UNIVERSITÉ DE GENÈVE
Département d'informatique FACULTÉ DES SCIENCES
Professeur Dr.Stéphane Marchand–Maillet
Département de radiologie et informatique médicale FACULTÉ DE MÉDECINE
Professeur Dr. Henning Müller

Use Case Oriented Medical Visual Information Retrieval & System Evaluation

THÈSE

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

par

Alba García Seco de Herrera

de
Madrid (Espagne)

Thèse N^{o} 4781

GENÈVE
2015

La Faculté des sciences, sur le préavis de Monsieur H. MÜLLER, professeur titulaire et directeur de thèse (Faculté de médecine, Département de radiologie et informatique médicale), Monsieur S. MARCHAND-MAILLET, professeur associé et directeur de thèse (Département d'informatique) et Monsieur D. L. RUBIN, professeur (Department of Radiology and Medicine, Stanford University, Stanford, California, U.S.A.), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 22 Mai 2015

Thèse - 4781 -

Le Doyen

Contents

Abstract

My original contribution to knowledge is done in the following two fields: *medical visual information retrieval* and *evaluation of retrieval systems.*

Large amounts of medical visual data are produced daily in hospitals, while new imaging techniques continue to emerge. In addition to this, many images are made available continuously via publications in the scientific literature. Scientific publications can be very valuable for clinical routine, research and education where up–to–date medical knowledge is needed. However, it is not always easy to find the desired information in this large amount of data and in clinical routine the time to fulfil an information need is often very limited. As a consequence, there is a requirement to manage and retrieve these documents/images in the most efficient and effective way. Retrieval systems are a useful tool to provide access to the biomedical literature related to information needs of medical professionals. Clinicians regularly use information retrieval systems, which benefits decision making and patient care.

To better design retrieval systems based on clinicians' real needs, this thesis explicitly defines and validates a use case associated with a specific evaluation task. The use case deals with retrieval mechanisms able to jointly exploit textual and visual information connected, in the medical domain. This thesis can potentially help clinicians make decisions about difficult diagnoses by developing a medical case–based retrieval system based on the defined use case. This system retrieves articles from the biomedical literature when querying a case description and attached images.

Another main contribution of this thesis consists of a multi–modal approach for medical case–based retrieval with special focus on the integration of visual information connected to text. Different fusion strategies are analysed to evaluate if multi–modal retrieval systems can achieve good performance. However, this is a challenging task and visual features do not always bring up enough information for the retrieval. Therefore, this thesis defines a query–adaptive multi–modal fusion criterion, which shows when visual features are suitable to be fused with text features. This criterion is based on synonym relations between text and visual information. Furthermore, an image modality classification approach is implemented to integrate image modality information in the retrieval step. A semi–supervised learning technique is developed to deal with uneven classes in training data. A crowdsourcing platform is then employed to obtain a more accurate image collection.

The final contribution of this thesis is an evaluation framework for medical retrieval systems. After an in–depth analysis of ImageCLEFmed benchmark in years previous to this thesis, the ImageCLEFmed evaluation campaign has been organised during this thesis oriented to the studied use–case. It includes the generation of a freely available database and ground truth following a meticulous prior preparation process. Lessons learned are also extracted from a careful evaluation and comparison of participants' systems.

Résumé

Ma contribution originale à la connaissance concerne deux domaines d'études: la *recherche d'information médicale visuelle* et l'*évaluation des systèmes de récupération* d'information.

Une grande quantité d'images est produite quotidiennement dans les hôpitaux. Beaucoup d'entre elles sont utilisées par la littérature scientifique et sont extrêmement précieuses pour la pratique clinique ordinaire, la recherche et l'éducation. Cependant, il n'est pas aisé pour les professionnels de la santé de trouver l'information désirée entre la quantité massive de données présentes et le temps limité disponible. Par conséquent, il est nécessaire de gérer et de récupérer des documents/images de manière efficace et efficiente.

Les systèmes de recherche d'information sont des outils utiles pour fournir un accès à la littérature biomédicale liées aux besoins des professionnels de la santé. Ces systèmes peuvent fournir une aide précieuse pour la prise de décision et les soins aux patients.

Afin d'améliorer la conception de ces systèmes conformément aux besoins réels du personnel de santé, cette thèse définit explicitement et valide un cas d'utilisation associé à une tâche d'évaluation. Ce cas d'utilisation traite des mécanismes de récupération capables d'exploiter conjointement des informations médicales textuelles et visuelles liées entre elles. En outre, cette thèse peut potentiellement aider les cliniciens à prendre des décisions pour les diagnostics difficiles, à travers le développement d'un système de récupération à partir de cas médicaux pour le cas d'utilisation définie.

Une autre contribution essentielle de cette thèse est l'approche multi–modale pour la recherche d'information basée sur des cas médicaux et qui se concentre sur l'intégration de l'information visuelle liée au texte. Différentes stratégies de fusion d'information sont analysées pour évaluer si ces systèmes peuvent obtenir de bons résultats. Cependant, ceci constitue une tâche difficile specifique, car les caractéristiques visuelles ne contiennent pas toujours suffisamment d'information pour améliorer la qualité des résultats. Cette thèse définit un critère de fusion multi–modale adaptative à la requête, et indique dans quelles circonstances les caractéristiques visuelles sont éligibles pour être fusionnées avec le texte. Ce critère est basé sur la synonymie entre l'information textuelle et les caractéristiques visuelles. De plus, il met en place une stratégie de classification des modalités d'imagerie pour être intégrée à l'étape de récupération. Une technique d'apprentissage semi-supervisée est développée conjointement avec une stratégie de "crowdsourcing" pour faire face à l'inégalité des classes dans l'ensemble d'entraînement et obtenir une collection d'images plus équilibrée.

La dernière contribution de cette thèse est un cadre pour l'évaluation des systèmes de récupération d'information médicale. La campagne d'évaluation ImageCLEFmed a été organisée en cours de la présente thèse suite à une analyse approfondie des standards antérieurs d'évaluation. Une base de données publique avec des données de validation sont générées pour l'évaluation et la comparaison des systèmes des participants.

Resumen

Mi contribución original al conocimiento radica en dos campos de estudio: la *recuperación de información médica visual* y la *evaluación de sistemas de recuperación.*

Una inmensa cantidad de imágenes es producida diariamente en los hospitales derivadas del diagnóstico a través de técnicas de imagen. Muchas de estas imágenes son distribuidas a través de la literatura científica, sumamente valiosas para la práctica clínica rutinaria, para la investigación y para la educación. Sin embargo, para el personal sanitario no es fácil encontrar la información deseada entre la enorme cantidad de datos disponibles y el tiempo limitado del que dispone. Por tanto es necesario gestionar y recuperar documentos/imágenes de manera efectiva y eficiente. Los sistemas de recuperación de información son herramientas muy útiles para proporcionar acceso a la literatura biomédica relacionada con las necesidades de los profesionales sanitarios, quienes asiduamente usan estos sistemas que benefician la toma de decisiones y la atención al paciente.

Para mejorar el diseño de estos sistemas basándolos en las necesidades reales del personal sanitario, esta tesis define explícitamente y valida un caso de uso asociado a una tarea de evaluación específica. Este caso de uso se ocupa de los mecanismos de recuperación capaces de aprovechar conjuntamente la información médica textual y visual relacionadas. Así mismo, esta tesis puede potencialmente ayudar al personal sanitario a tomar decisiones sobre diagnósticos difíciles, mediante el desarrollo de un sistema de recuperación basado en casos, fundamentado en el caso de uso definido.

Otra contribución esencial de esta tesis consiste en una estrategia multimodal para la recuperación de información basada en casos médicos, que se centra en la integración de la información visual relacionada con la textual. Diferentes estrategias de fusión de información son analizadas para evaluar si estos sistemas pueden obtener buenos resultados. Sin embargo, esta es una tarea difícil, ya que las características visuales no siempre contienen suficiente información para ayudar en la recuperación de información. Esta tesis define un criterio para la fusión multimodal adaptable a la consulta, que muestra cuándo las características visuales son apropiadas para ser fusionadas con el texto. Este criterio se basa en la sinonimia entre la información textual y las características visuales. Adicionalmente, se implementa una estrategia de clasificación de imagen en modalidades para ser integrada en la etapa de recuperación. Una técnica de aprendizaje semi–supervisado junto con una estrategia de “crowdsourcing” es desarrollada para lidiar con la desigualdad de las clases en el conjunto de entrenamiento y así obtener una colección de imágenes más precisa.

La última contribución de esta tesis es un marco para la evaluación de sistemas de recuperación de información médica. La campaña de evaluación ImageCLEFmed ha sido organizada durante esta tesis, tras un análisis de sus estándares previos de evaluación. Mediante un proceso meticuloso, se ha generado una base de datos pública así como los datos claves para la evaluación y comparación de los sistemas de los participantes.

Acknowledgements

> "El cerebro se inventó para salir de casa
> y la memoria para volver a casa."
>
> Jorge Wagensberg

I would like to thank all the people who contributed in some way to achieving this thesis. First and foremost, I would like to thank my advisors Professor Henning Müller and Professor Stéphane Marchand–Maillet for providing me with the opportunity to complete my Ph.D. thesis at the University of Geneva. I especially want to thank my advisor on the spot, Professor Henning Müller, who has been a tremendous mentor for me. I would like to thank him for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. I would also like to thank Professor Daniel L. Rubin very much for accepting to be in the jury for my thesis defence and taking the travel from Stanford to Sierre.

The members of the medGIFT group have contributed immensely to my personal and professional time during these four years at Sierre. The group has been a source of friendships as well as good advice and collaboration. I am especially grateful to Roger Schaer and Antonio Foncubierta–Rodríguez for their friendship and support, which made my thesis work possible. I want to thank Yashin Dicente Cid who has been my colleague, friend and neighbour in Amsterdam and Sierre. I want to thank present and past members of the group: Oscar Jiménez del Toro, Ranveer Joyseeree, Manfredo Atzori, Dimitrios Markonis, Antoine Widmer, Adrien Depeursinge and the numerous summer and rotation students who have come through the group.

I spent two exciting months at the Fudan University in Shanghai, and I would like to thank Professor Yuanyuan Wang for hosting me. I would also like to thank Yu Ma and all the members of the group for their continuing hospitality.

I thank Mete for doing part of the proofreading, which was very helpful for the quality of the English in this thesis. I also want to thank Stéphane and Adrien for correcting the French in the "résumé" of this thesis.

My research work was financed by University of Applied Sciences Western Switzerland. Part of this work was supported by the European Seventh Framework Programme in the context of the PROMISE (FP7–258191) and Khresmoi (FP7–257528) projects. WIDTH (PIRSES-GA-2010-269124) project funded the exchange visit in Shanghai. I would like to express my gratitude to the institution for its support.

My time at Sierre was made enjoyable in large part due to the many friends that became a second family. In addition to the friends from the medGIFT group already mentioned, I am grateful for the time spent with my friends: Sergio, Evelyne, Tania, Andrés, Alejandra, Visara, Stefano and little Emilia (who has grown up with this thesis).

The “chiceria” (Claudia B., Céline, Caroline, Sandra, Claudia P. and Lysiane) has also enriched my stay in Sierre, specially thanks to Lisa.

I particularly want to thank Stéphane for his faithful support, encouragement and patience during the final stages of this Ph.D.

No puedo olvidar en mis agradecimientos a la gente de mi querida España. Me gustaría expresar mi más profundo y sincero agradecimiento al Profesor Emanuele Schiavi quien fue mi fuente de motivación y curiosidad en mis primeros pasos en investigación. Quisiera hacer extensiva mi gratitud a mis compañeros del área de Neuroimagen de la Fundación CIEN, ya que con ellos he compartido despacho e incontables horas de trabajo y buenos ratos. Especialmente gracias a Pablo y Gonzalo por su amistad.

Quiero expresar mi agradecimiento a todos aquellos amigos con los que tanto he compartido. En especial quiero agradecer a Diego, Pablo, María, Cañas, Luis, Eva, Diana, Elvira y Andrés por recibirme con los brazos abiertos en cada una de mis visitas a Madrid. Igualmente, quiero agradecer a mis amigos de la“pandillita” porque siempre están ahí (Jose, Lurdes, Sofía, Almudena B., David, Maxi, Zana, Grande, Rebeca, Almudena R.,Cristina, Nany y Lorena). En especial gracias a mi querida Bea por tantos buenos momentos. No puedo olvidar a mis amigos de Cabeza del Buey a pesar que los he visto muy poquito debido a esta tesis pero con los que siempre es una alegría reencontrarse. También es un honor haber conocido en Amsterdam a grandes amigos, en especial a Ana María, Josep y Joselito.

El agradecimiento más profundo y sentido va para mi familia. Gracias a mis primas y primos, Natalia, Bárbara, Germán, Fernando, María, Inamar y María (última pero no menos querida), por hacerme disfrutar siempre que estoy con ellos. Gracias a mis tíos y tías, German, Rosa, Antonio, Felisa, Manolo y Pilar por haber estado siempre conmigo. Un agradecimiento muy especial merecen mis queridísimos abuelos, Antonio, Patro y Angelita, por la comprensión, el ánimo y el amor que siempre me han dado, incluso a pesar de la distancia. Quiero recordar también a mi querido abuelo Antonio, que estoy segura de que se sentiría muy orgulloso de mi.

Todo esto nunca hubiera sido posible sin el apoyo incondicional y el cariño que siempre obtengo de mi maravillosa famila, de mis padres, Jose Luis y Pilar, y mi hermano, Daniel. Ellos, que me enseñaron a amar la ciencia y que siempre han entenido mi ausencia y mis malos momentos. A pesar de la distancia siempre han estado a mi lado para saber cómo iba y para animarme a seguir adelante. Las palabras nunca serán suficientes para mostrar mi amor y mi agradecimiento.

A todos los que habéis hecho posible esta tesis...¡Gracias!

To all of you who have made this thesis possible...Thanks!

Chapter 1

Introduction

> "Everything you want is on the other side of the fear."
>
> Jack Canfield

This chapter gives a brief introduction to this Ph.D. thesis. The chapter begins describing the motivations for this research. Next an overview of the outline of this thesis is given. Finally the achievements of this thesis in the medical visual Information Retrieval (IR) and system evaluation fields are stated.

1.1 Motivations

Clinicians generally base their decisions for diagnosis and treatment planning on a mixture of acquired textbook knowledge and experience acquired through real–life clinical cases [195]. Therefore, in the medical field, two knowledge types are generally available [170]:

- *Explicit knowledge*– already well established and formalised domain knowledge, e.g., textbooks or clinical guidelines;
- *Implicit knowledge*– individual expertise, organizational practices and past cases.

When working on a new case that includes images, clinicians analyse a series of images together with contextual information, such as the patient age, gender and medical history as this data can have an impact on the visual appearance of the images. Since related problems may have similar solutions, clinicians use past situations similar to the current one to determine the diagnosis and potential treatment options, information that is also transmitted in teaching, where typical or interesting cases are discussed, and used for research [170, 249]. Thus, the goal of a clinician is often to solve a new problem by making use of previous similar situations and by reusing information and knowledge [4], also called case–based reasoning. The problem can be defined in four steps, known as the four "R's" [93, 170]:

1. retrieve the most similar case(s) from the collection;
2. reuse them, and more precisely their solutions, to solve the problem;
3. revise the proposed solution;

4. retain the current case in the collection for further use.

This thesis focuses on the retrieval step because the retrieval of similar cases from a database can help clinicians to find the necessary information [195, 213]. In the retrieval step a search over the documents in the database is performed using the formulation of the information need that can include text and images or image regions. Relevant documents are ranked depending on the degree of similarity to a given query case or the similarity to the information need. The most relevant cases are then proposed on the top of the list and can be used to solve the current problem [18].

Medical IR systems are increasingly complex: they need to satisfy diverse user needs and support challenging tasks. Their development calls for proper evaluation methodologies to ensure that they meet the expected user requirements and provide the desired effectiveness [181]. Large–scale worldwide experimental evaluations provide fundamental contributions to the advancement of state-of-the-art techniques through common evaluation procedures, regular and systematic evaluation cycles, comparison and benchmarking of the adopted approaches, and spreading of knowledge. In the process, vast amounts of experimental data are generated that beg for analysis tools to enable interpretation and thereby facilitate scientific and technological progress [236, 7].

Medical visual IR and its system evaluation comprise the main motivation of this thesis, taking into consideration that the medical literature currently constitutes an enormous knowledge base that includes visual as well textual information.

This thesis was carried out in the context of Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation (PROMISE)[1] and Khresmoi[2] projects. Both projects received funding support from the European Commission in the context of its European Seventh Framework Programme (FP7) and had a common interest and close cooperation on medical visual IR.

PROMISE is a Network of Excellence (NoE) funded by the FP7. PROMISE aimed at advancing the experimental evaluation of complex multimedia and multilingual information systems in order to support the decision making process of individuals, commercial entities and communities who develop, employ and improve such complex systems [63].

