively small amount of feedback regarding the exploitation of visual information for information search, indicating that the radiologists are not familiar with the concept, the state of the technology and research. However, there were interesting suggestions: the search for similar images and similar cases (mentioned by 5), search for similar regions of interest (3) and the possibility to search for similar images and have social judgements of other radiologists on the similarity (mentioned by two persons). The importance of not only visual information but the connection with other patient data was mentioned three times, full text search also three times. The possibility to have statistics on the diagnoses for similar images was mentioned once.

In total, 24 radiologists responded to the question about possible goals of automatic annotation. Anatomic region is mentioned 24 times as being important and modality 11 times. For research and teaching the modality is mentioned more often than for clinical work but otherwise differences between categories are small. Another annotation target mentioned was the quantification of the size of structures (6 persons). It was also mentioned that all the extracted information should be made available as free text for image search. Few radiologists (mostly the more experienced ones) mention to use systematic terminologies for image search or image descriptions. Of 20
persons who responded to the question, 7 mention to not use any terminology at all. Medical Subject Headings (MeSH) was mentioned most frequently (10). There is no major difference in terminology use between clinical work, teaching and research. Five participants mention to use RadLex and SNOMED CT was mentioned once. The dominant PACS system mentioned is AFGA IMPAX (8), followed by Fuji Synapsis (3) with one participant each mentioning GE, McKesson, CareStream, PACS/RIS and Siemens Syngo. One participant answered to use several systems.

The perfect system:

As the questions on how visual information can be exploited and how the perfect system should look like are of great importance for the design of medical image retrieval systems, we analyze the responses with a focus on the text given by the respondents. It is clear that imagining the perfect search system is hard when no example system is known. Still, besides the current image use and search behavior several persons added comments about the perfect search system. Most comments mentioned here in their raw form (the survey as well as the responses were entirely in English), not ordered by domain (teaching, clinical work, research):

- like Google but including DICOM images and text combined;
- structured information on a case including histopathology, images, structured data;
- confidence score in the diagnosis, e.g. backed by biopsy or other exams; search by diagnosis;
- simple to use, minimize need to play with 3D stuff;
- differential diagnosis, multiple views possible, feedback of others possible;
- pathology chosen by radiologists, search among images in a similarity cluster;
- like Google but with references to the literature such as Goldminer;
- marking the ROI in an image and search with this, search for anatomical structures, for normal and abnormal cases;
- keyword and image as input for search, selection of case and then search for further information on the found cases;
- quantification of the size of structures to search for;
- search for differential diagnosis, location, organs and particular conditions;
- search by diagnosis, the opposite to clinical reading;
- look for a certain pathology and find cases with it including images;
- having research databases and research PACS linked, search by keywords rather than by image;
- having the entire patient documentation searchable by keywords; yottalook is of good quality for this;
- search by image description, pathology and histology would be useful;
- "show me a similar image for which a final diagnosis is available";
- free text search in radiological reports;
- Be able to search by various key words. E.g., By diagnosis: "pneumonia" would retrieve many different types and appearances of pneumonia. You should be able to limit your retrieval by additional terms, e.g., "Klebsiella pneumonia" would limit the retrieval to pneumonia caused by that pathogen;
2.3. DISCUSSION

- By manifestation: "interstitial disease" would retrieve many different types and appearances of interstitial diseases. You should also be able to limit further using additional descriptors;
- By appearance: e.g., "round lung lesions" would retrieve all types of diagnoses that would fit that description;
- It should also have a "more like this" function;
- It should have a way of circling a region of interest and retrieve more images that are similar to the region of interest that was indicated.

The results show that perfect search systems are more concerned with structured data than they are with visual data, although several people mention search for similar cases or images. On the one hand, this could indicate the lack of knowledge about current research prototypes for visual retrieval, but on the other hand it highlights the importance of structured clinical data in the process even for an image–based domain such as radiology. This means that even for visual retrieval, combinations with clinical data are essential. Some text search systems such as Goldminer and Yottalook are known to the clinicians and used by them. Similar to search engines like Google used by all clinicians it can be important to base a new search tool on what is known and used at the moment and then add functionalities such as visual similarity search. This is not unproblematic since it includes the continuation of design problems of current systems, but it can avoid a rejection by the clinicians. Social interactions with others and comments of others are also mentioned to be important for the search.

2.3 Discussion

Many problems still exist in the current situation in terms of information retrieval. Correctly describing the pathology can be a difficult task but it is essential for retrieving usable results. Visual retrieval is not possible with the clinical systems currently in use. Books are not instantly available for all pathologies and sometimes have to be found in the library. Searching in books is easy when ideas about pathologies exist. The Internet is a commonly used resource. Besides standard medical databases (such as MedLine), search engines (e.g. Google) are primarily used for the search. Search engines are powerful but not designed for medical purposes and often return undesired and not trustworthy results.

Hardware limitations of the eye tracking system lead to a reduced experimental setup where only one monitor was tracked (instead of the two that are otherwise often used in radiology). Therefore, it was not possible to capture the precise workflow of radiologists using the specialized hardware. About 30 minutes of the process were recorded. When discussing the data it became apparent that it might be better to clearly select a number of reference cases for such a study and then compare this between radiologists. In general, a larger number of cases might have lead to more conclusive results but this first study was mainly aimed at obtaining first ideas. Substantial differences in viewing behaviour between image types were found. In some cases single areas of high concentration of fixation can be determined while other image types show broad scanning paths in viewing behaviour. Administrative difficulties such as the anonymization of videos when viewing patient data were found during this test session and these could help to improve the recordings for a future eye tracking session:

- tracking two screens at the same time lead to bandwidth problems with the eye tracking system.
- videos recorded had to be anonymized as well before they could be analyzed, so all patient names were removed using a low pass filter. In future (large scale) eye-tracking tests the anonymization process needs to be automated.

The analysis of the observations was mainly used for the construction of the survey. The eye tracking confirmed the belief that very small regions of interest are what clinicians really focus their activity on and thus search by regions constitutes one of the important requirements for medical CBIR systems.
Chapter 3

System design and development

Chapter 2 provided some insight on the current image search status in radiology. The results of radiologists failing one out of four times in image search, clearly shows that there are needs in radiology–related information retrieval that are currently unmet. Scenarios where search using keywords is not optimal were also identified. In this thesis, apart from using conventional methods like text–based retrieval, novel techniques such as CBIR are integrated into a medical information system to cover the unmet information needs.

Medical image retrieval systems exist both for general radiological information and in more specialized databases (see also Section 1.3.1). Systems that provide user–interaction techniques, such as relevance feedback, have been proposed [63, 152]. In [168] a physician–in–the–loop is included in the evaluation process. The Image Retrieval in Medical Applications (IRMA) system is reported in [64] to be integrated into a hospital PACS. The THESEUS MEDICO [165] system searches for similar cases after applying semantic annotation to images. For the semantic annotation it uses the RadLex Ontology\(^1\) and a human–based correction step.

While most of the systems mentioned above are designed to exploit internal resources of the medical institutions, a large amount of visual information can be found in the medical literature on the Internet. Radiology–related information search engines, such as Goldminer\(^2\) and Yottalook\(^3\) access this information using text–based search with great success. The OpenI system\(^4\) allows either text–based or image–based retrieval in the open access biomedical literature.

More detailed reviews of medical CBIR techniques and systems can be found in [6, 77].

3.1 The Parallel Distributed Image Search Engine

In the context of this thesis, the findings of the Chapter 2 survey are translated into a list of system requirements (Section 3.1.1). Then the design and the implementation of a novel image retrieval system, named Parallel Distributed Image Search Engine (ParaDISE) are described in detail (Section 3.1.2). The main application of ParaDISE, the KHRESMOI Radiology prototype, is presented along with two innovative human–computer interaction applications that use ParaDISE as backend (Section 3.1.3). Finally the methods used in the indexing and retrieval pipelines of KHRESMOI are described 3.2. The various components of ParaDISE are evaluated in Chapter 4.

3.1.1 Specifications and System Design

The observations of the workflow in the investigation of the image search behavior showed that the need for additional information during clinical duties occurred when the pathology of an abnormality found in a new case was unclear or unknown. Moreover, it was often mentioned in

---

\(^1\)http://www.radlex.org/
\(^2\)http://goldminer.arrs.org/
\(^3\)http://www.yottalook.com/
\(^4\)http://openi.nlm.nih.gov/
the survey that images or interesting cases were searched for lectures or presentations in academic work. Thus, the radiologist may or may not know some keywords to initiate the search. This dictates that a medical image retrieval system should support querying by text, by image example (e.g. for the cases where no pathology keywords are known) or a combination of the two (e.g. for cases that the user may have a hint but not certainty). Relevance feedback or term suggestion techniques could also help refine the search if the object of the search is not fully clear.

The Internet was mentioned as one of the main sources where radiologists where seeking for information. At the same time, the quality of the results and the case context associated with the image were mentioned among the most important criteria when judging the results quality. As peer-reviewed articles can be considered as a trustable source, indexing images from the medical literature on the Internet can achieve a high level of result quality and quantity. A search system should provide linking and easy navigation between the images and the case they belong to.

Linking of internal sources, such as PACS with the medical literature and personal archives was also considered important when searching for information. As these sources contain heterogeneous imaging data, different features and image representations need to be supported. Extending the search into multiple indices should be possible, as the ability to interconnect with other search systems.

The main reason for image search failure given by the participants of the survey was that the information sought was too rare. However it was believed that it should have been available somewhere but they could not find it. Moreover search needs to be fast as radiologists have very tight schedules. In order to provide quick access to new findings on a rapidly-growing scientific field, the system needs to have regular index updates and be scalable to millions of images and articles.

A first list of system requirements can be derived from this analysis:

- support of query by keywords, image example or combination of both;
- index of a trustable source, such as the peer reviewed medical literature on the Internet;
- linking of images and associated articles, easy navigation between the two;
- support of different visual features and image representations;
- support of search into multiple indices, interoperability;
- scalability and support of regular index updates.

As seen from the literature review, while certain text-based medical image search engines exist on the Internet, CBIR systems are less common and effective. This study focuses more on integrating the CBIR in an image search application for effective retrieval. The design and implementation of ParaDISE is described in Section 3.1.2.

### 3.1.2 Architecture

The design of ParaDISE was based on the following concepts: flexibility, expandability and scalability. How these concepts are achieved by the design choices will be discussed after the overview of the architecture of the ParaDISE backend and frontend.

- **Backend**
  The ParaDISE backend follows an object-oriented architecture and consists of basic elements, called Components. Each Component is associated with a Manager object. The Manager is responsible for selecting one out of the supported instances of the Component. The behavior of the selected instance is controlled by a Parameters object that contains the tunable values of the method implemented in the instance. The Components are the Extractor, the Descriptor, the Storer and the Fusor:

  - **The Extractor** undertakes the extraction of local descriptors.
3.1. THE PARALLEL DISTRIBUTED IMAGE SEARCH ENGINE

Figure 3.1: An overview of the ParaDISE backend. The four basic elements are combined to perform the indexing and search processes.

- **The Descriptor** is responsible for the mid-level features aggregating the local descriptors extracted by the Extractor. It also contains global descriptors, for which no local feature extraction is needed.
- **The Storer** is used to store the image representation vectors produced by the Descriptor during the indexing process. It is also responsible for accessing the index during online search.
- **The Fusor** undertakes the fusion of retrieved results lists. These can be either lists retrieved by multiple image queries or results retrieved using different features, indices and even other image retrieval systems.

The Components are combined to perform the two main operations for CBIR, offline indexing of the database images and online search using a set of image examples (Figure 3.1). These two processes are implemented in complex ParaDISE elements, called CompositeComponents: the Indexer and the Seeker. Again, a Manager is used to select an available Indexer or Seeker and a CompositeParameters object is used to control its behavior. The indexing and search processes are described in more detail in Sections 3.1.2 and 3.1.2 respectively.

The Component count was kept as low as possible to cover most CBIR approach pipelines without making the system architecture too complex. However, the addition of new Components (e.g. a Preprocessor or a Classifier) is relatively simple due to the component-based architecture. JAVA was chosen as the main programming language for the implementation of the ParaDISE backend.

- **Frontend**
  The ParaDISE frontend, namely the service layer, consists of multiple web services which use a REpresentational State Transfer (REST)-style architecture. Standard Hyper Text Transfer Protocol (HTTP) GET and POST requests are used to communicate with the web services. However, an offline version of ParaDISE frontend exists in the form of a JAVA library. This facilitates easy installation and usage of the engine for single-server applications, personal databases and small-scale research experiments.
As mentioned before, the design concepts of ParaDISE were flexibility, expandability and scalability.

Flexibility for such a system is crucial, in order to be usable for both research purposes and as an application. Evaluating image representations is really important in CBIR as different features perform better for different databases, depending on the content and the task. Moreover, state-of-the-art CBIR techniques usually include several steps and require a lot of parameter tuning [197]. The choice of component-based architecture for ParaDISE allows for combining local and mid-level features and the evaluation of single steps in the indexing and retrieval pipeline. The use of editable parameters of the ParaDISE components facilitates tuning parameters and experiment with different configurations of methods. Scientific software packages, such as Matlab can cope with most research tasks. They are, however, rarely used in practical applications due to their lack of efficiency. ParaDISE is programmed in JAVA and uses JSON as a data transfer protocol to enable interoperability and realistic application development. The frontend webservice-based architecture allows for the integration of ParaDISE into larger systems and a flexible hardware topology. The use of REST and HTTP request simplifies interaction between the system and various client applications (Web-based or desktop applications that can be written in any language capable of making HTTP requests).

With CBIR being an active research field, novel techniques emerge achieving faster and more precise performance. Thus, expandability is important to be able to add new components for specific steps or new algorithms for the existing components. The object-oriented and plugin-like architecture of ParaDISE allows for such expansions (e.g. 3D features, a Classifier component etc.). The late fusion techniques of the Fusor component can be used to expand the engine by combining it with other retrieval systems (e.g. text-based retrieval engines, such as Lucene).

Last but not least, scalability is a critical issue for many real-life applications and an active research field in CBIR [8, 92, 150, 166]. Indexing large image collections and storing the indices can be troublesome and resource-demanding. Updating such indices in regular time intervals should be taken into consideration when designing the indexing pipelines. Moreover, exhaustive search time in large indices is prohibitive in CBIR applications, since CBIR search constitutes of computing distances of image descriptor vectors. In ParaDISE, parallel indexing is supported using the MapReduce framework [52] (see Section 3.1.2). Efficient indexing methods to facilitate fast online search and binary descriptors to reduce memory storage are also supported (Sections 3.1.2, 3.1.2, and 3.1.2). The component-based architecture is dealing with scalability by allowing the use of distributed resources and expand when the amount of data and computations grows.

In order to represent the 2D images included in the biomedical literature, a large bank of visual feature extractors has been built into the ParaDISE system. This bank was used for finding the features or combination of features that better model the visual information. Already implemented available features were integrated into the feature bank and new, state-of-the-art features were implemented, to provide a large variety of features. These features are split into two categories, local features and global descriptors, and are presented in Sections 3.1.2 and 3.1.2.

ParaDISE Extractor

Local features have been used in CBIR for more than a decade [117], demonstrating state-of-the-art performance in many applications [30, 134]. They represent low-level visual characteristics of regions of the image, such as color, shape and texture. The local feature extraction takes place in the Extractor component of ParaDISE. The following local features are supported in the current version of ParaDISE (see also [108]):

- **Scale Invariant Feature Transform (SIFT)** [117]
  SIFT is one of the most commonly used local features, using gradient orientation histograms to model shape. The method of extracting the SIFT features is divided into two main parts: the detection of the interest points in different scales and the description of the neighbourhoods of these points in the appropriate scale. The first part uses the result of a Difference of Gaussians (DoG) applied in scale-space to a series of smoothed and resampled
3.1. THE PARALLEL DISTRIBUTED IMAGE SEARCH ENGINE

images to detect minima and maxima. The Difference of Gaussians operator can be seen as an approximation of the Laplacian operator:

\[ \text{DoG}(x, y; s) = L(x, y; s + \Delta s) - L(x, y; s) \approx \frac{\Delta s}{2} \nabla^2 L(x, y; s) \] (3.1)

These minima and maxima are candidates for interest points (keypoints). Low contrast candidate points and edge response points along an edge are discarded for robustness to noise. Finally, dominant orientations are assigned to localized keypoints to provide rotation–invariance.

In the second part, a keypoint descriptor is created by first computing the gradient magnitude and orientation at each sample point in a region around the keypoint location. These are weighted by a Gaussian window to give less emphasis to gradients that are far from the center of the descriptor, as these are most affected by misregistration errors. These samples are then accumulated into 8-binned orientation histograms \( H_i \) with \( i = 1 \ldots 16 \) summarizing the contents over \( 4 \times 4 \) subregions, taken by a \( 16 \times 16 \) array around the keypoint. This results in a \( 4 \times 4 \times 8 = 128 - d \) vector for each keypoint:

\[ f(x) = [H_1, H_2, \ldots, H_{16}] \] (3.2)

The implementation of the SIFT feature in the Fiji image processing package\(^5\) was used.

- **Speeded Up Robust Feature (SURF)** [18]
  This local feature is similar to SIFT, with the main difference found in the way of detecting interest points. SURF creates a stack without downsampling higher levels in the pyramid, resulting in images of the same resolution. Due to the use of integral images, SURF filters the stack using a box–filter approximation of second-order Gaussian partial derivatives. The implementation of the SURF feature in the Fiji image processing package was used.

- **RootSIFT** [14]
  A local feature that is based on SIFT and introduced in [14]. The idea behind this local descriptor is to use the Hellinger kernel to alleviate the effect of large bin values “dominating” the descriptor.

- **Lab local features** [188]
  These local features are used in the approach Bag–of–Colors (BoC) [188]. They produce a 3D feature vector in the CIELab space for the most frequent color of a region.

**ParaDISE Descriptor**

While local features perform well in object recognition, image classification and CBIR, they are inefficient for large scale tasks. For this reason statistical image representations have been used, also called mid–level features, with Bag–of–Visual–Words [170] being the most commonly used. Moreover, since there is no one–solution–fits–all in image retrieval applications, other global descriptors have been included in the feature bank. The following mid–level features and global descriptors are supported in the Descriptor component of ParaDISE:

- **Bag–of–Visual–Words (BoVW)** [170]
  One of the most common approaches for image description using local features in large datasets is the BoVW representation. A set of images is chosen and local descriptors are extracted from interest points of each image of this set. The descriptors are then clustered using a clustering method into \( k \) clusters and the centroids of the clusters are used as visual words. The visual vocabulary \( V \) represents all cluster centers:

\[ V = \{v_1, \ldots, v_k\}, \ v_i \in \mathbb{R}, \ i = 1, \ldots, k \] (3.3)

---

\(^5\)http://fiji.sc/
Then, local features are also extracted from all other images in the database and mapped to the cluster centers to create for each image a histogram of visual words. Images are thus indexed as histograms of the visual words (bag–of–visual–words) by assigning the nearest visual word to each feature vector.

The final image descriptor of image $I$, called Bag–of–visual–words, is defined as a vector $F(x) = \{v_1, \ldots, v_k\}$ such that, for each local feature vector $f(x)$ extracted from the image $I$:

$$v_i = \sum_{j=1}^{n_f} \sum_{l=1}^{k} g_j(f(x_l)), \quad \forall i = 1, \ldots, k$$

where $n_f$ is the number of local features extracted from the image and

$$g_j(f(x)) = \begin{cases} 1 & \text{if } d_\epsilon(f(x), v_j) \leq d_\epsilon(f(x), v_l) \forall v_l \in V \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

This mid–level feature, as well as its variants, can be used in combination with any of the low–level local features described in section 3.1.2. The following variants of BoVW are available:

- **Binary BoVW** [170]
  A compact binary representation of BoVW for large scale CBIR. In this descriptor, the elements of the final image descriptor are given by:

$$v_j = \begin{cases} 1 & \text{if } \exists i \in [1..n_f] \text{ such that } d_\epsilon(f(x_i), v_j) \leq d_\epsilon(f(x_i), v_l) \forall v_l \in V \\ 0 & \text{otherwise} \end{cases} \quad (3.5)$$

where $n_f$ is the number of local features extracted from the image.

- **Grid BoVW**
  A mid–level representation that splits the image into an $n \times n$ grid of subsections consisting of the concatenated BoVWs of these subregions, to include spatial information.

- **Spatial Pyramid Matching (SPM) BoVW** [110]
  This mid–level feature includes spatial information of the visual words, by employing pyramid matching. The final image descriptor is a weighted concatenation of BoVWs of the subregions produced by splitting the image into various grid sizes. The weights are inversely proportional to the cell width of the grid.

- **Vector of Locally Aggregated Descriptors (VLAD)** [93]
  VLAD is a recently introduced mid–level feature that has been shown to outperform the state–of–the–art BoVW representation in several computer vision tasks. The image descriptor $F(x) = \{v_1, \ldots, v_k\}$ is a concatenation of vectors $v_j$ with elements $v_{i,j}$ defined as:

$$v_{i,j} = \sum_{j=1}^{d} f(x_j) - v_{i,j}, \quad \forall f(x) \text{ such that } NN(f(x)) = v_i \quad (3.6)$$

where,

- $d$ the dimensionality of the feature space,
- $f(x)$ a local feature vector extracted from the image $I$,
- $v_i \in V$ the visual word,
- $v_{i,j}$ the $j$th element of $v_i$,
- $NN(x)$ the nearest visual word to x.

- **GIST** [148]
  The GIST descriptor is a global feature originally created to model the shape of scenes using spatial envelope properties which were estimated using spectral and coarsely localized information. The implementation provided in [148] was used.
3.1. THE PARALLEL DISTRIBUTED IMAGE SEARCH ENGINE

- **Riesz miniature**
  This descriptor represents the image as a single Riesz transform vector [53], which is a multidimensional extension of the Hilbert transform. A first step is the downsampling of the image to reduce the dimensionality of the descriptor. It uses a linear combination of N-th order Riesz templates at multiple scales. The weights of the linear combination are derived from one-versus-all support vector machines. Steerability and multi-scale properties of Riesz wavelets allow for scale and rotation covariance of the descriptor. Orientations are normalized by locally aligning the Riesz templates, which is carried out analytically. This approach has been used to model texture, shown to outperform state-of-the-art texture attributes in lung classification [53]. An adapted version of the implementation provided in [53] was used.

- **Histograms of Gradients (HoG) miniature** [60]
  This feature mainly represents the image as a single SIFT descriptor of a downsampled version of the image. A version of this descriptor in the RGB space is also implemented to combine color information.

- **Gabor Filters**
  Gabor filters have been known to be used in CBIR, modelling texture [174]. Usually, a set of several Gabor filters of different orientations and scales is applied in blocks over the image and histograms of mean filter outputs are used to represent the texture characteristics of the image.

- **Tamura** [177]
  This texture feature was created with regard to human visual perception. Six basic textural properties (i.e., coarseness, contrast, directionality, line-likeness, regularity, and roughness) are used to model texture. For the implementation of this feature, the Lucene Image Retrieval library (LIRe) was used [119].

- **Color and Edge Directivity Descriptor (CEDD)** [34]
  This global descriptor extracts the color information from regions of the image using a set of fuzzy rules and resulting in an HSV color space histogram. It includes texture information using the proposed MPEG-7 [32] Edge histogram Descriptor rules. Finally, it uses the Gustafson Kessel fuzzy classifier [81] to binarize the histogram. For the implementation of this feature, LIRe was used.

- **Fuzzy Color and Texture histogram (FCTH)** [35]
  Very similar to the CEDD feature, FCTH mainly differs in using Haar Wavelet transform to model texture information. For the implementation of this feature, LIRe was used.

- **Color Layout** [99]
  The extraction for this descriptor consists of four stages: image partitioning, dominant color selection, Discrete Cosine Transform (DCT), and non-linear quantization of the zigzag-scanned DCT coefficients. In the first stage, an input picture is partitioned into 64 blocks. The size of the each block is \(W/8 \times H/8\), where \(W\) and \(H\) denote the width and height of the image, respectively. In the second stage, the dominant color (e.g. the average color) is selected in each block to build an image of size \(8 \times 8\). In the third stage, each of the three components (Y, Cb and Cr) is transformed by \(8 \times 8\) DCT, and three sets of DCT coefficients are obtained. A few low frequency coefficients are then extracted using zigzag scanning and quantized. This descriptor is part of the MPEG-7 standard and the LIRe implementation was used.

- **Fuzzy Color histogram** [82]
  The Fuzzy Color histogram is defined as \(F(x)\)
  \[
  F(x) = \left[ f_1, f_2, \ldots, f_N \right] \tag{3.7}
  \]
  where \(f_i = \frac{1}{N} \sum_{j=1}^{N} \mu_{ij}, \forall i = 1\ldots N \) where \(\mu_{ij}\) is the membership of the \(j\)th pixel in the \(i\)th color bin. The fuzzy C means algorithm is used for the computation of the memberships. For the implementation of this feature, LIRe was used.
• **HSV Color histogram**
  A simple color histogram in the HSV color space. For the implementation of this feature, LiRe was used.

• **Singular Value Decomposition (SVD)**
  SVD has been used as a feature in image retrieval and also medical CBIR [50, 68]. The top \( k \) singular values are used to create an image descriptor.

**ParaDISE Storer**

Four different Storers are currently supported in ParaDISE:

• **CSV Storer**
  This Storer uses a Comma–Separated Values (CSV) file to store the index. It is mostly suitable for research evaluations and small image collections, as it is very inefficient for large–scale applications.

• **SQL Storer**
  The SQL storer stores the image descriptor vectors in a table in a MySQL database. It can be used for application use cases and can handle large datasets as well as image vectors of small dimensionality.

• **CouchDB Storer**
  A noSQL alternative of SQL storer for image vectors of high dimensionality, such as concatenated feature vectors or BoVW models with large vocabularies.

• **Cassandra Storer**
  Cassandra Storer stores the index in a column family of a Cassandra\(^6\) keyspace. Cassandra allows to have a parallel database with millions of columns. This makes it suitable for very large datasets and image vectors of very high dimensionality.

**ParaDISE Fusor**

The fusion rules supported in Fusor are:

• **CombSUM**

\[
S_{\text{combSUM}}(i) = \sum_{k=1}^{N_k} S_k(i) \tag{3.8}
\]

where \( S_k(i) \) is the score assigned to image \( i \) in retrieved list \( k \).

• **CombMNZ**

\[
S_{\text{combMNZ}}(i) = F(i) \ast S_{\text{combSUM}}(i) \tag{3.9}
\]

where \( F(i) \) is the number of times an image \( i \) is present in retrieved lists with a non–zero score.