To move from abstract benchmarking to more user–sensitive evaluation schemes, PROMISE formulated a set of use cases based on scenarios of use for multimedia and multilingual information access. This allows leveraging previous knowledge and to avoid re–treading previous erroneous tracks. One of the use cases is the “visual clinical decision support” which constitutes the focus of this thesis. The use case deals with visual information connected with text in the clinical domain in order to provide retrieval and access mechanisms able to jointly exploit textual and visual features.

PROMISE also facilitated management of the evaluation activities and offered access, duration, preservation, reuse, analysis, visualization and mining of the collected experimental data.

Khresmoi is an integrated project funded by the FP7. Khresmoi's goal was to develop tools for multilingual multi–modal search and access system for biomedical information

[1] Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation (PROMISE) Network of Excellence is a FP7 –funded research network focused on researching the evaluation of multimedia and multilingual information systems (see `http://www.promise-noe.eu/`).

[2] Khresmoi is a European Union project funded by the FP7 focused on researching tools for multi–modal multilingual search and access systems (see `http://khresmoi.eu/`).

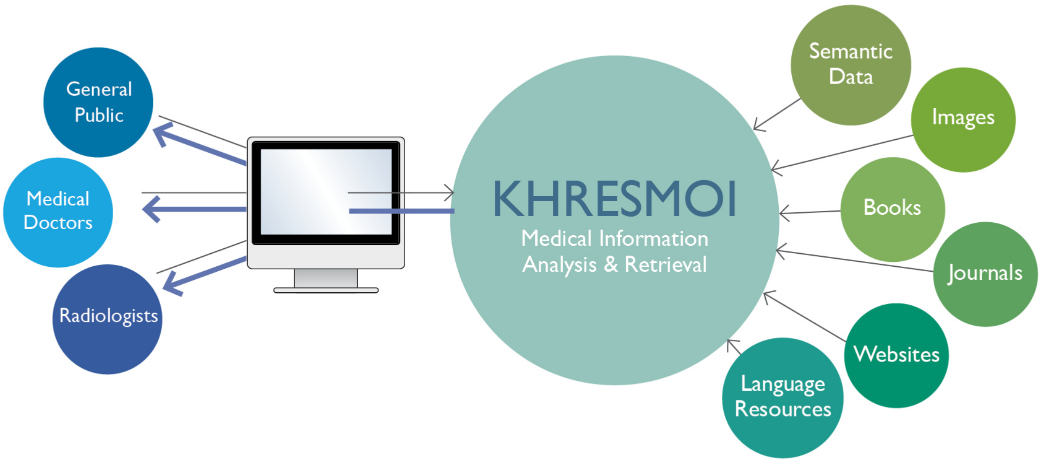


Figure 1.1: Overview of the Khresmoi project. Khresmoi combines multiple data sources and knowledge derived from various heterogeneous knowledge sources. The system allows users to access biomedical data.

and documents [10]. It addressed the challenges of searching through huge amounts of medical data, including general medical information available on the Internet, as well as radiology data in hospital archives. It allows text querying, in combination with image queries. It has three main end user groups: members of the general public, physicians and radiologists (a group of physicians for which image search is of immense importance). An overview of the Khresmoi concept is shown in Figure 1.1.

PROMISE and Khresmoi cooperated on the "visual clinical decision support" use case in order to achieve their respective objectives. They carried out joint evaluation activities by exploiting the PROMISE evaluation infrastructure to experiment with Khresmoi outcomes.

1.2 Thesis overview

This thesis deals with various aspects of medical visual IR, which are studied with a focus on system evaluation.

This first chapter gives a short introduction explaining the main motivation for the research described in this thesis. The principal scientific contributions of this thesis are also briefly listed at the end of this chapter.

Chapter 2 gives an overview of the biomedical visual IR background with a focus on medical case–based retrieval. It provides references for a number of biomedical IR systems. This chapter introduces various components and algorithms which are important throughout the multi–modal aspect of this thesis. Most importantly it includes information fusion techniques, query adaptive multi–modal fusion overview and integration of modality classification into the retrieval. Retrieval evaluation activities' history and retrieval evaluation methodology are reported in this chapter.

Chapter 3 defines and validates the "visual clinical decision support" use case, validating that the use case reflects a real–life problem for the clinicians.

Chapter 4 analyses the Cross–Language Retrieval in Image Collections (ImageCLEF)[3] evaluation campaign scholarly impact. The medical visual IR evaluation, *ImageCLEFmed*, organised in the context of this thesis is described in detail.

Chapter 5 contains a detailed description of the techniques applied to develop a medical case–based retrieval system. It uses the Parallel Distributed Image Search Engine (ParaDISE) system as a baseline and further components are included. This chapter focuses mainly on three features of the system: information fusion, query–adaptive multi–modal fusion and modality classification.

Chapter 6 contains the description of the experiment carried out and the results achieved thanks to the features described in Chapter 5. The ImageCLEFmed framework is used to evaluate the performance of the system. This chapter concludes by discussing the results of the experiment carried out.

Chapter 7 presents a web–based retrieval interface called *Shangri–La*. This interface integrates the multi–modal retrieval approach presented in Chapter 5. Features provided by Shangri–La are described and illustrated with screenshots of the application.

Chapter 8 concludes by revisiting the objectives and summarizing the contributions made in this thesis. It points our further research directions based on the findings of this thesis.

Appendix A contains the surveys and their answers carried out for the use case validation described in Chapter 3.

Appendix B presents most of the answers from the questionnaire filled by ImageCLEFmed organisers between 2011 and 2013. The analysis of this data is done in Chapter 4.

In addition to the main content, more sections are created to make reading the manuscript easier: table of contents; abstract of the contents of this thesis in English, French and Spanish; acknowledgement to everyone who has assisted me throughout my doctoral studies over the years; mathematical notation used in the text; glossary containing abbreviations that are used in the manuscript; list of figures and tables referred in the document; bibliography referring the literature used to write this thesis and an index to help find keywords in the text.

1.3 Scientific contributions of this thesis

The main scientific contributions are in the two fields of *medical visual IR* and *evaluation of retrieval systems*. The contributions can then be classified according to these two fields.

The main contributions of this thesis in the field of retrieval system evaluation and benchmarking are the following:

- definition and validation of the “visual clinical decision support” use case [116, 115];
- ImageCLEFmed benchmarking organization [122, 180, 79]; this includes the creation of freely available databases and ground truth, the evaluation of participant systems and comparison of techniques;

[3] The Cross–Language Retrieval in Image Collections (ImageCLEF) is part of the Conference and Labs of the Evaluation Forum (CLEF) and aims to provide an evaluation forum for the cross–language annotation and retrieval of images (see `http://imageclef.org/`).

- detailed study of the outcomes of the ImageCLEFmed evaluation activities, especially between 2011 and 2013 [194, 85, 119] as well as an assessment of the scholarly impact of ImageCLEF in previous years [236, 194];
- creation of an image database for evaluating image modality classification [76, 81].

Contributions to a medical case–based retrieval system include the following:

- a medical case–based retrieval approach implementation as well as a biomedical image modality classification approach [161, 80, 83, 164];
- an analysis of different fusion strategies to compare their performance [84, 86];
- a query–adaptive multi–modal fusion criterion implementation to decide when to use multi–modal (text and visual) or only text approaches in the retrieval step [77];
- modality classification approach implementation integrated into the medical case–based retrieval; a semi–supervised learning technique is also proposed to exploit unlabelled data and to expand the training set [76, 81];
- a web–based retrieval interface, called *Shangri–La.*

Other articles written on this project include [162, 10, 11, 82, 72, 78, 73, 164, 9].

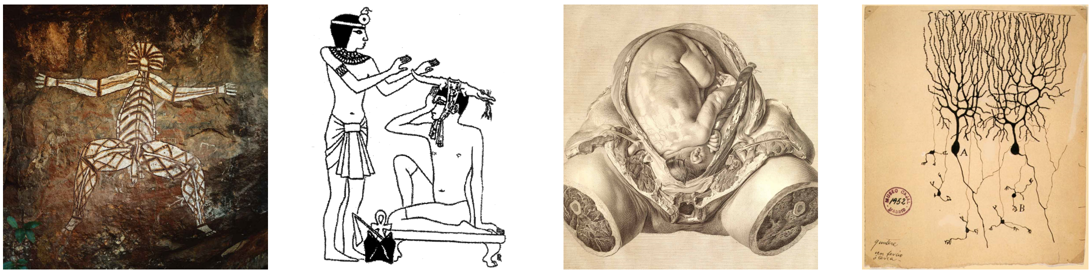
Chapter 2

Medical Visual Information Retrieval

> "En la sociedad de la información radica la solución a la generación de la inteligencia colectiva que necesitamos para seguir adelante."
>
> Gaspar Ariño Ortiz

Medicine has been represented in images since prehistoric times with early illustrations leaning toward symbolic representations. Illustrations have been developing from symbolism to greater realism (see Figure 2.1). Advances in medical technologies have changed the physicians vision and understanding of the human body. Different modalities of medical images, such as radiology or microscopy, show objective evidence of disease and decrease the dependence on patient's subjective descriptions. Figure 2.2 shows some examples of findings in medical images which help physicians in their work on patient cases.

Today, images are produced in hospitals in ever–increasing numbers [5] and provide crucial information for diagnosis, treatment planning and other tasks. A recent European report estimates that 30% of the global digital storage is occupied by medical image



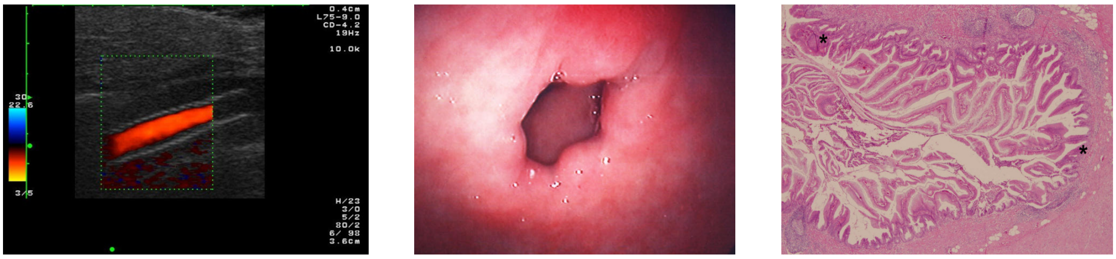
(a) Rock painting, 6000 B.C. Aboriginal "X–ray style" figure. Kakadu National Park, Northern Territory, Australia.

(b) The Ebers Papyrus, 1200 B.C. Egyptian papyrus which describes a therapy for migraine.

(c) Copperplate engraving of a woman who died near the end of term by William Hunter, 1774. National Library of Medicine.

(d) Drawing of Purkinje cells and granule cells from pigeon cerebellum by Santiago Ramón y Cajal, 1899. Instituto Santiago Ramón y Cajal.

Figure 2.1: Examples of historical medical illustrations.



(a) Findings on colour Doppler after endovascular treatment (stenting) in a 52-year-old woman suffering from recurrent transient ischemic attacks.

(b) A complete healing at the polypectomy site on an endoscopy after a 12–week course of proton pump inhibitor therapy.

(c) Hematoxylin and eosin stain on the appendix tissue reveals villous adenoma with moderate to severe dysplasia located suppurative appendicitis.

Figure 2.2: Examples of medical images that help in the diagnosis and treatment planning of cases.

data [3]. Besides clinical settings, images are also made available via biomedical publications. The number of biomedical articles published grew at a double–exponential pace between 1986 and 2006 according to [106]. For example, the biomedical open access literature of PubMed Central (PMC)[4] alone contained almost 2 million images in 2014.

Many physicians have regular information needs during clinical work, teaching preparation and research activities [99, 179]. Therefore there is a need for searching through the immense collection of images in institutions and on the World Wide Web, making the data accessible for reuse. Studies showed that the time for answering a clinical information need using IR systems is around 30 minutes [101], while clinicians state to have approximately five minutes available [103]. Finding relevant information quicker is thus an important task to bring search into clinical routine [167].

Retrieval and classification of medical images have been explored to get additional information for reading and interpretation of medical cases [241] when open questions remain and thus help clinicians in their daily work.

Although text queries are commonly used, the visual information of the images can enrich the search. Images represent an important part of the content in many publications and searching for medical images has become common in retrieval applications, particularly for radiologists. Image retrieval has been shown to be complementary to text retrieval approaches and images can well help to represent the content of scientific articles, particularly in applications using small interfaces such as mobile phones [60]. Furthermore, medical case–based retrieval taking into account several images and potentially other data of the case has also been proposed by other authors over the past 7 years [186, 249].

2.1 Components of a retrieval system

IR systems search for relevant documents and information within the contents of a specific database. In this section, the components needed to develop a rudimentary IR system that retrieves documents are first described. Figure 2.3 puts together all the basic components to outline a complete IR system. The architecture of the IR system consists

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

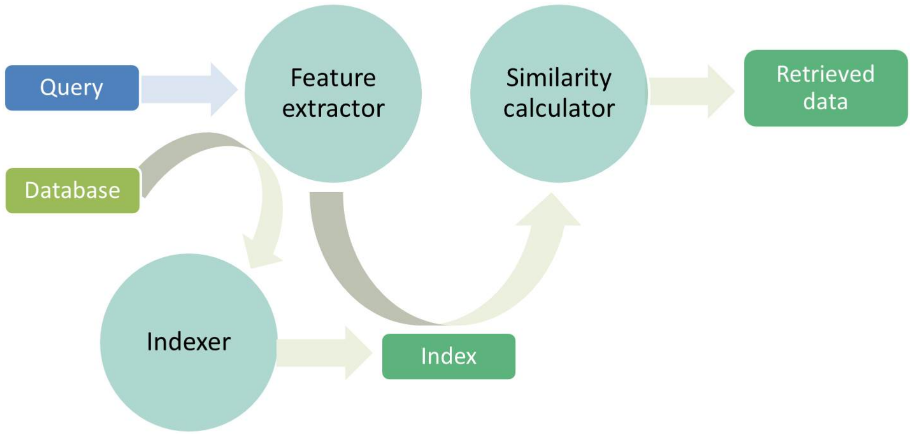


Figure 2.3: Outline of the basic elements of a complete retrieval system.

of the following three components:

1. *Feature extraction* – the system describes the *query* as a set of features to handle the *index*;

2. *Indexing* – the system builds an *index* of the document descriptors to record and maintain the *database* information;

3. *Similarity calculator* – the system retrieves documents that are relevant to the *query* from the *index* and displays the *retrieved data* to the user.

This thesis focuses on the visual information integration of the medical retrieval systems. Therefore, this section presents an overview of text and visual information extraction and describes several methods to improve the retrieval precision using multi–modal approaches.

2.1.1 Information sources and retrieval

Text retrieval has been successfully used in various medical fields from lung disease through cardiology, eating disorders and diabetes to hepatitis [235] and Alzheimer's disease [147].

Text in the anamnesis is often the first data available and based on the initial analysis other exams are ordered. Most biomedical search engines, also systems searching for images, have been based on text retrieval only. Sources of biomedical information can be scientific articles and also reports from the patient record [217]. The various parts of the text such as title, abstract and figure captions can then be indexed separately. Some examples for general search tools that have also been used in the biomedical domain are the Lucene, Essie or Terrier IR tools. Lucene[5] is an open source full–text search engine. The advantage of Lucene is its simplicity and high performance [166]. Essie [109] is a phrase–based search engine with term and concept query expansion and probabilistic relevancy ranking. It was also designed to use terms from the Unified Medical Language

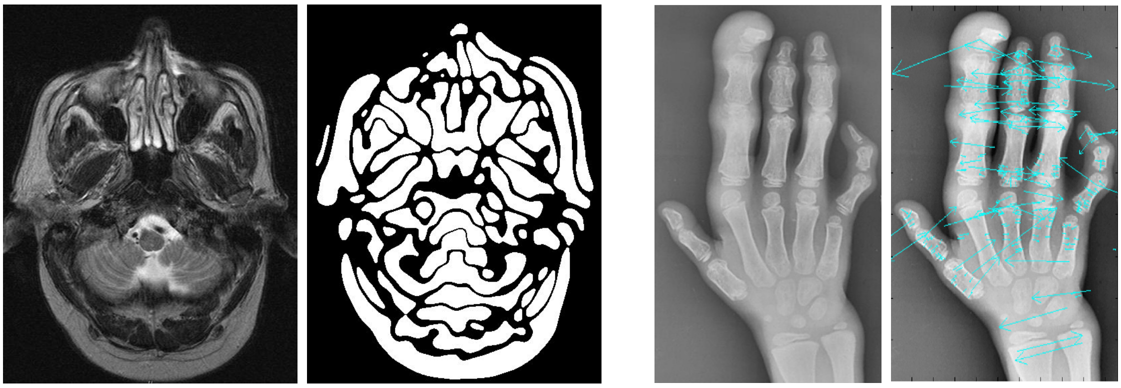
[5] The Apache Lucene is a project that develops open–source search software including indexing and search technology (see `http://lucene.apache.org/`).

System (UMLS). Terrier[6] is also an open source platform for research and experimentation in text retrieval developed at the University of Glasgow. It supports most state of the art retrieval models such as Dirichlet prior language models, Divergence from Randomness (DFR) models or Okapi BM25.

In addition to the text in the anamnesis, another initial data source for diagnosis are the images [249]. Users of biomedical sources are often interested in images for biomedical research or medical practice [193], as the images carry an important part of the information in articles. Rather than using text queries, in Content–Based Image Retrieval (CBIR) systems, visual features are extracted from the images and, based on them, images are retrieved. This allows the use of visual information to find images in a database similar to examples given or with similar regions of interest.

Visual retrieval for medical applications has also become an important research area over the past 15 years [213]. The most commonly used features for visual retrieval can be grouped into the following types [12, 107]:

- *Colour* – several colour image descriptors have been proposed [34] such as simple colour histograms, a colour extension to the Scale Invariant Feature Transform (SIFT) [242] or the Bag of Colours (BoC) [82];
- *Texture* – texture features have been used to study the spatial organization of pixel values of an image like first order statistics, second order statistics, higher order statistics and multiresolution techniques such as wavelet transform [212];
- *Shape* – various features have been used to describe shape information, including moments, curvature or spectral features [257].



(a) In the right, regions detected by a key–region detector from the image in the left.

(b) The arrows, in the image in the left,represent the centre, scale and orientation of the key points detected in the image in the right, by the SIFT algorithm.

Figure 2.4: Information can be extracted from the visual content of the images.

Figure 2.4 shows examples of the visual information that can be extracted from the images. The extraction of multiple visual features often enhance the retrieval performance. Multiple features have been explored, most frequently SIFT variants [37, 80, 52, 252, 219], Local Binary Patterns (LBP) [37, 219], edge and colour histograms [57, 37, 252, 219, 223, 39]

[6] Terrier is an open source search engine, readily deployable on large–scale collections of documents (see `http://terrier.org/`).

and grey value histograms [252]. Several texture features have also been explored such as Tamura [37, 252, 219, 223], Gabor filters [252, 219, 223], Curvelets [37], a granulometric distribution function [39] and spatial size distribution [39]. In recent years, visual words [234] have become the main way of describing images with a variety of basic features such as SIFT [159] and also texture or colour measures.

2.1.2 Information fusion

The combination of various single search modalities (such as text and visual image features) makes it possible to use cross–modal relationships and thus improve the performance beyond the performance of single components [254]. However, the improvement of the performance of these multi–modal systems has long been considered difficult due to the richness of multimedia [95, 141] and the complexity of extracting meaningful information from visual documents in a large domain automatically [197]. Fusing the retrieval results of visual and textual resources into a final ranking is a popular approach for multi–modal retrieval. Fusion can either be performed early or late, creating a unified data representation or fusing after each data type is analysed independently [61, 65].

Several fusion models are described in the literature to combine multi–modal sources. Already in 1998, La Cascia et al. [148] presented a CBIR system which combined visual and textual information directly in the feature vector space representation. Textual information is extracted using latent semantic indexing. In addition, visual information is captured in color and orientation histograms. More recently, Pham et al. [192] combine text and visual features by normalizing and concatenating them to generate the feature vectors. Traditionally, the most common method followed for data fusion is to search the modalities separately and fuse their results (ranked lists) with methods such as linear combination [8]. Methods to obtain suitable weights for linear combinations are reviewed by Wu [253]. Furthermore, Kludas et al. [140], Atrey et al. [12] and Depeursinge [61] provide an overview of the different fusion methods that have been used for multimedia analysis and IR.

In terms of medical cases, images are always associated with either text or structured data and this can then be used in additional to the visual content analysis for retrieval. Most often text retrieval has much better performance than visual retrieval, describing the context in which the images were taken. Poorest performance of visual techniques are achieved when applied to databases with a wide spectrum of image modalities, anatomies and pathologies [196]. However, there is evidence that the combination or fusion of information from textual and visual sources can improve the overall retrieval quality [157, 87, 79].

The most common approach to get the final result is the result combination of visual and text retrieval. Cao et al. [38] represent the features from different modalities as a multi–dimensional matrix and incorporate these feature vectors using an extended Latent Semantic Analysis (LSA) model. Gkoufas et al. [88] increase the retrieval performance by applying linear methods to combine visual and textual sources of images. Classical approaches such as the maximum combinations (combMAX), the sum combinations (combSUM) and the multiplication of the sum and the number of non-zero scores (combMNZ) are studied by Zhou et al. [259] showing that fusing visual and text runs outperforms single modality runs. Mourão et al. [173] introduce a new fusion technique, Inverted Squared Rank (ISR), a variant of the Reciprocal rank fusion (RRF).

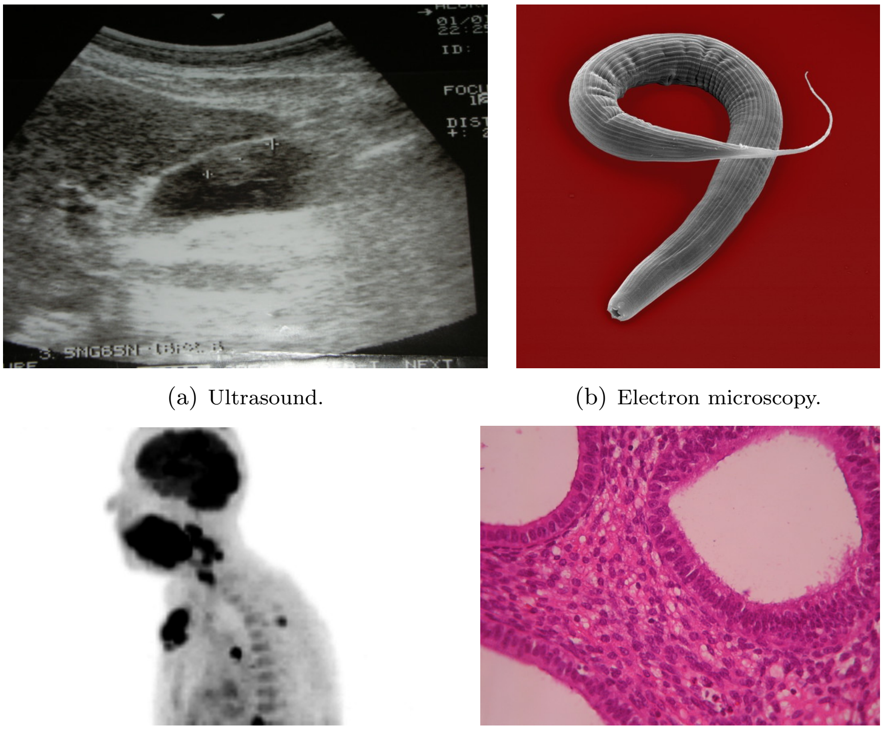
Furthermore, some reranking methods have also been explored [89, 105] for fusion visual and text information. However, strategies that reorder top–ranked documents limit the margin of improvement due to their use on a limited number of documents [187]. Martínez Fernández et al. [165] reorder the results from the CBIR using text–based retrieval. Viswa [243] uses visual information to rerank text–based image retrieval. The relevance of the images is linked to their initial rank position to relax the assumption that the top–ranked images in the text–based results are equally relevant.

2.1.3 Query–adaptive multi–modal fusion

Section 2.1.2 investigates techniques to fuse visual and text information to improve the precision of the retrieval. However, fusion does not always lead to better results and can even decrease the performance of the retrieval [83, 220, 174]. Therefore, to combine multi–modal retrieval two fundamental aspects should be studied: when and how multiple retrieval models can be combined to obtain better performance than individual models [157]. How to fuse multi–modal systems has been explored by studying multiple fusion techniques. These methods are particularly suitable under different settings and are studied in detail in this thesis. When to fuse multiple retrieval models, such as text or visual retrieval models, is a complicated topic. Different models used in a fusion process can provide complementary or contradictory information [12]. Hence, applying a single standard retrieval method for all possible queries is inadequate [155]. Recently, *adaptive query* retrieval has been an emerging trend as a solution to this problem [62]. Adaptive query techniques aim to associate individual queries with specific retrieval strategies [135]. Kennedy [135] reviews the methods proposed for adapting retrieval strategies according to the intentions of the user. Several strategies have been proposed, such as the prediction of the quality of each available tool based on statistical measures of the returned results or the adaptation strategies based on the user context. However, most of the techniques are based on query classification using Natural Language (NL) analysis of the query.

NL analysis is used in IR to translate potentially ambiguous NL queries and documents into unambiguous internal representations for retrieval [158]. Text retrieval techniques commonly use terminologies for query expansion [55, 215]. The queries can be expanded automatically with synonyms from such a terminology, for example. Díaz Galiano et al. [64] consider terms associated with Medical Subject Headings (MeSH) descriptors as synonyms and use these to expand queries. More recently Dramé et al. [66] explore the use of term synonyms to expand queries. However, visual retrieval techniques cannot apply these methods directly for synonym extraction because visual information cannot be directly represented as words. Nevertheless, language modelling techniques can be extended easily to visual techniques [71].

In order to efficiently use multi–modal retrieval systems some efforts have been made to find a relation between images and text. Recently, Simpson et al. [218] review the techniques applied to deal with image content and its semantic meaning in terms of NL. A method based on global feature mapping is also presented. Kurtz et al. [145, 146] propose annotating the images with semantic terms extracted from a given ontology to build a vector of terms representing the image. Lacoste et al. [149] represent the images and the text in the same way, as vectors of concepts, building a conceptual index. However, most of the approaches use joint probabilistic models to find relationships between multi–modal features [153, 16, 68, 171, 202, 22]. Additionally, some approaches are based on image region categorization [58, 150].



(a) Ultrasound.

(b) Electron microscopy.

(c) Positron Emission Tomography (PET).

(d) Light microscopy.

Figure 2.5: Examples of images of various modalities that can be found in the biomedical literature.

2.1.4 Modality classification

Finally, it is also possible to use image analysis and classification to extract relevant information from the images (such as modality types, anatomic regions or the recognition of specific objects in the images such as arrows) to filter results lists or rerank them. In the biomedical literature images can be of several types, some of which correspond to medical imaging modalities such as ultrasound, Magnetic Resonance Imaging (MRI), X–ray and Computer Tomography (CT) (see examples of images from various modalities in Figure 2.5). In user–studies [163], clinicians have indicated that modality is one of the most important filters that they would like to be able to limit their search by [56]. Previous studies [123, 56] have shown that imaging modality is an important piece of information relating to the image for medical retrieval. Image categories can be integrated into any retrieval system to enhance or filter its results [233], benefiting both in speed and precision of the search [120] by reducing the search space to a set of relevant categories [199, 83]. Furthermore, classification methods can be used to offer adaptive search methods [247, 20]. Automatic modality classification is thus an important part of the performance and usability of modern medical retrieval systems. However, image modality is typically extracted from the caption. Caption information can help if captions are well controlled like in the radiology domain but the more general biomedical literature makes it hard to find the modality information in the caption. Studies have shown that the modality can be extracted from the image itself using visual features [191, 151, 114]. Visual image classification techniques have other shortcomings as some modalities can easily be mixed up when categorising automatically such as CT and MRI. In these cases text information of the captions can be used as additional cues to disambiguate the two.

A big variety of visual classification techniques have been explored. Csurka et al. [57] use a Fisher Vector representation of the images built on low level features. Kitanovski et al. [139] use a spatial pyramid in combination with dense sampling using an opponentSIFT descriptor for each image patch. Support Vector Machine (SVM) with χ^2 kernel is then used as a classifier. Classifiers employed range from simple k–Nearest Neighbours (k–NN) [161, 70, 80, 83, 260] or logistic regression model [39] to Genetic Programming (GP) [70] or SVM [216, 37, 252, 219, 223, 139, 220, 260, 226, 20].

An overall system which uses the predicted modality within a retrieval system consists of the following steps: the modality is extracted from the query; the usual retrieval step is performed; the predicted modalities of the document are integrated into the search. Information about image types can be used in various ways in the retrieval. The following approaches have been explored to integrate the classification into the results [233]:

- *Filtering* – discarding the images of which the predicted type is different to the query. Thus, when filtering using the image type only potentially relevant results are considered;

- *Reranking* – reranking the initial results with the image type information. The goal is to improve the retrieval ranking by moving relevant documents towards the top of the list based on the categorization;

- *Score fusion* – fusing a preliminary retrieval score S_R with an image classification score S_M using a weighted sum: $\alpha \cdot S_T + (1-\alpha) \cdot S_M$, where S_R and S_T are normalised. This approach allows for adjusting the parameter α to emphasise the retrieval score or the categorization results.

Sometimes, the training set contains labelled data that are rare and some classes are under–represented. This scenario is often met in medical image analysis, where accurate labelling of big datasets is difficult and expensive to obtain. Therefore training data can be augmented with additional examples to improve the classification, which has also been explored in [37, 80, 223]. Semi–supervised learning [41] uses a small number of labelled instances and a large amount of unlabelled data for training the classifier. Methods of semi–supervised learning have been applied to handwritten text recognition [36] and biological networks [255]. Related to this work, in [57] semi–supervised classification is applied to medical image classification to expand the training set. The confidence scores for the unlabelled data are given by SVM classifiers using multi–modal (visual and textual) information. Moreover, the expansion of the training set by visual retrieval is explored.

2.2 Example systems

Due to the many challenges in biomedical retrieval, research has been attracting increasing attention, and many approaches have been proposed [157]. This section presents a few retrieval systems that use multi–modal information for the search. A more detailed overview on platforms specialised on biomedical search can be found in Gottlieb et al. [92].

Well–known free retrieval systems such as ARRS Goldminer[7] or Yottalook[8] retrieve

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

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

images and articles from peer–reviewed biomedical journals but only based on text queries. On the other hand, systems such as Image Retrieval in Medical Applications (IRMA)[9] or img(Anaktisi)[10] provided only CBIR. Regarding multi–modal retrieval systems, the Center of Informatics and Information Technology group CITI presented the NovaMedSearch[11] as a medical multi–modal search engine that can retrieve either similar images or related medical cases [172]. The National Library of Medicine (NLM)[12] provides Open–i[13] [59], a service to search and retrieve abstracts and images from the open source literature and biomedical collections.

Furthermore, as described in Section 2.1.4, to improve retrieval quality a successful classification of images into types (e.g. X–ray, ultrasound, CT, etc) can be applied to filter out irrelevant images [199]. Already many web–accessible search systems such as Goldminer or Yottalook allow users to limit the search results to a particular modality [180] as this is a feature often requested by end users [163]. However, they extract the modality information only from the text and not from the visual features of the images.

2.3 Retrieval evaluation activities

Systematic and quantitative evaluation activities using shared tasks on shared resources have been instrumental in contributing to the success of IR as a research field and as an application area in the past few decades. Evaluation campaigns have enabled the reproducible and comparative evaluation of new approaches, algorithms, theories, and models, through the use of standardised resources and common evaluation methodologies within regular and systematic evaluation cycles.

2.3.1 History

In 1955, a criterion of relevance and measures for the evaluation of *text* IR systems was proposed for the first time by Kent et al. [136].

In the 1960s, the Cranfield tests [45] were pioneering evaluating text retrieval technology comparing the effectiveness of the different indexing techniques. Many other research groups reused the Cranfield test collection for evaluating their systems [245]. The Cranfield studies set the importance of creating test collections and using these for comparative evaluation of IR systems. After these first benchmarks, several large–scale evaluation campaigns have been established at the international level, with major initiatives in the field of textl IR [210].

Starting also in the 1960s and through the 1990s, the SMART IR project at Cornell University investigates the effectiveness and efficiency of automatic text retrieval methods [31, 29]. This project emphasises completely automatic approaches to retrieve large

[9] Image Retrieval in Medical Applications (IRMA) is a project at the Aachen University of Technology (RWTH Aachen) that aims to develop and implement high–level methods for CBIR with prototypical application for medical tasks on a radiologic image archive (see `http://ganymed.imib.rwth-aachen.de/irma/`).

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

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

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

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

quantities of text. It offers a basic framework for research on the vector space and related models of IR [32].

In the 1990s, the Text REtrieval Conference (TREC)[14] started a consolidation to allow comparing results across the same data using the same evaluation methods [245]. TREC has provided large collections and uniform scoring procedures over the years [96]. TREC developed a research tool for evaluating retrieval methods: trec_eval [33]. This tool has become the primary method used in research for retrieval evaluation to calculate the same measures using the same implementation.

Since 1999, the NII Testbeds and Community for Information access Research (NTCIR)[15] placed emphasis on IR with Japanese or other Asian languages and cross–lingual IR [131, 124, 125, 126, 127, 128, 129, 130, 207, 132]. NTCIR aims to advance in information access technologies including IR shifting from document retrieval to IR using information in the documents. NTCIR has also investigated evaluation methods for information access developing the tool NTCIREVAL [206].