• **CombMAX**

\[
S_{\text{combMax}}(i) = \max_k S_k(i) \tag{3.10}
\]

where \( S_k(i) \) is the score assigned to image \( i \) in retrieved list \( k \).

\(^6\)http://cassandra.apache.org/
3.1. THE PARALLEL DISTRIBUTED IMAGE SEARCH ENGINE

• **CombMIN**

\[ S_{\text{CombMIN}}(i) = \min_k S_k(i) \quad (3.11) \]

where \( S_k(i) \) is the score assigned to image \( i \) in retrieved list \( k \).

• **Linear Weighting**

\[ S_{\text{linear}}(i) = \sum_{k=1}^{N_k} w_k S_k(i) \quad (3.12) \]

with \( w_k \in [0, 1] \) and \( \sum_{k=1}^{N_k} w_k = 1 \).

• **Borda Count**

\[ S_{\text{Borda}}(i) = \sum_{k=1}^{N_k} \frac{1}{R_k(i)} \quad (3.13) \]

where \( R_k(i) \) the rank of the image in retrieved list \( k \).

• **Reciprocal Rank**

\[ S_{\text{RRF}}(i) = \sum_{k=1}^{N_k} \frac{1}{c + R_k(i)} \quad (3.14) \]

where \( c \) a constant and \( R_k(i) \) the rank of the image in retrieved list \( k \).

**ParaDISE Indexer**

The indexing of the visual content of the image collection is an offline operation. As mentioned in Section 3.1.2, the Indexer CompositeComponent is responsible for this task in ParaDISE. Apart from serial indexing, parallel indexing is also supported using the MapReduce framework. Below follows the description of the two currently supported methods:

• **Serial Indexer**

The serial indexing pipeline uses the basic ParaDISE Components (see Figure 3.2). First, the local features of each image are extracted by the Extractor, if needed. Then, the image descriptor is created by the Descriptor, either integrating the local feature vectors into a mid-level representation or using a global descriptor. The Storer inserts the image descriptor vector into the index. The direction in the decision nodes is decided by the values of the Indexer Parameters.

After the index is created, a weighting can be applied to the index. The following weighting methods are supported:

– **Term Frequency – Inverse Document Frequency (TF-IDF)**

The TF-IDF weighting is widely used in text-based information retrieval. The rationale behind this weighting is that words that are found often in a document contain more information. At the same time, words that are found often in the document collection are not that informative. The mathematical expression of TF-IDF is the following:

\[ tfidf = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \quad (3.15) \]

where \( n_{id} \) is the number of occurrences of word \( i \) in document \( d \),
\( n_d \) is the total number of words in the document \( d \),
\( n_i \) is the number of occurrences of word \( i \) in the whole database and \( N \) is the number of documents in the whole database.
CHAPTER 3. SYSTEM DESIGN AND DEVELOPMENT

Figure 3.2: The indexing pipeline of ParaDISE Indexer.
3.1. THE PARALLEL DISTRIBUTED IMAGE SEARCH ENGINE

Figure 3.3: An overview of the HES–SO in–house network of computers.

It can be used in CBIR in combination with BoVW approaches. For documents containing no visual words ($n_d = 0$) then tfidf = 0 ∀i. For visual words not present in the database ($n_i = 0$), the TF–IDF of the term $i$ is not relevant and the word can be excluded from the vector comparison. However, for compatibility reasons the maximum TF–IDF weight is given:

$$t f i d f = \frac{n_{id}}{n_d} \log N$$

– Frequent Item Selection [185]

This weighting uses only the top $k$ TF–IDF values per image, called frequent items, to provide a compact image representation. The images are then ranked according to the number of common shared frequent items with the query image.

Finally, the indexer can create an Approximate Nearest Neighbour (ANN) index structure to facilitate fast retrieval. Currently, serial and parallel versions of Euclidean Locally Sensitive Hashing (E2LSH) [10] ANN method are supported. This algorithm uses families of hashing functions to partition the index feature space and thus limit the search into the subspace that a query falls into.

• Hadoop Indexer

The Hadoop [190] implementation of MapReduce was used for the parallelization of the indexing, since it is an easily parallelizable task. The pipeline is identical to the one shown in Figure 3.2 except for the fact that the blocks in the frame are executed in parallel. This is achieved by splitting the image collection into small groups of images. Each group is indexed by a different map task.

Either an in–house or a cloud Hadoop cluster can be used for this indexing method, since the implementation is fully parametrizable. An in–house visual cluster was created in the institute for the needs of the prototype, consisting of 13 workstations, 2 servers and 5 virtual machines. This resulted in a 20 node cluster with a computational capability of 99 concurrent map tasks (Figure 3.3) for the needs of the study. It should be noted that this architecture allows for easy expansion of the cluster. The cluster described in this study was initially
setup having 10 nodes and 42 concurrent map tasks and also grown after the conduction of the study to more than 120 tasks.

The background of the framework and the details of the implementation of the cluster are described in Section 3.2.5.

Once the index is stored, the index parameters are saved in JSON format in a configuration file. This way, the ParaDISE Seeker can use the same configuration for extracting the visual features of the query images when searching within the specific index.

**ParaDISE Seeker**

As mentioned in Section 3.1.2, the Seeker Composite Component is responsible for CBIR search in ParaDISE. As required by CBIR, the ParaDISE Seeker allows similarity search using image examples as queries. Multiple query images and negative examples are also supported. From the user side, this allows for iterating the search using relevance feedback [157]. The relevance feedback can be handled in various ways. In ParaDISE Seeker the following algorithms are supported for handling relevance feedback:

- **Rocchio Seeker**
  This Seeker uses the Rocchio algorithm [157] to handle multiple images of positive or negative relevance. The Rocchio formula is given by:

  \[
  \overline{q}_m = \alpha \overline{q}_o + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j
  \]  
  \[ (3.16) \]

  where \( \alpha, \beta \) and \( \gamma \) are weights,
  \( \overline{q}_m \) is the modified query,
  \( \overline{q}_o \) is the original query,
  \( D_r \) is the set of relevant images and
  \( D_{nr} \) is the set of non-relevant images.

  The search pipeline of this method is shown in Figure 3.4. The Seeker reads the index Parameters from the configuration file of the index it tries to access (see Section 3.1.2). According to these Parameters, it transforms the images to the appropriate vector representations. The Rocchio formula is then executed, producing a single merged vector. If an ANN index exists for the accessed visual index then a shortlist of the vectors existing in the same subspace as the merged vector is returned. In this case, Storer searches within the returned shortlist otherwise the whole index is searched. The similarity search uses a distance metric or a similarity measure to rank the images. The following distances/similarities are supported in ParaDISE:

  - **Euclidean distance** (L2 norm)
    \[
    d_e(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}
    \]  
    \[ (3.17) \]

  - **Manhattan distance** (L1 norm)
    \[
    d_{manhattan}(p, q) = \sum_{i=1}^{n} |p_i - q_i|
    \]  
    \[ (3.18) \]

  - **Canberra distance**
    \[
    d_{canberra}(p, q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}
    \]  
    \[ (3.19) \]
Figure 3.4: The search pipeline of the Rocchio Seeker.
34  

CHAPTER 3. SYSTEM DESIGN AND DEVELOPMENT

- $\chi^2$ distance

$$d_{\chi^2}(\vec{p}, \vec{q}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{p_i + q_i}$$  \hspace{1cm} (3.20)

- Jeffrey divergence

$$d_{jd}(\vec{p}, \vec{q}) = \sum_{i=1}^{n} \left( \log \frac{2p_i}{p_i + q_i} + \log \frac{2q_i}{p_i + q_i} \right)$$  \hspace{1cm} (3.21)

- Histogram intersection

$$s_{hi}(\vec{p}, \vec{q}) = \sum_{i=1}^{n} \min(p_i, q_i)$$  \hspace{1cm} (3.22)

- Cosine similarity

$$s_{\text{cosine}}(\vec{p}, \vec{q}) = \frac{\sum_{i=1}^{n} (p_i \times q_i)}{||\vec{p}|| \times ||\vec{q}||}$$  \hspace{1cm} (3.23)

where $\vec{p}, \vec{q} \in \mathbb{R}^n$. Also special similarity measures are supported for specific approaches:

- Hamming Distance
For binary vectors $\vec{p}, \vec{q}$, the hamming distance $d(\vec{p}, \vec{q})$ is defined as the number of ones of $p \oplus q$. It can be used for comparing binary representations, such as binary BoVW.

- Frequent Item Selection Distance
This similarity is used in combination with the Frequent Selection weighting (see Section 3.1.2). The similarity score is equal to the number of common shared frequent items.

- LateFusion Seeker
The pipeline of this Seeker is demonstrated in Figure 3.5. It is similar to Rocchio Seeker pipeline but instead of producing a single merged query vector it initiates a different search for each positive query image. In the end the Fusor ParaDISE Component is used to fuse the retrieved lists. Negative query image examples are ignored.

3.1.3 Applications

Apart from promoting research in CBIR, the ParaDISE was designed to be easy to integrate into real applications as well. This section describes in detail KHRESMOI Radiology (Section 3.1.3), which is the main system that uses ParaDISE as a backend component, as well as other innovative applications, such as the Shambala interface (Section 3.1.3) and SearchParaDISE (Section 3.1.3).

KHRESMOI Radiology

As mentioned in Chapter 1 KHRESMOI\footnote{http://www.khresmoi.eu/} is a project that aims at creating a multilingual, multi-modal search and access system. One of the main functionalities is to allow efficient access to the visual information available in electronic records and the open access medical literature on the Internet. Some of the techniques described in the indexing and retrieval pipelines below, have been developed and evaluated in the context of this thesis and will be described in Section 3.2.

The user interface of KHRESMOI Radiology is based on ezDL [21]. A more detailed description can be found in [20]. A screenshot of the main 2D image search interface is given in Figure 3.6. The basic elements are the Query View, the Results View and the Detail View. The user can use the Query view to add text or positive and negative image examples and initiate a search. Restricting the search with a specific image modality (or a group of modalities) is also supported.
3.1. THE PARALLEL DISTRIBUTED IMAGE SEARCH ENGINE

Figure 3.5: The search pipeline of the LateFusion Seeker.
Once a search has been initiated, the results are presented in the Result View in either ranked list or grid format. Results found in this list can be added in the query to initiate a new search iteration through relevance feedback. Filtering the results by modalities and media type is also supported.

By selecting a result, its associated information appears in the Details view. For articles this means the full title, the abstract and the images included. Search for similar images can be initiated from this view. For image results this means the full size image, the caption and link to the corresponding article. Basic image manipulation is available to allow for better image content inspection.

More tools, such as the Personal Library and collaborative tools are available and described in more details in [20].

The indexing and retrieval pipelines that are based on ParaDISE follow below. In Figure 3.7 the full pipeline of 2D image indexing is presented. In the beginning, the images are downloaded to the server for faster access and caption–image pairs are created. Lucene is used to index the captions of the images. An info table with the various image information, such as the corresponding article URL, the image URL and the caption of the image, is created during that step.

The next step is to classify the images according to their image modality. The compound figures are separated and their subfigures are saved as new images and are reclassified. The info table is then updated, including the modality information and the subfigure URLs. The method presented in [72] and described in Section 3.2.2 was used for the modality classification. The modality hierarchy used in ImageCLEF 2013 medical challenge was used. The method proposed in [38] and described in Section 3.2.3 was used for the compound figure separation. Hadoop was used for the parallelization of this task.

After a new image list is created with the inclusion of subfigures, ParaDISE undertakes the task of visual indexing. For the visual indexing, BoVW and BoC representations were used as shape and color features of the images. E2LSH was used as an ANN indexing method.

A new round of caption indexing is performed, this time on the subfigures captions. The compound figures are then removed from the indices.

A dataset of 1.2 million images from 500,000 articles of PubMed Central (accessed in 2013) has been indexed using this pipeline, resulting in 1.7 million images after subfigure indexation.

The development of a new Seeker was dictated by the requirements for the KHRESMOI system,
Figure 3.7: The KHRESMOI indexing pipeline of 2D images.
such as search by modality. The object-oriented implementation of ParaDISE facilitated this and ModalityFilter Seeker was created. This Seeker extends the Rocchio Seeker and accepts as inputs a list of modalities which uses to filter the results. The weights used for the Rocchio algorithm are $\beta = 0.6$ and $\gamma = 0.4$, chosen after empirical tests. Query and relevant vectors were considered as the same set of vectors.

The backend 2D image search pipeline is presented in Figure 3.8. Once the web service is called, the call arguments dictate the behavior of the workflow. Query Images can be automatically classified to produce a list of target modalities or specific target modalities can be passed as arguments. If text is included in the query then the text search pipeline is enabled (in the left frame). Image captions can also be used in relevance feedback iterations. RadLex terms can be extracted by the captions of the query images using the Ontotext disambiguation service [132] and can be added to the query string. Captions of negative query image examples have their terms (the ones not present in positive ones) added using the NOT boolean operator.

The next step is the visual similarity search. For each visual index needed to be accessed there is a concurrent search using modality filtering. The histogram Intersection similarity measure is used. If there is no text included in the query, the ANN index is used to build the shortlist to be searched. Otherwise, the top results returned by the text query constitute the shortlist for the visual search. The ParaDISE Fusor is then used to fuse the retrieved lists from the visual indices. The CombMNZ rule is used for this fusion as it is the best performing fusion rule in the empirical evaluation experiments (Chapter 4).

The next step consists of the fusion of the text and visual search results, using the Fusor and Reciprocal Rank fusion rule. Finally, image information existing in the info table is added to the results.

KHRESMOI Radiology 3D image search system

KHRESMOI Radiology is designed to search into both internal databases and external sources. In the previous section the external part, based on ParaDISE, is described in detail. The external part was integrated with the 3D image search and the complete system was evaluated in the user tests presented in Chapter 5. Thus, a brief description of the functionalities of the 3D prototype is presented here.

As it was mentioned before, the 3D Image Search prototype was responsible for search into internal image databases of hospitals. It uses automatic anatomy localization and search by ROIs to find volumes that contain similar pathologies with the marked ROI. The retrieved results can be viewed along with their associated case. RadLex terms are automatically identified and highlighted in the radiological reports. More information can be found in [123]. The user interface used for the 3D prototype consisted of two perspectives.

As starting point the (1) index perspective shown in Figure 3.9 displays all the available cases in the database. The user is able to filter by any field in order to search for specific subsets and it is also possible to filter the medical findings of the datasets. If a volume is selected by left clicking on it in the index view, its center slice is displayed on the right side together with the medical finding and marked ROIs (if any). Clicking on the slice initiates the loading of the volume which is then displayed on top of the current perspective (Figure 3.10), where the user can manipulate the volume by brightness, contrast, pan and zoom.

Adapting brightness and contrast is done by left clicking on the volume, which enables the so called windowing function, where the user can change the brightness by moving the mouse forward/backward and the contrast by moving the mouse left/right. By right clicking on a slice of the volume the ROI marking mode can be enabled. A search can be initiated by clicking on the search button.

The system now switches into the (2) result perspective, which is divided into three windows as shown in Figure 3.11. On the left the query volume is displayed, while in the center the result thumbnails of the best matching slice and report summary are shown. In the center window three additional tabs are created containing the results of the integrated 2D prototype (based on ParaDISE). These contain results matching a generated query string (visible in top left of the result perspective) in terms of article search, 2D image search and RadioWiki search. On the right
Figure 3.8: The KHRESMOI search pipeline for 2D image retrieval.
CHAPTER 3. SYSTEM DESIGN AND DEVELOPMENT

Figure 3.9: The index view of the 3D Image Search prototype.

Figure 3.10: The ROI marking (full-volume-view) overlayed over the index view of the 3D Image Search prototype.
3.1. THE PARALLEL DISTRIBUTED IMAGE SEARCH ENGINE

side the details of the selected result are presented. If a volume is selected in the center window, its thumbnail and medical finding are displayed in this window. The user can now load the full volume and the generated overlays by clicking on request images.

Clicking on the perspective menu in the top left allows the user to manually switch between the different perspectives.

Shambala

Whereas ezDL, used in the KHRESMOI system, is a more complex interface, Shambala is an alternative search interface mainly targeted towards simple navigation. It is built on 3 principles: it should be easy to use (not overwhelming the user with options and parameters), interactive (allow the user to drag&drop images, update search results in real–time) and modern (use of HTML5 and Leap Motion™ technology). Leap Motion allows the user to interact with the interface using only hand and finger gestures, without needing to use a mouse or keyboard.

Figure 3.12 depicts the user interface of Shambala (in this case on a small screen). The user interface is composed of 3 main areas in vertical direction. The central area presents the set of images returned by ParaDISE in a space–efficient layout. The left and right areas consist of the drop zones for keywords and images that are relevant and irrelevant respectively to the current image search. These keywords are used for multi–modal relevance feedback, described in Section 3.2.4. To create an intuitive and engaging interaction, a hand and finger motion sensor (Leap Motion(tm)) tracks the user gestures (Figure 3.12). Nevertheless, a keyboard is still used in the current version to enter keywords for relevant and irrelevant terms. Multi–modal relevance feedback makes the use of text possible without a keyboard. In theory, speech–to–text techniques could be used to make the need of a keyboard obsolete.

More details about gesture interaction with Shambala can be found in [192].

SearchParadise

Another application prototype which uses ParaDISE, called SearchParadise, was created for the Google Glass™, a wearable computing device developed by Google. The application allows

---

8http://leapmotion.com/
9http://www.google.com/glass/
CHAPTER 3. SYSTEM DESIGN AND DEVELOPMENT

Figure 3.12: The web-based interface of Shambala.

Figure 3.13: Gesture interaction with Shambala.
the user to perform a search query based on a photo taken by the Glass camera and see the results on the display located in front of the right eye (Figure 3.14. This can allow physicians to get additional information on situations or lesions that they have and does not interrupt the contact with the patient.

The scenario of the application is as follows:

1. The user starts the application by saying ”’ok, Glass, search ParaDISE”’.
2. Google Glass starts the application and activates the camera showing the preview in the prism.
3. When ready, the user can take a photo by tapping on the touchpad on the side of Glass.
4. Once the photo is taken, the user can add spoken keywords if he wishes.
5. The spoken keywords are transform to text and are transmitted to the ParaDISE webservice while the photo is uploaded to ParaDISE server.
6. The request string including keywords, the link to the uploaded image and the number of results (currently 10) is formed and sent to the ParaDISE service.
7. The ParaDISE service searches images that match visual features contained in the uploaded photo and images associated to captions including the keywords in the request.
8. The 10 most relevant results from ParaDISE are sent to Glass. Glass creates visual cards containing the result images with their captions. The user can navigate among them by swiping the touchpad forward or backward as illustrated in Fig. 3.15.

More details about the SearchParadise application can be found in [191].

### 3.2 Additional components

In this section, the techniques developed to facilitate the indexing and retrieval pipelines described in Section 3.1.3 are described.
3.2.1 Density–based visual vocabulary creation

For the creation of the visual vocabulary a common practice is to use K–means as clustering technique. This algorithm is an attractive solution as it uses a single parameter (K — the number of the clusters). However, it also suffers from several drawbacks. First, the divergence of the algorithm is not guaranteed as it can get trapped in local minima. It works well if the data set is a mixture of Gaussian distributions, which is difficult to decide in high dimensional spaces. Moreover, it is not robust to noise and outliers, nor can it detect arbitrarily shaped clusters. K–means has a time complexity of \( O(kNd) \) per iteration where \( N \) is the data set size and \( d \) its dimensionality. While this is acceptable for small vocabularies, the optimal vocabulary size can vary from \( O(10^3) \) to \( O(10^6) \) depending on the image data set. After taking these points into account another approach was followed for the clustering step in this study. A density–based clustering method called DENCLUE [88] was investigated, which is designed to deal with high dimensional data and large data sets. The main idea behind this method is to cluster the data using the local minima of a function that represents their density in the d–dimensional space. Before presenting the algorithm, some definitions of basic concepts are provided.

**Definition 1** A kernel function \( K : \mathbb{R}^d \rightarrow \mathbb{R}, K(x) \geq 0 \), which has
\[
\int_{\mathbb{R}^d} K(x)dx = 1
\]

The density function is defined as the sum of the kernels of all data points. Given \( N \) data objects described by a set of \( d \)–dimensional feature vectors \( D = x_1, \ldots, x_N \subset \mathbb{R}^d \) the density is defined as
\[
f^D(x) = \frac{1}{Nh^d} \sum_{i=1}^{N} K\left(\frac{1}{h}(x - x_i)\right)
\]

**Definition 2** A point \( x^* \in \mathbb{R}^d \) is called a density attractor of a density function \( f^D \), iff \( x^* \) is a local maximum of \( f^D \). A point \( x \in \mathbb{R}^d \) is density attracted to a density attractor \( x^* \) iff a hill–climbing procedure started at \( x \) converges to \( x^* \).

**Definition 3** The local density \( \hat{f}^D(x) \) is
\[
\hat{f}^D(x) = \frac{1}{Nh^d} \sum_{x' \in \text{near}(x)} K\left(\frac{1}{h}(x' - x)\right)
\]
where \( \text{near}(x) = x' : \text{dist}(x_1, x) \leq \delta_{\text{near}} \)

DENCLUE consists of two steps. The first step approximates the density function. This can be done by using a cubeMap data structure [88]. After the local density has been computed for each point, the points with \( \hat{f}^D < \xi \) where \( \xi \) is one of the parameters of the algorithm, are considered noise. For the remaining points, density attractors are determined using a hill–climbing procedure, guided by the local density gradient \( \nabla \hat{f}^D(x_i) \).

\[
x = x^0, x^{i+1} = x^i + \delta \frac{\nabla \hat{f}^D(x^i)}{||\nabla \hat{f}^D(x^i)||}
\]

where \( \delta \) is a parameter that controls the speed of the hill–climbing procedure. This procedure stops at \( k \in \mathbb{N} \) if \( \hat{f}^D(x^{k+1}) < \hat{f}^D(x^k) \) and takes \( x^* = x^k \) as a density attractor. When using averaged shifted histograms for the approximation of the local density [89], the time complexity is \( O(d^2N \log(N)) \) for the worst case scenario that every point is in a different hypercube. Note that its complexity is independent of the number of clusters, something that for large vocabularies is an important speed–up. The parameters \( h, \xi, \delta \) of the method were empirically chosen. However, the speed–up that DENCLUE provides allows even a grid search for the optimal parameters. Algorithms to detect the optimal parameters are proposed in [65, 87].

### 3.2.2 Modality classification

Journals in the medical literature contain images of various types depending on the scope of the journal and the subject of the article. Automatic categorization of image types can help in various ways when using large image data sets. Radiologists have a preference for searching for specific radiology modalities as it is often mentioned in Chapter 2. Moreover, filtering out graphs and diagrams when searching for images could improve the retrieval performance and efficiency, since the largest majority of radiology users only want to get diagnostic images.

The ImageCLEF2012 dataset that was used in the evaluation contains over 300,000 images of 75'000 articles of the biomedical open access literature. This is a subset from the PubMed Central database containing over one million images. This set of articles contains all articles in PubMed that are open access but the exact copyright for redistribution varies among the journals. Thus, the secondary use is permitted for this set but no distribution rights are given. 2000 labelled images were split in two sets of 1000 images that served as training and test sets. A more detailed description of the ImageCLEF2012 setup is given in [138].

In this dataset some image types in the training data were poorly represented (e.g. the training set containing very few images of those types) hurting the automatic classification accuracy. A way to overcome this problem is to expand the training set by adding (possibly) noisy data [96]. This is a semi–supervised learning [33] technique. Semi–supervised learning uses a small number of labelled instances and a large amount of unlabelled data for the training of the classifier. It is used in cases where labelled data are small and classes are under–presented in the training set. This scenario is often met in medical image analysis, where accurate labelling of big datasets is difficult and expensive to obtain.

Methods of semi–supervised learning have been applied to handwritten text recognition [29] and biological networks [200]. Related to this work, in [45] semi–supervised classification is applied in medical image classification to expand the training set. The confidence scores for the unlabelled data are given by Support Vector Machine (SVM) classifiers using multi–modal (visual and textual) information. Moreover, the expansion of the training set by visual retrieval is explored. The method proposed in this work uses multi–modal retrieval to expand the training set. This section discusses the details of the method implementation and its evaluation using the ImageCLEF2012 dataset. Results are compared to submitted runs, outperforming kNN classifiers that use the original training set and single–modal semi–supervised kNN classifiers, similar to [45].

This paragraph describes the general algorithm for the training set expansion. The ImageCLEF2012 modality classification training set will be denoted as the set of labelled examples.
In practice, because of the removal of instances contained in more than one classes, the size of the expanded training set is smaller than \( l + l \times k_r \).

Two methods of expanding the training set were examined. In the first expansion technique, \( s \) images taken randomly of each class were used as queries. As in the original training set the number of images per class varies from 5 to 50, by choosing \( s = 5 \), \( k_r = 20 \) we can theoretically obtain a relatively balanced training set (105–150 per class) of 4,100 images. In the second, all the training images were queried, resulting in a larger non–balanced training set. E. g. for \( k_r = 20 \) an expanded training set of 21,000 images can be obtained theoretically.

For a run to qualify as visual, we considered that the expanded training set used in this run needs to be created only by visual means. This means that the queries on the full data set used only visual features for the retrieval. Similarly, this was repeated using mixed (visual and textual) queries for the mixed runs. This resulted in a final number of 5 training sets (2 balanced, 2 non–balanced and the original training set).

A kNN classifier using weighted voting was used to classify the test images. For the choice of the classifier parameters the results of [72] were taken into account and \( k = 11, k = 7 \) were used for the visual runs. However, since the non–balanced expanded training set was significantly larger, double the value \( k = 14 \) was also tested for this case. The inverse of the similarity score of the \( k \)–nn images was used to weight the voting.

### 3.2.3 Compound figure separation

Compound figures (figures consisting of several subfigures, as shown in Figure 3.16) constitute a very large portion (between 40% and 60%) of the images found in the biomedical literature, according to [138]. When data of articles are made available digitally, often the compound figures are not separated but made available in a single block. This can be problematic, since subfigures of different imaging modalities can be grouped together (for example a graph and a Computed Tomography image), making an accurate description of the image content more difficult. Information retrieval systems for images should be capable of distinguishing the parts of compound figures that are relevant to a given query. Compound figure separation is therefore a required first step to retrieving focused figures. Several approaches have been published for separating figures.
3.2. ADDITIONAL COMPONENTS

from text in scanned documents [40] and more specifically, for separating compound figures in the biomedical literature [13, 37, 38]. A task to compare compound figure separation approaches was run in the ImageCLEF benchmark [69].