Since 2000, the Conference and Labs of the Evaluation Forum (CLEF)[16] have organised a series of evaluation labs designed to bring different aspects of mono– and cross–language IR systems following TREC–style [26]. CLEF have support the development of an evaluation framework for IR systems operating in both monolingual and cross–language contexts including the creation of reusable data for benchmarking purposes.

In 2002, the INitiative for the Evaluation of XML retrieval (INEX)[17] organised the first workshop. The main goal of INEX has been to promote the evaluation of structural information (XML elements) to yield focused retrieval and identify relevant parts of relevant documents [75]. In 2013 and 2014, INEX run as a lab of CLEF.

Following the success of these evaluation campaigns, in 2008, the Forum for Information Retrieval Evaluation (FIRE)[18] proposed a retrieval benchmark to deal with South Asian languages.

Similar evaluation exercises have also been carried out in the field of *visual* IR. In the 2000s, the Benchathlon[19] initiative tried to set up a common framework for the evaluation of CBIR systems. Unfortunately this initiative did not organised an evaluation campaign.

In 2001, TREC Video Retrieval Evaluation (TRECVid)[20] organised a track as part of TREC. TRECVid has encouraged video IR [221] and it became an independent benchmarking initiative.

[14] The Text REtrieval Conference (TREC) aims to support research within the IR community by providing the infrastructure necessary for large–scale evaluation of text retrieval methodologies (see http://trec.nist.gov/).

[15] The NII Testbeds and Community for Information access Research (NTCIR) is an evaluation forum which aims at promoting research in information access technologies (see http://research.nii.ac.jp/ntcir/index-en.html).

[16] The Conference and Labs of the Evaluation Forum (CLEF) is a self–organised body whose main mission is to promote research, innovation and development of information access systems with an emphasis on multi–lingual and multi–modal information (see http://www.clef-initiative.eu/).

[17] The INitiative for the Evaluation of XML retrieval (INEX) provides an IR test collection in order to measure the performance of a search engine (see https://inex.mmci.uni-saarland.de/).

[18] The Forum for Information Retrieval Evaluation (FIRE) aims to encourage research in South Asian language information access technologies (see http://www.isical.ac.in/~fire/).

[19] Benchathlon aimed to set up a favourable environment for sharing CBIR resources (see http://www.benchathlon.net/).

[20] The TREC Video Retrieval Evaluation (TRECVid) evaluation meetings are an on–going series of workshops focusing on a list of different IR research areas in content–based retrieval and exploitation of digital video (see http://trecvid.nist.gov/).

Similarly, MediaEval[21] started in 2008 as a lab of the CLEF campaign, VideoCLEF [152]. MediaEval became an independent benchmarking initiative in 2010. It has focused on the social and human aspects of multimedia access and retrieval.

ImageCLEF was offered for the first time in 2003 as one of the CLEF labs including media data such as images. This lab has aimed to compare CBIR systems and to determine how associated cross–language text can be used in combination with CBIR, which is language independent, to improve retrieval performance [119].

In the *biomedical* field, retrieving large amounts of data is an important issue in the clinical routine. In the 1990s, OHSUMED[22] provided a clinically–oriented MEDLINE subset covering all references from 270 medical journals over a five–year period (1987–1991). The references include the title, abstract, MeSH indexing terms, author, source, and publication type. Moreover, novice physicians generated 106 queries [98]. However, OHSUMED did not provide standardised evaluation measures.

More recently, in 2011 and 2012, TREC organised the Medical Records track. This track examined the problem of retrieving relevant clinical reports from free–text fields [246]. Moreover, in 2014, TREC proposed the Clinical Decision Support track to retrieve biomedical articles relevant for answering generic clinical questions about medical records.

Although the medical information usually contains masses of free text [143] it also contains images. In 2004, the ImageCLEF lab introduced a medical task: *ImageCLEFmed* [50]. The tasks organised over the years by ImageCLEFmed have provided an evaluation forum and framework for evaluating the state of the art in biomedical image retrieval. This thesis focuses on the campaigns from 2011 to 2013 when the provided repositories had evolved to be close to real world in theirs size and scope [119]. Chapter 4 gives a detailed description of the evolution of ImageCLEFmed over the years.

Following the interest created by ImageCLEFmed, the Visual Concept Extraction Challenge in Radiology (Visceral)[23] is organizing a retrieval benchmark to find cases with similar anomalies based on large–scale sets of 3D radiology images in 2015 [117].

These evaluation campaigns have been widely credited with contributing tremendously to the advancement of IR by providing access to infrastructure and evaluation resources that support researchers in the development of new approaches, and encouraging collaboration and interaction between researchers from both academia and industry [119].

2.3.2 Evaluation process

A typical evaluation cycle is depicted in Figure 2.6. Each evaluation activity can have a different cycle time, e.g., the CLEF cycle operates over one year although some other evaluation campaigns operate over a longer period [48], such as NTCIR which operates over 18 months.

This section gives an overview of each step of the cycle described in the Figure 2.6.

[21] MediaEval is a benchmarking initiative dedicated to evaluating new algorithms for multimedia access and retrieval (see `http://www.multimediaeval.org/`).

[22] OHSUMED is test collection proposed for research (see `http://ir.ohsu.edu/ohsumed/ohsumed.html`).

[23] The Visual Concept Extraction Challenge in Radiology (Visceral) is a project supported by the European Commission under the Information and Communication Technologies (ICT) theme of the FP7 for research and technological development (see `http://www.visceral.eu/`).

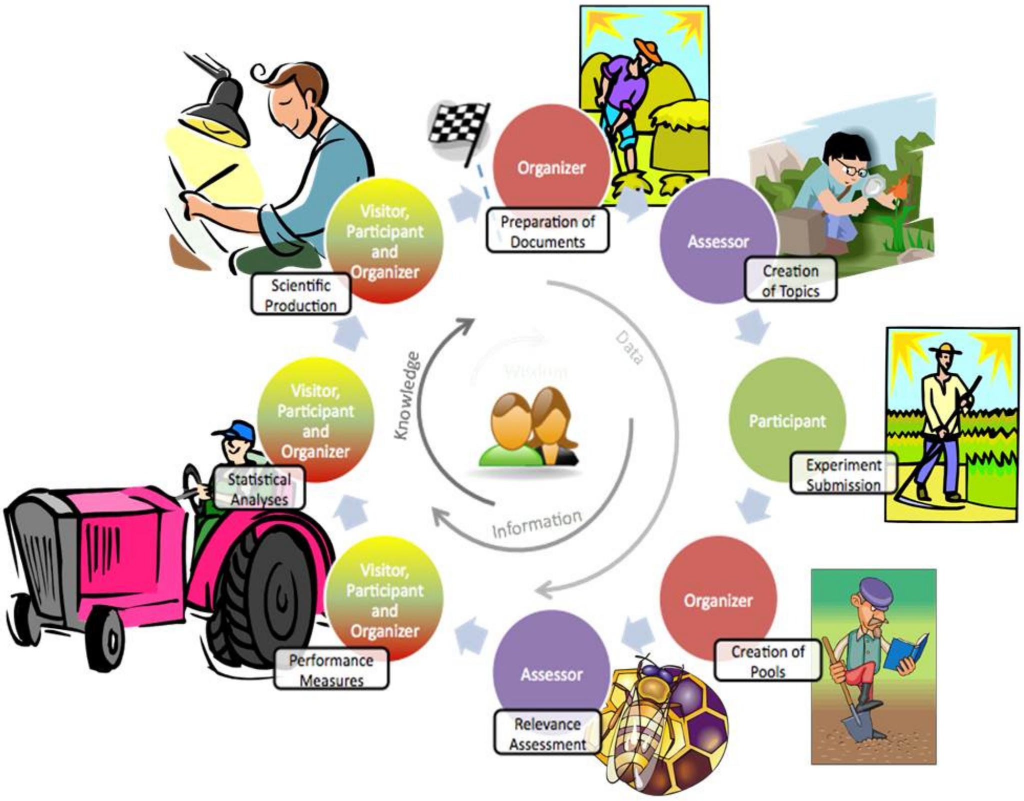


Figure 2.6: Cycle of activities in an evaluation campaign described in the context of the PROMISE project.

Preparation of documents

The cycle begins with organisers preparing the *test collections*. As a response to a user query an IR system retrieves documents. Therefore, the test collection must contain a static set of documents which reflects the use case of the chosen domain [49]. The collection should be static to allow its reusability. Sanderson et al. [209] address some practices to create test collections such as carefully considering the purpose of the evaluation or the resources available.

Creation of query topics

A set of information needs, called *query topics*, is used to test the ability of retrieval systems to retrieve an accurate and complete ranked list of documents in response to them. The query topics should be created with the help of experts in the domain, the *assessors*, to be realistic and representative of the use case. Often, the log of queries submitted to search engines is employed as representative query topics [209].

Experiment submission

The participants use the created datasets (containing the test collection and the query topics) to run their experiments and produce system outputs in standard format, called *runs*, which are then submitted to be evaluated [48]. Buckley et al. [33] defined a run for a retrieval task as a ranked list of documents for each set of query topics in a test collection.

Relevance judgements & pool creation

The evaluation of the submitted runs is based on *relevance assessments*, also called *qrel*, for each query topics performed on the test collection. Obtaining the relevance assessments for large–scale test collections requires a large investment of time and human resources [49]. Therefore it is common to judge a subset of the collection instead of the whole collection. This is known as *pooling*. This approach selects the top n–retrieved documents for each query topics within the set of submitted runs. Therefore the *pool* does not contain documents that were not retrieved by the systems under test and the combination of results may provided a higher quality pool [209]. Documents outside the pool are assumed to be not relevant [144]. Furthermore, resulting judgement set can becomes a biased sample of the complete judgement set and thus systems might not be fairly compared if the collection size grows [30].

The assessors then go through each document in the pool and make relevance judgements.

Crowdsourcing Relevance assessments can be collected by multiple assessors using crowdsourcing [49]. Crowdsourcing allows dividing the problem of relevance assessments into microtasks that can be solved in a short amount of time by users familiar with the domain [91]. Crowdsourcing has recently emerged as a tool in biomedical sciences because it can improve the quality, cost and speed of manually processing large amounts of data [200]. This methodology allows dividing a data processing problem into manual simple human micro–tasks [91].

Crowdsourcing has also been used for collecting and analysing health and medical research data or to create pre–clinical medical study material [24]. In particular several challenges for image annotation have been proposed, such as generating models of proteins for successful molecular replacement and subsequent structure determination [137], classification of retinal fundus photography [168] or evaluating medical pictograms [256].

Performance Measures & statistical analysis

The relevance assessments are used to quantify system effectiveness [49]. Dozens of *evaluation measures* can be calculated to assess the runs' performance based on the number of relevant documents found [33]. Different measures evaluate different aspects of the system performance but always they can be used to compare different runs over the same collection. Submitted runs are released and analysed using the chosen measures.

Scientific production

Finally, the activities and results can be published or shared to transfer the experience and ideas that have been learned across the evaluation activity [48]. Therefore, the evaluation activities trigger sharing insights, knowledge and ideas to develop a common understanding [119].

2.4 Summary

This chapter gives a short introduction of the use of medical images in the clinical routine and its use to enrich physicians' information searches. It shows the basic architecture of an IR system and describes in more detail the techniques commonly used to integrate multi–modal information sources such as text and image. Particular attention is paid to information fusion, specially to adaptive–query methods, which define strategies to fuse the information depending on the query. Integrating modality classification methods into the retrieval step has also been explored.

This chapter also covers existing biomedical retrieval systems. Finally, an overview of the retrieval evaluation activities that have been proposed is presented as well as a detailed description of the evaluation campaign process.

Chapter 3

Use Case Description

"El mundo es único pero, además, es diverso."

Carlos San Juan

During PROMISE a study was performed to better understand the user needs and its role in the retrieval application. The goal was to evaluate a use case framework. A detailed description of the features that a medical retrieval system should have to satisfy the real needs users was obtained.

Typically, search engines output a ranked list of search results in response to a query. System evaluation has focused on assigning a score to such a ranked list based upon the relevance of each returned document to the information need underlying the query.

However, this kind of evaluation has limitations. The tasks evaluated are abstractions of real tasks. Assumptions about typical end users, their tasks, goals, local environment and social context are often not made explicit. But even if they are not, every test collection has an underlying user and task model. Every decision regarding query topics, relevance assessments and metrics chosen reflects certain assumptions about a typical end user. For example, in ad hoc TREC campaigns, the end user is assumed to issue informational queries [27, 203] (which intended to find information about a query topic), to have liberal relevance criteria [224], and to find duplicates of already seen relevant information still relevant.

In PROMISE [116] a use case framework was developed for explicitly describing the use case associated with an evaluation task: the (desired) functionality of systems under scrutiny, typical end users, their tasks, goals, local environment and social context.

Ad hoc search evaluation tasks like the one described above can work well to establish the usefulness of systems with respect to human activities if the activities in question fit this implicit use case. As it is uncertain that this specific use case would cover a large enough part of human information seeking activities to motivate evaluation based solely on it, it would make sense to look into other kinds of use cases too. The use case framework developed by PROMISE allows for describing very different use cases, broadening the scope of the traditional ad hoc evaluation.

An important approach to evaluation discussed is to conduct user studies [133]. User studies are a very powerful way of controlling variables to isolate those variables that contribute to user satisfaction, task completion time or task accuracy. For example, Turpin and Scholer [239] show that ranking quality in terms of Mean Average Precision (MAP) does not necessarily correlate with task based measures such as task completion time or task accuracy. Smith and Kantor [222] show that user adaptation may play a crucial role here, end users can obtain good results with bad systems by changing their interaction strategies. However, user studies are very expensive in terms of labour, and are typically conducted with only a small number of users and systems. The fact that people are very different and display unexpected behaviour becomes a challenge and limits repeatability of such experiments. A systematic description of underlying use cases and the way they informed choices in the setup of evaluation experiments can help bridge the gap between user studies and benchmarking.

3.1 What is a use case?

Use cases are a well–established system development methodology. A use case is a relatively informal or semi–formal description of a system's behaviour and usage which is intended to capture all the functional requirements of a system by describing the interactions between outside actors and the system to reach the goal of the primary actor [110, 111, 51, 190]. In other words, a use case is a system with its primary actor (the user) captures and organises the functional requirements of the system defining the goal of the primary actor, the outside actors that the system relies on to achieve its goals, and the sequence of actions between the system and the actors. The actions of the primary actor, as formalised in the use case, are mapped onto system components and system development objects, most often using the Unified Modelling Language (UML)[24] [35], for system development and evaluation.

Use cases are typically organised around a main success scenario that describes the simplest path through the use case, the one in which everything goes right and the goal is reached without difficulty. Also all the other scenarios, both those leading to success (possibly through recovery) and those leading to goal abandonment (failure) are described. Each scenario is an instance of the use case, a possible path through it. Usually several scenarios are needed to describe all the required system functionality (with respect to that use case). Also additional information such as the priority and the frequency of the use case and related higher or lower level use cases may be described.

3.1.1 Use case for information retrieval

In the context of Cranfield–style [208] IR evaluation actors are typically not separated properly from the system proper that is to be evaluated. The systems are treated as black–boxes, where different components (e.g., query and document representation and matching mechanisms, language or image processing components) are not considered as separate actors having their own use cases and deserving their own evaluations. The evaluation consists of assessing the ranked output of the system against the input request. Consequently IR evaluations produce a single figure as a result for complicated interaction effects of several components, where the gain or loss in performance becomes difficult to localise and explain. Primary actors and their goals and interaction with the system are rarely explicitly discussed in Cranfield–style studies, but are implicitly included in the experimental design as representation of focused, active and well–spoken users working on

[24] The Unified Modelling Language (UML) is a general–purpose modelling language, and the way the world models not only application structure, behaviour, and architecture, but also business process and data structure (see `http://www.uml.org/`).

topical, well–defined, static and exhaustive retrieval tasks. Essentially, this kind of study can work well to establish the usefulness of systems with respect to activities that fit this narrow use case. If the activities do not fit, evaluations will fail to establish success criteria. As information access technology has moved from this current prototypical domain of topical text retrieval, it has become less (and less) motivated to focus the research efforts on this implicit use case alone. The advent of multimedia as a large information carrier may be the most obvious example, as multimedia is different, used differently, by different users, and for different reasons than text. Thus, to capture the most important criteria for success for a variety of information access systems benchmarking should change to accommodate a variety of users with a variety of needs and goals and searching under varying conditions in varying contexts.

This is where use cases show promise for being useful tools for evaluation of new generations of information access systems. They can be a practical tool to bridge the divide between benchmarking and validation and they can guide the design of benchmarking efforts by requiring the evaluation design to make explicit the intended usage of the evaluated system, and how it provides value for its users.

The PROMISE project has developed a framework for writing use cases for information access evaluation. The goal was to build a resource that could support experimental design in the field of information access by making explicit the user–related functional system requirements and their connections to benchmarking mechanisms. The framework is based on the use case methodology, but the structure was modified somewhat. All of the central components of use cases are in place, but they have been specified to a quite detailed level through identifying several features related to them that can affect the design and evaluation of information access systems. The structure of the use case framework is presented in Figure 3.1.

The framework begins with a summarizing description of the use case. After that the system features are presented, followed by features related to the primary actor. Finally, the features related to interaction between the primary actor and the system are discussed in the session features section. The features related to each of the sections can be found in PROMISE deliverables 2.2 [116] and 2.4 [115].

The PROMISE use case framework presents an elaborate protocol for discovering, identifying, and modelling real life problems that the retrieval system needs to solve.

3.2 Use case description

In this section one of the three main domains under study in the PROMISE project is developed: the medical domain [116]. Therefore, the "visual clinical decision support" use case is defined. The task studied in this use case is to find medical cases/images similar to the one under observation for supporting a clinician's decision making during medical diagnosis using medical images and text describing the case under observation as queries in biomedical literature. To this end the PROMISE use case framework presented in Section 3.1.1 is used. To get an idea of a typical situation, a hypothetical scenario is first described. After that, there is a discussion about the system features, user features, session features, evaluation and UML use case diagram in some detail.

- USE CASE DESCRIPTION
  - Use case name
  - Use case Summary
  - Usage Narrative
- SYSTEM FEATURES
  - System boundary
  - Secondary actors
    - Repository
    - Service provider
    - Morphology processing
    - ...
  - Utilities
    - Devices
- USER FEATURES
  - Primary actor (user)
  - Task Context
  - Local context
- SESSION FEATURES
  - Goal
  - Elements of the interaction pattern
    - Search
      - Query
      - Browsing & Navigation
    - Inspect/assess
    - Manipulate
    - Export
  - Main Success Scenario
- POSSIBLE EVALUATION

Figure 3.1: Structure of the PROMISE use case framework.

3.2.1 Usage narrative

Alonso, a medical graduate, is currently a second year intern in the radiology department of a large university hospital. The clinician supervising Alonso has asked him to perform a medical diagnosis on a patient and has provided him with the patient's latest MRI scans and medical record. Unable to reach a decision as he is not 100% sure about the diagnosis and potential co–morbidities, Alonso decides to search the literature for similar cases by using as queries the MRI scans and also text that describes the medical case under observation. A successful end would be for Alonso to find articles in the literature that help him decide on a medical diagnosis.

As part of his training, Alonso has become quite familiar with medical cases and images, but he does not yet have substantial experience in searching the PubMed [25] collection for

[25] PubMed is a free resource developed and maintained by the National Center for Biotechnology Information (NCBI) at the NLM (see `http://www.pubmed.gov/`).

locating similar cases in the literature. He has used search systems before (e.g., the Web search engines), but he has no knowledge of the internal techniques of IR systems (i.e., he is IR illiterate). Although his mother tongue is not English, his language skills allow him to formulate English queries.

3.2.2 System features

The system under discussion is a biomedical literature retrieval system.

The platform being used can be a desktop or a laptop or a tablet computer, or a cell phone without display size restrictions. The input can be provided through typing or clicking, while a keyboard, mouse or touchscreen would be ways for interacting with the system.

The repository, a biomedical literature collection, typically contains millions of scientific articles published in biomedical journals and other venues such as conferences and workshops. These articles are mostly written in English and a large number of them contain images and graphs. They are high-quality and trusted sources since they are peer–reviewed with a known provenance. Such collections are updated in regular intervals (e.g., weekly) with timely additions of recent scientific articles and are expected to be maintained for the foreseeable future. Their coverage of the literature published in the field is generally very comprehensive.

Such collections and retrieval systems are typically maintained by organizations that provide access to biomedical libraries and tools, such as the NCBI of the NLM in the USA. These are highly trusted service providers that follow a no–cost business model.

3.2.3 User features

The primary actor is a clinical practitioner searching the biomedical literature to find information relevant to a medical case under observation on the basis of the patient's medical imaging exams and medical record; this primary actor has the role of a consumer of the information access system.

The primary actor is typically a single user with a higher level of education, but with varying levels of domain and collection expertise (ranging from medical students and interns to professors of medicine) and also of system expertise (ranging from novices to clinicians with significant experience in using such medical IR systems). However, the primary actors have no knowledge of the internal techniques of IR systems, i.e., they are IR illiterate. Furthermore, the language skills of the primary actor with respect to the information sources, i.e., the biomedical literature, which is mostly written in English, are typically at the very least adequate and very often excellent. Finally, the demographic variables cover a wide spectrum in terms of age (ranging from young medical graduates to older experienced clinicians) and of socio–economic and geospatial variables (ranging from clinicians working in a small hospital in a rural area to those employed by a large university hospital in a metropolitan area).

The task context for this use case is the medical domain. Since the information sources used are scientific articles published in the literature, there are no confidentiality problems. The database is potentially accessible to all clinicians.

During their daily work routine, the clinicians need the information access system to decide on a medical diagnosis for a specific patient given the patient's medical exams and in particular medical images and the overall medical record. This is a complex task since there is a large amount of information to handle and there is also a need to work with multi–modal information (structured data, text and images). This task is highly important as it can be lifesaving for the patients under observation.

The clinicians are highly motivated to use the system because the online access is free and the system supports them in their decision making. The response time should be fast, as clinicians should find relevant information quickly to prevent frustration and time-loss. The typical location is a hospital during daily clinical routine. However, since online access is provided, clinicians may further use the system after work to continue their research on the particular case.

3.2.4 Session features

The goal is clinical decision support for the medical diagnosis of a specific patient under observation on the basis of evidence from their medical imaging exams and medical record. This is an informational task where the aim is to get advice, ideas or suggestions from scientific articles describing medical cases similar to the one under observation and containing images similar to the ones from the current case.

This section looks at elements of the interaction patterns relating to searching, the queries, browsing and navigation, inspecting and assessing results, and exporting or saving searches and/or results. Then, one concrete example of a successful interaction with the system follows.

The main type of search is querying through either a simple or an advanced query interface. Support for browsing and navigation should also be provided, together with support for changing between the different types of search.

Queries are formulated both through specification and also through providing examples and include multiple modalities (structured data, text or images). Advanced query support functionalities improve the effectiveness of the performed searches. The next step of the person is to search for similar cases. This is an example of a multi–modal query with a complex structure.

Navigation support can be performed through filtering the search results (or even the whole collection) based on various features, e.g., the modality acquisition of medical images, the patient age and sex, and also metadata, e.g., the author names, journal titles, or MeSH terms of the articles in the biomedical literature.

Search results are presented as a list sorted by relevance. It is desirable for each result to be presented with its title, a snippet with some text relevant to the query (possibly with the query terms highlighted), and thumbnails of the images it contains. Additional information, such as the MeSH terms under which it is classified or the number and types of images it contains, can also be displayed.

Saving past queries, possible with the whole list of results, or individual search results would be desirable.

3.2.5 One example of a successful flow of interaction

1. The clinician chooses to use the biomedical search engine to find similar cases to the one he is diagnosing.

2. The clinician formulates the query (using text, example images or regions, structured data).

3. The system retrieves the results according to the defined criteria.

4. The clinician peruses the first result page and clicks on a few of the results to read the articles in more detail.

5. Every time a result is clicked, the system presents the full article from the biomedical literature together with its metadata and the images it contains.

6. END: success, the clinician finds the images and articles that help him make a decision on the medical diagnosis of the case under observation.

3.2.6 Evaluation task: image based and case–based retrieval

An evaluation task based on the "visual clinical decision support" for medical diagnosis use case should evaluate several aspects: effectiveness remains the most important, together with efficiency, whereas evaluation of the usability of the user interface to maximise the clinicians' satisfaction with the full system should also be considered.

A medical image–based task was running at ImageCLEF since 2004. The focus is on the retrieval of similar images for a precise information need. In 2009, a medical case–based retrieval task was introduced. The goal of the case–base retrieval task is to retrieve articles from the biomedical literature that might be relevant for a differential diagnosis of the provided case description (which includes images).

For the image–based retrieval task, textual queries with some sample images for each query were given to the participants. In contrast, for the medical case–based retrieval, a case description, with patient demographics, limited symptoms and test results including imaging studies, was provided.

The evaluation focuses on the effectiveness of the medical image and case retrieval, with MAP being the main evaluation metric.

3.2.7 UML use case diagram

Figure 3.2 shows the UML use case diagram which represents the visual clinical decision support use case. The participants involved in the UML use case diagram are users and people involved in the realization and maintenance of the system as well as in its evaluation.

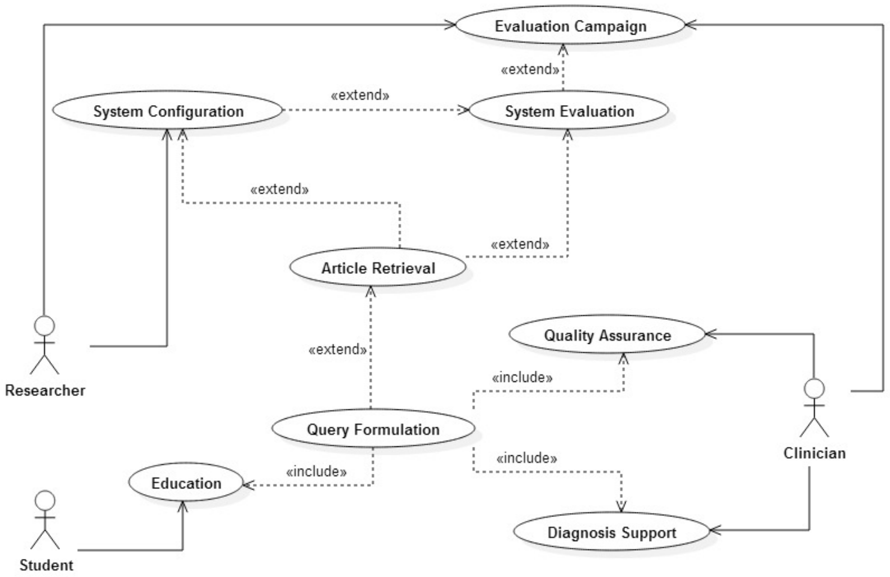


Figure 3.2: UML use case diagram of the visual clinical decision support use case.

3.3 Validation

The validation sought to determine whether the use case specified in Section 3.2 covers the requirements of the stakeholders of the biomedical literature retrieval system and whether it provided realistic descriptions of this system usage and behaviour (see PROMISE deliverable 2.4 [115]).

The validation was an iterative process: user requirements and system usage and behaviour informed the initial specification of use cases and thus the development of the use case framework. The goal of this final round of validation was to further validate and, if necessary, improve the realism, accuracy and coverage of the use cases before the final specification of the use cases and evaluation tasks.

The final validation was carried out by interviewing a group of use case stakeholders, who had not been involved in the previous use case requirement analysis phase. Each interviewer was presented with the use case description (from Section 3.2) and then asked to evaluate it by answering a structured questionnaire. The questionnaire can be found in Appendix A.

The questionnaire contained 5 parts: use case description; system features; user features; session features and evaluation task. The first four parts discuss the use case and were divided into three subsections of questions concerning the realism, accuracy, and coverage of the use case section. The final part asked for the stakeholders view of the usefulness of the evaluation tasks defined, based on the use cases: do the tasks target interesting issues and measure them in a reasonable way?

All the survey contained 52 multiple choice questions and 2 open ended questions– one for general thoughts concerning the use case, and one for the evaluation task. Each use case had their own questionnaire form and collected their own data separately.

Attracting stakeholders not previously involved in the development of the use cases proved difficult. Also, interviewing stakeholders not familiar with the use cases, PROMISE and CLEF required extensive explanations of many concepts before and during the interviews, making the interviews laborious. These factors limited the number of participants.

There were four respondents for the visual clinical decision support for the medical diagnosis use case who were either medical doctors or researchers in the medical imaging domain. They filled in the questionnaire online and did not report any problems with it.

3.3.1 Use case description

The use case description was very positively scored and considered to describe a realistic situation and a complete sequence of events without too many simplifications. However, the possible variations of the flow of interactions and the points of interaction where they may occur are not adequately described.

3.3.2 System feature description

The system feature description was also very positively rated. It was found to be a realistic and accurate system description, identifying correct secondary actors and system utilities. The only (weak) negative answers considered the definition of system boundaries and coverage of all necessary system features but even only one respondent was critical, while others were very positive. None of the respondents found that there were simplifications made in the system description (even if the coverage was slightly criticised by one).

3.3.3 User feature description

The description of the user features was evenly positively scored. The respondents agree that correct users are described realistically and at an appropriate level of detail. The only (weak) negative answer was concerning the simplifications made in description of user features: one of the four respondents indicated that simplifications were made. This result does not concur with the results for the accuracy or coverage of the system and session feature descriptions.

3.3.4 Session feature description

Most problems were identified with the description of the session features. It was indicated that the system–user interaction was not accurately described and not at an appropriate level of detail. Also, some simplifications were identified in the description of the session features. Thus, further information seems to be required concerning the interaction. However, this problem was only mentioned by one of the respondents, while the three others were generally very positive. The description of user goals was found very realistic.

3.3.5 Evaluation task

Only one respondent keeps track of the evaluation task. The problems, technologies, and user groups targeted by the evaluation are relevant according to all respondents. All the participants believe that the case–based retrieval task is the most relevant. They are equally interested in the evaluation of mature and of new and experimental technologies. The participants also agree that the document collection contains realistic data and most of them understand how the ground truth is created for the test collection. Finally, they think that the measurement of clinical accuracy in the search is missing as well as the query times and index sizes.

3.3.6 Conclusions

The overall scores were very positive although there are many disagreements in the answers. This indicates that the use case is well described although the interaction sequences, including their variations, could be described in more detail. Despite the limitations the use case was considered very realistic. Furthermore, all the respondents found that the medical case–based retrieval task is the most relevant. It therefore emphasises the need for a well–resolved framework to tackle this problem. Consequently, this thesis will focus on the medical case–based retrieval task.

3.4 Summary

This chapter describes and validates the use case under study consisting of finding medical cases/images similar to the one under observation for supporting a clinician's decision–making during medical diagnosis. A short introduction is given of the PROMISE use case framework which is used to describe the use case. The main features of the use case are explained and a hypothetical scenario is described. After that, the use case is validated by interviewing stakeholders and end users. The goal is to verify that it reflects usage by real end users, of real systems owned by real service providers (stakeholders). Since the overall scores were very positive the use case can be used to design a well–defined evaluation framework for a retrieval system. Following the suggestions of the people surveyed, the medical case–based retrieval task will be developed as the most realistic problem to be solved. This task aims to retrieve articles from the biomedical literature that might help in the diagnosis of a given case including images.

Chapter 4

Evaluation Framework: ImageCLEFmed

> "What does a fish know about the water in which he swims all his life? "
>
> Albert Einstein

The CLEF conference contributes to the continued evolution of IR by providing access to infrastructure and evaluation resources that support researchers in the development of new approaches, and encouraging collaboration and interaction between researchers both from academia and industry. CLEF has evolved to a self–sustaining and independent annual conference on experimental evaluation with research presentations, panels, poster and demo sessions and laboratory evaluation workshops interleaved during three and a half days of intense and stimulating research activities.

As in 2010, CLEF 2011–2013 were organised in the framework of PROMISE and consisted of an independent conference on a broad range of questions in the fields of multilingual and multi–modal information access evaluation and a set of labs that continued the CLEF tradition of community–based evaluation. The CLEF conference was hosted by the University of Amsterdam in The Netherlands, the University La Sapienza in Italy and the Technical University of Valencia in Spain in September 2011[26], 2012 [27] and 2013[28] respectively.

ImageCLEF is one of the CLEF labs. It is a benchmarking activity on the cross–language annotation and retrieval of images, running since 2003. The main goal of ImageCLEF continues to be promoting multi–modal IR by combining a variety of media including text and images for more effective IR. *ImageCLEFmed* is the medical task offered in ImageCLEF which was added in 2004 and has been held every year since, apart from 2014 [119]. In its 10th edition in 2013, ImageCLEFmed was organised outside of Europe for the first time at the annual American Medical Informatics Association (AMIA)[29] meeting.

[26] For more information about CLEF 2011, see `http://clef2011.clef-initiative.eu/`.

[27] For more information about CLEF 2012, see `http://clef2012.clef-initiative.eu/`.

[28] For more information about CLEF 2013, see `http://clef2013.clef-initiative.eu/`.

[29] The American Medical Informatics Association (AMIA) is a professional scientific association that have sponsored meetings, education, policy, and research programs related to biomedicine, health care and science (see `http://www.amia.org/`).