Separating a compound figure into its subparts was achieved by using the technique originally described in [38]. To summarise, the technique is based on the systematic detection and analysis of uniform space gaps. A recursive algorithm is used in order to determine a series of horizontal and vertical separator lines inside an image, allowing the extraction of the subfigures delimited by these separators.

Once the separation of all the compound figures was performed and all the necessary information (visual descriptors of the children, modality of the children, links between children and parents) was stored, the next step is to fully take advantage of this extra information.

A first step is to provide a visual distinction between results originating from separated figures and standard results. In this way, the user knows that the image exists in a context which can be interesting to him. Figure 3.17 shows a prototype where a coloured border and an icon are added to separated results.

The second step is displaying the image in its original context, preferably while highlighting the selected child in the parent image. Thanks to the information stored in CouchDB, we have access to the coordinates of the child, making this easily feasible. Figure 3.18 shows a prototype where a given subfigure is highlighted in its corresponding parent image.
Other navigation possibilities can be imagined, such as switching back and forth between a parent image and a more detailed view of its children.

### 3.2.4 Multi-modal relevance feedback

Relevance feedback allows the user to mark results returned in a previous search step as relevant or irrelevant to refine the initial query. The concept behind relevance feedback is that though users may have difficulties in formulating a precise query for a specific task, they generally see quickly whether a returned result is relevant to the information need or not. This technique was used in image retrieval particularly with the emergence of CBIR systems [173, 178, 194]. Following the CBIR mentality, the visual content of the marked results is used to refine the initial image query. With the result images represented as a grid of thumbnails, relevance feedback can be applied quickly to speed up the search iterations and refine results. Recent user-tests with radiologists on a medical image search system also showed that this method is intuitive and straightforward to learn [122].

Depending on whether the user manually provides the feedback to the system (e.g. by marking results) or the system obtains this information automatically (e.g. by log analysis) relevance feedback can be categorized as explicit or implicit. Moreover, the information obtained by relevance feedback can be used to affect the general behaviour of the system (long-term learning). In [143] a market basket analysis algorithm is applied in image retrieval for long-term learning. A recent review of short-term and long-term learning relevance feedback techniques in CBIR can be found in [112]. An extensive survey of relevance feedback in text-based retrieval systems is presented in [160] and for CBIR in [159].

In the medical informatics field, [36] applies CBIR with relevance feedback on mammography retrieval. In [152], an image retrieval framework using relevance feedback is evaluated on a dataset of 5000 medical images that uses support vector machines to compute the refined queries.

In this study various explicit, short-term relevance feedback techniques using visual content or text for medical image retrieval have been evaluated. A technique that combines visual and text-based relevance feedback is proposed that achieves a competitive performance to the state-of-the-art approaches.

One of the most well known relevance feedback techniques is Rocchio’s algorithm [157]. Its mathematical definition is given in Section 3.1.2: Rocchio’s algorithm is typically used in vector models and also for CBIR. Intuitively, the original query vector is moved towards the relevant vectors and away from the irrelevant ones. By giving a weight to the positive and negative parts a problem of CBIR can be avoided that when more negative than positive feedback exists that also many relevant images disappear from the results set.

Another technique that showed potential in image retrieval [73] is late fusion. Late fusion [55] is used in information retrieval to combine result lists. It can be applied for fusing multiple features, multiple queries and in multi-modal techniques. The concept behind this method is to merge the result lists into a single list while boosting common occurrences using a fusion rule.

Most of the techniques use vectors either from the text or the visual models. However, it has been shown that approaches that use both text and visual information can outperform single-
3.2. ADDITIONAL COMPONENTS

3.2. ADDITIONAL COMPONENTS

modal ones in image retrieval. In this thesis the use of multi–modal information when applying
relevance feedback is proposed to enhance the retrieval performance. This is, to the extend of our
knowledge, the first time that such a technique is proposed in medical image retrieval. As late
fusion is applied on result lists, it is straightforward to use for combining results from visual and
text queries.

3.2.5 Parallel image indexing

As seen in Section 3.1.2 parallel computing was used for the visual indexing of the dataset.
This is necessary for regularly updating the indices as the extraction of visual features and the
representation of an image is a computational intensive task. For example, an average time for the
extraction of SIFT local features is around 3.5 seconds. For indexing serially the available datasets
from the medical literature that amount to millions of images, this would mean more than 7 weeks
of indexing time.

In order to cope with these potential bottlenecks to large–scale image analysis and data flexible
and scalable infrastructures that are able to cope with the exponential growth of visual data need
to be used. Several approaches exist:

- Single Host (e.g., a server with 12 physical processor cores with hyper–threading\textsuperscript{11} & 96 GB
  of RAM (Random Access Memory)),

- Small local cluster (e.g. 8 hosts with 2–4 physical cores with hyper–threading and 16 GB of
  RAM, plus the above–mentioned server),

- Alternative infrastructures, such as GPUs (Graphical Processing Units),

- Cloud computing infrastructures, such as Amazon’s Elastic Cloud Compute (EC2).

In the context of this study, the first two solutions are tested and compared, in order to determine
the best possible option for a given analysis task. A brief presentation of the related work in
parallel computing in image retrieval is given in the following paragraphs.

Due to the growth in the data size and the development of new computationally intensive
algorithms, research has been performed in creating parallel processing algorithms \cite{75} and de-
veloping cloud computing systems \cite{28,156}. MapReduce \cite{52} proposed initially by Google\textsuperscript{12}
has become one of the most popular distributed computing frameworks, due to its simplicity in setup
and programming. Hadoop \cite{190} is a very popular open source implementation of MapReduce
with a large community of users. Although being a powerful computational tool, MapReduce
should not be seen as a “one–fits–all” solution \cite{111}. For example, as demonstrated in \cite{115} on
a data warehousing use–case using astrophysical data sets MapReduce is outperformed by database
management systems (DBMS). In \cite{175}, it is stated that MapReduce should be seen as an “extract–
transform–load” f(ETL) tool and complement DBMS in tasks that require both data warehousing
and intensive processing.

In the field of image processing, apart from cloud computing, parallelization using GPUs \cite{149}
is often used. However, as stressed in \cite{149}, GPU hardware architectures differ much form other
architectures and should be taken into consideration when designing parallel processing algorithms.
MapReduce has recently been used for large–scale image annotation \cite{7,113} and efficient image
description and analysis \cite{189}. In \cite{7} a parallel SVM \cite{184} algorithm is proposed for automatic
image annotation, while in \cite{113} a dataset of 30 million images are automatically labelled using
the MapReduce framework, although no details on the parallel implementation are given. Several
image analysis algorithms (e.g., image feature extraction, local descriptors clustering and image
registration) adaptation to the MapReduce framework is discussed in \cite{189}. Content–based image
retrieval is another field that combines large amounts of data with computationally intensive tasks.

\textsuperscript{11}Hyper-threading: technology from Intel\textsuperscript{R} which makes each physical processor core appear to the
Operation System as 2 virtual cores, which improves the execution of multi–threaded code. See
http://en.wikipedia.org/wiki/Hyperthreading (as of 29 August 2014) for more information.

\textsuperscript{12}http://research.google.com/archive/mapreduce.html/
For these reasons, indexing large image datasets using visual features is expected to be a well–suited task for the MapReduce framework. However, a few studies have also used MapReduce for the online part of the retrieval [119, 198, 202]. To overcome its inherent limitations, data warehousing and storage tools like Hive [180] and HBase (an implementation of Google’s BigTable abstraction [31]) that are built on top of Hadoop are often used. In [198] a CBIR system, NIR, on top of Hadoop, Nutch [101] and LIRe (Lucene Image Retrieval) [119] is presented. However, a very small data set is used for evaluation of the retrieval time. Another system called Distributed Image Retrieval System (DIRS) based on LIRe and HBase is described in [202]. Data sets of up to 100,000 images are used for testing the query times. When using datasets above 20,000 images, the retrieval times reported are restrictive for online use even though they are faster than without Hadoop use. Other approaches to deal with these challenges have included implementing efficient indexing schemes such as an inverted index [116] or locality sensitive hashing [176] on top of the Hadoop file system (HDFS) [167]. While [176] reports promising results our belief is in accordance with [111, 175] that for online tasks parallel DBMS should be the first choice.

In the medical field, cloud computing is also starting to find use. CCMedII [91] is a proposed medical information file exchanging and sharing system built on top of Hadoop. Medical Image File Access System (MIFAS) [196] is an access system for images using HDFS. MapReduce has also been used in fluorescence image analysis tasks in [201] and as a framework of the Hadoop–GIS query system in [5] for analytical pathology imaging. These systems use HBase and Hive for data storage and warehousing respectively. Moreover, Hadoop has been used for anatomical landmark detection [171] and medical image registration [102].

In this study, the MapReduce framework was used to speed up SIFT feature extraction [117] and Bag–of–Visual–Words indexing of large image datasets. A parallel computing environment was set up in our network using Hadoop13, an open–source implementation of Google’s MapReduce framework. Detailed descriptions of MapReduce and Hadoop are provided in Sections 3.2.5 and 3.2.5. The "real–world" application mentioned above was converted to Hadoop programs, allowing performance testing and comparison as well as identification of bottlenecks and other problems.

**MapReduce**

MapReduce is a programming model and an associated implementation developed by Google for processing large datasets. Typically the computation runs in parallel on a cluster of machines of up to several thousand nodes (usually commodity personal computers) in order to finish the processing task in a reasonable amount of time. Users define the required computation, which must be embarrassingly parallel, in terms of a map and a reduce function with the underlying system automatically distributing it across the cluster. The system itself manages machine failures and inter–machine communication to ensure efficiency of the network and disks. This approach is a reaction to code complexity by hiding the messy details of fault tolerance, data distribution and load balancing in a library. The computation can as well be run across multiple cores of the same machine(s) [52, 190].

Google’s implementation of MapReduce runs on top of the Google File System (GFS), a scalable distributed file system for large data–intensive–applications providing fault tolerance. Files on GFS are typically split into chunks of 64 MB and distributed across chunkservers with a default replication factor of three [76]. MapReduce cannot solve every problem, but being a general data–processing tool there is a wide range of algorithms that can be expressed such as machine learning algorithms, graph–based problems and image analysis [190]. A typical MapReduce program is split into a Map phase and into a Reduce phase. The Map function has a key/value pair as input and produces a set of intermediate key/value pairs as output:

\[(k_1, v_1) \rightarrow \text{list}(k_2, v_2).\]  

(3.24)

After the map phase the MapReduce library groups all intermediate values for the same intermediate key I. The Reduce function accepts an intermediate key I with its set of values from the map

13http://hadoop.apache.org/
output (supplied via an iterator) as input and merges these values together to produce a possibly smaller set of values as output. Normally the number of output values per reduce invocation is zero or one:

\[(k_2, \text{list}(v_2)) \rightarrow \text{list}(v_2).\] (3.25)

It is worth mentioning that the map input keys and values are related to a different domain than the keys and values of the intermediate and reduce output [52].

**Hadoop**

Hadoop was created by Doug Cutting, the creator of Apache Lucene\(^{14}\). The origins of Hadoop are found in Nutch\(^{15}\) (Lucene subproject), an open source web search engine supposed to scale to billions of pages. However, realizing that it was not possible with their architecture at that time, Cutting and his partner Mike Cafarella, inspired by the publication of the GFS paper in 2003 [76], decided to write an open source implementation named Nutch Distributed File System (NDFS)\(^{16}\). In early 2005, after the publication of the Google paper that introduced MapReduce to the world in 2004 [51], the Nutch developers presented a working implementation of MapReduce. As NDFS and the MapReduce implementation in Nutch were considered as potentially useful to a broader field of application they were moved out of Nutch and became an independent subproject of Lucene called Hadoop. Shortly later, Cutting joined Yahoo!\(^{17}\), which provided a dedicated team just for the extension of Hadoop. This makes Yahoo! the largest contributor of Hadoop. Confirmed by its success, Hadoop turned into an own top–level project at Apache in 2008 [190]. Since then, various large companies such as Amazon\(^{18}\), Facebook\(^{19}\), Microsoft\(^{20}\) have started using Hadoop [1].

The Apache Hadoop Common library is written in Java and consists of two main components: the MapReduce framework and HDFS\(^ {21}\), which implements a single–writer, multiple reader model [2, 167]. However, Hadoop does not solely support HDFS as an underlying file system. It also provides a general–purpose file system abstraction making it possible to integrate other storage systems, such as Amazon S3\(^ {22}\), which targets the Amazon Elastic Compute Cloud\(^ {23}\) server–on–demand infrastructure. In our own Hadoop environment, we exclusively make use of HDFS as file system. Currently, the Linux operating system is the only officially supported Hadoop production platform [3, 190].

The purpose of HDFS is to store large datasets reliably and to stream them at high bandwidth to user applications. HDFS has two types of nodes in the schema of a master–worker pattern: a namenode, the master and an arbitrary number of datanodes, the workers [190]. The HDFS namespace is a hierarchy of files and directories with associated metadata represented on the namenode. The actual file content is split into blocks of typically 64MB where each block is typically replicated on three namenodes. The namenode keeps track of the namespace tree and the mapping of file blocks to datanodes. An HDFS client wanting to read a file has to contact the namenode for the locations of data blocks and then reads the blocks from the closest datanode since HDFS considers short distance between nodes as higher bandwidth between them. In order to keep track of the distances between datanodes HDFS supports rack–awareness. As soon as a datanode registers with the namenode, the namenode runs a user–configured script to decide which rack (network switch) the node belongs to. Rack–awareness also allows HDFS to have a block placement policy that provides a trade–off between minimizing write cost and maximizing data reliability, availability and aggregate read bandwidth. For the creation of a new block, HDFS places the first replica on the datanode hosting the writer and the second and third replicas on two different datanodes located in a different rack [167].

\(^{14}\)http://lucene.apache.org/
\(^{15}\)http://nutch.apache.org/
\(^{16}\)http://wiki.apache.org/nutch/NutchDistributedFileSystem
\(^{17}\)http://www.yahoo.com/
\(^{18}\)http://www.amazon.com/
\(^{19}\)http://www.facebook.com/
\(^{20}\)http://www.microsoft.com/
\(^{21}\)http://hadoop.apache.org/hdfs/
\(^{22}\)http://aws.amazon.com/s3/
\(^{23}\)http://aws.amazon.com/ec2/
A Hadoop MapReduce job, a unit of work that the client wants to be performed, consists of the input data (located on the HDFS), the MapReduce program and configuration information. Native Hadoop MapReduce programs are written in Java, however Hadoop also provides the Hadoop Streaming API which allows writing map and reduce functions in languages other than Java by using Unix standard streams as the interface between Hadoop and the user program. In Hadoop there are two types of nodes that control the job execution process: one job tracker, and an arbitrary number of task trackers. The job tracker coordinates a job run on the system by dividing it into smaller tasks to run them on different task trackers, which in turn transmit reports to the job tracker. In case a task fails, the job tracker is able to automatically reschedule the task on a different available task tracker. In order to have a task tracker run a map task the input data needs to be split into fixed-size pieces. Hadoop runs one map task for each split with the user-defined map function processing each record in the split. As soon as a map task is accomplished its intermediary output is written to the local disk. After that the map output of each map task is processed by the user-defined reduce function on the reducer. The number of map tasks running in parallel on one node is user-configurable and heavily dependent on the capability of the machine itself, whereas the number of reduce tasks is specified independently and is therefore not regulated by the size of the input. In case there are multiple reducers, one partition per reducer is created from the map output. Depending on the task to accomplish, it is as well possible to have zero reduce tasks in case no reduction is desired [190].
Chapter 4

Empirical evaluation

In this Chapter the methods that are described in Chapter 3 are empirically evaluated. The evaluation focus on the core aspect of image description, and the capability of different approaches regarding most important aspects of the indexing and retrieval pipelines. Mean average precision (mAP) is used as an evaluation measure for retrieval tasks while accuracy is used for classification tasks. Time in seconds is used for measuring time efficiency for indexing and retrieval.

4.1 Datasets

The datasets from the ImageCLEF initiative\(^1\) were used for the evaluation of image retrieval, modality classification and compound figure separation. Specifically the ImageCLEF2012 and ImageCLEF2013 datasets were used. They are based on the image collection of 300,000 images and 75,000 articles of the biomedical open access literature. It is considered as a realistic dataset of medical visual data available in the medical literature as it is a subset of PubMed Central\(^2\) containing over 1.5 million images and increasing daily. The distributed PubMed subset contains only articles allowing redistribution.

For the modality classification task the hierarchy depicted in Figure 4.1 was used. It contains 2000 labelled images out of which 1000 are used as training set and 1000 as test set. As discussed in Section 3.2.2 the distribution of the samples among the classes are quite uneven, making learning of distinctive classifiers a challenging task. The sample images presented in Figure 4.2 demonstrate the visual diversity of the classes of the data set.

For the compound figure separation a subset of 2982 manually classified compound figures was used. These images were extracted randomly from the ImageCLEF database by manually selecting just the compound figures. The ground truth was constructed in a database for several subclasses of images (e.g. 2x1, 1x2, etc). It was found that uniform–space–separated compound figures form a major part of compound figure images, making up a vast majority of the database used in this work. However, there are classes of compound images that are either separated by rectangular boxes or are without any actual separation line or space, which are not considered in this study.

4.2 Local Features

An evaluation on how well local visual features (commonly–used in object recognition and scene classification) perform in medical image retrieval was run using ParaDISE. A subset of ImageCLEF2012 of 10,000 images was used for this purpose. First the local features’ retrieval performance was evaluated using the BoVW representation for 4 distance measures (histogram

\(^1\)http://www.imageclef.org/
\(^2\)http://www.ncbi.nlm.nih.gov/pmc/
Figure 4.1: Modality hierarchy of the ImageCLEF 2012 modality classification task.

(a) Graph  (b) Digital light microscopy  (c) Endoscopy

(d) Non-clinical Photo  (e) Radiological image with combined modalities (PET–CT)

Figure 4.2: Sample images from ImageCLEF2012 medical data set
Table 4.1: Average mAP scores of BoVW representations of local features over six vocabularies for different distance measures.

<table>
<thead>
<tr>
<th>Run</th>
<th>Histogram intersection</th>
<th>Euclidean distance</th>
<th>Cosine Similarity</th>
<th>$\chi^2$ distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>0.0225</td>
<td>0.01087</td>
<td>0.0194</td>
<td>0.0129</td>
</tr>
<tr>
<td>SURF</td>
<td>0.0124</td>
<td>0.0062</td>
<td>0.0081</td>
<td>0.0078</td>
</tr>
<tr>
<td>RootSIFT</td>
<td>0.0207</td>
<td>0.0118</td>
<td>0.0204</td>
<td>0.0140</td>
</tr>
<tr>
<td>Lab</td>
<td>0.0135</td>
<td>0.0127</td>
<td>0.0134</td>
<td>0.0107</td>
</tr>
</tbody>
</table>

Table 4.2: mAP scores of the fusion of best-performing runs: SIFT ($k = 20$), SURF ($k = 40$), RootSIFT ($k = 30$) Lab ($k = 100$).

<table>
<thead>
<tr>
<th>fusion Rule</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CombMNZ</td>
<td>0.0223</td>
</tr>
<tr>
<td>CombSUM</td>
<td>0.0216</td>
</tr>
<tr>
<td>Reciprocal rank</td>
<td>0.0206</td>
</tr>
<tr>
<td>Borda count</td>
<td>0.0198</td>
</tr>
</tbody>
</table>

intersection, euclidean distance, cosine similarity and $\chi^2$ distance and 6 vocabularies of different sizes (10, 20, 30, 40, 50, 100). The BoVW vectors were $l2$ normalized and the Rocchio Seeker was used for the fusion of multiple query images. The average mAP over the 6 vocabularies is given in Table 4.1.

The best performing runs were combined using 4 fusion rules (CombMNZ, CombSUM, Reciprocal rank fusion and Borda Count) to investigate if they contain complementary information (Table 4.2). The histogram intersection was used for the similarity comparison.

The features were also assessed in 4 different visual vocabulary-based image representations (BoVW, VLAD, SPM and GridBoVW) using the histogram intersection similarity measure (except for the VLAD representation that can have negative values, so cosine similarity was used) (Table 4.3). Small-sized vocabularies were chosen as the dimensionality of VLAD is $k \times d$ where $k$ is the number of clusters and $d$ the dimensionality of the feature, so larger vocabularies would result to representations of dimensionality inefficient for quick retrieval.

The best performing local feature is SIFT using all of the distance measures, except cosine similarity where RootSIFT performed slightly better (Table 4.1). It can be seen that the distance metric is very crucial for the retrieval performance. Similarity measures perform better, with histogram intersection achieve the best results in all the local features. The fusion of the best performing runs is not providing better results than the best performing local feature (SIFT) (Table 4.2). This indicates that the evaluated features model the same visual information.

Regarding the local feature representations, SPM appears to enhance the BOVW representation, modelling the spacial information (Table 4.3). Grid spatial modelling degrades performance of BOVW for all features except Lab. VLAD achieves the worst overall performance, however it is mainly caused by the fact that cosine similarity had to be used instead of histogram intersection.

Table 4.3: Average mAP scores of local features for different image representations.

<table>
<thead>
<tr>
<th>Run</th>
<th>BoVW</th>
<th>SPM BoVW</th>
<th>Grid BoVW</th>
<th>VLAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>0.0225</td>
<td><strong>0.0227</strong></td>
<td>0.0166</td>
<td>0.0181</td>
</tr>
<tr>
<td>SURF</td>
<td>0.0124</td>
<td>0.0123</td>
<td>0.0102</td>
<td>0.0081</td>
</tr>
<tr>
<td>RootSIFT</td>
<td>0.0207</td>
<td>0.0214</td>
<td>0.0158</td>
<td>0.0151</td>
</tr>
<tr>
<td>Lab</td>
<td>0.0135</td>
<td>0.0157</td>
<td>0.0155</td>
<td>0.0050</td>
</tr>
</tbody>
</table>
Table 4.4: mAP scores of image representations for different distance measures.

<table>
<thead>
<tr>
<th>Run</th>
<th>histogram intersection</th>
<th>Euclidean distance</th>
<th>Cosine Similarity</th>
<th>(\chi^2) distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoVW SIFT k20</td>
<td>0.0268</td>
<td>0.0107</td>
<td>0.0208</td>
<td>0.013</td>
</tr>
<tr>
<td>SPM BoVW SIFT k40</td>
<td>0.0245</td>
<td>0.0122</td>
<td>0.0109</td>
<td>0.0124</td>
</tr>
<tr>
<td>CEDD</td>
<td>0.0216</td>
<td>0.010</td>
<td>0.020</td>
<td>0.0073</td>
</tr>
<tr>
<td>FCTH</td>
<td>0.0218</td>
<td>0.0095</td>
<td>0.0207</td>
<td>0.009</td>
</tr>
<tr>
<td>Fuzzy Color histogram</td>
<td>0.0144</td>
<td>0.0034</td>
<td>0.0152</td>
<td>0.0032</td>
</tr>
<tr>
<td>Color Layout</td>
<td>0.0189</td>
<td>0.0134</td>
<td>0.018</td>
<td>0.0093</td>
</tr>
<tr>
<td>ColorHoG</td>
<td>0.0051</td>
<td>0.0063</td>
<td>0.005</td>
<td>0.0046</td>
</tr>
<tr>
<td>GIST</td>
<td>0.0097</td>
<td>0.0014</td>
<td>0.0068</td>
<td>0.0019</td>
</tr>
<tr>
<td>CombMNZ of 5 best</td>
<td>\textbf{0.0296}</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

4.3 Global image descriptors

For the image representations evaluation, 8 descriptors were used. The two best performing aggregated local feature representations (Section 4.2), two global multi–feature descriptors (CEDD, FCTH), two color descriptors (Color layout, Fuzzy color histogram) and two miniature–based descriptors (ColorHoG, GIST). The results over 4 different distance measures are presented in Table 4.4. The 5 best performing runs (BoVW, SPM BoVW, CEDD, FCTH and Color layout) were combined using CombMNZ to investigate if they contain complementary information. histogram intersection was used for this run.

Judging from results of Table 4.4 the local feature aggregated vectors (BoVW and SPM BoVW using SIFT) achieve the best performance. Multi–feature descriptors (CEDD and FCTH) come second in performance with Color layout descriptor being the best color histogram. The miniature–based representations seem to have less consistent mAP even though they perform very well in certain topics. The distance measure is again shown to be very important in terms of retrieval performance, with histogram intersection and cosine similarity outperforming the Euclidean and \(\chi^2\) distance. The fusion of the best performing runs achieves the highest mAP, indicating this way that the features are complementary.

4.4 Relevance feedback

The image dataset, topics and ground truth of ImageCLEF 2012 medical image retrieval task [138] were used in this evaluation. The image captions were accessed by the text–based runs and indexed with the ParaDISE text search component, based on Lucene. The vector space model was used along with tokenization, stopword removal, stemming and Inverse document frequency–Term frequency weighting. The Bag–of–visual–words model described in [71] and the Bag–of–Colors model appearing in [72] were used for the visual modelling of the images. In multi–modal runs, the fusion of the visual and text information is performed only for the text 1000 top results as in the evaluation of ImageCLEF only the top 1000 documents are taken into account in any case.

For evaluating the relevance feedback techniques the following experimental setup was followed: The \(n\) search iterations were initiated with a text query in iteration 0. The relevant results from the top \(k\) results of iteration \(i\) were used in the relevance feedback formulae of the iteration \(i + 1\) for \(i = 0 \ldots n – 2\).

Five techniques were evaluated in this study:

1. \textbf{text}: text–based RF using the vector space model. Word stemming, tokenization and stop word removal is performed in both text and multi–modal runs.
4.4. RELEVANCE FEEDBACK

2. **visual_rocchio**: visual RF using Rocchio to fuse the relevant image vectors and CombMNZ fusion to fuse the original query’s results with the visual ones.

3. **visual_lf**: visual RF using late fusion (and the CombMNZ fusion rule) to fuse the relevant image results and the original query results with the visual ones.

4. **mixed_rocchio**: multi-modal RF using Rocchio to fuse the relevant image vectors and CombMNZ fusion to fuse the original query results with the relevant caption results and relevant visual results.

5. **mixed_lf**: multi-modal RF using late fusion (and the CombMNZ fusion rule) to fuse the relevant image results and the original query results with the captions’ results and relevant visual results.

The evaluation of the five techniques was performed for $k = 5, 20, 50, 100$ and $n = 5$. Results of the mAP of each technique per iteration are shown in Figures 4.3, 4.4, 4.5, 4.6.