This chapter first carries out a scholarly impact analysis of the ImageCLEF benchmark up to 2010 (before the beginning of this thesis). The analysis pays particular attention to the medical tasks. Afterward the evolution of the ImageCLEFmed evaluation campaign is chronicled. This thesis focuses on ImageCLEFmed between 2011 and 2013, when it was organised in the context of this thesis.

4.1 ImageCLEF impact analysis (2003–2010)

The contribution of the evaluation campaigns to the field is mainly indicated by the research that would otherwise not have been possible, i.e., research that heavily relies on the use of resources they provide. It is then reasonable to consider that their success can be measured to some extent by the scientific and possibly the economic impact of the research they foster.

The scientific impact of research is commonly measured by its scholarly impact, i.e., the publications derived from it and the citations they receive, and by additional indicators, such as filed patents, whereas its economic impact can be measured, for example, by the technology transfer efforts that result in commercial products and services or by the technological balance of payments and high–technology trade [90]. Other aspects, such as the scientific impact of the increased quality in evaluation methodologies or the economic impact of the time saved by researchers, who now reuse evaluation resources, rather than create them from scratch, are harder to assess. Investigations have reported on the scholarly impact of TRECVid [232] and on the economic impact of TREC [205]. Building on this work, this section presents a study on assessing the scholarly impact of ImageCLEF before this thesis. To this end, it performs a citation analysis on a dataset of publications derived from ImageCLEF. Furthermore, the impact of publications in the medical domain studied in ImageCLEF is investigated.

More details of this study on assessing the scholarly impact of ImageCLEF and CLEF can be found in [236] and [237], respectively.

4.1.1 ImageCLEF tasks

ImageCLEF has organised a number of tasks within two main domains:

- medical image retrieval (ImageCLEFmed);
- general (non–medical) image retrieval from historical archives, news photographic collections, and Wikipedia pages.

These tasks can be broadly categorised as follows:

- *Ad hoc retrieval* – this simulates a classic document retrieval task: given a statement describing a user's information need, find as many relevant documents as possible and rank the results by relevance. In the case of cross–lingual retrieval, the language of the query is different from the language of the metadata used to describe the image. Ad hoc tasks have run since 2004 for medical retrieval and since 2003 for non–medical retrieval scenarios;
- *Object and concept recognition* – although ad hoc retrieval is a core image retrieval task, a common precursor is to identify whether certain objects or concepts from a pre–defined set of classes are contained in an image (object class recognition),

assign textual labels or descriptions to an image (automatic image annotation) or classify images into one or many classes (automatic image classification). Such tasks, including a medical image annotation and a robot vision task, have run since 2005;

- *Interactive image retrieval* – in 2003 and 2004, a user–centred task was run as a part of ImageCLEF and eventually followed by the Interactive CLEF (iCLEF) lab in 2005. Interaction in image retrieval can be studied with respect to how effectively the system supports users with query formulation, translation (for cross–lingual IR), document selection and examination.

Table 4.1 summarises the ImageCLEF tasks that ran between 2003 and 2010 and shows the number of participants for each task along with the distinct number of participants in each year. The number of participants and tasks offered by ImageCLEF has continued to grow steadily throughout the years, from four participants and one task in 2003, reaching its peak in 2009 with 65 participants and seven tasks. The number of participants, i.e., research groups that officially submit their runs, is typically much smaller than the number of groups that register and gain access to the data; e.g., in 2010, 112 groups registered but only 47 submitted runs. Given its multi–disciplinary nature, ImageCLEF participants originate from a number of different research communities, including (visual) IR, cross–lingual IR, computer vision and pattern recognition, medical informatics, and human-computer interaction. Further information can be found in the ImageCLEF book [176] describing the formation, growth, resources, tasks, and achievements of ImageCLEF.

Table 4.1: Participation in the ImageCLEF tasks and number of participants by year.

Task	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
General images								
Photographic retrieval	4	12	11	12	20	24	19	–
Interactive image retrieval	1	2	*2*	*3*	–	*6*	*6*	–
Object & concept recognition	–	–	–	4	7	11	19	17
Wikipedia image retrieval	–	–	–	–	–	12	8	13
Robot vision	–	–	–	–	–	–	7	7
Medical images								
ImageCLEFmed	–	12	13	12	13	15	17	16
Medical image annotation	–	–	12	12	10	6	7	–
Total (distinct)	4	17	24	30	35	45	65	47

4.1.2 Bibliometric analysis method

Bibliometric studies provide a quantitative and qualitative indication of the scholarly impact of research by examining the number of scholarly publications derived from it and the number of citations these publications receive. The most comprehensive sources for publication and in particular for citation data are:

- *Thomson Reuters Web of Science*[30] – established by Eugene Garfield in the 1960s;

[30] Thomson Reuters Web of Science provides a destination to access reliable, integrated, multidisciplinary research (see `http://thomsonreuters.com/thomson-reuters-web-of-science/`).

- *Scopus*[31] – introduced by Elsevier in 2004;
- *Google Scholar*[32] – freely available since 2004, developed by Google.

In addition to publication and citation data, Thomson Reuters Web of Science and Scopus also provide citation analysis tools to calculate various metrics of scholarly impact, such as the h–index [102], a robust metric of scientific research output that has a value h for a dataset of N_p publications, if h of them have at least h citations each, and the remaining $(N_p - h)$ publications have no more than h citations each. Google Scholar on the other hand is simply a data source and does not have such capabilities; citation analysis using its data can however be performed by the Publish or Perish (PoP)[33] system, a software wrapper for Google Scholar.

Each of these sources follows a different data collection policy that affects both the publications covered and the number of citations found. Thomson Reuters Web of Science has a complete coverage of more than 10,000 journals going back to 1900, but its coverage of conference proceedings or other scholarly publications, such as books, is very limited or non–existent. For instance, in the field of computer science, Thomson Reuters Web of Science only indexes the conference proceedings of the Springer Lecture Notes in Computer Science and Lecture Notes in Artificial Intelligence series. The citations found are also affected by its collection policy, given that in its General Search, Scopus aims to provide a more comprehensive coverage of research literature by indexing nearly 18,000 titles from more than 5,000 publishers, including conference proceedings and "quality web sources".

Google Scholar, on the other hand, has a much wider coverage since it includes academic journals and conference proceedings that are not Thomson Reuters Web of Science– or Scopus–listed, and also books, white papers,theses, preprints, abstract and technical reports, which are sometimes highly cited items as well. However, the choice of Google Scholar to include documents is not always clear.

As is evident, these differences in coverage can enormously affect the assessment of scholarly impact metrics, although the degree to which this happens varies among disciplines [15, 97]. For computer science, where publications in peer–reviewed conference proceedings are highly valued and cited in their own right, without necessarily being followed by a journal publication, Thomson Reuters Web of Science greatly underestimates the number of citations found [198, 15], given that its coverage of conference proceedings is only very partial, and thus disadvantages the impact of publications. For example, Harzing [97] examined the effect of using different data sources for citation analysis across different disciplines and she found that for a particular case of an established computer science academic, Scopus found 62% more publications and 43% more citations than Thomson Reuters Web of Science. Scopus' broader coverage can however be hindered by its lack of coverage before 1996, but this is not a problem in this case since the ImageCLEF evaluation campaign started in 2003. Google Scholar offers an even wider coverage than Scopus and thus benefits citation analyses performed for the computer science field [198, 97]. As a result, this study employs both Scopus and Google Scholar (in particular its PoP wrapper)

[31] Scopus provides access to peer-reviewed research (see `http://www.scopus.com/`).

[32] Google Scholar provides a simple way to broadly search for scholarly literature (see `http://scholar.google.com/`).

[33] Publish or Perish (PoP) is a software program that retrieves and analyses academic citations (see `http://www.harzing.com/pop.htm`).

for assessing the scholarly impact of ImageCLEF until 2010. Scopus and Google Scholar were also employed in the examination of the TRECVid scholarly impact [232], where emphasis was however mostly given on the Google Scholar data.

It should be noted that the reliability of Google Scholar as a data source for bibliometric studies is being received with mixed feelings [15], and some outright scepticism [112, 113]. This is due to its widely reported shortcomings [198, 112, 113, 15], which mainly stem from its parsing processes. In particular, Google Scholar frequently has several entries for the same publication, e.g., due to misspellings or incorrectly identified years, and, therefore, may deflate its citation count [198, 112]. This however can be rectified through Google Scholar and through PoP which allow for the manual merging of entries deemed to be equivalent. Inversely, Google Scholar may also inflate the citation count of some publications, since it may group together citations of different papers, e.g., the conference and journal version of a paper with the same or similar title or its pre–print and journal versions [198, 112]. Furthermore, Google Scholar is not always able to correctly identify the publication year of an item [112]. These deficiencies were taken into account in this analysis and addressed with manual data cleaning when possible, although the validity of the citations in Google Scholar were not examined.

4.1.3 The dataset of the ImageCLEF publications

CLEF's annual evaluation cycle culminates in a workshop where participants of all CLEF labs present and discuss their findings with other researchers. This event is accompanied by the CLEF workshop proceedings, known as *working notes*, where research groups publish, separately for each lab, notebook papers that describe the techniques used in their participation and results. In addition, the organisers of each lab (and/or each task within each lab) publish overview papers that present the evaluation resources used, summarise the approaches employed by the participating groups, and provide an analysis of the main evaluation results. The papers in the CLEF working notes are available online on the CLEF website and while they are not refereed, the vast majority of participants take the opportunity to publish there. Since 2014, CEUR[34] contains CLEF publications including previous publications,

After the workshop, participants are invited to publish more detailed descriptions of their approaches and more in–depth analyses of the results of their participation, together with further experimentation, if possible, to the *CLEF proceedings*. These papers go through a reviewing process and the accepted ones, together with updated versions of the overview papers, were published in a volume of the Springer Lecture Notes in Computer Science series in the year following the workshop and the CLEF evaluation campaign. That means that the CLEF proceedings of the CLEF 2005 evaluation campaign were published in 2006. This publication scheme was followed until 2009; in 2010 the format of CLEF changed and the participants' and overview papers were only published in the CLEF working notes, i.e., there were no post conference CLEF proceedings.

Moreover, CLEF participants may extend their work and publish in journals, conferences, and workshops. The same applies for research groups from academia and industry that, while not official participants of the CLEF activities, may decide at a later stage to use CLEF resources to evaluate their approaches. These *CLEF–derived* publications are a good indication of the impact of CLEF beyond the environment of the evaluation

[34] The CEUR Workshop Proceedings is a free open–access publication service (see `http://ceur-ws.org/`).

campaign. Furthermore, researchers directly involved with the development of CLEF evaluation resources and/or the coordination of labs and tasks also publish elsewhere detailed descriptions of the applied methodologies, analyses of the reliability of the created resources, and best practices. These *CLEF resources* publications can be seen as complementary to the overview papers in the CLEF proceedings and working notes.

To assess the scholarly impact of ImageCLEF, bibliometric analysis can be applied to the dataset of publications that contains:

- the ImageCLEF–related publications in the *CLEF working notes*;
- the ImageCLEF–related publications in the *CLEF proceedings*;
- papers describing *ImageCLEF resources* (typically written by ImageCLEF organisers/coordinators);
- *ImageCLEF–derived* publications where ImageCLEF datasets are employed for evaluating the research that is carried out.

Although publications in the CLEF working notes do attract citations, given that Scopus does not index them, they are excluded from this analysis. Moreover *ImageCLEF–derived* publications are also excluded because locating all publications that use ImageCLEF data is a hard task. One may assume that such papers would cite the overview article of the corresponding year of ImageCLEF but often only the URL of the benchmark is mentioned; or that such papers are written by researchers having access to the data.

4.1.4 Analysis of the ImageCLEF publications

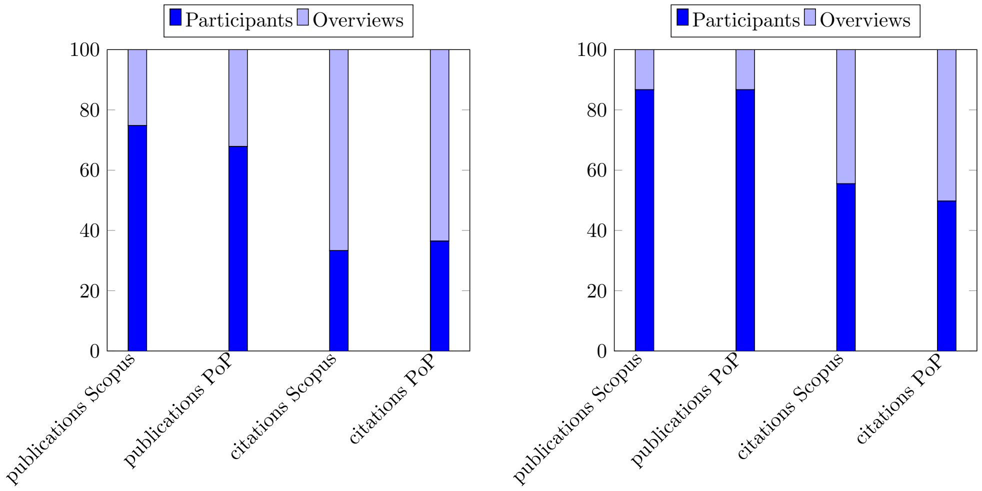
The results of the study to assess the scholarly impact of ImageCLEF, presented in Table 4.2, show that there were a total of 195 ImageCLEF–related papers in the CLEF proceedings published between 2004 and 2010. Over the years, there is a steady increase in such ImageCLEF publications, in line with the continuous increase in participation and in the number of offered tasks (see Table 4.1). The coverage of publications regarding ImageCLEF resources varies greatly between Scopus and Google Scholar, with the former indexing a subset that contains only 57% of the publications indexed by the latter. These publications peak in 2010, which coincides with the year that ImageCLEF organised a benchmarking activity as a contest in the context of the International Conference for Pattern Recognition (ICPR). This event was accompanied by several overview papers describing and analysing the ImageCLEF resources used in the contest, published in the ICPR 2010 [108] and ICPR 2010 Contest [240] proceedings.

The number of citations varies greatly between Scopus and Google Scholar. For the publications in the CLEF proceedings, Google Scholar finds almost nine times more citations than Scopus. Apart from the wider coverage of Google Scholar, this is also partly due to its inability to distinguish in some cases publications with the same or similar title published in different venues, as is sometimes the case with papers published in the CLEF working notes and in the CLEF proceedings. Differentiating between the citations of two such versions of a CLEF paper requires extensive manual data cleaning that examines the list of references in the citing papers, a task which is beyond the scope of this study. Nevertheless, the inclusion of the citations to the CLEF working notes versions of some CLEF proceedings papers is considered acceptable in the context of this analysis, since they are still indicative of ImageCLEF's scholarly impact. When examining the distribution of citations over the years, Scopus indicates a variation in the number of citations, while Google Scholar shows a relative stability from 2005 onwards. For publications regarding ImageCLEF resources, Google Scholar finds almost five times more citations than Scopus. These peak for papers published in 2006 and 2004, mainly due to three publications that describe the creation of test collections that were used extensively in ImageCLEF in the following years, and thus attracted many citations. Overall, Google Scholar indicates that the total number of citations over all 249 publications in the considered dataset is 2,147, resulting in 8.62 average cites per paper. This is comparable to the findings of the study on the scholarly impact of TRECVid [232], with the difference that they consider a much larger dataset of publications that also includes TREC–derived papers.

Table 4.2: Overview of ImageCLEF publications 2004–2010 and their citations. #P and #C are the number of papers and the number of citations, respectively.

		CLEF proceedings			**ImageCLEF resources**			**All**		
	Year	#P	#C	*h*–index	#P	#C	*h*–index	#P	#C	*h*–index
Scopus	2004	5	13	2	4	31	3	9	44	4
	2005	20	50	4	–	–	–	20	50	4
	2006	25	24	3	3	28	1	28	52	3
	2007	27	25	2	6	29	2	33	54	3
	2008	29	18	3	5	22	2	34	40	3
	2009	45	14	2	2	4	1	47	18	2
	2010	44	38	4	11	7	2	55	45	4
	Total	195	182	6	31	121	5	226	303	9
Google Scholar	2004	5	65	3	5	105	4	10	170	6
	2005	20	210	8	5	47	4	25	257	10
	2006	25	247	7	8	144	5	33	391	9
	2007	27	259	7	10	76	4	37	335	9
	2008	29	249	7	7	73	5	36	322	9
	2009	45	284	7	7	53	4	52	337	9
	2010	44	259	7	12	76	6	56	335	10
	Total	195	1,573	18	54	574	13	249	2,147	22

Next, the distribution of citations over different types of papers is analysed. First, a comparison of the participants' papers in the CLEF proceedings with overviews describing ImageCLEF resources published both in the CLEF proceedings and elsewhere is carried out. Figure 4.1 compares the relative number of papers with the relative citation frequency for these publication types. While participants' papers account for a substantial share of the publications, namely 74.8% for Scopus and 67.9% for Google Scholar, they receive around 35% of the citations. Even when considering only the CLEF proceedings, i.e., when excluding the ImageCLEF resource papers published elsewhere so as to limit the bias towards overview papers that comes from including this dataset in the analysis, Figure 4.1 indicates that while participants' publications constitute 86.7% of the total, they attract around 50% of the citations. These results indicate the significant impact of the ImageCLEF overview papers.