Table 4.5 gives the best mAP scores of each run. The numbers in parentheses are the number of the iteration when this score was achieved. For scores that were the same in multiple iterations of the same run, the iteration closer to the first is used.

![Figure 4.3](image-url): Mean average precision per search iteration for result list size $k = 5$.

![Figure 4.4](image-url): Mean average precision per search iteration for result list size $k = 20$.

All of the evaluated techniques improve retrieval after the initial search iteration. This demonstrates the potential of relevance feedback for refining medical image search queries.

Relevance feedback using only visual appearance models, even though improving the retrieval performance after the first iteration, performed worse than the text–based runs in most cases. Visual features still suffer from the semantic gap between the expressiveness of visual features and our human interpretation. Still, this shows their usefulness in image datasets where no or little text meta–data are available. Moreover, when combined with the text information in the proposed method, they improve the text–only baseline.
Figure 4.5: Mean average precision per search iteration for result list size $k = 50$.

Figure 4.6: Mean average precision per search iteration for result list size $k = 100$.

The proposed multi-modal runs provide the best results in all the cases except for case $k = 5$. Surprisingly, the visual runs perform slightly better than the text and the multi-modal approaches for this case. However, assuming independent and normal distributed average precision values the significance tests show that the difference is not statistically significant.

We consider the case $k = 20$ as the most realistic scenario since users do not often inspect more than 2 pages of results. Especially for grid-like result interface views, where each page can contain 20 to 50 results, we consider $k = 20$ more realistic than $k = 5$. In this case the proposed methods achieve the best performance with 0.2606 and 0.2635 respectively. Again, the significance tests do not find any significance difference between the three best approaches. However, applying different fusion rules for combining visual and text information (such as linear-weighting) could further improve the results of the mixed approaches.

It can be noted that as the $k$ increases, the performance improvement also increases, highlighting the added value of relevance feedback. Larger values of $k$ were not explored as this scenario was judged as unrealistic.

In the visual runs using Rocchio for combining the visual queries is performing worse than late fusion. This comes in accordance with the findings in [71]. The reason behind this could be that the large visual diversity of relevant images in medicine and the curse of dimensionality cause the

<p>| Table 4.5: Best mAP scores of RF techniques. |</p>
<table>
<thead>
<tr>
<th>Run</th>
<th>$k = 5$</th>
<th>$k = 20$</th>
<th>$k = 50$</th>
<th>$k = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>0.197 (1)</td>
<td>0.2544 (4)</td>
<td>0.3107 (3)</td>
<td>0.3349 (4)</td>
</tr>
<tr>
<td>visual_if</td>
<td>0.2099 (2)</td>
<td>0.2243 (3)</td>
<td>0.2405 (4)</td>
<td>0.2553 (3)</td>
</tr>
<tr>
<td>visual_roc</td>
<td>0.2096 (2)</td>
<td>0.2187 (2)</td>
<td>0.2249 (3)</td>
<td>0.2268 (2)</td>
</tr>
<tr>
<td>mixed_if</td>
<td>0.1971 (3)</td>
<td>0.2606 (4)</td>
<td>0.3079 (4)</td>
<td>0.3487 (3)</td>
</tr>
<tr>
<td>mixed_roc</td>
<td>0.1947 (1)</td>
<td>0.2635 (4)</td>
<td>0.3207 (4)</td>
<td>0.3466 (4)</td>
</tr>
</tbody>
</table>
4.5. MODALITY CLASSIFICATION

Table 4.6: Modality classification runs.

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Techniques</th>
<th>fusion Rule</th>
<th>Training Set</th>
<th>k</th>
<th>Run Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>mc1</td>
<td>BoVW</td>
<td>n/a</td>
<td>original</td>
<td>11</td>
<td>Visual</td>
</tr>
<tr>
<td>mc2</td>
<td>BoVW + BoC</td>
<td>combMNZ</td>
<td>original</td>
<td>7</td>
<td>Visual</td>
</tr>
<tr>
<td>mc3</td>
<td>BoVW + BoC</td>
<td>combMNZ</td>
<td>visual non-balanced</td>
<td>7</td>
<td>Visual</td>
</tr>
<tr>
<td>mc4</td>
<td>BoVW + BoC</td>
<td>combMNZ</td>
<td>visual non-balanced</td>
<td>14</td>
<td>Visual</td>
</tr>
<tr>
<td>mc5</td>
<td>BoVW + BoC</td>
<td>combMNZ</td>
<td>visual balanced</td>
<td>7</td>
<td>Visual</td>
</tr>
<tr>
<td>mc6</td>
<td>BoVW + BoC + Captions</td>
<td>Reciprocal</td>
<td>original</td>
<td>7</td>
<td>Mixed</td>
</tr>
<tr>
<td>mc7</td>
<td>BoVW + BoC + Captions</td>
<td>Reciprocal</td>
<td>mixed non-balanced</td>
<td>7</td>
<td>Mixed</td>
</tr>
<tr>
<td>mc8</td>
<td>BoVW + BoC + Captions</td>
<td>Reciprocal</td>
<td>mixed non-balanced</td>
<td>14</td>
<td>Mixed</td>
</tr>
<tr>
<td>mc9</td>
<td>BoVW + BoC + Captions</td>
<td>Reciprocal</td>
<td>mixed balanced</td>
<td>7</td>
<td>Mixed</td>
</tr>
</tbody>
</table>

Table 4.7: Modality classification results.

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Visual</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>mc1</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>mc2</td>
<td>38.1</td>
</tr>
<tr>
<td></td>
<td>mc3</td>
<td>41.8</td>
</tr>
<tr>
<td></td>
<td>mc4</td>
<td>42.2</td>
</tr>
<tr>
<td>best run</td>
<td>mc5</td>
<td>69.7</td>
</tr>
</tbody>
</table>

modified vector to behave as an outlier in the high dimensional visual feature space. In the mixed runs the difference between the two methods is not statistically significant with Rocchio performing slightly better than the late fusion.

Irrelevant results were ignored, as they often have little or no impact on the retrieval performance [141, 161]. More importantly, the ground truth of the dataset used contains a much larger portion of annotated irrelevant results than relevant ones. This was considered to potentially simulate an unrealistic scenario, as users do not usually mark many results as negative examples. Having too many negative examples could also cause the modified vector to follow an outlier behaviour. Preliminary results confirmed this hypothesis, where the use of negative results for relevance feedback can decrease performance after the first iteration.

It should be noted that this is an automated relevance feedback experiment of positive only feedback and that in selective relevance feedback situations the retrieval performance is expected to perform even better. A larger number of steps could be investigated but this might be unrealistic, given the fact that physicians have little time and stop after a few minutes of search [129]. Often users will only test a few steps of relevance feedback at the most.

4.5 Modality Classification

As discussed in Section 3.2.2, the ImageCLEF2012 dataset was used. Table 4.6 gives the details of the runs evaluated. The bag-of-visual-words representation (Section 3.1.2) using SIFT (denoted in the following of this section as BoVW) and Lab features (BoC), (both local features are described in Sections 3.1.2) were used for the visual description of the images. The fusion rules (CombMNZ and Reciprocal) descriptions can be found in Section 3.1.2.

Table 4.7 presents the classification accuracy. The runs mc8, mc6, mc7 achieved the three best accuracies in the mixed run category in the ImageCLEF2012 challenge. The visual runs achieved an average performance with the inclusion of BoC as a global descriptor to improving the classification accuracy. It can be observed that the runs mc4, mc8 using the non-balanced expanded training sets and \( k = 14 \) are outperforming the runs mc6, mc2 that use the original training set. These runs also perform better than the runs mc3, mc7 that use \( k = 7 \), confirming our hypothesis that using a larger \( k \) for the expanded training set can improve results.
4.6 Compound figure separation

The methods proposed in this study were implemented in MATLAB R2012a and were tested with a computer running on Intel Core i7 processor of 2.3GHz. In order to evaluate each separator, the image subfigure boundaries were tested with 7% tolerance margin relative to the image dimensions, i.e., separator positions that differed by more than 7% were concluded to be false separators. The dataset was split into two parts for training and testing. The training was done manually to tune the thresholds in the rule based classification. In the rule–based analysis of separators, slightly different conditions were used for the horizontal and vertical separators in order to tune for the best output with the training set. Although, our configuration supports classification of separators with other supervised classifiers, we have selected the rule-based classifier for its accuracy. The rules and their parameters/thresholds used for analysis have been designed by studying the false detections, and tuned to achieve a good performance on a training set of 1444 images. The test and train image datasets were also made available for other research groups in ImageCLEF2013.

The results on the testset of the complete database show a good performance. Evaluation of separation lines show a sensitivity of 87.74% and precision of 84.36%. The result analysis was also done on different subclasses of images based on the number of subfigures and their arrangement. The results are obtained for each subfigure separator rather than individual compound-figure images. This is summarized in Table 4.8. Figure 4.7 shows examples of successful subfigure detection results.

![Figure 4.7: Examples of successful separation results](image-url)
4.6. COMPOUND FIGURE SEPARATION

<table>
<thead>
<tr>
<th>Class</th>
<th>Sensitivity (%)</th>
<th>PPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2x1</td>
<td>93.17</td>
<td>67.02</td>
</tr>
<tr>
<td>1x2</td>
<td>84.45</td>
<td>71.10</td>
</tr>
<tr>
<td>2x2</td>
<td>92.25</td>
<td>90.42</td>
</tr>
<tr>
<td>2x3</td>
<td>87.45</td>
<td>91.93</td>
</tr>
<tr>
<td>3x1 + 1x3</td>
<td>91.72</td>
<td>88.09</td>
</tr>
<tr>
<td>3x2</td>
<td>87.43</td>
<td>86.54</td>
</tr>
<tr>
<td>3x3</td>
<td>84.85</td>
<td>95.71</td>
</tr>
<tr>
<td>4x1</td>
<td>91.02</td>
<td>84.21</td>
</tr>
<tr>
<td>4x2 + 2x4</td>
<td>82.45</td>
<td>94.84</td>
</tr>
<tr>
<td>4x3</td>
<td>84.04</td>
<td>94.97</td>
</tr>
<tr>
<td>5x1</td>
<td>85.19</td>
<td>52.17</td>
</tr>
<tr>
<td>5x2</td>
<td>81.72</td>
<td>93.42</td>
</tr>
<tr>
<td>3-figures</td>
<td>77.54</td>
<td>79.31</td>
</tr>
<tr>
<td>5-figures</td>
<td>77.61</td>
<td>85.26</td>
</tr>
<tr>
<td>Other</td>
<td>84.72</td>
<td>86.08</td>
</tr>
<tr>
<td>Total</td>
<td>87.74</td>
<td>84.36</td>
</tr>
</tbody>
</table>

Table 4.8: Sensitivity and Positive Predictive Value (PPV) for each compound figure subclass.

With the results obtained from the MATLAB implementation of the algorithm, our two part approach of detection and analysis for subfigure separation shows good results. The detection was handled with and without preprocessing of the subfigures. The best results were obtained with preprocessing done only when certain conditions were satisfied or when direct application of the method failed. As the majority of the image classes are simple enough to be handled without preprocessing, this approach is computationally efficient as well as accurate.

It can be seen from the results that the sensitivity of separator detection can still be improved. Although most of the difficult types of subfigure segmentation have been addressed by our algorithm, some specific types of compound figures are not currently handled, such as compound figures with no borders. Subfigures that do not have a continuous uniform space between them are not detected by the methods discussed in this article. Apart from that, difficulties arise due to the particular sequence of detecting horizontal separators followed by vertical separators. Also, some subfigures images exist that cannot be separated by a rectangular area. The failed separation results for these kinds of images are shown in Figure 4.8.

Comparison between the work done on subfigure separation is difficult as there is not yet a common dataset used for the purpose. Furthermore, there is also no concrete guidelines for defining compound figure or subfigures. Nevertheless, the work presented in this article show that the results are at least comparable to those presented by others. Antani et al. [11] have reported a sensitivity of 95.78% and 94.38% in segmenting single panel and multi panel subfigures respectively. Cheng, Antani et al. [37] used two different methods based on the modality to improve the algorithm but they tested their implementation on a different dataset. Apostolova et al. have presented their results on a subset of the open ImageCLEF benchmark but the actual dataset has not been made available to other researchers.

Here it is necessary to note that the results of separation are sometimes subjective due to the lack of strict guidelines for defining compound figures, as previously mentioned. Although, the subfigure labels can provide a useful guide to decide on the number of actual subfigures in the image, this is often not the case in a significant number of images. Frequently, subfigures carrying independent information are grouped by a single label and on the contrary, subfigures sharing common information are given different labels. Often important content such as captions and legends are shared by subfigures, that puts doubt in the necessity or benefits of segmenting such subfigures. Deciding whether further separation is needed is usually the matter of maximal allowable information loss when the subfigures are separated. In other words, the number of separations expected for an image also depends on the purpose/application of subfigure separation. Our manual classification was done by analyzing whether separation in each case produced significant negative
impact on the information carried by subfigures. For this task, the following points were taken into account to resolve most of the ambiguities:

- Subfigures that are graphs or illustrations should have separate axes label/information if present.
- Subfigures may or may not have different labels or captions.

(a) subfigures stitched without spaces between them
(b) subfigures arranged in an order not separable when horizontal separators are searched before vertical
(c) subfigures not separable by any row or column

Figure 4.8: Examples of separation failures.

4.7 Indexing and retrieval efficiency

4.7.1 Indexing

Three subsets of the ImageCLEF 2011 medical task dataset [96], containing 1,000, 10,000 and 100,000 images respectively, were used as test datasets for evaluating the efficiency of the Hadoop indexer of ParaDISE. Two approaches were used; 1) a component–based approach and 2) a monolithic approach. The first one consists of two components; the feature extractor and the BoVW indexer. In this approach 1), the intermediate output of the extractor is stored on the disk before the indexer uses it as an input. This is a common practice for component–based evaluation and parameter tuning. The monolithic approach 2) is not storing any intermediate output and performs the whole pipeline each time it is initiated.

We consider the component–based approach as being an input–output (IO)–intensive task due to the writing and reading of the outputs of the extractor component. Runs using multi–threading techniques with 1, 2 and 3 threads were performed on a single machine to serve as baseline.

Preliminary results showed that even for the smaller datasets of 1,000 and 10,000 images, the multithreading run using 3 threads outperformed the MapReduce tasks in the 42–task cluster, taking about half the time for the extractor phase. For 100,000 images, the multithreaded runs scale linearly for 1–3 threads. It requires approximately 7 hours with 3 threads for the whole pipeline (4.9 hours for the feature extractor and 2.7 hours for the BoVW indexer). In comparison, MapReduce required more that 14 hours using 42 concurrent map tasks (13 hours for the feature
extractor and 1.5 hours for the BoVW index). Although the MapReduce framework reached performance comparable to the single machine by increasing the number of reducers, the amount of resources used to achieve such a performance showed that MapReduce may not be suited for IO-intensive tasks.

For the monolithic runs 2), two approaches were taken. The first one uses 6, 12 and 24 simultaneous tasks on a single node (i.e., the PowerEdge R710 Server) for the indexing MapReduce run, while the second uses the same number of tasks distributed among the nodes of the Hadoop cluster. A text file containing a list of the image file paths was given as input and the same number (50) of images per map task, was given in all runs. A single reducer was used in all runs. The results are shown in Figure 4.9.

![Comparison of runtimes using a single node and the distributed Hadoop cluster.](image)

**Figure 4.9:** Comparison of runtimes using a single node and the distributed Hadoop cluster. For the datasets of 10k and 100k images, the single machine run does not scale linearly with more than 12 simultaneous tasks.

Two approaches for content-based image indexing were compared and implemented in the MapReduce framework: component-based versus monolithic indexing. The former is convenient to separately optimize feature extraction and the BoVW indexer because it does not require to run the whole pipeline for each optimization. However this costs time and storage space when you perform the whole pipeline due to the output writes and reads. This was observed with an unexpectedly long runtime for the feature extractor with the MapReduce framework in the component-based approach. The process was slowed down because it requires to write the features to a very large CSV (Comma-Separated Values) file of approximately 100 Gb for 100,000 images. The result is consistent with previous work that showed that the MapReduce framework was not performing well with IO-intensive tasks [115]. The monolithic strategy showed to be well-suited to the MapReduce framework, which allowed indexing 100,000 images in about one hour using 24 concurrent tasks (see Fig. 4.9 (c)). The performances of the monolithic approach when running on a single node versus distributed nodes is compared. Whereas runtimes using 6 concurrent map tasks are similar (see Fig. 4.9 (a) and (b)), in the case of 12 and 24 nodes the results for large datasets differ. In contrast to the multi-node approach, the single-node times do not scale linearly. This illustrates the advantage of using several nodes, which benefits from the distributed memory and disk access of each separated node.
4.7.2 Retrieval

Image retrieval using modality classification

As seen in Chapter 2 filtering by modality is feature that was commonly requested by radiologists. In this section the effect of the modality filtering on the retrieval performance is assessed. This evaluation was submitted to ImageCLEF2013 challenge [73] as part of the medGIFT participation. For this purpose, the ImageCLEF2013 image dataset was classified using the method of the best mixed run of the 2012 modality classification task [71]. The query images of each topic were also classified and a set of query modalities was produced. Images among the 1,000 top images retrieved by the retrieval methods that were classified into one of these modalities were placed in top of the other retrieved images.

Three approaches of modality filtering were tested. In the first one named "exact" only the modality detected by the kNN classifier for each query image of the topic was put into the query modality set. The second named "close" puts all the modalities detected by the kNN classifier of any query image into the topic. The third one named "prefix", is similar to the first but the broadest modality (diagnostic, general, compound) was used instead of the exact modality for boosting the image score in the retrieved set. Due to the limited number of submissions the "exact" approach was not submitted as it had a low performance in preliminary tests on the ImageCLEF 2012 collection.

Multiple features were used for visually indexing the dataset. Apart from the features used last year (SIFT–based BoVW and BoC features), more features were added for the image retrieval task. The CEDD and FCTH descriptors were also used, due to their good performance on ImageCLEF 2012 challenge. BoVW and BoC features that contain spatial information were used. For BoVW an SPM approach was used [110] while an \( n \times n \) spatial grid was used for BoC. The chosen pyramid depth level was \( L = 1 \) and the grid size was selected to be \( n = 3 \), after tuning on the ImageCLEF 2012 benchmark.

For the text runs, a late fusion of full text search with caption search was followed. Moreover, for each of these searches three different queries were fused for each topic. The first one queried the topic query for exact matching. The second connected the query terms with 'AND', while the third used 'OR' as a connector.

In the mixed runs, two different approaches were submitted. In the first approach, a linear weighted late fusion was used. The weights were tuned using the ImageCLEF 2012 benchmark. The second one used a late fusion (CombMNZ) of the visual and text runs to rerank the images in the result set that was retrieved by the text run, similar to [42]. Preliminary experiments had shown that weighted fusion had the best performance among all other fusion rules and for this reason it was chosen for the comparison with the reranking runs. Below the characteristics of each of the 10 runs submitted are presented:

- **Run13–medgift_visual_nofilter**: visual run that uses the 6 features (BoVW, BoC, SPM_BoVW, Grid_BoC, CEDD and FCTH) and CombMNZ fusion. No modality filtering is used.

- **Run14–medgift_visual_close**: same as Run13 but the "close" modality filtering approach is used.

- **Run15–medgift_visual_prefix**: same as Run13 but the "prefix" modality filtering approach is used.

- **Run16–medgift_text_nofilter**: text run using caption and fulltext search with CombMNZ fusion. No modality filtering is used.

- **Run17–medgift_text_close**: same as Run16 but the "close" modality filtering approach is used.
4.7. INDEXING AND RETRIEVAL EFFICIENCY

- Run18–medgift_text_prefix: same as Run16 but the "prefix" modality filtering approach is used.

- Run19–medgift_mixed_rerank_nofilter: mixed run fusing the methods used in Run13 and Run16 to rerank the top 1,000 results of Run16. No modality filtering is used.

- Run20–medgift_mixed_rerank_close: same as Run19 but the "close" modality filtering approach is used.

- Run21–medgift_mixed_rerank_prefix: same as Run19 but the "prefix" modality filtering approach is used.

- Run22–medgift_mixed_weighted_nofilter: mixed run using linear weighted fusion for Run13 and Run16. Weights were set to be for visual: 0.2, text: 0.8. No modality filtering is used.

The results are presented in Table 4.9. There are situations when "nofilter" or "close" filter perform better depending on the type of run (visual, textual or combined). Second best results were achieved in the visual runs using the baseline run ("nofilter"). On mixed techniques medGIFT was the second best group and on the textual runs medGIFT was the fourth group when using "close" filtering. In several queries the average precision was strongly improved by the modality filtering, while other queries had much lower performance, for example when the modality was not correctly detected. This indicates that more accurate modality classification should further improve retrieval performance. Results also show that the use of reranking in place of weighting for the fusion achieves better results. This is important for large-scale retrieval as visual search on the whole dataset is computationally costly. By having text retrieval as a first step, the image-based retrieval search subspace is significantly (magnitude of two orders) smaller.

Table 4.9: Ad-hoc image retrieval results in ImageCLEF2013.

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Run type</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best ImageCLEF run</td>
<td>Visual</td>
<td>0.0185</td>
</tr>
<tr>
<td>medgift_visual_nofilter</td>
<td>Visual</td>
<td>0.0133</td>
</tr>
<tr>
<td>medgift_visual_close</td>
<td>Visual</td>
<td>0.0132</td>
</tr>
<tr>
<td>medgift_visual_prefix</td>
<td>Visual</td>
<td>0.0129</td>
</tr>
<tr>
<td>Best ImageCLEF run</td>
<td>Textual</td>
<td>0.3196</td>
</tr>
<tr>
<td>medgift_text_close</td>
<td>Textual</td>
<td>0.2478</td>
</tr>
<tr>
<td>medgift_text_nofilter</td>
<td>Textual</td>
<td>0.2281</td>
</tr>
<tr>
<td>medgift_text_prefix</td>
<td>Textual</td>
<td>0.2262</td>
</tr>
<tr>
<td>Best ImageCLEF run</td>
<td>Mixed</td>
<td>0.3196</td>
</tr>
<tr>
<td>medgift_mixed_rerank_close</td>
<td>Mixed</td>
<td>0.2465</td>
</tr>
<tr>
<td>medgift_mixed_rerank_nofilter</td>
<td>Mixed</td>
<td>0.2375</td>
</tr>
<tr>
<td>medgift_mixed_rerank_weighted_nofilter</td>
<td>Mixed</td>
<td>0.2309</td>
</tr>
<tr>
<td>medgift_mixed_rerank_prefix</td>
<td>Mixed</td>
<td>0.2271</td>
</tr>
</tbody>
</table>
Chapter 5

User–oriented evaluation

User–centered evaluation is an important part of user–centered design, which needs to be performed in the early stages of the development [90] and is seen as an iterative process throughout the development cycle [47, 95]. It is often performed in the form of empirical usability tests, which include having a number of target users to interact with the system. Usability of the system is assessed with factors such as learnability, efficiency, effectiveness, memorability and satisfaction [90]. A survey on common usability assessment techniques and tools is given in [16]. Various methods exist for conducting these tests, including thinking aloud, direct or recorded observation of the interaction, survey forms and log analysis. A more detailed description of aspects to be taken into account when designing a usability test can be found in [100].

An important aspect when designing a usability test is the number of participants required. Early studies have discovered that a single person is not able to detect all usability problems but 3-4 are sufficient [147]. In [146] it is suggested that five users are enough, while studies have questioned this choice [172, 195]. The exact number of participants remains an open question, though in [145] it is explained that five participants are indeed enough for each iteration of an iterative user–centered evaluation. The key elements of UCD are described in the ISO standard for the Human–centered design for interactive systems (ISO 9241-210, 2010).

In this chapter, the design choices, the setup and the results of the user–centered evaluation of the KHRESMOI search engine for the radiology use case are presented. As described in Chapter 3, the system combines text and CBIR search to retrieve images and articles from internal and external data sources. The KHRESMOI development cycle included two main evaluation rounds. One at year 2 assessing the first version of the system and one at year 4 for the final prototype. The general research questions that the evaluation tries to answer are:

• Does the KHRESMOI system improve current search for information in radiology (which is mainly patient–centered or using Google on the Internet)?

• Does it cover unmet information needs and to what extent?

• Which functionalities are more useful and which tools need to be improved, changed or added?

While the contribution of this thesis lies more on the full user–oriented development cycle of the 2D image search subsystem of KHRESMOI, its evaluation is strongly connected with the 3D image search subsystem. Moreover the same study protocol, designed in the context of this thesis, was used for both subsystems. Thus the full evaluation of the KHRESMOI Radiology prototype is also presented in detail.

5.1 First evaluation round

At year 2 of the KHRESMOI project the first user–oriented evaluation round in the development cycle was run. Apart from the general research questions, this step aimed to get initial insight on
aspects of the system that need improvements and the effectiveness of the tools. Getting more concrete feedback on image search requirements in addition to the survey described in Chapter 2 was also one of the main goals. Performing information tasks on a real system was thought to assist the participants in having clearer ideas about additions that would be of interest to them.

5.1.1 Methods and Materials

This section describes the methodology followed for designing, setting up and running the first round of user tests. The datasets used for performing the tasks and the materials used for capturing the user interaction with the system are presented.

The evaluation process followed an iterative approach and different preliminary steps were taken for each system. The 3D image search system was designed to mainly provide information search in the hospital image data during the clinical duty. Before running the user tests, interviews with radiologists were arranged in order to design an interface that would fit in the clinical workflow.

Pilot user tests were performed to evaluate the basic aspects of the interface and the system’s functionalities. This also helped detecting shortcomings of the user tests design and refining the study protocol.

User study protocol

In order to investigate the research questions described in the beginning of this chapter, the following aspects were taken into account:

1. Success of information finding by radiologists using KHRESMOI.
2. Time to find relevant information using KHRESMOI.
4. Usability of the KHRESMOI system.
5. Missing useful functionalities in the current system.

In this user study, the methods of the above mentioned evaluation aspects needed to be decided. The final selection of methods, after being refined by the preliminary step of pilot user tests, is presented below:

- Participants were asked to perform information retrieval tasks for which at least one of the results is known. Therefore aspect no.1 could be evaluated.
- The time taken to fulfil each task was measured. For tasks whose time was fixed, the time taken to find the first relevant result was measured, instead. This method evaluated aspect no.2.
- Participants were asked to fill a questionnaire about their experience of using the system. This allowed to evaluate user satisfaction (aspect no.3) and detect usability problems found by the participants. Questions were included that requested feedback and propositions for system improvement (aspect no.4).
- Participants were observed and video recorded while using the system. Possible system flaws or usability problems that were not consciously detected by participants were identified through this technique (aspect no.4).