(a) All.

(b) CLEF proceedings.

Figure 4.1: Relative impact of ImageCLEF publication types.

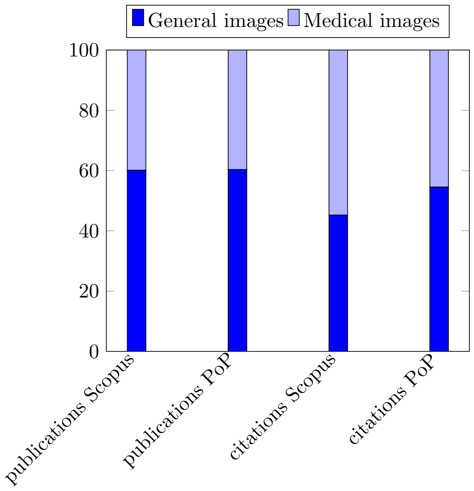


Figure 4.2: Relative impact of ImageCLEF publication in the two domains.

Publications on the ImageCLEF medical domain

Figure 4.2 compares the relative number of publications with the citation frequency for the domains. It should be noted that some publications examine both domains at once, e.g., participants' papers presenting their approaches in ImageCLEF tasks that represent both domains or overview papers reporting on all tasks in a year. Overall, the publications in the medical domain appear to have a slightly higher impact.

A total of 249 publications were analysed obtaining 2,147 citations in Google Scholar and 303 in Scopus. With the proceedings covering almost 230 papers and the non–reviewed working notes a larger number, 500 articles have already been published in this context. Taking into account the derived work, over 1,000 articles can be expected to be based on ImageCLEF data.

4.2 The ImageCLEFmed evaluation campaign before this thesis (2004–2010)

2004 was the beginning of the medical image retrieval and classification tasks at ImageCLEF [47]. The collection used for this task was a subset of the CasImage collection [204], a dataset of anonymised medical images and associated notes from the University Hospital of Geneva. These textual annotations, in English or French, consisted of a number of fields including diagnosis, clinical presentation, keywords, title and unstructured description and were associated with a case that can include multiple images. Not all fields were populated for all cases and the annotations that were present may have had problems typical of real–life clinical notes such as abbreviations, spelling errors, and other linguistic problems as well as challenges with multilingual collections such as incorrect French accents. The query tasks were selected by a radiologist and were made available to participants in the form of a sample image. Thus, this was a query by example task and the goal was to retrieve similar images, where similarity was based on modality, anatomical locations and imaging protocols. Participants could use purely visual techniques (CBIR) as well as text retrieval techniques based on the notes associated with the sample image. A radiologist, a medical doctor and a medical computer scientist performed the relevance assessments on pools created from the submissions. Images were judged using a ternary scale as relevant, partially relevant or not relevant. Based on these assessments, relevance sets used for the judging were created in a number of ways. These include deeming an image to be relevant only if all 3 agree (most strict), relevant if all three judges say that the images were relevant or partially relevant, relevant if at least 2 judges say that the image is relevant, relevant if any of judges say that the image is at least partially relevant (most lenient).

The size of the dataset was greatly increased for the 2005 medical retrieval task from the 6,000 images in 2004. In addition to the CasImage collection, images from Pathology Education Instructional Resource (PEIR)[35] , images from the Mallinckrodt Institute of Radiology (MIR) and the PathoPic[36] collection were also made available. The PEIR collection of about 33,000 pathology images included annotations in English associated at the image level, the MIR dataset consisted of about 2,000 nuclear medicine images and had English annotations at the case level and the Pathopic collection consisted of about 9,000 images with extensive German annotations and incomplete English translations. Thus, this large and diverse collection of over 50,000 images contained images from radiology, nuclear medicine and pathology with annotations in English, French and German that were associated with the images at either the images level or the case level where a single annotation could apply to multiple images. Twenty–five query topics were defined based on a user survey conducted at Oregon Health and Science University (OHSU) and developed along the following axes: anatomy, modality, pathology or disease and abnormal visual observation. Twelve of the 25 query topics were thought to be best suited for visual systems, eleven for mixed systems while a couple were semantic query topics where visual features were not expected to improve performance [46]. Relevance assessments were performed by 9 judges, most of whom were clinicians while one was an image–processing specialist. Pools were created using the top 40 results from each run resulting in pools of approximately 900 images. A ternary scale was used again and relevance sets were created

[35] The Pathology Education Instructional Resource (PEIR) is a multidisciplinary public access image database for use in medical education (see `http://peir.path.uab.edu/library/`).

[36] PathoPic is a public access image database providing images of high quality for use in medical education and public health information (see `http://alf3.urz.unibas.ch/pathopic/`).

in a few different ways from most strict to most lenient.

The same dataset was used again in 2006 [177]. However, the query topics were selected based on search logs of a medical media search engine created by the Health On the Net (HON) foundation [175]. Thirty search topics were generated with ten each expected to be amenable to visual, textual and mixed search methods. Seven clinicians from OHSU performed the relevance assessments.

In 2007, in addition to the dataset used in 2005 and 2006, two more datasets were added [178]. These included the myPACS dataset of about 15,000 primarily radiology images annotated in English at the case level and about 1,500 images from the Clinical Outcomes Research Institute (CORI) dataset of endoscopic images annotated in English at the image level. This combined dataset of more than 66,000 images had annotations in English, French and German and images of a variety of modalities. Thirty query topics from PubMed log files were selected that sought to cover at least two of the axes (modality, anatomy, pathology and visual observation) and again 30 search topics were created. The top thirty images from each run were combined to create the pools with an average pool size of about 900. Judges were clinicians that were also students in the OHSU biomedical informatics graduate program.

A new database was used in 2008 but the task remained essentially the same as in 2007 [185]. The Radiological Society of North America (RSNA) had made available a set of about 66,000 images published in two radiology journals (Radiology and Radiographics). These images were a subset of the images used by the Goldminer search engine [118]. The high quality annotations associated with the images were the figure captions published in the journal. However, the images were primarily radiology focused unlike in previous years where pathology and endoscopic images were also represented. The query topics were selected from the query topics previously used between 2005 and 2007. Training data was also made available. This consisted of the images and annotations as well as the query topics, sample images and relevance judgements ("qrel"). The judges again were clinicians who were students in OHSU's biomedical informatics training program.

In 2009, the size of the image dataset increased to over 74,000 [183]. These images again were provided by RSNA (similar to 2008) and were part of the Goldminer database. The 2009 search topics were selected from a set of queries created by clinicians participating in a user study of medical search engines. In addition to "ad hoc" search topics, case–based query topics were introduced for the first time. These case–based query topics are meant to more closely resemble the information needs of a clinician in a diagnostic role. Teaching files in CasImage were used to create five query topics. A textual description and a set of images were provided for each case but the diagnosis was withheld and only given to the judges for assessment.

The RSNA dataset of about 77,500 images was used in 2010 [184]. The 16 image–based search topics were selected, as in 2009, from query topics that had been searched for in the above–mentioned user study. Additionally, fourteen case–based query topics were provided. Based on research that had demonstrated the improvements in early precision obtained in filtering out images of non–relevant modalities [17], a modality classification sub–task was added in 2010. The goal of this subtask was to classify an image into one of 8 classes (CT, MRI, nuclear medicine, PET, ultrasound, X–ray, optical and graphics). A training dataset of 2,390 images was provided and the test set had 2,620 images.

As seen in Figure 4.3, the number of images in the collections has grown from 6,000 to over 300,000 over 10 years. As seen in Figure 4.4, the number of groups submitting runs generally increased from ten during the first year to about 17. The total number of runs submitted fluctuated over the years as seen in Figure 4.5 depending on the number of sub–tasks being organised. The number of registrations has increased strongly from about 10 in 2004 to around 70 in 2012.

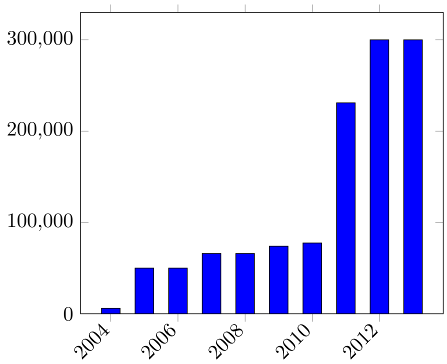


Figure 4.3: Each bar represents the number of images in the ImageCLEFmed collection over the years.

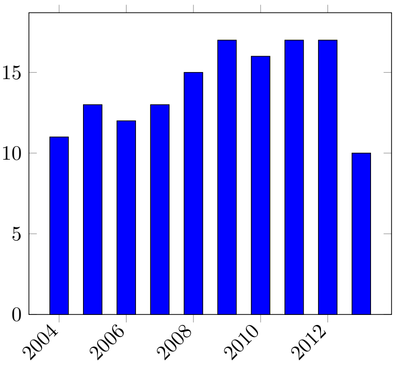


Figure 4.4: Each bar represents the number of groups submitting runs for the ImageCLEFmed task over the years.

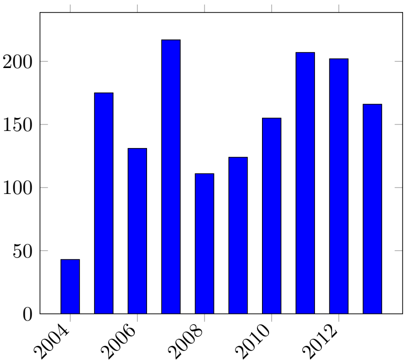


Figure 4.5: Each bar represents the total number of runs submitted for the ImageCLEFmed task over the years.

Many groups then do not feel confident about the results but often continue working and publishing on the data well after the collections. Participation in tasks is often seen as good when sure to obtain good results even though the workshop highlighted talks from interesting techniques and not necessarily the best performing techniques.

4.3 ImageCLEFmed during this thesis (2011–2013)

This section reports an overview of the ImageCLEFmed campaigns that have been organised and taken place during this thesis (2011–2013). Further material can be found in the ImageCLEF overview articles [122, 180, 79].

4.3.1 Database

A new database was created to allow for new challenges in ImageCLEFmed 2011. The database is a subset of 231,000 images from the PMC database containing in total over 1,700,000 images in 2014. PMC contains all articles in PubMed that are open access but the exact copyright for redistribution varies among the journals. The subset chosen includes all journals of BioMed Central, as these allow redistribution of the data. A set of imaging oriented journals that also allow redistribution were taken in addition to this.

In ImageCLEFmed 2012, a larger subset of PMC than 2011 was provided. The database contains over 300,000 images of 75,000 articles of the biomedical open access literature that allow free redistribution of the data. In 2013, the same database as in 2012 was supplied to the participants. Figure 4.6 contains several examples of images from the biomedical literature.

The ImageCLEFmed 2013 database is used in this thesis to carry out the experimental analysis of the proposed retrieval system (see Chapter 5).

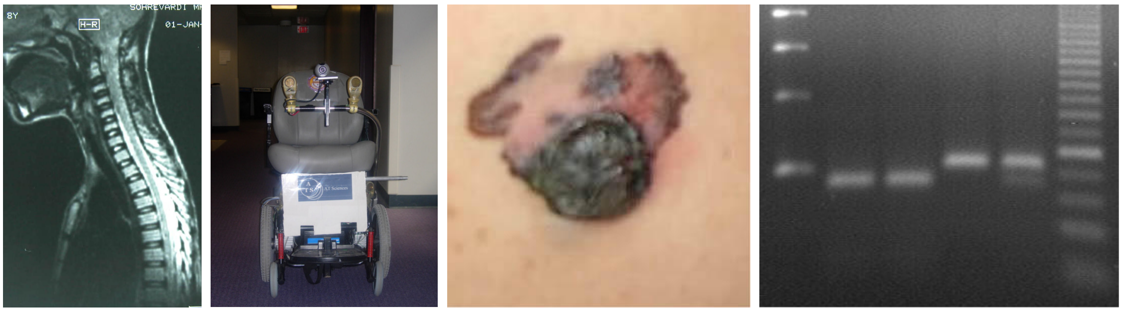


Figure 4.6: Examples of images found in the ImageCLEF databases.

4.3.2 Tasks

ImageCLEFmed has proposed several sub–tasks over the years since 2004, always in an end-user oriented way based on surveys or log files analysis. After the validation of the use case description in 2011 presented in Chapter 3, the evaluation tasks proposed were developed in a way that the problem to solve fit the use case. Therefore, four types of sub–tasks were conducted by ImageCLEFmed since 2011:

- image–based retrieval;
- medical case–based retrieval;

- modality classification;
- compound figure separation.

In this section each of the mentioned sub–tasks is described. The details concerning the set–up of the tasks are presented. Figure 4.3 shows the details about the ImageCLEFmed collections between 2011 and 2013.

Table 4.3: Overview of the data distributed by ImageCLEFmed between 2011 and 2013.

Task	**2011**		**2012**		**2013**	
	# images	# topics	# images	# topics	# images	# topics
Image–based retrieval	231,000	15	300,000	22	300,000	35
Case–based retrieval	231,000	15	300,000	22	300,000	35
	# images	# classes	# images	# classes	# images	# classes
Modality classification	2,000	18	2,000	31	5,483	31
Compound fig. separation	–	–	–	–	2,967	–

Image–based retrieval

The image–based retrieval task is the classical medical retrieval task, similar to those organised each year since 2005 with the target unit being the image. The goal is to retrieve similar images where similarity is based on the relevance of the retrieved images.

Query topics In 2011, the query topics for the image–based retrieval task were a selection of query topics that had been used in the past based on [100, 179]. They were generated from a variety of real–world Internet medical search engine logs. In 2012 and 2013, the query topics were created based on a selection of queries from search logs of the Goldminer radiology image search system [238]. Only queries occurring 10 times or more (about 200 queries) were considered as candidate query topics for this task. A radiologist assessed the importance of the candidate query topics, resulting in 50 candidate query topics that were sure to occur at least a few times in the database. A subset of the resulting queries were then distributed among the participants and example query images were selected from a past collection of ImageCLEFmed [100].

15, 22 and 35 query topics were given to the participants in 2011, 2012 and 2013 respectively. The 22 query topics used in 2012 were part of the 35 query topics used in 2013. All the query topics contain text (in English, French, German and, since 2012, also Spanish) with 1–7 sample images for each query.

The query topics were classified as “textual”, “mixed” or “semantic”, based on the methods that are expected to yield the best results.

Figure 4.7 shows one of the distributed query topics to the ImageCLEFmed participants.

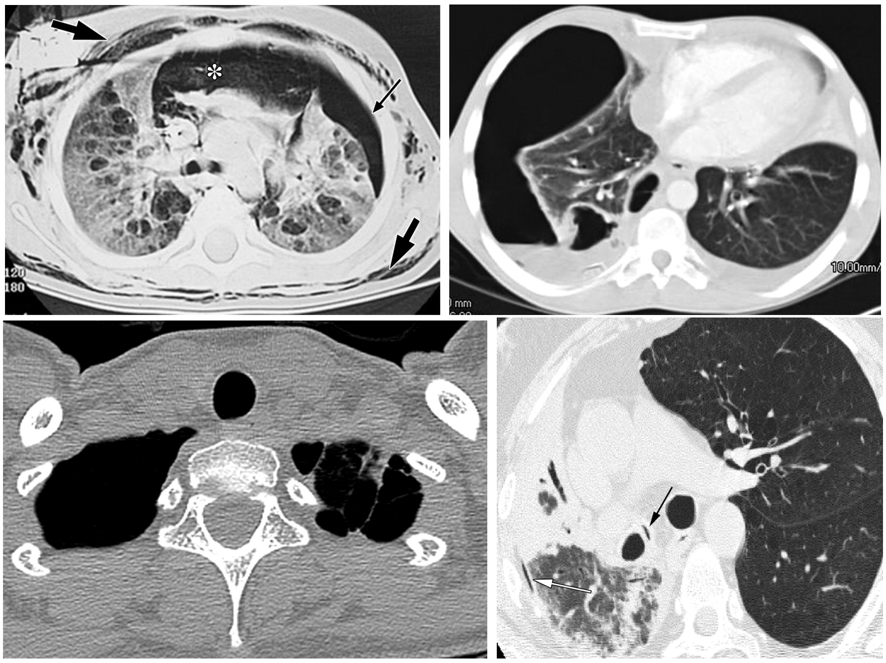


Figure 4.7: Images from one of the query topics in the image–based retrieval task of ImageCLEFmed 2013. They correspond to the textual query "pneumothorax CT images" that is also expressed in French, German and Spanish.