The concept of usability is highly correlated with the concept of user satisfaction. In this study, the term user satisfaction is used for the explicit feedback provided by the participants and the term usability is used for the implicit insights obtained by the observation of the participants’ behavior. The user tests were conducted in the format of one-to-one sessions, one participant performing the tasks and one observer to facilitate the user test. The details of the session were also refined after the pilot tests by including and removing tasks, as well as modifying the time limitations. The final session outline is presented below:
5.1. FIRST EVALUATION ROUND

1. Introduction to the KHRESMOI project, the existing search system and the user test goals (5 minutes).

2. Tutorial video on the system tools and functionalities (5 minutes).

3. Demographic survey (5 minutes).

4. Introductory task, simple use of the tools (5 minutes).

5. Guided user tests in clear scenarios (30-40 minutes).

6. Survey on the satisfaction with the tools and functionalities (10 minutes).

7. Free possibility to use the system (5+ minutes).

8. Survey on the satisfaction with the system, free discussion (10 minutes).

The introduction by the test facilitator intended to help the participant understand the concept of the system and motivate to do the test. Then, the video demonstration of the system introduced the tools offered by the application. The introductory task was introduced after the pilot user tests because the video tutorial alone did not contain enough information for the user to get familiar with the tools available. Throughout the session, the participant was being test facilitator by the observer to identify potential shortcomings of the system. The observer was instructed to have a neutral attitude and was allowed to help only when the participant was blocked and could not proceed with a task.

The setup of the session included hardware and software preparation but also training sessions of the observer to get familiar with the recording tool and the study purpose. The hardware used in each session included two Windows computers - one for the participant and one for the observer. The KHRESMOI client was downloaded to the participant’s computer and the recording software was installed on both computers.

At the end of each session the file containing the recordings, the answers to the surveys and the observer’s notes were acquired. The details of preparing, setting up and running a session were added into a document to ensure that the experiment can be reproduced under the same conditions. This document of instructions can be found in Appendix B.

Tasks and datasets

As mentioned in section 5.1.1, the user was requested to perform several information seeking tasks during a session. The design of the tasks took into account that they need to use most of the system tools and functionalities. They had to describe realistic scenarios that appear in clinical and academic workflows. Depending on the tasks and the subsystem used, different data sources were required.

For the evaluation of the KHRESMOI system in terms of radiology–related information search into external sources, the ImageCLEF2012 medical data set was used [138]. It represents a relatively realistic source for a medical literature search and especially for an initial test on the system’s scalability and performance.

The total anatomical dataset for the 3D image search prototype consisted of a total of 7936 MRI and CT Volumes with a total size of 470GB including 5817 radiological descriptions as reports. A subset of 117 Lung CT volumes was used for the pathology dataset. The labelled pathologies for these volumes were PE (panlobular emphysema) and ZE (centrilobular emphysema).

Two groups of information retrieval tasks were used: Three 2D image search tasks and two article search tasks. A subset of the ImageCLEF2012 medical image–based and case–based retrieval task topics was used respectively. The topics for the image–based task were selected after the log analysis of queries to a radiology image search engine [182], while case-based topics consisted of cases included in an educational database [138]. The full task descriptions can be found in Appendix C.
The guided scenarios of the user tests were based on these information retrieval tasks and included use of the various tool of the system, such as query–by–text, query–by–image–example, the personal library, the tray and others.

The tasks for the 3D image search were defined in order to answer distinct questions that allow quantifying the quality and features of the prototype.

1. How do the new features and functionalities added with the prototype compare to the standard tools of radiology. More exactly, how the addition of searching for similar images compares to only using tools like scrolling and changing of brightness and contrast.

2. This question is about the quality of the retrieval. Does it actually retrieve the desired pathologies and therefore help the radiologists in answering the question of finding similar diseases in the database. The radiologists can also use the radiological report of the found volumes.

3. Another question concerns the educational aspect of the prototype. The goal is to find out if it is possible for radiologists to find good examples for certain pathologies using the new features. This would allow them to create better cases and visual examples for demonstration purpose.

4. The quality of the anatomy retrieval is also an important part of the prototype. Therefore a question is designed where radiologists are asked to find similar anatomical regions based on some random examples.

5. Finally the participants should have time to freely use the system. This could give some new perspectives and additional information that could be used for further development of the prototype.

For the first user tests the goal was to decide whether the way the prototype is designed would be useful for their daily routine. This includes the design of the user-interface and the representation of the query image, the result list as well as the detail view. Therefore only a subset of the tasks was necessary in the beginning, which also reduced the time taken for each user-test. For this purpose the chosen tasks were (1), (2) and (5). The final tasks are described in detail in Appendix C.

Session setup and recording material

The setup of the session included hardware and software preparation but also training sessions of the observer to get familiar with the recording tool and the study purpose. The hardware used in each session included two Windows-based computers one for the participant and one for the observer. The KHRESMOI client was downloaded to the participants computer and the recording software was installed on both computers. For observation and recording, the commercial software Morae mentioned in [16] was used. This software allows screen and face video recording of the participants (Figure 5.1) and also remote online observing on a different computer (Figure 5.2). Upon start, Morae guides the user through the steps of the session, having all additional material, such as survey forms, task descriptions and instructions integrated and displayed on the participants computer screen (Figure 5.3). All the surveys’ answers, observer’s notes and recordings are saved in a digital format which is compatible with commonly used statistical packages for result analysis and presentation.

Survey forms

Four survey forms were used in this study. The initial demographics survey form was used to get information on medical experience and computer use of the participants. Two survey forms were used to evaluate the subsystems’ tools and functionalities usability and one to evaluate user satisfaction with the global system.

A combination of modified versions of the System Usability Scale (SUS) [26] and the Questionnaire for User Interaction Satisfaction (QUIS) [39] was used for the user satisfaction and usability
5.1. FIRST EVALUATION ROUND

Figure 5.1: The interface of the Morae Recorder software, which was installed to the participant’s computer. A common study configuration file is created for all the users and the recording starts by pressing red button.

Figure 5.2: The interface of the Morae Observer software, which was installed to the observer’s computer. This tool allows the observer to take notes on the timeline of the recording and use different markers (e.g. found bug, participant comment, participant blocked etc.) The program in the screenshot is waiting for the recording to begin.
survey forms. Open questions for providing comments on specific aspects of the system and suggestions for improvements were added. To get preliminary answers to the research goals, questions about the novelty, usefulness and intention of use of the tools were included. The final survey forms, after the refinement during the pilot user tests, can be found at the Appendix D.

Interviews and Interface design

The goal of the KHRESMOI retrieval system is to add content based retrieval functionality to the clinical radiologist workflow. We do not aim at replacing existing image management systems, but instead want to offer additional functionality. In the beginning we were interested in the workflow radiologists perform and during which they could imagine that KHRESMOI functionality would be helpful. A few draft layouts for the user interface as shown in Figure 5.4 were presented to the physicians that would allow the radiologist (1) to work with a query image (typically this is the image the radiologist is analyzing at the moment) and (2) to browse through retrieval results. Retrieval results are images that are identified by the retrieval engine, and share information such as pathological features, local appearance, and anatomical location with the query image.

The goal of the interviews was to get insights into the various ways radiologists can imagine integrating the tool into their workflow and what specific functionality would be necessary to do this. Optimally interviews should be a conversation, in which the needs of the radiologists and the know-how about what is possible should be discussed. The information extracted from the first interviews is discussed in Section 5.1.2.

Based on the output of the interviews the user interface for the first 3D image search prototype was designed. This prototype was applied for the user tests, where the users have to complete specific tasks. The tasks were designed in a way that makes it possible to extract relevant information about the usability, quality and improvement options for the prototype.

5.1.2 Results

Two sets of user tests were run in the context of KHRESMOI first round of user–centered evaluation. The first set was a pilot user study that aimed at finding the most significant bugs and inconsistencies as well as the user study’s protocol shortcomings. Then the full user tests were performed after refining the protocol using the results of the pilot study.

The user tests took place at the University hospitals of Geneva and the Medical university of Vienna. Twelve persons (3 females, 9 males) participated in the full user tests round. This number does not include the participants in the pilot user study and the interviews. They were all below 40 years old, with eight of them being below 30 and three below 35.
5.1. **FIRST EVALUATION ROUND**

Five persons were interns, four were residents, one associate professor in radiology, one attending and one with no radiology background. Among the radiology specializations (participants could choose more than one field) the most common was thorax (3), radio oncology (3) and bone(2) while other chosen fields were echocardiography, neuroradiology, cardiac, pediatric, general and emergency radiology. All of the participants reported frequent computer use (more than once a day) and search for medical information (7 reported more than once a day, 3 once a day and 2 once a week). Due to technical difficulties, bandwidth problems and development schedule, not all of the participants were able to perform all the tasks.

User satisfaction results over key general aspects of the system are presented in Figure 5.5. The median for the question about intention to use the system frequently was 4. The same median was obtained for easiness to use, the ability to use the system without technical support and feeling confident when using the system. The median for easiness to learn and using the system without prior training was 5 for both and the general feeling about the system’s consistency was 3. In order to assess the global satisfaction of each participant the mode over the general satisfaction questions was taken, measuring the most frequent grade given (Figure 5.6). Also, for measuring the consistency of this satisfaction, the frequency of mode was given (Figure 5.7).

In the following sections a detailed description of the results of each step of each subsystem prototype evaluation is given.
CHAPTER 5. USER-ORIENTED EVALUATION

Figure 5.5: Median values of measuring general user satisfaction about the system in Likert scale (1=strongly negative, 5=strongly positive).

Figure 5.6: Mode values for each participant over the global satisfaction questions in a Likert scale.

Figure 5.7: Mode frequency for each participant over the global satisfaction questions.
5.1. \textit{FIRST EVALUATION ROUND}

Figure 5.8: Pilot study: Median of measuring user satisfaction over specific system aspects in a Likert scale (1=strongly negative, 5=strongly positive).

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure5_8.png}
\caption{Pilot study: Median of measuring user satisfaction over specific system aspects in a Likert scale (1=strongly negative, 5=strongly positive).}
\end{figure}

\textbf{2D Image/article search subsystem}

Five persons (2 females, 3 males) participated in two sets of parallel sessions. All were below 30 years old, with two of them being below 25. Two participants had radiology background (one specializing in bones), one was a non–radiology intern and two were final year students in medicine. All participants declared frequent computer use. Three persons answered to search for medical info more than once per day, one once per day, and one answered once per week. The recruitment of participants was done via personal contacts and people who volunteered to take part in the study at the radiology department of the University hospitals of Geneva.

The mean time for retrieving the first relevant result during the 2D image search tasks was 158 seconds. This time included choosing image examples, investigating the results and judging a result as relevant. This time includes only the cases when a relevant result was found. For case–based retrieval tasks the respective mean time was 179 seconds. The mean number of results selected as relevant was 5 for the 2D image search tasks and 2.6 for the case–based search. One participant (one still studying medicine) did not select any relevant results for any of the tasks.

User satisfaction on the specific system aspects was measured on a Likert scale where 1 was strongly negative and 5 was strongly positive. Results are given in Figure 5.8. Questions about the user’s use intention in academic, research and clinical work respectively obtained medians of 4. Finally a question regarding the practical usefulness of the novel features of the system obtained a median of 5 out of 6 due to a design error. This was excluded from the global user satisfaction evaluation. User satisfaction results over general aspects of the system are presented in Figure 5.9.

In order to assess the global satisfaction of each participant the mode over the general satisfaction questions was taken, measuring the most frequent grade given (Figure 5.10). Also, for measuring the consistency of this satisfaction, the frequency of mode was given (Figure 5.11).

All open responses were classed into similar comments. Redundant comments were removed and all comments were transmitted to the development team. Frequent comments include:

- complaints about CBIR performance were frequent as often several irrelevant results were ranked highly;
- Zooming into images and basic manipulation such as level/window settings were considered important but are currently not possible in the interface;
- Displaying more information about the images in the result lists was also requested;
- Other propositions about functionalities such as backspace usage, radiology related functionalities (contrast adjusting etc.) were given;
CHAPTER 5. USER-ORIENTED EVALUATION

Figure 5.9: Pilot study: Median values of measuring general user satisfaction about the system in Likert scale.

Figure 5.10: Pilot study: Mode values for each participant over the global satisfaction questions in a Likert scale.

Figure 5.11: Pilot study: Mode frequency for each participant over the global satisfaction questions.
5.1. **FIRST EVALUATION ROUND**

![Figure 5.12: Median of measuring user satisfaction over specific 2D image search prototype aspects in a Likert scale (1=strongly negative, 5=strongly positive).](image)

Below, some of the comments are given in their raw form (translated from French):

- The search for associated articles is interesting at this stage, the search by images would also be useful if visual results were more relevant.

- As a student, search results have to be extremely relevant because we do not have the knowledge to exclude bad images on our own.

- It seems reliable more or less, I feel like it has difficulties distinguishing CT scan images from MRIs.

- More information on the description of images could be interesting to narrow down searches. A zoom in on an image in the 'details' section would be useful.

- There is no text below images in the list of result. It would be good to see the description when you hover the mouse over a result image. Difficult to get a good idea of the image at a glance when they are so small.

- The tool reacts very well to its use, no delay, no bug, tasks we are asked to do are rapidly performed.

- Takes 15 minutes to be comfortable.

Eleven out of the twelve persons that participated in the user study tested the 2D image and article search subsystem in the full user tests. However in two cases the participants did not performed all the tasks or answered all the questions due to technical difficulties. This resulted to 31 performed 2D image search tasks out of 33 (11 participants × 3 tasks) and 19 article search tasks out of 22 (11 participants × 2 tasks).

The mean success rate was 80.65% (25/31) for image search tasks and 78.95% (15/19) for article search tasks. Every task that the user found at least one relevant result was considered as successful. The mean time for finding the first relevant result during the 2D image search tasks was 106 seconds. This time included choosing image examples, investigating the results and judging a result as relevant. It includes only the cases when a relevant result was found. For case–based retrieval tasks the respective mean time was 150 seconds. The mean number of results selected as relevant was 4 for the 2D image search tasks and 3.1 for the case–based search. This numbers include also the cases that no relevant result was found by the user. User satisfaction on key aspects of the 2D image search prototype and intention of use was measured on a Likert scale where 1 was
strongly negative and 5 was strongly positive. Results are given in Figure 5.12. The median for system response time was 5 (mean 4.5). A median of 3 was reported for system reliability (mean 3.6). In terms of results quality and presentation the median was 3.5 (mean 3.2) and 4 (mean 4.1) respectively, while ability to correct mistakes and system design to be used by all levels of users both obtained a median of 4 (means 4.1 and 4).

Question about the user’s use intention in academic work obtained a median of 5 (mean 4.2), while the respective questions for research activity and clinical duty obtained medians of 4 (mean 3.9 for both). The question regarding the novelty and practical usefulness of the features of the system obtained a median of 4 (mean 4.5).

The same procedure with the pilot study was followed for the open responses. The comments were classified into Frontend and Backend-related. The most common comments can be summarized in the following points:

- **Frontend**
  - Querying, such as advanced text querying and relevant/non-relevant marking of images; available options should be more explicit and easy to use;
  - Basic and radiological–based image manipulation of the selected and query images should be available;
  - Results presentation and views; images should be presented in grid as default and articles as lists.

- **Backend**
  - Complains about CBIR returning many non relevant results; Non relevant marking didn’t produce the desired results;
  - Modality filtering requests;
  - Propositions about Finding articles using images only or using example article (“Find similar articles”);

Below, some of the comments are given in their raw form:

- **Frontend – Querying**
  - When the mesh system is on, and is researching for auto completing, it is really hard to position the cursor in the middle of a term to refine it;
  - Non-relevant marking not intuitive. It would be useful to be able to mark images as non-relevant directly from the results list;
  - It is not very obvious when you are able to drag and drop the images to the query zone (multiple images, results list, personal library, details view etc.);
  - Want to see the detail of an image from the query zone (images from the query zone drag n droppable also to tray);
  - Re do “enter” when modifying query by for example drag and dropping images is not very comfortable;
  - When searching with images in the query zone, it would be useful to have a contextual command (right click) to launch the search easily;
  - Globally, the use of AND OR NOT, of text in bracket, and of non relevant marking of images should be more explicit (also what is default?)

- **Frontend – Details View**
  - In the detail view, the participant would like to have the possibilities: scroll (CTs), scale, contrast, brightness;
5.1. **FIRST EVALUATION ROUND**

- Didn’t have time to read all the article: better summarization could maybe spare time to users;
- Abstract translation to German is useful;

• Frontend – Results View

- Thumbnails would be better a little bigger, so that we can already start a quick pre-analysis at this stage;
- Couldn’t it be that we could resize thumbnail as we want them to be?
- When you select ”Text” from the media type displaying options for results, we still find images in the first place. Couldn’t that be changed, so that when you select one media type, this type appears first?
- Thumbnails presentation is not great. Everything is a bit mixed up and unclear, there is no organization, we don’t know about modality, anatomy, or pathology;
- The button to switch mode in the results list (list or grid) is a bit confusing because it makes you think it is switch between images and articles;
- Articles: would be nice to have the whole title when hovering with the mouse;
- Multiple selection (shift or ctrl) would be nice in the result list (to drag n drop them all at once to the query zone);
- The button of grid mode in the result list is not reverted back after some clicks;
- Usually expect results as grid if image and as list if articles.

• Backend

- Marking graphs as non-relevant gives still a lot of graphs in result. This comment is the same with other types of images: e.g.: marking a given chest x-ray image as non–relevant does not even exclude the result itself from the resulting results list;
- It would be great to be able to specify the type of image that we want in result list (x-ray, MRI, CTs, graphs, statistics etc.);
- If i m using std x ray for visualizing a pulmonary parenchyma, then I probably want std x rays and CTs but no ultrasound or graphs etc.;
- Globally the system is more relevant with textual queries (with or without images) than with image queries only. This is the case for both article results and image results. Osteoporosis does not find many images, and only from source 2;
- It is impossible to return article with queries with only image-search. This is maybe not adapted and we should always be able to search for articles;
- The search over images only does not return anything that I could use. As soon as I add terms to my query (with or without images), I can find some more relevant results. Some results are even focused on the same anatomical area, but usually regards different pathologies;
- I would like to have something like search for similar articles, as it is now for images;
- It would be nice to be able to check ”Only search articles” before the search, so that he does not have to do one more step by filtering results. Maybe it could spare a bit of execution time too, which is great.

• General

- Performing the is highly dependent on the user’s skills to analyze de radiological image, and therefore depends a lot on experience;
- Consistency and integrity of the program, and execution times are good. Real problem is result relevancy;
The learning curve of the system is quite rapid;
During the 2D test, the high number of windows in the perspective is confusing;
I can trust the image results because they come from scientific articles, in contrast to images I find from general purpose search engines;
I understand the concept behind the system. I find the connection of images and articles useful and this is not currently available with the existing tools. It will be more useful if the results are better.

3D Image search subsystem

This section covers the specific results of the 3D image search prototype. First information about the design and integration of a 3D image search system was acquired with a series of interviews. Based on the results the 3D prototype was designed. After removing initial bugs and refinement of the user test protocol the full user tests were performed. Note that at this stage, the evaluation focused on the workflow and user interaction. Only a preliminary version of the image search algorithms was deployed in the system.

The interview results are grouped into (1) workflow, (2) 3D frontend and (3) organization. Workflow results present what is important for the radiologists in order to quickly and efficiently use the system. Furthermore, they describe what information needs to be available at each time and how it should be presented. This also includes whether or not to display the associated report. The 3D frontend examples are shown in Figure 5.4. In this paragraph the comments about which layout would work best, what is missing and what should be changed is summarized. Information concerning the organization and setup of the user tests is outlined in the last paragraph.

(1) Workflow: For the physicians it is important to quickly judge relevant cases. Therefore they need quick browsing through the thumbnail images. In order to reduce the number of thumbnail results to a smaller number of relevant cases, text search for the corresponding reports should be possible. Additional filters like gender, age or modality would also reduce the number to more relevant search results. Furthermore the reports could also be used for making differential diagnosis.

Optional improvements regarding the thumbnails were identified: enlargement of the thumbnails or having scrollable thumbnail to further facilitate relevance judgement of the results. Optional thumbnail size would also be a possibility, because this can depend on the modality and pathology that is currently worked on.

In the detail view more extra information would contribute to the decision making of the physicians, like: gender, age, slice id, modality, sequence information, windowing information (brightness/contrast), number of images, anatomical position, patient preparation, confirmation tests (lab tests, biopsy etc).

Keywords and ontology terms extracted from the reports could be shown in either the detail view or thumbnail view. A search history would be appreciated by the physicians where they can easily and quickly look up the latest volumes they looked at. Adjusting brightness and contrast, as well as zooming in and out of the image are very important for the daily routine of a radiologist. The method called windowing is widely used in different radiology systems and is an essential tool for any physician. The current settings of the query window should also be automatically transferred to the detail view in order to make the two volumes instantly comparable.

(2) 3D frontend: Figure 5.4 shows six drafts named from (A) to (F) for the first UI design. The desired thumbnail size tends to be as in (A) and (B). Reports are important especially during the detail view, as in (C), (E) and (F), although the report should not be overlaid with the image (E). In this way they can look at the volume and the report simultaneously. A short description of the report added to the thumbnails view is also appreciated (B). The presentation of the report to the user should be clearly structured for easy readability including highlighting of important parts (F).

Marking a region of interest should be as simple as possible and the visualization of the result ROIs (F) could vastly reduce the time needed for the tasks. Indicating the relevance of the
5.1. **FIRST EVALUATION ROUND**

Figure 5.13: Median of measuring user satisfaction over specific 3D image search prototype aspects in a Likert scale (1=strongly negative, 5=strongly positive).

matching ROIs (e.g. by brightness) could improve and speed up the interpretation of the results. The physician should be able to easily mark and browse one or more slices of the query volume.

If different layouts are used, fast switching between them is very important. During browsing the thumbnails additional information should be available, for example: age, gender, modality, sequence information, number of images, resolution. Primarily the axial view is preferred by the physicians, options to switch to one of the other two planes can be included.

**(3) Organization:** A handout for the physicians should be prepared, where the project and its goals are briefly described. Information concerning the user tests and the analysis of its result should be mentioned. The hardware and software used for the tests should also be defined, such that results from different locations are consistent and comparable.

5.1.3 **Discussion**

A total number of 17 persons participated in the user tests of the KHRESMOI radiology prototype. The sample of users was relatively young and had a varying level of medical experience, with the participants being involved into various radiology specializations. Recruiting radiologists was a difficult task as radiologists are usually on a busy schedule, with a lot of clinical and academic. The number of participating users, even though it was less than what was aimed at, is mentioned as sufficient in the related literature, in regard to the current development stage, the specialized type of the system and the goals of the study. More participants would potentially be an overspend of human resources without resulting to significantly better insight.

Main tendencies on user satisfaction could be identified (Figure 5.5), which can be used to guide further technical development on improving certain aspects. In regard to the global aspects of the system, users found the system easy to use without the use of technical support. They felt confident using the system and are positive towards using it frequently. They seemed strongly positive about the system being easy to use without any prior training, despite the new tools offered (e.g. CBIR, relevance feedback and ROI marking). They were less satisfied with the consistency of the system, which is a logical outcome considering the current stage of the system’s development. The main tendency of the users seems to be strongly positive with the majority of the users (6 persons) giving a mode grade of 5 with a frequency above 0.5 over the general satisfaction questions.

**2D Image/article search subsystem**

One of the main outcomes of the pilot study was that a video tutorial alone was not enough and that a user required exploring himself the new functionalities before proceeding to complex
information search tasks. This can limit the effectiveness of information finding during the early tasks and makes them less appropriate for performance comparison (text search vs. visual plus text search). For this purpose, a guided tutorial task after the video was included in the full tests, where it was asked from the user to perform very simple tasks using the tools (see Appendix C).

Some task descriptions and questions of the survey were not completely clear and this caused misunderstanding results retrieved by the participants. It was also observed that participants did not read the tasks in full detail and often performed slightly differing actions than the ones the task asked. This led to rewriting the task descriptions to be shorter and more clear. Moreover an oral description was given in the full tests, pointing out the important parts of each task.

Concerning the use of a commercial recording and observation software such as Morae, both advantages and drawbacks were found. All information that the participant needs for performing the test can be found on his screen and no transition to paper is needed. This helps the user concentrate on the tasks and facilitates an uninterrupted flow of the tests. It provides results in a unified digital format that is easy to transfer to statistical packages, to analyze and present in a meaningful way. It allows for indirect observation (as the observer can remotely observe the user’s screen and face), which removes some of the subject’s stress of being observed and the incorrect feeling that he is being evaluated.

On the other hand, the use of such a tool increases the hardware (e.g. every session uses two computers with Internet connection instead of one) and software requirements, adds extra complexity to the setup of the tests and is prone to software crashes. Moreover, purchasing a commercial product depends on the available resources. It needs to be noted though, that all parts of the user test can be performed without the use of such software but require additional manual work. Overall, it is a helpful solution but it would be advised to also have a paper version of the user test material available as a backup plan if the software fails.

A general feeling expressed by a few participants was that they felt they were being evaluated instead of the system. This feeling can affect the subject’s behavior, performance and answers, so this aspect was more explicitly clarified when the purpose of the study was explained in the beginning of each full tests session.

This pilot study was considered as partly internal because participants were chosen through personal contacts. For this reason, user satisfaction measurements were taken with scepticism, while feedback on improvements and proposed additions continue to be fully valid. Main satisfaction tendencies of the system could be observed.

Overall system satisfaction was high as it can be seen in Figures 5.8 5.9 and 5.10, with the majority of the participants having a mode above neutral and mode frequency above 0.5. However there was a clear drop in satisfaction about certain aspects, such as the results quality and presentation (with median 2 and 3 out of 5 respectively) and the development team concentrated on improving these aspects before the full user tests round.

In the full user tests, the overall success in finding relevant images (80.65%) using the KHRESMOI 2D image search prototype indicates an improvement over the percent (75%) that was reported in [129] as self assessment of radiologists about their image finding success rate using current tools. Case-based retrieval, was shown to be a more challenging task (78.95% success rate) which was expected by the results reported in KHRESMOI deliverable document D2.3 [109].

The average time over the successful tasks for the participants’ to select a relevant result was less than 3 minutes for both types of tasks (1 minute 46 seconds for image retrieval and 2 minutes 30 seconds for article retrieval). This is also below the average estimated time reported in [129] (between 5 and 10 minutes) and indicates an added value in terms of time efficiency when using the KHRESMOI system.

Regarding user satisfaction over the basic aspects of the 2D image search prototype (Figure 5.12), participants seemed strongly positive about the system response time. They had a positive opinion on the ability to correct mistakes and on the fact that the system can be used by all levels of users. The modification done to the results presentation and retrieval performance after the pilot study seem to have worked in a positive way, shifting the medians of satisfaction from 2 and 3 to 3.5 and 4.0 respectively. Participants were neutral over the system reliability, which can be explained of the presence of bugs and inconsistencies in the prototype.
5.1. FIRST EVALUATION ROUND

The users seemed to find the system novel and useful in practice giving a positive to strongly positive grade on this aspect (median 4 with a mean of 4.5). The activity that they gave a preference in intention of use was on academic work which goes along the design purposes.

A lot of feedback was given by the participants on the open questions, post-test discussions and spoken comments while performing the tests. Some confirmed the outputs of the pilot user study while many new comments and propositions were introduced.

On the graphical user interface aspect, the main comments were related to the image use, either requesting basic image manipulation features (which was also identified in the pilot tests but was not yet implemented for the full tests) or were about the image inconsistencies (e.g. drag and drop not being available on all views, detail views not being available for query images, non relevant marking being non-intuitive). Advanced text querying seemed to not be straightforward and several participants either used advanced queries or at least asked about the availability. These facts may indicate that a more comprehensive interface would be useful for radiologists.

Regarding the functionalities offered by the prototype, CBIR performance was again one of the common complains. Moreover, filtering out images of irrelevant modalities that appeared in the results was also a common request and relevance feedback results were questioned. Propositions about finding similar articles and finding articles using image examples were made by a couple of participants.

Overall, the system’s concepts were appreciated, such as the connection of articles and images and the trustworthiness of the results. An improvement over the results quality would result in a system with even more practical use. Moreover, even though most of the tasks were successfully performed the quantity and quality of resources returned in several scenarios was considered insufficient.

3D Image search subsystem

From the interviews already a lot of information was gathered. Due to time and work power constraints not all of them could be included in the first prototype version. Therefore the decision ones made to implement only one few. The user comments were prioritized and resulted in a final list of points that were included in the prototype:

- Marking one or more ROIs per slice and volume and options to remove the latest or all ROIs from the volume. Matching ROIs are overlaid in the detail view.
- Changing of brightness and contrast in the query and detail view.
- Options for changing between coronal, axial and sagittal pane.
- Corresponding report and extracted Radlex terms in the detail view.
- Information about zoom level and current slice ID in the query and detail view.
- A search history is added such that the user is possible to switch to previous detail views that were cached within the session.

Figure 5.13 presents the user satisfaction over basic aspects of the 3D image search prototype. System response time and results quality were the least satisfactory aspects while the participants were positive about the ability to correct mistakes, the results presentation and the system reliability.

The users seemed to strongly agree about the system’s novelty and practical usefulness in practice. The activity that they gave a preference in intention of use was academic work. A reason behind this choice could be the slow response times, as in clinical duty time saving is critical.

Using the preliminary search engine backend, the most time consuming process is querying the database with a volume and a ROI. It takes about 40 seconds until the thumbnail results are shown in the result view and another 15 seconds for a selected detail view to load. The lack of quality was the retrieval of matching ROIs, which currently did not satisfy the users needs.
Since some features that were mentioned in the interviews could not be added for the first prototype, the full user tests also pointed to some of those points indicating their necessity for the physicians. The following list shows the most important features and changes that were requested by the physicians during the interviews and user tests that should be implemented in the future version of the 3D image search prototype:

- Increase thumbnail size by either making them 4 times larger, or by letting the user choose the thumbnail size.
- Show the most significant slice according to the query ROI as the result thumbnail instead of always the center slice.
- Include more case relevant information in the result view by adding indicators/text for gender, age, modality, study/series description and a report summary or important report keywords. Additional information like patient preparation and confirmation tests (laboratory tests, biopsies etc.) would be beneficial.
- Additionally indicate whether the results are from the same user/study as the query volume.
- Change of brightness/contrast using the windowing method.
- Automatically setting the brightness/contrast values for the detail view to the one from the query view.
- The overlaid ROIs of the detail view should represent the estimated precision of the algorithm (e.g. matchings with higher probability should have a higher color saturation)
- Implement a drag and drop feature for the result view.
- Include full text search for reports, highlighting important key words.
- Make scrolling through volumes with a large number of slices more user-friendly by either including a scroll bar or by implementing the hold-left-mouse-button scroll function.
- Improve the response time of loading a result in the detail view.
- Improve the quality of the retrieval engine.

5.2 Second evaluation round

The first round of the user tests (Section 5.1) provided feedback for the further development of the system. The results indicated that the ideas behind the system design were appreciated by young radiologists. Novel functionalities offered, such as the linking of images and articles, were found to be helpful, while others, such as CBIR were reported to need improvements. The trustworthiness and the quality of the external sources that were indexed by the system were satisfactory. However, at the same time there were indications that the amount of indexed data was insufficient.

The findings of that evaluation round were used to redesign and improve the system. The final prototypes included improvements such as the indexing of a larger external resources corpus, the inclusion of automatic image modality classification and the revision of the CBIR algorithms [61]. 3D retrieval implementation achieved response times less than 5 seconds and the 2D image search prototype included basic image manipulation similarly to the 3D prototype. Integration of the two prototypes was obtained by extending 3D searches to external sources based on the consensus of the top results [123].

In this section, the final round of the user-centered evaluation of the KHRESMOI Radiology prototype is presented.
5.2. SECOND EVALUATION ROUND

5.2.1 Methods and Materials

The study protocol of the first evaluation round (described in Section 5.1.1) was slightly modified to focus more on assessing the newly added features. This section describes only these changes in methodology, as a full description would be redundant.

From the feedback of the first round of evaluation it became apparent that the ImageCLEF2012 indexed data was not fully adequate for the information seeking needs of radiologists. For this reason the PubMed Central dataset, consisting of 500,000 articles and more than 1,700,000 images, was used for the evaluation of the KHRESMOI system in terms of radiology–related information search as external sources.

The focus of the final user tests for the 3D prototype was on the pathology retrieval. Therefore the anatomical dataset was discarded and the pathology dataset was increased from 117 to 1163 Lung CT volumes. Also the number of labeled pathologies was increased to atelectasis, bullae, emphysema and ground-glass.

In order to decrease the run–time of a session the number of tasks was reduced from 10 (which was the first evaluation round tasks number) to 7.

In the 2D prototype, an image search task was created in order to evaluate an experimental feature, that included image search using the semantic web. The guided scenarios of the user tests were based on these information retrieval tasks and included use of the newly added features, such as filtering by modality and image manipulation.

In order to evaluate the quality of the 3D prototype the users were given two similar tasks. For each of them a case was randomly chosen from six example cases. Both of the tasks included the following:

- The unique ID and the pathology specific slice of the volume.
- We assume that there is no report available for this volume.
- If you know one or more of the pathologies visible in the given slice try to verify it/them using the image search feature. Try to find a better visual representation of the pathology that could be shown as an example.
- If you don’t know what the visible pathology is, use the image search feature to find out what it might be and try to find a better visual representation of the pathology that could be used as an example.
- Store results that were useful for you in the personal library.

The focus of the first task was to also take a specific look at the consensus and non-consensus suggestions of the algorithm which were indicated in the result view in green and white respectively. This was done in order to identify if the automatic division is useful for the physicians.

In the second task the users were advised to specifically look at the additional 2D/Article search tabs as the task was also about finding out if the integration of the 2D prototype was useful for the physicians.

The six example cases were given by the unique ID of the volume and a specific slice and contained the following pathologies:

- Centrilobulary emphysema
- Bullae emphysema
- Fibrosis
- Bronchiecstasis
- Ground-glass
- Honeycombing
The variety and randomization of pathologies ensured that the user satisfaction was not based on one specific pathology retrieval.

Finally, the survey forms were modified accordingly to include questions on the new system features and user study tasks.

5.2.2 Results

The final user tests took place at the University hospitals of Geneva, Switzerland, the Medical University of Vienna, Austria, the University Hospital of Freiburg, Germany and the University Hospital of Larissa, Greece. In total 26 users participated in the tests, split into 17 male and 9 female. The age distribution shown in Figure 5.14 ranges from 20 to 60 years. There are 7 different native languages within the participants, with a majority of eighteen speaking German, while two speak Greek and one for each of the following languages: Hungarian, Romanian, Macedonian, Croatian, French. They self assess their English skills to a median of 4 on a scale of basic-1 to native-5 as shown in Figure 5.15. Sixteen of the participants were residents, two where consultants, two associate professors and one for each of the following positions: general physician,
junior consultant, vice chairman, head of unit and attending radiologist. The distribution of their years working in radiology is presented in Figure 5.16. Almost half of the participants specialized in the fields of thorax, emergency radiology or intervention radiology. The full distribution is shown in Figure 5.17. The demographic information about usage of computer and internet is summarized in Figure 5.18. All of the participants use a computer in their day-to-day life as well as for their job or educational related tasks. Also search engines (Google) are used more than once a day. The Internet is used to search for health related information at least once a day and by 75% for more than once per day. Most of the participants (eighteen) use the Google image search at least once a day and eight of the participants do not use any social network.

The user satisfaction is analyzed over key general aspects of the system and is presented in Figure 5.19. Results are compared to the one of the previous prototype evaluation where possible. The median for the question about intention to use the system frequently increased from 4 to 5. The same median was obtained for straightforward design, the easiness to learn how the prototype works and that there is no special training required. The consistency of the prototype increased from a median of 3 to 4. A median of 4 was also obtained for the remaining satisfaction criteria.

In order to assess the global satisfaction of each participant the mode over the general satisfaction question was measured and combined with its frequency in Figure 5.20.

The average run time and the time consumption of a user test is visualized in Figure 5.21. The left hand side shows that the average total runtime of one user test was about 65 minutes with the colors indicating the connection of the part to the introduction in blue, the surveys in green, the 2D prototype in orange and the 3D prototype in yellow. As expected the most time consuming parts include the 3D prototype taking almost 40% of the total runtime, shown on the right side of Figure 5.21.

In the following sections a detailed description of the results of each subsystem prototype evaluation is given.

2D Image/article search subsystem

The results of the 2D Image Search subsystem evaluation are presented in this section. The average time for finding the first result for the 2D image task was 102 seconds. For the task using the Semantic PImage Search it was 86 seconds. The article search task took on average 159 seconds to find the first relevant result. The success rates for the 2D image task, semantic image task and article search were 100%, 100% and 85% respectively. The respective average numbers of found
CHAPTER 5. USER-ORIENTED EVALUATION

Figure 5.17: Field in radiology of the participants.

Figure 5.18: Demographic information about computer and internet usage of the participants.
5.2. SECOND EVALUATION ROUND

Figure 5.19: Median values of measuring general user satisfaction about the system in Likert scale (1=strongly negative, 5=strongly positive) compared to the median values of the first prototype evaluation.

Figure 5.20: Mode and mode frequency values for each participant over the global satisfaction questions in Likert scale and % of questions.
results were 5.2, 4.5 and 3.0.

The user satisfaction over specific aspects of the system collected from the questionnaires is given in Figure 5.22 in comparison to the results of the previous prototype. It is graded on a Likert scale where 1 was strongly negative and 5 was strongly positive. The results satisfaction and presentation increased by a scale value of 2, to 4 and 5 respectively. Also with the response time and error correction support the users were more satisfied with an increase from 4 to 5. To assess the global satisfaction of each participant, the mode and its frequency over the questions is given in Figure 5.23.

Below follow the comments about the 2D Image Search subsystem, obtained by the open questions and the observations of the task performance. They are organized in three main categories, according to the part of the system that they refer to and their type: Frontend, Backend and Bugs. The Frontend list contains comments regarding the interface of the system. The Backend list contains comments regarding the tools functionality (e.g. search results, modality filtering etc.). The Bugs list contains irregularities causing erroneous behavior of the system. A number at the end of the line indicates the number of participants having this comment, while general observations by the observers do not have these numbers.

- Frontend
  - Drag and drop, participants expected to directly drag and drop items without selecting them. - all
  - More clear workflow structure, current order: top left, top/mid right, mid/bottom left, bottom right.
  - Windowing function in 2D needs to be more sensible.
  - After making a selection in a drop-down menu, continuing is not intuitive (it overlies the search button, users have to click somewhere else or exactly at the top of the drop down menu to close it).
  - Participants would like to use keyboard to jump to the words starting with specific letter in the drop-down menu, scrolling through is frustrating. - 5
  - Extraction of semantic terms should be automatic when typing and be given as a suggestion -2
  - Result images should be larger, show e.g. 2 rows with 3 images each. - 5
  - In the detail window participants always have to scroll down to view (larger) images. - 5
  - Participants have a hard time selecting from the drop down menu (semantic image tool), they have to click several times until its marked as selected, maybe the click area is too small? - 5
5.2. SECOND EVALUATION ROUND

Figure 5.22: User satisfaction survey on the 2D/article search subsystem compared to the previous prototype.

Figure 5.23: Mode and mode frequency of overall satisfaction on the 2D/article search subsystem per participant.
– Difficulty finding the modality search filter - 2
– Sorting results by modality was found to be very helpful - 6

• Backend

– When selecting multiple modalities the participants expected that OR is used (instead of AND, where no results showed up) - 6
– Participants expected more results to show up, especially for the first task with only about 12 osteoporosis images. - 8
– Results are too rare/specialized and would not help during clinical routine, would like more general results - 5
– Include Radiographics as additional source - 1
– Users are irritated when the exact same images show up that were marked as non-relevant. - 2
– Semantic Image Search missing anatomy brain as option, should accept terms that are not listed. - 2
– When searching for a specific modality, images from other modalities show up (e.g. CT → x-ray results). - 4
– Negative results should affect the modality filtering - 3

• Bugs

– When adding multiple images the placement order of dragged 2D images seems to be inconsistent.
– Using right click, add as non-relevant in Semantic Image Search causes the system to switch back to the 2D Image Search.
– The Select Sources tab sometimes does not show up when trying to open it in the Image Search Perspective.
– Query with only non-relevant images is not working. - 4
– Using images only and no text does not give any results. user does not know why (text is required, e.g. do not allow search when textbox is empty!) - 1
– For some time no images were returned for a query, but it worked again after about a minute. - 2
– Images did not load immediately with search - 1
– The image in the search query was on the far left and not visible (only its boundary), and there was no scroll option.
– Clicking image search after article search enables additional undesired resources in the image search.
– Drag and drop from article detail view was not working, had to press add to query - 3
– Personal library image was zoomed in and caused the system to slow down - 10
– cannot use keyboard arrows to move between personal library items. - 1
– In the personal library participants do not realize that they are looking at an image with a different tag. - 3
– Personal library was not available. - 7
– Reset content does not remove tabs from DETAIL view.
– The result tab showed up in the query tab.
This section covers the specific results of the 3D Image Search prototype. The 3D prototype was evaluated by 25 participants, as during one user test the 3D system became unavailable. The user satisfaction over specific aspects of the system collected from the questionnaires is given in Figure 5.24, where the current results are shown in red and the results from the previous user test using the first prototype in blue. The biggest difference is the increase of 2 scale points in terms of response speed and result satisfaction, reaching median values of 5 and 4 respectively. The use of the prototype for research work also increased from 4 to 5, while the other results did not change. To assess the global satisfaction of each participant, the mode over the questions together with its frequency is given in Figure 5.25.

The average time until the first matching result is found using the 3D prototype is 259 seconds, while the users spent an average of 448 seconds in total for a 3D task. The success rate of finding matching results is at 86%.

Below follow the comments about the 3D Image Search subsystem, obtained by the open questions and the observations of the task performance. The organization is the same as Section 5.2.2.

- **Frontend**
  - Make images larger, fill the unused space, especially in the ROI marking view (volume-only-view). - 5
  - Cannot stop / cancel loading volume in the result view - 4.
  - When using the advanced filter option, also mark the matching term in the detail-view report - 2
  - In the article detail view only mark the searched terms, not the rest - 2.
  - Participants would like to use multiple filters in the result view (e.g. other patient + consensus + age) - 4
  - Participants did not realize whether marking mode was enabled or not - 4
  - Make the switching between perspectives more easy, 1 button instead of having to use menu. - 6
– There is no progress information for loading the volume. - 7
– There is no progress information about if the integrated 2D/Text-tabs will show up. - 6
– There is no information in the detail view if the volume is from the same patient or not that can be immediately noticed - 5
– Participants would like to be able to search directly from the marking view in index perspective
– Using the result view for marking ROIs is difficult, brightness/contrast is slower, ROI overview is missing - 4
– The slice of the query volume in the result view should be the slice where a ROI was marked in - 10
– User would like to modify drawn ROI by moving, changing - 3
– Users would expect to adjust window levels when moving mouse to the whole screen (out of the volume) - 8
– Zoom in/out and pan buttons should be in volume-only-view - 1
– Participants wanted to use double click to open volume, but this causes different behaviour (changes perspective) and they got confused. - 9
– Use the windowing setting of the query volume for result list - 5
– Predefined/preset windowing options (bone, lung, ) - 9.
– Show windows level details in numbers. - 1
– Participants tried to use window function on thumbnails, would be nice to have - 7
– It would be nice to have more slices in the result thumbnails - 2
– Filtering the results using the advanced filter is very useful - 5
– Less color in the text, only mark searched terms. - 2
5.2. SECOND EVALUATION ROUND

- Color coding was helpful - 1
- Blue and green hard to see, red-green blind - 2
- Better representation and structuring of the report. - 6
- Search button was expected to be at top left due to common habit (current PACS system in Vienna) - 7
- In the result view the number of slices would be interesting. - 1
- Labeling of the 2D/Article/BigData tabs is confusing - 1
- Limit search to patient returns nothing if there are no other recordings, participant thought its not working. - 1

• Backend

- Participants selected multiple pathologies the first time. - 3
- The system should be able to find more specific/higher resolution features, only large bullae but not smaller ones were detected.
- There are different demands of features and resolution based on modalities and organs.
- Location of the pathology in the lung important (e.g. upper lobe) - 2
- Only search within similar recordings, e.g. do not return MIPs. - 5
- All results (also text search) only anatomy specific, e.g. if searched in lung do not return rib fractures, aorta, - 2
- Textsearch, automatic text query for 2D is not so good, using own words is very useful - 7
- Change order of detail view, s.t. results are in the middle, and detail on the right - 13.
- Having consensus of organs different from the anatomy where the ROI was in is strange - 2.
- Use DICOMs instead of 16bit JPEGs to improve the windowing function. - 2
- Exam protocol should be included in the result view.

• Bugs

- Volume view different from radiologists, they look from the feet up (and not from the top down as implemented).
- Zooming is problematic, participants cannot use windowing and cannot mark ROIs, also zooming in is slowing down the system. - 5
- The loaded volume did not match the selected one. updated a few times until the correct volume was finally shown. (assumably caused by clicking on a lot of volumes, without loading them) - 2
- When clicking/selecting an index image using a filter-field, right view does not update accordingly, cannot compare directly if its the one i selected in index view - 12
- Rapidly clicking onto several volumes in the index view made the system slow, loading volume did not work, loading volume overview on the right of index view did not work anymore - 1
- The same patient was not indicated as same patient, although the query volume was the same. - 4
- Prototype ran out of memory, no notification for the user, images just did not load. had to restart application - 8
CHAPTER 5. USER-ORIENTED EVALUATION

5.2.3 Discussion

A total number of 26 persons participated in the final user tests of the KHRESMOI radiology prototype. The majority of users were young and mid-age with ages between 20 and 40 years and were involved into various radiology specializations. The level of experience in radiology also spread the whole spectrum of 0 to > 10 years. As we already realised at the previous user-tests recruiting radiologists was a difficult task, as they usually are on a busy schedule. Nevertheless we were able to recruit 20 radiologists in Vienna for the user tests. More recruitments were found in hospitals not associated with the project, to get broader feedback from radiologists with different backgrounds. Thus, 2 participants were recruited in Germany, 2 in Greece and 2 in Geneva, resulting in the already mentioned total of 26 participants, which is about the participants number that was aimed at. All participants were able to complete the 2D prototype tasks and all but one were also able to complete the 3D ones (due to network issues the 3D prototype became unavailable once). It is interesting to note that all the participants use the internet and internet search engines at more than once a day, either private or in their job/education.

Main tendencies on the user satisfaction could be identified in Figure 5.19. Compared with the previous prototype the user satisfaction to use the system frequently increased from 4 to 5. The consistency was rated with a value of 4 which is also one value higher than previously. Regarding the global aspects of the system, users found that the system’s design is straightforward and that is also easy to learn, without prior training. However, from the comments and the tester point of view a tutorial session would be beneficial. The main tendency of the users stayed strongly positive, with 10 persons giving a mode of 5 with frequency of at least 0.5 over the general satisfaction questions, which is also shown in the normalized histogram over all the given ratings shown in Figure 5.26.

In the following the results of the 2D and the 3D subsystems are discussed.

2D Image/article search subsystem

The average time taken to complete the 2D image–based tasks were 102 and 86 seconds respectively. This is an improvement over the 106 seconds reported in the first round of the user–centered evaluation. The success rates of 100% for both image search tasks are higher than the reported 80.65% in year 2. Note that the times and percentages reported in the first evaluation round were already better than the self assessments of the image search survey (Chapter 2). The time taken in information finding in the article search task is a little higher than year 2 (159 seconds compared to 150). However, the success rate of 85% is higher than the previous 78.95%. The mean number of relevant image found was slightly bigger than the year 2 prototype (5.2 and 4.5 compared to 4) even

![Normalized histogram of the general satisfaction questionnaire.](image)
5.2. SECOND EVALUATION ROUND

though the indexed dataset was significantly larger. This may have been caused by users stopping to search for additional results after the first relevant one was found, due to the time constraints of the participating physicians, albeit the instructions to search for as many as possible.

Regarding the user satisfaction, the most satisfactory aspects of the system seem to be the response time, the results presentation and the ability to undo errors, all of them achieving a median of 5. The other parts of the system also obtained medians above average (4) on the Likert scale. The novelty of the tools is also recognised from most of the participants, getting a median of 5. The use of the system in academic work was found more probable than use in research work or clinical duty achieving a median of 4.5 compared to 4 and 4 respectively.

The newly added features, such as modality filtering, semantic image search and image manipulation were satisfactory as well achieving mean grades of 4. From the feedback received in the open questions and the free discussion after the session, it was often mentioned that the semantic image search tool should use the autocomplete feature instead of dropdown menus for each field. Also regarding the modality filtering, even though very useful should be more accurate, as several images where wrongly classified. There was a request that the relevance feedback should affect the modality filtering which is currently not taken into account. The interface was often found crowded or badly structured and a cleaner, bug–free and better structured version should be considered for a real world application.

Interesting suggestions for additions to the system were given, such as the ability to find similar articles when querying articles or images. The option to make the link of internal and external sources operate in both ways, was also mentioned. Currently the system offers the ability to have a search in internal sources extended to external sources. However, when an interesting case is found in the medical literature, searching for similar cases into the hospital records can also be of interest.

3D Image search subsystem

With a success rate of 86% most of the users found a satisfying result within an average time of 260 seconds. The differences to the Y2 prototype in Figure 5.24 show that the performance of the two worst performing points of the user tests could be drastically improved. The decrease in response time for a query decreased from 15 to 3 seconds and the loading time of the volume details from 45 to 17 seconds, resulting in a new satisfaction value of 5. Also the quality of the retrieval improved, now showing a satisfaction value of 4. The increase for use of the 3D prototype for research work from 4 to 5 may come from the integration of the 2D prototype, as these tabs give more information concerning research of a certain topic. While the overall response to the 3D prototype was very positive the physicians had some additional suggestions that could help improve the work-flow and efficiency in using the prototype.

In general the unused space should be used to increase the size of the thumbnails and volumes for improved visibility. In the ROI marking mode the volume could be shown in full-screen, as this is the only thing visible at this point.

After pressing the search button, the query volume shown in the result view should by default show the slice where the ROI was marked. Manually searching for the correct slice is redundant and time consuming.

The users mentioned again that presets for brightness and contrast windows would be very useful. Also the set brightness and contrast window of the search volume should be automatically transferred to the result view and all the volumes shown there, this would even remove the necessity of being able to change contrast and brightness for thumbnail images which is currently not implemented. The brightness and contrast manipulation mode should be available even when moving the mouse cursor out of the volume, as this was expected by the users.

Another important comment of the users was to give them information about the progress of loading a volume and whether the integrated 2D search tabs will give any results or not.

The coloring of the report satisfied some of the users, but dividing it into its basic structure (technical information, medical finding, diagnosis) would be more helpful to them.

In terms of retrieval quality the users suggested to include information about the location of the pathology. Most important only return results for matching anatomies and second look at the
local information of the pathology within the anatomy (e.g. search for pathology in the upper lobe of the lung). It would be also useful to limit the search to recordings with similar properties. Showing results from maximum intensity projections is not useful when searching for an normal Lung CT.

The users also mentioned additional scenarios where the system could be useful. For example if the head of radiology is currently not available, or if they would want to show students examples of certain pathologies (e.g. emphysema).

During the tutorial and free use of the 3D prototype participants searched for a variety of other pathologies. Most frequent and interesting ones were searches for granuloma, metastasis and pleura effusion. These could be specifically looked at in order to improve the quality of the retrieval engine.
Chapter 6
Conclusions and Perspectives

Image search is a regularly performed task for clinicians, especially in image–based domains such as Radiology. The ever–growing amount of visual data in institutions and the medical literature need to be indexed in an efficient manner in order to be beneficial for radiologists. Visual information needs to be easily accessible and strongly linked to the cases it is associated with.

In this context, medical information systems need to follow a user–oriented design in order to have an impact in clinical practice. User needs need to be fully explored so that the implementation of the system provides tools that address them. End users need to be included in the development cycle and assess the functionality of the tools.

One of the main goals of this thesis was to assist the medical community in their clinical, teaching and research activities. It is a study in the field of information retrieval, improving current information search capabilities and providing tools that can be used to access trustworthy information in an effective and time efficient manner. Using such tools, clinicians can have quick access in valuable information for their diagnostic and academic work.

In more details, this thesis presented the full development lifecycle for a medical image retrieval system. Initial insights about the current information requirements of radiologists are given. Novel image retrieval techniques were assessed in the medical image domain and complex but efficient indexing and retrieval pipelines were designed. A protocol for a user–centered evaluation of medical image retrieval systems that support CBIR was designed and carried out.

6.1 Achievements and limitations

A summarization of the achievements of this study, with a critical view on its limitations and weaknesses is given in this section.

6.1.1 Image use and search behavior survey

The results of this survey are a first step to better understand the requirements of radiologists in handling images and searching for visual data that can help them in daily tasks. It has to be mentioned that many radiologists are not familiar with visual retrieval, so being able to show them prototypes and having them work with the prototypes was a necessary step in order for them to understand problems and potential and to make it easier to formulate desires for a perfect search system.

Obtaining a very large number of responses from often busy and overloaded radiologists was difficult. 34 radiologists responded to the questions asked within the three months of the online and paper survey. Most persons responding were junior (below 30 years) and with less than 5 years experience. This has the advantage to have persons who grew up with the Internet and digital image handling but the inconvenience that they might not question current practices and might have had fewer situations where they were lacking crucial information in clinical work. The current Internet generation is also plagued by the problem that they often believe to be competent
CHAPTER 6. CONCLUSIONS AND PERSPECTIVES

information searchers, which does not always correspond to reality [183] — particularly in terms of how to use the information found.

Here, the most important aspects are listed to give a complete picture.

• **Role of image search**: The search for images and similar cases is an essential part in the radiology workflow. During the assessment of clinical data they use information from other images obtained from multiple sources: reference books, communication with other radiologists, personal files, the hospital database, and the Internet (both specialized databases such as PubMed, or general search engines such as Google). Radiologists allocate a significant amount of time to searching but fail in a substantial number of cases (around 25%).

• **How and what to search for?** Keyword search is currently the dominant search modality, including Internet search, and access by patient ID in clinical records, or the oral communication with colleagues. During result selection, experience plays a dominant role when analysing and choosing images. This indicates that substantial prior knowledge is necessary to perform efficient and successful search. Communication among colleagues is used to share knowledge not only during training but also in clinical practice. Past cases store experience of other colleagues and can make this experience available in a systematic way. Trust in the information found and evidence for a particular diagnosis are important. This can be more easily confirmed in communication with colleagues than when searching other sources such as the Internet. The scientific literature has an advantage over general Internet sources. Visual retrieval is little known although first prototypes exist such as IRMA (Image Retrieval in Medical Applications) and MedGIFT (Medical GNU Image Finding Tool).

• **Limits of current search**: There is clearly room for improvement considering the allocated time and the success rate of current image search. This is consistent with the perception of radiologists who conclude that the obstacle for finding images is not the availability but the limits of search technology or novelty of the data. Keyword search is perceived to have limits as an accurate prior assessment of the present case is required before formulating a query. A tedious selection of results based on individual inspection of potentially ambiguous candidate images is then necessary. This is limiting in the case of rare diseases, where search and comparison with other examples might be most relevant but little prior assessment is feasible. A related limitation is the lack of comprehensive keyword assignments in reference databases. Keywords are ambiguous and only using terminologies can help in this respect, which is currently uncommon. Many radiologists build their own personal reference databases to compensate for searchability. In many institutions the lack of institutional archives requires this. Another way is storing patient IDs of interesting cases with short textual annotations in files that allow finding cases. The dominant role of experience, the emergence of scattered personal reference databases and the culture of communication among colleagues suggest that facilitating the sharing of knowledge and removing requirements of prior assessment and keyword identification can have significant impact on the radiologists’ clinical work, teaching and research.

• **Wishes for future systems**: Suggestions for future search are consistent with limits of current search. Radiologists name search for pathology as a goal. However, they value images equally to keywords as potential query inputs and suggest the use of ROIs to obtain more specific search results. Other functionalities include limiting the search to modalities and to include textual data into the visual retrieval, which could be achieved by faceted filtering of the search results. The idea of trust or confidence in a diagnosis of a case was also mentioned to be important. An important aspect is to link search results and cases with the literature. The peer-reviewed literature offers a level of trust.

6.1.2 System design and empirical evaluation

New decision support tools are required to assist radiologists in their daily work and help them cope with the increasing amount of data that they need to analyze daily. Medical CBIR is a
promising technique that can be used for a large variety of scenarios from keyword search to visual search of full images and regions or volumes of interest.

An image retrieval system combining CBIR and text–based search was developed in this thesis for accessing images from the medical literature and integrated into the KHRESMOI system. Complex but efficient indexing and retrieval pipelines were implemented to cover the system requirements as those were dictated by the user information needs. Empirical assessments of the most important components of these pipelines were made on publicly available datasets.

It became evident from both the empirical and the user–oriented evaluation that visual information is not enough to achieve a satisfactory retrieval performance. One of the biggest challenges in CBIR is the “semantic gap”. The term “semantic gap” is used to address the difficulty in using low–level visual features, such as color, shape and texture, to describe high–level concepts, such as a pathology or an anatomic location. Techniques emerging from fields such as pattern recognition and machine learning can be used to bridge this “gap”. These techniques require manually annotated data sets for training and evaluating such a CBIR system. Thus, medical CBIR performance could greatly benefit from collaborations between medical experts and computer scientists for creating such data sets.

Secondly, visual information does not always contain all the necessary information for its interpretation. That being said, additional clinical context such as demographics and lab tests is often required in combination with the visual data to reach a potential diagnosis. Thus, the use of multi–modal information should lead to a more complete description of a relevant medical case. Moreover, abnormalities can be very subtle and include visual differences that many mathematical models ignore. CBIR systems have to take this into consideration when trying to meaningfully represent the visual content of the medical images.

The size of the visual data available is also a challenge that is encountered by CBIR systems. With the development of ParaDISE this study attempts to assist the CBIR research field by providing a platform for conducting large–scale evaluations of visual feature extraction, image representations, indexing schemas and retrieval pipelines.

6.1.3 User–centered evaluation

Most medical CBIR systems are usually developed in computer science labs and only few reach the clinicians. Development of such systems needs to be user–oriented and evaluation in real datasets needs to be carried out in order to prove effective in real life scenarios.

An iterative approach of user–centered evaluation throughout the development lifecycle can provide more diverse results on the system usability and keep the system development on track with the original requirements. Results show that radiologists quickly feel comfortable in using new search tools, such as CBIR, relevance feedback and querying using complex statements. More importantly the systems concepts were appreciated, such as the connection of articles and images and the indexing of trustable sources.

An improvement over the retrieval and classification quality would result in a system with even more practical use. Moreover, even though most of the tasks were successfully performed, the quantity and quality of resources returned in several scenarios was considered insufficient.

Radiologists have a tight time schedule and are difficult to recruit. This resulted in having a relatively low number of users for quantitative analysis. For this reason, absolute quantitative results need to be taken with caution and serve mostly as indicators and as relative comparison. The user tests were performed in a lab room of a hospital and not in a room with diagnostic activity and standard viewing stations. This makes it difficult to assess the impact of the system on the actual clinical workflow of the radiologists. Moreover, because of the early stage in the system development and the low number of participants, no A/B testing was performed to compare the system with other current solutions.
6.2 Perspectives

This section provides some insight for future directions in development and user–centered evaluation of medical document image retrieval systems.

6.2.1 Radiology information needs

The results of the user–centered evaluation study are useful as additional specifications for medical image retrieval system design and can assist in avoiding potential pitfalls. Insights in the methodology for conducting more meaningful user–centered evaluations are also provided. The feedback obtained should be taken into account on the development of future medical retrieval systems.

As discussed in Section 6.1.3 the performed usability study consisted of a short one hour session on specific tasks in a lab room. In order to assess the long–term impact of a system, it needs to be integrated into a clinical environment for a long period. In this manner, it will be possible to evaluate also the organizational and social impact of the system. Query log analysis and regular interviews with the end users can potentially give a more complete view on the shortcomings of the system and bring into the light additional tools needed. For such an integration, first the evaluated system has to reach a high level of performance.

For general purpose image retrieval systems such as ParaDISE that provide a large bank of features, it maybe difficult to clinicians setting up a pipelines suitable for their medical applications. Automatic feature selection and feature attribute selection may address such issues and alleviate the system setup complexity without harming the flexibility. Graphical user interfaces that will explicitly allow the user to decide which visual characteristics are more important for a each specific query can be a next step towards better image retrieval performance.

After the integration of a system in a clinical environment, long term machine learning could be applied to automatically select or weight feature depending on the query. Graph theory could also be used to map visual features to semantic ontologies through the interaction of users with a visual retrieval system.

6.2.2 Future of medical document image retrieval

One of the main shortcomings of medical CBIR can be addressed by combining it with text information. The addition of semantic search is also of interest, to take advantage of the structured knowledge of the medical ontologies such as RadLex (Radiology Lexicon) and MeSH (Medical Subject Headings). These ontologies of offer links between medical terms and mapping visual features to them could also potentially help bridging the semantic gap.

The visual representation of the images can benefit of the recent prevalence of deep learning. Learning metrics, features and image representations are some potential applications for this. Modality classification can also be improved by creating more broad and representative training datasets for machine learning. Classification of extracted subfigures could also provide more information.

Parallel architectures are necessary in indexing very large image collections and regularly updating them. Fortunately, the indexing of image visual features is a highly parallelizable task. Easy–to–deploy computer clusters can be more widely used in research labs by using available workstations in MapReduce implementations such as Hadoop.

Compact representations and efficient indexing structures are required for achieving fast retrieval in a reasonably sized index, with respect to the available hardware resources of medical institutions. New coding algorithms such as Compact KdTrees and Product Quantizers that provide a good performance vs. speed tradeoff could be applied in medical content based image retrieval. ParaDISE could be used as a platform to implement these techniques and evaluate them on medical datasets.

ParaDISE will be publicly released so that other researchers can develop, finetune and evaluate their approaches in large–scale datasets. An open–source community evolved around ParaDISE could accelerate the development of techniques that can better model the multi–modal information...
6.2. PERSPECTIVES

of medical cases. In this way, medical information retrieval research could result into tools that will better assist clinicians with their decision making.
Appendix A

Survey questionnaires
Radiology activity Questionnaire

This survey will help us understand how radiologists use reference images and literature during the assessment and reporting of medical imaging data in their daily routine work. The goal is to find points where we can help improve the efficiency and effectiveness of the search for those other images.

Within the European Khresmoi-Project it is our goal to improve the usability of the Radiology Software using State of the Art Computer Science technologies. We would therefore appreciate your help by answering the following questions. This questionnaire is anonymous and all the information provided in this form will only be used for statistical purposes.

Information retrieval by using the visual image content is one of the goals of the project such as finding similar image series or similar regions of interest in, for example, the medical literature, teaching files or eventually the entire PACS. Such retrieval can be on internal searches such as personal directories, the PACS/RIS or on the Internet. By combining clinical data with images, similar cases can be found from a patient record and compared to a new case.

Please take a few minutes of your time to help us gather important information, which will be used to improve usability of existing and develop new State of the Art Radiology Software to make your future daily Radiology Routine easier and simpler.

The form is divided into following questions: General (7) and Activity (A - Clinical work, B - Teaching, C - Research (10 each))
A. General

1. Age: □ <25 □ 26-30 □ 31-35 □ 36-40 □ 41-55 □ >55

2. Gender: □ Female □ Male

3. Specialization: …………………………………………………

4. Country of radiology specialization:
………………………………………………………………………………

5. Activity Area:

□ Public Hospital □ Private Clinic □ other:
………..

6. Years of experience in radiology:

□ <2 □ 3-5 □ 6-10 □ 11-15 □ 16-20 □ >21

7. Activity focus distribution in percent:

B: Clinical work …….%

C: Teaching …….%

D: Research …….%

As it is also important to analyze trends and developments we would appreciate your help to further improve our research by participating in future (yearly) surveys. If you want to help us please enter your e-mail address here:

e-mail: ……………………………….
B. Activity – Clinical work

1. Please, list the most important tasks where you need to find images other than the ones of the patient that you are treating (such as searching for reference cases/images in books or teaching files, or finding images on the Internet). For each of these tasks, give an example of the kind of images you would try to find.

<table>
<thead>
<tr>
<th>Task</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. When performing each of these tasks, where exactly do you look for the images (Internet search engines, your personal files)? When looking for these images, how do you look for them (key word search, personal structure, patient name, asking a colleague…)?

<table>
<thead>
<tr>
<th>Where</th>
<th>How</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. When you find images, how do you decide whether one (or more) are suitable for your needs?
4. How often does your search for relevant images fail? (choose one)

(rarely) □ <20%  □ 21-40%  □ 41-60%  □ 61-80%  □ >81%

(often)

Why do you think you did not find any relevant images? (choose one)

□ probably not available
□ should be available, but could not find them
□ I don’t know
□ other: ........................................................................................................

In which situations do you not find the images you were searching for?

5. How much time does it take before you stop searching for relevant images? (in minutes) (choose one)

□ <1 min  □ 3 min  □ 5 min  □ 10 min  □ 15 min  □ >15 min

When you find relevant images, how much time does it on average take you? (in minutes) (choose one)

□ <3 min  □ 5 min  □ 10 min  □ 15 min  □ >15 min
4. What would be useful additions for an image search system (a sort of Google for medical images) ...

   and useful functionalities? (search for...)
   □ modality
   □ patient demography
   □ pathology
   □ similar images
   □ other: .................................................................

5. What would a perfect search system for images look like?

6. How could an automatic system exploit visual information for information search?

7. Tools for an automatic annotation of images using only the visual information in the images exist; what kind of annotation would be most useful? (choose one)
   □ Modality
   □ Anatomic region
   □ Quantification of size
   □ other: .................................................................
4. Are there any radiology/medical terminologies that you use for searching or describing the images/cases?

☐ RadLex  ☐ MeSH  ☐ none
☐ other: …………………………………………………………………………

C. Activity – Teaching

1. Please, list the most important tasks where you need to find images other than the ones of the patient that you are treating (such as searching for reference cases/images in books or teaching files, or finding images on the Internet). For each of these tasks, give an example of the kind of images you would try to find.

<table>
<thead>
<tr>
<th>Task</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. When performing each of these tasks, where exactly do you look for the images (Internet search engines, your personal files)? When looking for these images, how do you look for them (key word search, personal structure, patient name, asking a colleague...)?

<table>
<thead>
<tr>
<th>Where</th>
<th>How</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. When you find images, how do you decide whether one (or more) are suitable for your needs?

2. How often does your search for relevant images fail? (choose one)
   
   (rarely) □ <20% □ 21-40% □ 41-60% □ 61-80% □ >81%

   (often)

   Why do you think you did not find any relevant images? (choose one)
   
   □ probably not available
   □ should be available, but could not find them
   □ I don’t know
   □ other: .................................................................

   In which situations do you not find the images you were searching for?

3. How much time does it take before you stop searching for relevant images? (in minutes) (choose one)
   
   □ <1 min □ 3 min □ 5 min □ 10 min □ 15 min □ >15 min
When you find relevant images, how much time does it on average take you? (in minutes) (choose one)

☐ <3 min  ☐ 5 min  ☐ 10 min  ☐ 15 min  ☐ >15 min

1. What would be useful additions for an image search system (a sort of Google for medical images) ...

and useful functionalities? (search for...)

☐ modality  ☐ patient demography  ☐ pathology

☐ similar images  ☐ other: ......................................................

2. What would a perfect search system for images look like?


3. How could an automatic system exploit visual information for information search?


4. Tools for an automatic annotation of images using only the visual information in the images exist; what kind of annotation would be most useful? (choose one)

☐ Modality  ☐ Anatomic region  ☐ Quantification of size

☐ other: ...........................................................................
1. Are there any radiology/medical terminologies that you use for searching or describing the images/cases?

☐ RadLex  ☐ MeSH  ☐ none

☐ other: ........................................................................................................

D. Activity – Research

1. Please, list the most important tasks where you need to find images other than the ones of the patient that you are treating (such as searching for reference cases/images in books or teaching files, or finding images on the Internet). For each of these tasks, give an example of the kind of images you would try to find.

<table>
<thead>
<tr>
<th>Task</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. When performing each of these tasks, where exactly do you look for the images (Internet search engines, your personal files)? When looking for these images, how do you look for them (key word search, personal structure, patient name, asking a colleague...)?

<table>
<thead>
<tr>
<th>Where</th>
<th>How</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. When you find images, how do you decide whether one (or more) are suitable for your needs?

4. How often does your search for relevant images fail? (choose one)

* rarely
  - □ <20%
  - □ 21-40%
  - □ 41-60%
  - □ 61-80%
  - □ >81%

* often

Why do you think you did not find any relevant images? (choose one)

* □ probably not available
* □ should be available, but could not find them
* □ I don’t know
* □ other: ________________________________________________________________
In which situations do you not find the images you were searching for?

4. How much time does it take before you stop searching for relevant images? (in minutes) (choose one)
   - [ ] <1 min  
   - [ ] 3 min  
   - [ ] 5 min  
   - [ ] 10 min  
   - [ ] 15 min  
   - [ ] >15 min  

   When you find relevant images, how much time does it on average take you? (in minutes) (choose one)
   - [ ] <3 min  
   - [ ] 5 min  
   - [ ] 10 min  
   - [ ] 15 min  
   - [ ] >15 min  

5. What would be useful additions for an image search system (a sort of Google for medical images) ...

   and useful functionalities? (search for...)

   - [ ] modality
   - [ ] patient demography
   - [ ] pathology
   - [ ] similar images
   - [ ] other: ..........................................................
4. What would a perfect search system for images look like?

5. How could an automatic system exploit visual information for information search?

6. Tools for an automatic annotation of images using only the visual information in the images exist; what kind of annotation would be most useful? (choose one)

- [ ] Modality
- [ ] Anatomic region
- [ ] Quantification of size
- [ ] other: .................................................................

7. Are there any radiology/medical terminologies that you use for searching or describing the images/cases?

- [ ] RadLex
- [ ] MeSH
- [ ] none

- [ ] other: .................................................................
This questionnaire is anonymous. All information provided in this form is confidential and will only be used for statistical purpose.

Thank you!

Contacts: info@khresmoi.eu
www.khresmoi.eu
www.cir.meduniwien.ac.at

The form is also available online:
www.tinyurl.com/ImproveRadiology
or
www.tinyurl.com/RadiologySearch
Appendix B

User Tests protocol
APPENDIX B. USER TESTS PROTOCOL

Before a session the investigator should perform the following actions:

1) Obtain 2 Windows-based computers (one for the participant and one for the investigator) with internet connection (preferably not wifi).

For the participant’s computer:
- Make sure that it has a conventional mouse, reasonably large, good resolution screen, a web camera and a microphone.
- Download the demo videos (2D/article and 3D) from here [link] save it on the desktop.
- Make sure that all Firewalls are disabled.

Morae:
- Install Morae Recorder.
- Download the Study Configuration File (the zip with the two files) from here [link] and save it to the desktop. **Important**: Do not make any changes to the file!
- Run Morae Recorder. Choose between the Study Configuration Files that you downloaded. If the participant ID number is even choose 2D-article-3D order. If the participant ID number is odd choose 3D-article-2D order. See point j) for details.
- Make sure that the web camera is adjusted to record the participants face and test the microphone.

ezDL 2D:
- Download ezDL client from here [link] and save it on the desktop.
- Start the ezDL client (Java 1.7 has to be installed for the ezDL client to run). Use the participant’s account ID to login.
- The account IDs that were created for the user tests are in the form username/password: xxxxx/xxxxx
- Check here [link] for available accounts that have not been used yet. Remember to update the list with the accounts that you use.
- Deselect the TextSearch from the Source Selection Tool (Tools->Source Selection). Only MedSearch and ParaDISE should be selected as sources.
- Create a custom perspective for the user tests. Start with the Image search perspective

and add the Personal Library and Tray tools. The layout should look like this:

- Save the perspective

ezDL 3D:
- Start the ezdl client from here [link] by clicking on the ezDL Icon
- The account IDs that were created for the user tests are in the form username/password: xxxxx/xxxxx
- Make sure that the Image 3D source is selected (Tools -> source selection)
r. Create and save a custom perspective that looks like this and save it (for better viewing the users are able to enlarge/minimize the query and detail views)

For the investigator’s computer:
   a. Make sure that all Firewalls are disabled.

Morae:
   a. Install Morae Observer.
   b. Run Morae Observer.
   c. Connect to the participant’s computer by using its ip address (you can find it the ip address of a computer by running cmd in the start menu of windows and then ipconfig – the ip address needed is the IPv4)

2) Download the consent form document from here link and print it.
3) Make sure there is a way(e.g. a timer) to keep track of the duration of the tasks.

Note: It would be really useful for the investigator to have watched the demo video and played with the system before the session so he can better help the participant and understand misuses of the system. It would be also useful to have a colleague performing the tasks and play the role of the observer to get familiar with the note taking using the markers of the Morae Observer tool.

The 2D article demo video includes among others, demonstration of the following features:
   a. Features/Functionalities of Query by text to retrieve articles.
   b. Features/Functionalities in the Detail view of articles.
   c. Features/Functionalities of Query by text and or/image to retrieve images (Relevance Feedback).
   d. The Facet functionalities and the results presentation options in the Results view.
   e. Features/Functionalities in the Detail view of images.
   f. Personal library and Tray drag drop use.

The 3D article demo video includes among others, demonstration of the following features:
   a. Features/Functionalities of the Query image view.
   b. Features/Functionalities of the Detail image view.
   c. Features/Functionalities of the Result view.

Steps of a Session
APPENDIX B. USER TESTS PROTOCOL

The steps of a session should be the following, in the following order:

1) The investigator should make an introduction to the Khresmoi project and the purpose of the user study. Key points should be
   a. What are the Khresmoi project goal in radiology use case,
   b. in which stage of the project we currently are (middle of the project – first prototype of the system, not finished)
   c. why we are doing the user tests (evaluate/redesign the tool, not the radiologists)

2) At this point, the participant should be asked to sign the consent form.

3) The participant should be asked to watch the video demonstrating the system features. He is allowed to pause and rewatch if he wants, as well as ask questions to observers.

4) Once the video is finished and the user is ready to start, he should be asked to press the red button in Morae Recorder. The recording will then start.

5) The Introduction of the system will be shown in a prompt box and the completion of the demographics survey form should start.

6) Once the user has finished the completion of the demographics survey form he should be asked to maximize the Khresmoi system window (ezdl client) to start with performing the tasks.

7) During each task, the investigator should note all required information and observation notes using the Observer tool on his computer and call for the end of a task if the time is up before the participant completes the task. The observer should have a neutral attitude. He can talk to the participant (which will also encourage him to think out loud) but should not do operations for him. He may help him when he sees he is blocked and can’t proceed. Observations can be of different types: free comments from the participant, free observations from the observers about the system or the user study itself, bugs and inconsistencies of the system. When an observation is noted into Morae, the observer should define its type in field next to where he entered his text. He can also define a score for each observation, but this step is not mandatory.

8) Once all the tasks are finished and questionnaires are filled in by the participant, the investigator could have an informal interview (off record) with the participant about the overall feeling of the use of the system, thoughts etc. The observer should take note of what the participant says during this phase of the test, since it is not recorded into Morae.

9) Two files need to be saved and be collected. The investigator should find the recording file on the participant’s computer and save it as well as the observer log on the investigator’s computer. The recording file should be saved as radioXy.rdg and the log file should be radioXy_observationLog.txt (e.g. radioV1.rdg, radioV1_observationLog.txt)
B.1 CONSENT FORM

I hereby give my consent to participate in a scientific study conducted as part of the EU-Project KHRESMOI by the Health on the net Foundation. I have been informed about the content, the purpose and the extent of this study. I have been given sufficient time to reflect on my participation in the study. All my questions have been answered. I understand that my participation is voluntary and can be terminated at any time, without any specified reason. I understand that participation in this study does not influence or affect my health status. As there are no health-related risks involved, I agree with the absence of an insurance coverage.

The experiment will take about 1 hour and 30 minutes. It consists of two questionnaires and a number of recorded computer related tasks related to search some specific topics with a search engine prototype. The principal investigators, as well as all involved project members, pledge to use the collected data only in anonymized form and with the utmost discretion. They are required to treat the data and observations confidentially. No personally identifiable data is passed or sold to third parties.

I give permission for the involved project members of KHRESMOI to have access to my anonymized data collected during the study. I agree with my data being used for approved research to help advance insights on medical information retrieval.

___________________________________
Name organization and researcher

___________________________________  ______________________
Date and signature Date and signature

Participant ID – provided by the researcher

____________________________________
Place

This form should be signed in two copies. One is retained by a participant, another by the investigator.

In case of questions, please, contact the principal investigators at

e-mail1 and e-mail2 // 0041 22 372 62 50
Appendix C

User Tests tasks
TUTORIAL TASK (FOR 2D/ARTICLE SEARCH AND 3D SEARCH)

During this session, you will be asked to fulfill tasks using both the 2D/article search and the 3D search tools. This task contains simple actions so that you get familiar with the functionalities of the two tools.

In the 2D/article search tool:
1. Enter the term “brain” in the query zone and start the search.
2. Select one image from the results view and find its details in the Details view.
3. Drag and drop this image to the query zone.
4. Add 1 positive and 1 negative example image to the query zone from the results view to refine your query.
5. Note: Negative examples can be given by right clicking an image in the query zone and choosing “mark as non relevant”
6. Drag and drop an image from the results view to the Tray.

In the 3D search tool:
1. Right click somewhere in the search zone, and select Load a volume. Use the following ID: ID_8006000001239002_3_1
2. When the volume is loaded, try to scroll through it, play with contrast and brightness.
3. Select one or more regions of interest (ROIs) by selecting an area on a slice of the volume with the mouse.
4. Try to select another ROI on a different slice and to remove other ROIs that you might have selected.
5. When you are satisfied with your selection, hit the Search button.
6. Browse through results in the results view and select a relevant volume.
7. Play with the volume loaded in the details view, contrast, brightness.

2D IMAGE SEARCH TASKS

Task 1 (max 5 minutes):
1. Click on "Clear all content" from the File Menu.
2. Find 3 osteoporosis x-ray images.
3. Only use textual queries.
4. Place the images you find into the Tray.

Task 2 (max 5 minutes):
1. Click on "Clear all content" from the File Menu.
2. Find 3 osteoporosis x-ray images.
3. Use the example images under the tag "image_task_1" in your Personal Library to complement the query.
4. You may mark the results as relevant or non relevant to relaunch the search.
5. Place any relevant results into the Tray.

Task 3 (max 5 minutes):
1. Click on "Clear all content" from the File Menu.
2. Find 3 images that share the same diagnosis with the example images under the tag "image_task_2" in your Personal Library.
3. You may mark results as relevant or non-relevant to relaunch the search.
4. Place the image you find into the Tray.
ARTICLE SEARCH TASKS

Task 1 (max minutes):
1. Click on "Clear all content" from the File Menu.
2. You have the following:
3. "A 43-year-old man with painless, gross hematuria. Abdominal CT scan revealed a large left renal mass with extension into the left renal pelvis and ureter."
4. You can find the images associated with the case are placed in your Personal Library under the tag: "case_retrieval_task_1".
5. Find 3 relevant articles to the case above and place them into the Tray.

Task 2 (max 7 minutes):
1. Click on the "Clear all content" from the File Menu.
2. You have the following:
3. "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."
4. you can find the images associated with the case are placed in your Personal Library under the tag: "case_retrieval_task_8".
5. Find 3 relevant articles to the case above and place them into the Tray.

3D SEARCH TASKS

Task 1 A (Max 7 min):
• Give a radiological description for the patient using only the given volume with simple image processing features (zoom, contrast, brightness).
• Data: ID_8006000001239002_3_1

Task 1 B (Max 10 min)
• Give a radiological description for the patient using one volume and the pathology image search including ROI
• Data: ID_8006000001239002_3_1

Task 2 (Max 10 min)
• Verify one or more of the given pathological descriptions using the image search feature, e.g. find similar pathologies
• Data: ID_8006100001247763_6_1
• Honey Combing, Emphysem, Milkglass (Milchglasstrübung), ...

FREE USE OF KHRESMOI

*In this task you may use the systems freely (tool for 3D search or 2D/articles search, or both).

In the 2D tool:
You may search for images or articles, using whichever of the functionalities available.

In the 3D tool:
• Find similar anatomical region using example Data or any of the search Results (enter ID Manually from the Details-View)
APPENDIX C. USER TESTS TASKS

- Example Data:
  - ID_RA10001171211450_2_1
  - ID_8004900001184356_2_1
  - ID_RA10001173454330_603_1
  - ID_RA10001180048880_16_1
  - ID_RA10001181474240_901_1
Appendix D

User Tests questionnaires

D.1 DEMOGRAPHIC SURVEY
APPENDIX D. USER TESTS QUESTIONNAIRES

1. Are you?
   - Male
   - Female

2. How old are you?
   - < 20
   - 20-30
   - 30-40
   - 40-50
   - 50-60
   - >60

3. What is your native language?
   - English
   - French
   - German
   - Spanish
   - Czech
   - Other

4. If you have chosen « other » please specify:

5. Your skills in English are? <Scale>
   - Basic, can comprehend simple issues
   - Native language

6. What is the highest position you have had in a medical service other than radiology?

7. What is the highest position you have had in a service of radiology?
   If non-applicable, enter N/A.

8. If you have a work experience in a radiology service, how long have you been working in radiology?
   - N/A
   - 0-3 y
   - 4-6 y
   - 6-10 y
   - >10 y

9. If you have a work experience in a radiology service, what field in radiology are you specialized in?
   - Bone
   - Nuclear radiology
   - Radio-oncology
   - Emergency radiology
   - Thorax
   - Interventional radiology
   - Echography
   - Other

10. If you have checked « other » please specify:

11. Do you use a computer in your day-to-day life?

D.2 USABILITY SURVEY
12. Do you use a computer for job or education related tasks?

13. Do you use Google search? <Scale + free text>

14. If you use other search engines, please specify below:

15. Do you search the Internet for health related information?

16. If yes, please specify the websites you use below:

17. Do you use Google image search? <Scale + free text>

18. If you use other image search engines, please specify below:

19. Do you use Facebook?

20. If you use other social media network, please specify below:
APPENDIX D. USER TESTS QUESTIONNAIRES

Usability of the Software

1. I would like to use this system frequently.
   - Strongly disagree
   - Disagree
   - Neutral
   - Agree
   - Strongly agree

2. I found the system unnecessarily complex.
   - Strongly agree
   - Disagree
   - Neutral
   - Agree
   - Strongly disagree

3. The system was easy to use.
   - Strongly disagree
   - Disagree
   - Neutral
   - Agree
   - Strongly agree

4. I would need the support of a technical person to be able to use this system.
   - Strongly agree
   - Disagree
   - Neutral
   - Agree
   - Strongly disagree

5. The various functions in this system were well integrated, that is, the program works in a harmonious way which is logical to me.
   - Strongly agree
   - Disagree
   - Neutral
   - Agree
   - Strongly disagree

6. There was too much inconsistency in this system, that is, the program react in a way that I was not expecting and surprised me.
   - Strongly agree
   - Disagree
   - Neutral
   - Agree
   - Strongly disagree

7. I would imagine that most radiologists would learn to use this system very quickly.
   - Strongly agree
   - Disagree
   - Neutral
   - Agree
   - Strongly disagree

8. I found the system very awkward to use.
   - Strongly agree
   - Disagree
   - Neutral
   - Agree
   - Strongly disagree

9. I felt very confident on what I was doing, using the system.
   - Strongly agree
   - Disagree
   - Neutral
   - Agree
   - Strongly disagree

10. I needed to learn a lot of things before I could get going with this system. That is, the program requires a lot of training before an adequate use.
   - Strongly agree
   - Disagree
   - Neutral
   - Agree
   - Strongly disagree

11. Are there any tools that need to be improved/changed? If yes, how would you like them to be changed so that they will be more useful to your searches?

12. Are there any new functionalities/tools that would like this search system to have?
**D.2. USABILITY SURVEY**

**SCREEN PRESENTATION**

13. Reading characters on the screen
   - difficult
   - easy

14. Add free comments

15. Presentation of images (e.g. size, position, additional information provided)
   - poor
   - excellent

16. Add free comments

17. Quality of translation
   - poor
   - excellent

18. Add free comments

19. Performing task is straightforward
   - never
   - always

20. Add free comments

**SYSTEM CAPABILITIES (FOR 2D/ARTICLE SEARCH AND 3D SEARCH)**

21. Does the system respond quickly to your requests? Are the results delivered quickly enough?
   - too slow
   - fast enough

22. Add free comments

23. Do you find the system reliable? Does it react the way you expect it to?
   - unreliable
   - reliable

24. Add free comments
25. Are the results satisfactory? Do they match the queries you formulated?

- Unreliable
- Reliable

26. Add free comments

27. Are the results well presented?

- Dislike how results are presented
- Like how results are presented

28. Add free comments

29. How easy is it to correct your mistakes, that is, undo, redo tasks?

- Difficult
- Easy

30. Add free comments

31. I think the system is appropriately designed for all levels of user (e.g. containing both simple and more advanced features in tools, for beginners and advanced in radiology respectively).

- Strongly disagree
- Strongly agree

32. Add free comments

33. I think the system provides some tools and features that can be helpful in my work/research that are not available in the current tools I use.

- Strongly disagree
- Strongly agree

34. Add free comments

35. I would use the 2D image and article search for academic work (preparation of lectures etc.).

- Strongly disagree
- Strongly agree

36. Add free comments
37. I would use the 2D image and article search for research activities.

- Strongly disagree
- Strongly agree

38. Add free comments

39. I would use the 2D image and article search during clinical work.

- Strongly disagree
- Strongly agree

40. Add free comments
Notation

**DoG**(x, y; s) the Difference of Gaussians
d*canberra* the Canberra distance
d*ε* the Euclidean distance
d*jd* the Jeffrey divergence distance
d*manhattan* the Manhattan distance
d*χ*² the *χ*² distance
D*ᵣ* the set of relevant images
D*nr* the set of non–relevant images
F the image representation vector
f the local descriptor vector
*f*D(x) the density function
*f*D(x) the local density function
H*ᵢ* the orientation histogram of the local descriptor
K the kernel function
L(x, y; s) the Laplacian operator
μ*ij* the membership of the *j*th pixel in the *i*th color bin in the fuzzy color histogram
n*f* number of local features
N the number of documents in the whole database when applying a TF–IDF weighting
NN(x) the nearest visual word to *x*
n*sid* the number of occurrences of word *i* in document *d*
n*d* the total number of words in the document *d*
n*i* the number of occurrences of word *i* in the whole database
q*m* the modified query after application of Rocchio algorithm
q*o* the original query
R*k*(i) the rank of the result in retrieved list *k*
S*combSUM*(i) the new score of result *i* after CombSUM fusion
S*combMNZ*(i) the new score of result *i* after CombMNZ fusion
S*combMax*(i) the new score of result *i* after CombMAX fusion
S*combMin*(i) the new score of result *i* after CombMIN fusion
S*linear*(i) the new score of result *i* after Linear fusion
S*Borda*(i) the new score of result *i* after Borda Count fusion
S*RBF*(i) the new score of result *i* after Reciprocal rank fusion
S*_k*(i) the score assigned to result *i* in retrieved list *k*
s*hi* the Histogram intersection similarity
s*cosine* the Cosine similarity
t*idf* the TD–IDF weight
XX the expanded training set
X the set of labeled and unlabeled images
x* the density attractor
V the visual vocabulary
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Approximate Nearest Neighbour</td>
</tr>
<tr>
<td>BoC</td>
<td>Bag-of-Colors</td>
</tr>
<tr>
<td>BoVW</td>
<td>Bag-of-Visual-Words</td>
</tr>
<tr>
<td>CAD</td>
<td>Computer-aided diagnosis</td>
</tr>
<tr>
<td>CEDD</td>
<td>Color and Edge Directivity Descriptor</td>
</tr>
<tr>
<td>CPOE</td>
<td>Computerized Physician Order Point</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-Separated Value</td>
</tr>
<tr>
<td>DBMS</td>
<td>Database Management Systems</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DIRS</td>
<td>Distributed Image Retrieval System</td>
</tr>
<tr>
<td>DoG</td>
<td>Difference of Gaussians</td>
</tr>
<tr>
<td>E2LSH</td>
<td>Euclidean Locally Sensitive Hashing</td>
</tr>
<tr>
<td>EBM</td>
<td>Evidence-based Medicine</td>
</tr>
<tr>
<td>EC2</td>
<td>Elastic Cloud Compute</td>
</tr>
<tr>
<td>EMR</td>
<td>Electronic Medical Records</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract-Transform-Load</td>
</tr>
<tr>
<td>FCTH</td>
<td>Fuzzy Color and Texture Histogram</td>
</tr>
<tr>
<td>HDFS</td>
<td>Hadoop Distributed File System</td>
</tr>
<tr>
<td>HIS</td>
<td>health information systems</td>
</tr>
<tr>
<td>HoG</td>
<td>Histograms of Gradients</td>
</tr>
<tr>
<td>HRCT</td>
<td>High Resolution Computed Tomography</td>
</tr>
<tr>
<td>GFS</td>
<td>Google File System</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphical Processing Units</td>
</tr>
<tr>
<td>HTTP</td>
<td>Hyper Text Transfer Protocol</td>
</tr>
<tr>
<td>IRMA</td>
<td>Image Retrieval in Medical Applications</td>
</tr>
<tr>
<td>KHRESMOI</td>
<td>Knowledge Helper for Medical and Other Information Users</td>
</tr>
<tr>
<td>kNN</td>
<td>k-Nearest Neighbour classifier</td>
</tr>
<tr>
<td>LiRe</td>
<td>Lucene Image Retrieval</td>
</tr>
<tr>
<td>mAP</td>
<td>mean Average Precision</td>
</tr>
<tr>
<td>MeSH</td>
<td>Medical Subject Headings</td>
</tr>
<tr>
<td>MIFAS</td>
<td>Medical Image File Access System</td>
</tr>
<tr>
<td>MR</td>
<td>Magnetic Resonance</td>
</tr>
<tr>
<td>NDFS</td>
<td>Nutch Distributed File System</td>
</tr>
<tr>
<td>PACS</td>
<td>Picture Archive and Communication Systems</td>
</tr>
<tr>
<td>ParaDISE</td>
<td>Parallel Distributed Image Search Engine</td>
</tr>
<tr>
<td>QUIS</td>
<td>Questionnaire for User Interaction Satisfaction</td>
</tr>
<tr>
<td>RAM</td>
<td>Random Access Memory</td>
</tr>
<tr>
<td>RCTs</td>
<td>Randomized Controlled Trials</td>
</tr>
<tr>
<td>REST</td>
<td>Representational State Transfer</td>
</tr>
<tr>
<td>RF</td>
<td>Relevance Feedback</td>
</tr>
<tr>
<td>ROI</td>
<td>region of interest</td>
</tr>
<tr>
<td>mAP</td>
<td>mean Average Precision</td>
</tr>
<tr>
<td>MIFAS</td>
<td>Medical Image File Access System</td>
</tr>
<tr>
<td>MR</td>
<td>Magnetic Resonance</td>
</tr>
<tr>
<td>NDFS</td>
<td>Nutch Distributed File System</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>PACS</td>
<td>Picture Archive and Communication Systems</td>
</tr>
<tr>
<td>ParaDISE</td>
<td>Parallel Distributed Image Search Engine</td>
</tr>
<tr>
<td>QUIS</td>
<td>Questionnaire for User Interaction Satisfaction</td>
</tr>
<tr>
<td>RAM</td>
<td>Random Access Memory</td>
</tr>
<tr>
<td>RCTs</td>
<td>Randomized Controlled Trials</td>
</tr>
<tr>
<td>REST</td>
<td>Representational State Transfer</td>
</tr>
<tr>
<td>RF</td>
<td>Relevance Feedback</td>
</tr>
<tr>
<td>ROI</td>
<td>region of interest</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale Invariant Feature Transform</td>
</tr>
<tr>
<td>SPECT</td>
<td>Single-Photon Emission Computed Tomography</td>
</tr>
<tr>
<td>SPM</td>
<td>Spatial Pyramid Matching</td>
</tr>
<tr>
<td>SURF</td>
<td>Speeded Up Robust Feature</td>
</tr>
<tr>
<td>SUS</td>
<td>System Usability Scale</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TF–IDF</td>
<td>Term Frequency – Inverse Document Frequency</td>
</tr>
<tr>
<td>UCD</td>
<td>User-centered design</td>
</tr>
<tr>
<td>VLAD</td>
<td>Vector of Locally Aggregated Descriptors</td>
</tr>
</tbody>
</table>
List of Figures

1.1 The unified development lifecycle model proposed in [162]. ................................. 4
1.2 The block diagram of a state–of–the–art CBIR pipeline, Bag–of–Visual–Words [170]. Local features from a training group are clustered and produce a “visual vocabulary”. The query image and the database images are then expressed as distributions of “visual words”. ............................................. 5

2.1 Heat map images for knee MRI (left) and a chest x–ray (right), showing that often a particular region attracts the main attention as in the knee MRI or several distinct regions as with the chest x–ray. .................................................. 13
2.2 (a) Information sources used for finding images. (b) Defining relevance or suitability of images ................................................................. 14
2.3 Self–assessment of the success rate of image searches, (a) the overall percentages and (b) a comparison between clinical work and teaching. ..................... 15
2.4 Reasons given for unsuccessful image search. .................................................... 15
2.5 Time taken to find relevant images (a) and times before an unsuccessful search is abandoned (b) .................................................................. 16
2.6 Time taken for successful (a) and failed (b) image search compared for clinical work, teaching and research. ......................................................... 16
2.7 Desired (a) input and (b) output data for a medical image search. ......................... 17

3.1 An overview of the ParaDISE backend. The four basic elements are combined to perform the indexing and search processes. .......................... 23
3.2 The indexing pipeline of ParaDISE Indexer. ......................................................... 30
3.3 An overview of the HES–SO in–house network of computers. ............................... 31
3.4 The search pipeline of the Rocchio Seeker. ......................................................... 33
3.5 The search pipeline of the LateFusion Seeker. .................................................... 35
3.6 The 2D image retrieval interface. .................................................................... 36
3.7 The KHRESMOI indexing pipeline of 2D images. ............................................. 37
3.8 The KHRESMOI search pipeline for 2D image retrieval. .................................. 39
3.9 The index view of the 3D Image Search prototype. ............................................ 40
3.10 The ROI marking (full-volume-view) overlayed over the index view of the 3D Image Search prototype. ......................................................... 40
3.11 The result view of the 3D Image Search prototype. .......................................... 41
3.12 The web–based interface of Shambala. ............................................................. 42
3.13 Gesture interaction with Shambala. ................................................................. 42
3.14 The Google Glass and their main components. ................................................. 43
3.15 List of results images as seen in the Google Glass. ......................................... 44
3.16 Graphs and radiology images in a single figure. .............................................. 47
3.17 Differentiation between separated and standard search results using a coloured border and an icon. ......................................................... 47
3.18 Highlighting a subfigure in its parent image. ................................................... 48

4.1 Modality hierarchy of the ImageCLEF 2012 modality classification task. .......... 54
LIST OF FIGURES

4.2 Sample images from ImageCLEF2012 medical data set .................. 54
4.3 Mean average precision per search iteration for result list size $k = 5$. 57
4.4 Mean average precision per search iteration for result list size $k = 20$. 57
4.5 Mean average precision per search iteration for result list size $k = 50$. 58
4.6 Mean average precision per search iteration for result list size $k = 100$. 58
4.7 Examples of successful separation results .............................. 60
4.8 Examples of separation failures ......................................... 62
4.9 Comparison of runtimes using a single node and the distributed Hadoop cluster. For the datasets of 10k and 100k images, the single machine run does not scale linearly with more than 12 simultaneous tasks ...................... 63

5.1 The interface of the Morae Recorder software, which was installed to the participant’s computer. A common study configuration file is created for all the users and the recording starts by pressing red button. .......................... 71
5.2 The interface of the Morae Observer software, which was installed to the observer’s computer. This tool allows the observer to take notes on the timeline of the recording and use different markers (e.g. found bug, participant comment, participant blocked etc.) The program in the screenshot is waiting for the recording to begin. ......... 71
5.3 Screenshot of filling in a survey form using the Morae software. .... 72
5.4 3D Mockups generated at MUW, and used during the interviews. (A) Query case and grid view on query results. (B) Query case and list view on the query results. (C) Query case, chosen case and given grid view on results. (D) In detail view on query case and one result case. (E) Same as D with Report displayed. (F) Same as C, but with visualized query ROI and ROIs in the result view that match the query. 73
5.5 Median values of measuring general user satisfaction about the system in Likert scale (1=strongly negative, 5=strongly positive). .................. 74
5.6 Mode values for each participant over the global satisfaction questions in a Likert scale .................................................... 74
5.7 Mode frequency for each participant over the global satisfaction questions. ........ 74
5.8 Pilot study: Median of measuring user satisfaction over specific system aspects in a Likert scale (1=strongly negative, 5=strongly positive). 75
5.9 Pilot study: Median values of measuring general user satisfaction about the system in Likert scale. ........................................ 76
5.10 Pilot study: Mode values for each participant over the global satisfaction questions in a Likert scale. ................................. 76
5.11 Pilot study: Mode frequency for each participant over the global satisfaction questions. 76
5.12 Median of measuring user satisfaction over specific 2D image search prototype aspects in a Likert scale (1=strongly negative, 5=strongly positive). 77
5.13 Median of measuring user satisfaction over specific 3D image search prototype aspects in a Likert scale (1=strongly negative, 5=strongly positive). 81
5.14 Age distribution of the user test participants. ........................... 86
5.15 Self assessment of English skills by participants. ....................... 86
5.16 Time of service in the field radiology. .................................. 87
5.17 Field in radiology of the participants. ................................. 88
5.18 Demographic information about computer and internet usage of the participants. 88
5.19 Median values of measuring general user satisfaction about the system in Likert scale (1=strongly negative, 5=strongly positive) compared to the median values of the first prototype evaluation. ....................... 89
5.20 Mode and mode frequency values for each participant over the global satisfaction questions in Likert scale and % of questions. ........ 89
5.21 On the left side the average run time of one user test together with the time each part took. On the right side the time distribution between introduction, tasks and surveys. ......................... 90
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.22</td>
<td>User satisfaction survey on the 2D/article search subsystem compared to the previous prototype.</td>
<td>91</td>
</tr>
<tr>
<td>5.23</td>
<td>Mode and mode frequency of overall satisfaction on the 2D/article search subsystem per participant.</td>
<td>91</td>
</tr>
<tr>
<td>5.24</td>
<td>Results of the user satisfaction survey on the 3D system.</td>
<td>93</td>
</tr>
<tr>
<td>5.25</td>
<td>Mode and mode frequency of overall satisfaction per participant.</td>
<td>94</td>
</tr>
<tr>
<td>5.26</td>
<td>Normalized histogram of the general satisfaction questionnaire.</td>
<td>96</td>
</tr>
</tbody>
</table>
List of Tables

1.1 The categorization of medical informatics, as it is proposed in [104]. 1

4.1 Average mAP scores of BoVW representations of local features over six vocabularies
for different distance measures. 55

4.2 mAP scores of the fusion of best–performing runs: SIFT ($k = 20$), SURF ($k = 40$),
RootSIFT ($k = 30$) Lab ($k = 100$). 55

4.3 Average mAP scores of local features for different image representations. 55

4.4 mAP scores of image representations for different distance measures. 56

4.5 Best mAP scores of RF techniques. 58

4.6 Modality classification runs. 59

4.7 Modality classification results. 59

4.8 Sensitivity and Positive Predictive Value (PPV) for each compound figure subclass. 61

4.9 Ad–hoc image retrieval results in ImageCLEF2013. 65
Bibliography


abnormality, 12
Abstract, vii
Acknowledgements, ix
ANN, 31, 32, 36, 38
Bag–of–Colors, 25, 56
Bag–of–Visual–Words, 25, 50
Bags–of–Visual–Words, 5
bottleneck, 49, 50
BoVW, 25, 26, 28, 31, 34, 36, 53, 55, 56, 59, 62–64
CBIR, 5–7, 9, 14, 19, 21–28, 31, 32, 34, 48, 50, 67, 75, 78, 81, 83, 84
CEDD, 27, 56, 64
Chapter five, 67
Chapter four, 53
Chapter one, 1
Chapter six, 99
Chapter three, 21
Chapter Two, 9
clinicians, 1–3, 6, 9, 11, 13–15, 19, 99
CombMNZ, 28, 38, 55–57, 59, 64
CombSUM, 28, 55
Component, 22, 23, 34
CompositeComponent, 23
Contents, iv
dataset, 7, 25, 28, 36, 38, 45, 46, 48–51, 53, 56, 57, 59–65, 68, 69, 85, 97, 101, 102
density–based clustering, 7, 44
descriptor, 22–25
difference of Gaussians, 24, 25
distance, 5, 24, 32, 34, 51, 53, 55, 56
evaluation, 2–4, 6, 7, 9, 10, 21, 24, 28, 45, 50, 53, 56, 57, 60, 67–69, 72, 73, 75, 80, 84, 85, 87, 86, 99–102
Evidence–based Medicine, 1, 12
Extractor, 22–24, 29
FCTH, 27, 56, 64
function, 10, 17, 19, 31, 34, 38, 44, 45, 49, 50, 52, 67–70, 72, 75, 81, 83, 84, 90, 94, 95
fusion, 23, 24, 28, 38, 48, 49, 55–59, 64, 65
Fusor, 22–24, 28, 34, 38
Gaussian, 44
Glossary, 139
Hadoop, 31, 36, 49–52, 62, 63
histogram, 24–28, 34, 38, 45, 53, 55, 56
image use, 6, 10, 18, 83
ImageCLEF, 45–47, 53, 54, 56, 59–62, 64, 65
Index, 159
Indexer, 23, 29–31
indexing, 5–7, 21–24, 29–31, 34, 36, 37, 43, 49, 50, 53, 63, 64, 99, 101, 102
Introduction, 1
k–Nearest Neighbour classifier, 7
KHRESMOI, 6, 21, 34, 36–39, 41, 67–70, 72, 81, 82, 84, 85, 96
kNN, 7, 45, 46, 64
large–scale, 5, 49, 65, 101, 102
LIRE, 27, 28, 50
List of Figures, 140
List of Tables, 143
local descriptors, 5, 22, 23, 25, 49
local features, 7, 23–26, 29, 49, 53, 55, 59
Manager, 22, 23
mAP, 53, 55–58
MapReduce, 24, 29, 31, 49, 50, 62, 63
Mean average precision, 53, 57, 58
medical image retrieval, 7, 18, 22, 48, 49, 53, 56, 99
mid–level feature, 23–25
modality filtering, 7, 38, 64, 65, 90, 92, 97
multi–modal, 7, 34, 41, 45, 48, 49, 56–58, 101
Notation, 137
Observation, 9, 19, 67, 70, 82, 90, 93
PACS, 4, 10, 12, 18, 21, 22, 95
parallel, 24, 28, 29, 31, 36, 49, 50, 52, 75, 102
Parameters, 22, 29, 32
pathology, 6, 21, 22, 50, 100
160
INDEX

performance, 1–4, 24, 45, 48–50, 53, 55–60, 63–65, 68, 69, 75, 82, 83, 90, 93, 97, 101, 102

prototype, 19, 21, 31, 38, 40, 41, 47, 67, 69, 70, 72, 73, 77, 80–84

questionnaire, 11, 13, 68, 90, 93, 96

Résumé, v

References, 145

region of interest, 5, 19, 80

relevance feedback, 6, 7, 21, 22, 32, 38, 41, 48, 49, 56–59, 161

relevance feedback, 21

representation, 5, 7, 22–26, 29, 31, 32, 34, 36, 49, 53, 55, 59, 70, 85, 95, 101, 102

Representational State Transfer, 23

Rocchio, 32, 33, 38, 48, 55, 57–59

RootSIFT, 25, 55

Seeker, 23, 32–36, 38, 55

semantic gap, 57, 101, 102

Semi–supervised learning, 45

Semi–supervised machine learning, 7

Shambala, 34, 41, 42

SIFT, 24, 25, 27, 49, 50, 55, 56, 59, 64

Standard Hyper Text Transfer Protocol, 23

Storer, 22, 23, 28, 29

survey, 2, 6, 7, 9–11, 13, 18, 19, 21, 22, 48, 67–69, 96

SVD, 28

Table of contents, iv

text–based retrieval, 21, 24, 48

text–based search, 6, 21, 101

TF–IDF, 29, 31

Thanks, ix

Title, i

UCD, 4, 5

usability, 3–7, 67, 68, 70, 72, 101, 102

user satisfaction, 9, 68, 70, 74–77, 81–83, 86, 87, 89, 90, 93, 96, 97

user tests, 7, 68–70, 72, 77, 80–82, 84–86, 96, 97, 101

user–centered design, 4, 7, 67

visual vocabulary, 5, 25, 44, 55

VLAD, 26, 55

workflow, 10, 12, 15, 19, 21, 68, 69, 72, 80, 90