Relevance judgements Hersh et al. [100] describe in detail the procedure to perform the relevance judgements. Physicians who were also current or former students of the OHSU biomedical informatics graduate program[37] were paid an hourly rate to judge the images. The pools for relevance judging were created by selecting the top ranking images from all submitted runs. The actual number selected from each run varied by year but was usually about 30–40, with the goal of having pools of about 800–1,200 images in size for judging. Judges were instructed to rate images in the pools as definitely relevant, partially relevant, or not relevant. Besides a short description for the judgements, a full document was prepared to describe the judging process including what should be regarded as relevant. Many query topics were judged by two or more judges to explore inter–rater agreements and its effects on the robustness of the rankings of the systems.

Evaluation The results were computed with the trec_eval software[38] (version 9.0) following the ImageCLEFmed practice. The trec_eval software is available to the retrieval research community at large, so organizations can evaluate their own retrieval systems at any time. This software computes a large array of measures including the ones used for ImageCLEFmed: MAP; Geometric Mean Average Precison (GMAP); bpref; precision 10 (P10) and precision 30 (P30). Details of the measure definitions are found below:

- *Precision n (Pn)* – for each query topic, precision is the percentage of retrieved items

[37] Oregon Health and Science University (OHSU) biomedical informatics graduate program trains future professionals, researchers, and leaders in the broad area of biomedical and health informatics (see `http://www.ohsu.edu/xd/education/schools/school-of-medicine/departments/clinical-departments/dmice/educational-programs/`).

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

which are relevant. Pn is the precision after n items have been retrieved:

$$Pn = \frac{m}{n} \tag{4.1}$$

where m judged relevant documents have been retrieved at rank n;

- *Average Precision (AP)* – is the average of the precision after each relevant item is retrieved, for a query topic:

  $$AP = \sum_{n} \frac{Pn}{N} \tag{4.2}$$

  where n is the rank of each judged relevant document, N is the number of judged relevant documents and Pn is the precision of the top-n retrieved documents;

- *MAP* – is the mean of the AP scores over all of the query topics;

- *GMAP* – is the geometric mean of per–topic AP:

  $$GMAP = exp\frac{1}{n}\sum_{n} logAP_n \tag{4.3}$$

  where n is typically 50 for with the trec_eval;

- *bpref* – is based on the relative ranks of judged documents only:

  $$bpref = \frac{1}{N}\sum_{n}(1 - \frac{|m \text{ ranked higher than } n|}{min(N, M)}) \tag{4.4}$$

  where N is the number of judged relevant documents, M is the number of judged irrelevant documents, n is a relevant retrieved document, and m is a member of the first N irrelevant retrieved documents.

MAP has been chosen as a lead metric although the measures cited above have been also analysed. Since MAP is the mean of the APs for all the query topics, it favours systems that return more relevant documents at the top of the list. For a single query topic, the AP approximates the area under the uninterpolated precision–recall curve, therefore, the MAP is approximately the average area under the precision–recall curve for a set of queries. However, the maximum MAP that a system can achieve is limited by its recall, and systems can have very high early precision despite having low MAP [121].

When using web–based interfaces, users are interested on how many good results there are on the first page or the first three pages. Precision measures, such as P10 or P30, the ability of a system to present only relevant items. GMAP measures improvements for low–performing query topics by weighting preferentially query topics with very low AP. The bpref measure is designed for situations where relevance judgements are known to be incomplete. It computes a preference relation of whether judged relevant items are retrieved ahead of judged irrelevant items. When the judgements are complete bpref and MAP are very highly correlated. However, if the judgements are incomplete, rankings of systems by bpref correlate highly to the original ranking, whereas rankings of systems by MAP do not.

For more details on the measures chosen see [244].

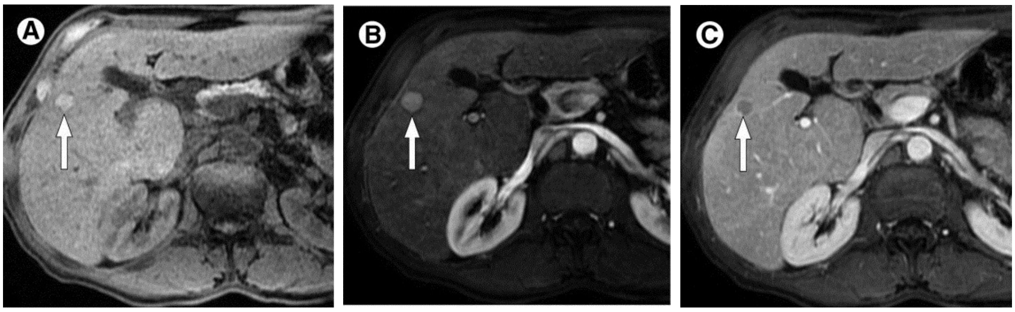


Figure 4.8: Images from one of the query topics in the medical case–based retrieval task of ImageCLEFmed 2013. They correspond to the textual query "A 56–year–old woman with Hepatitis C, now with abdominal pain and jaundice. Abdominal MRI shows T1 and T2 hyperintense mass in the left lobe of the liver which is enhanced in the arterial phase".

Medical case–based retrieval

The medical case–based retrieval task was first introduced in 2009 using a smaller database [183] than between 2011 and 2013. The goal was to move image retrieval potentially closer to clinical routine by stimulating the use case of a clinician who is in the process of diagnosing a difficult case. This is a more complex task but one that is considered closer to the clinical workflow (see Chapter 3). In this task, a case description is provided. The goal is to retrieve articles from the biomedical literature that are useful in differential diagnosis. Unlike the ad hoc task, the unit of retrieval here is a case, not an image.

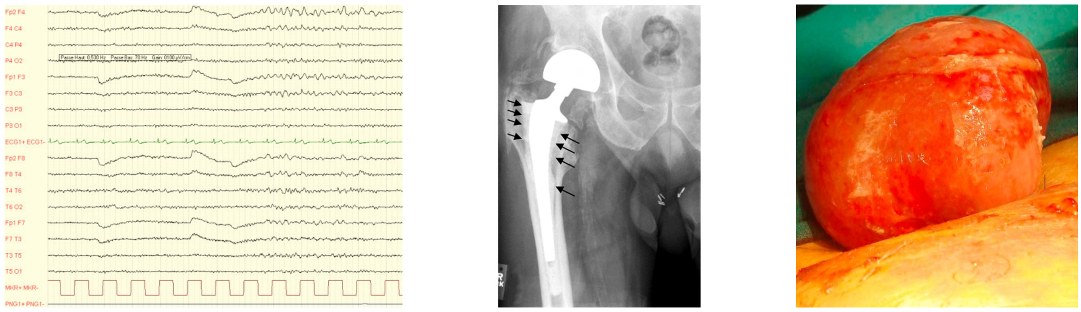
Query topics Initially, the query topics were created based on cases from the teaching file CasImage [204]. More query topics were created at NLM by physicians based on their experience. Each query topic consists of a case description with patient demographics, limited symptoms and test results including imaging studies (but not the final diagnosis). Each of the query topics was accompanied by one to three images. An example of a query topic can be seen in Figure 4.8. 15 query topics were given to the participants in 2011 and 22 in 2012. The 22 query topics used in 2012 were a subset of the 35 query topics used in 2013.

Relevance judgements The relevance judgements were performed using the same system as for the image–based query topics. The system was adapted for the case–based query topics showing the article title and several images appearing in the text. Each article in each pool was judged to be "relevant", "partly relevant" or "not relevant" for a differential diagnosis. An article was judged as partly relevant if the assessor could not define if it was relevant or not.

Evaluation The evaluation was performed using the the trec_eval software. The same measures as for the image–based task were considered.

Modality classification

The modality classification task was first introduced in 2010. The goal of this task is to classify the images into medical modalities and other image types, such as CT, X–ray or general graphs (see Figure 4.9).



(a) Printed signals, waves: electroencephalography.

(b) Radiology: x–Ray, 2D Radiography.

(c) Visible light photography: other organs.

Figure 4.9: Images from some modalities in the modality classification task of ImageCLEFmed 2013.

In 2011, an ad hoc hierarchy was created with 18 classes in the sections of radiology, microscopy, photography, graphics and other based on the ImageCLEFmed 2011 database (see Section 4.3.1). Figure 4.10 shows this hierarchy and its class codes with descriptions ([Class code] Description).

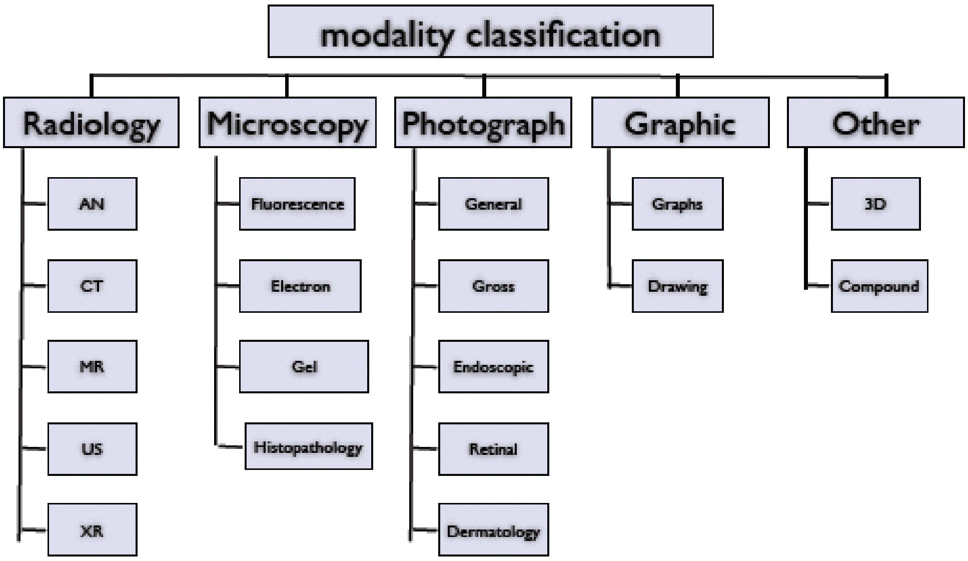
An improved ad hoc hierarchy with 31 classes in the sections of compound or multipane images, diagnostic images and generic biomedical illustrations was created based on the existing data set [182] in 2012. The hierarchy shown in Figure 4.11 was used for the modality classification, more complex than the classes in ImageCLEF 2011. The class codes with descriptions are shown in Figure 4.11 ([Class code] Description).

In 2013, the same hierarchy as in ImageCLEFmed 2012 was used. A larger number of compound figures than in ImageCLEFmed 2012 were provided in the training and test data sets. The current distribution corresponds to that in the PMC data set, much closer to reality than in previous years.

In 2011 and 2012, 1,000 training images and 1,000 test images were provided to the participants while in 2013 they were 2,582 training images and 2,901 test images. Labels for the training images were known whereas labels for the test images were distributed after the results submission only.

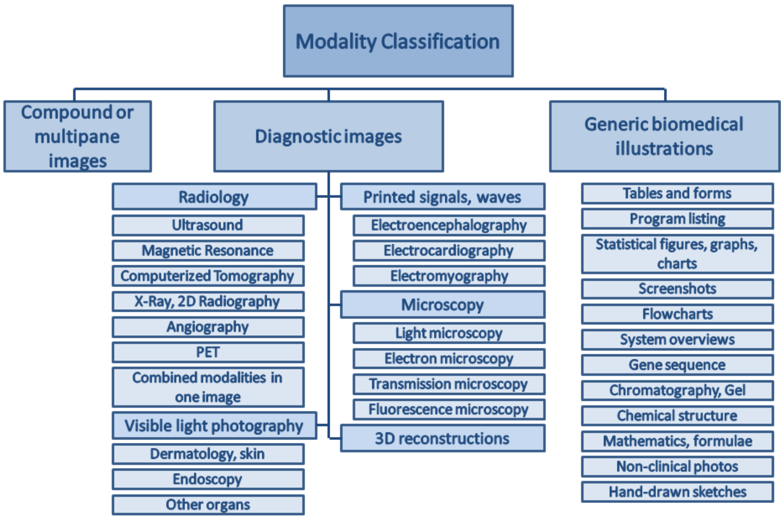
Ground truth generation Foncubierta–Rodríguez et al. [74] describe the crowdsourcing process carried out to generate the ground truth for the modality classification task. The ground–truthing task was divided into several steps that were executed in an iterative way. The results of an initial crowdsourcing round classifying 1,000 images into the set of image categories were manually controlled by domain experts. Then, the complete ImageCLEFmed set of images was automatically classified using a visual words approach and the training set as reference. The automatic classification results were then used for a second crowdsourcing task to manually confirm or refuse the automatic classes. This allowed for a faster annotation of correctly classified images and reduced the amount of images that need to be reclassified.

Evaluation The evaluation of this task is done in terms of classification accuracy, which is the proportion of images for which the classifier can correctly predict the class. In other words, the prediction of the classifier is compared with the actual class of the images. The



- [*3D*] 3D reconstruction
- [*AN*] Angiography
- [*CM*] Compound figure (more than one type of image)
- [*CT*] Computed tomography
- [*DM*] Dermatology
- [*DR*] Drawing
- [*EM*] Electron Microscopy
- [*EN*] Endoscopic imaging
- [*FL*] Fluorescence
- [*GL*] Gel
- [*GX*] Graphs
- [*GR*] Gross pathology
- [*HX*] Histopathology
- [*MR*] Magnetic resonance imaging
- [*PX*] General photo
- [*RN*] Retinography
- [*US*] Ultrasound
- [*XR*] X–ray

Figure 4.10: The image class hierarchy that was developed in 2011 for document images occurring in the biomedical open access literature and its class code.



- [$COMP$] Compound or multipane images (1 category)
- [$Dxxx$] Diagnostic images:
  - [$DRxx$] Radiology (7 categories):
  - [$DRUS$] Ultrasound
  - [$DRMR$] Magnetic Resonance
  - [$DRCT$] Computerized Tomography
  - [$DRXR$] X–Ray, 2D Radiography
  - [$DRAN$] Angiography
  - [$DRPE$] PET
  - [$DRCO$] Combined modalities in one image
- [$DVxx$] Visible light photography (3 categories):
  - [$DVDM$] Dermatology, skin
  - [$DVEN$] Endoscopy
  - [$DVOR$] Other organs
- [$DSxx$] Printed signals, waves (3 categories):
  - [$DSEE$] Electroencephalography
  - [$DSEC$] Electrocardiography
  - [$DSEM$] Electromyography
- [$DMxx$] Microscopy (4 categories):
  - [$DMLI$] Light microscopy
  - [$DMEL$] Electron microscopy
  - [$DMTR$] Transmission microscopy
  - [$DMFL$] Fluorescence microscopy
- [$D3DR$] 3D reconstructions (1 category)
- [$Gxxx$] Generic biomedical illustrations (12 categories):
  - [$GTAB$] Tables and forms
  - [$GPLI$] Program listing
  - [$GFIG$] Statistical figures, graphs, charts
  - [$GSCR$] Screenshots
  - [$GFLO$] Flowcharts
  - [$GSYS$] System overviews
  - [$GGEN$] Gene sequence
  - [$GGEL$] Chromatography, Gel
  - [$GCHE$] Chemical structure
  - [$GMAT$] Mathematics, formulae
  - [$GNCP$] Non–clinical photos
  - [$GHDR$] Hand–drawn sketches

Figure 4.11: The image class hierarchy that was developed in 2012 for document images occurring in the biomedical open access literature and its class code.

proportion of correct classifications as an estimate of the accuracy of the classifier is the measure used to evaluate this task.

Compound figure separation

In the ImageCLEFmed 2012 data set [180] between 40% and 60% of the figures are compound or multipane figures. Making the content of the compound figures accessible for targeted search can improve retrieval accuracy. For this reason the detection of compound figures and their separation into subfigures is considered an important task. Examples of compound figures can be seen in Figure 4.12. The goal of this task is to separate the figure into subfigures using separation lines (see Figure 4.13).

The data set used in the ImageCLEF 2013 compound figure separation task are all figures of the ImageCLEFmed 2013 dataset (see Section 4.3.1). 2,967 compound figures were selected from the complete data set after a manual classification of images into compound and other figures. This subset was randomly split into two parts: a training set containing 1,538 images and a testing set with 1,429 images.

A more complicated compound figure separation task will be held in ImageCLEFmed 2015[39]. In 2015, the task will try to separate the compound images if possible and/or attach labels about the content to the subparts. More details can be found in Chapter 8.

Ground truth generation The ground truth for the dataset was generated in a semi–automatic way, using a two–step approach: first, an automated separation process (using the technique described in [44]) was run on both image sets in order to obtain a general overview of the subfigures. The automatic results were then manually corrected.

Missing lines were added and incorrect lines removed, although often the lines were only slightly changed. Separating lines rather than bounding boxes were used to separate subfigures.

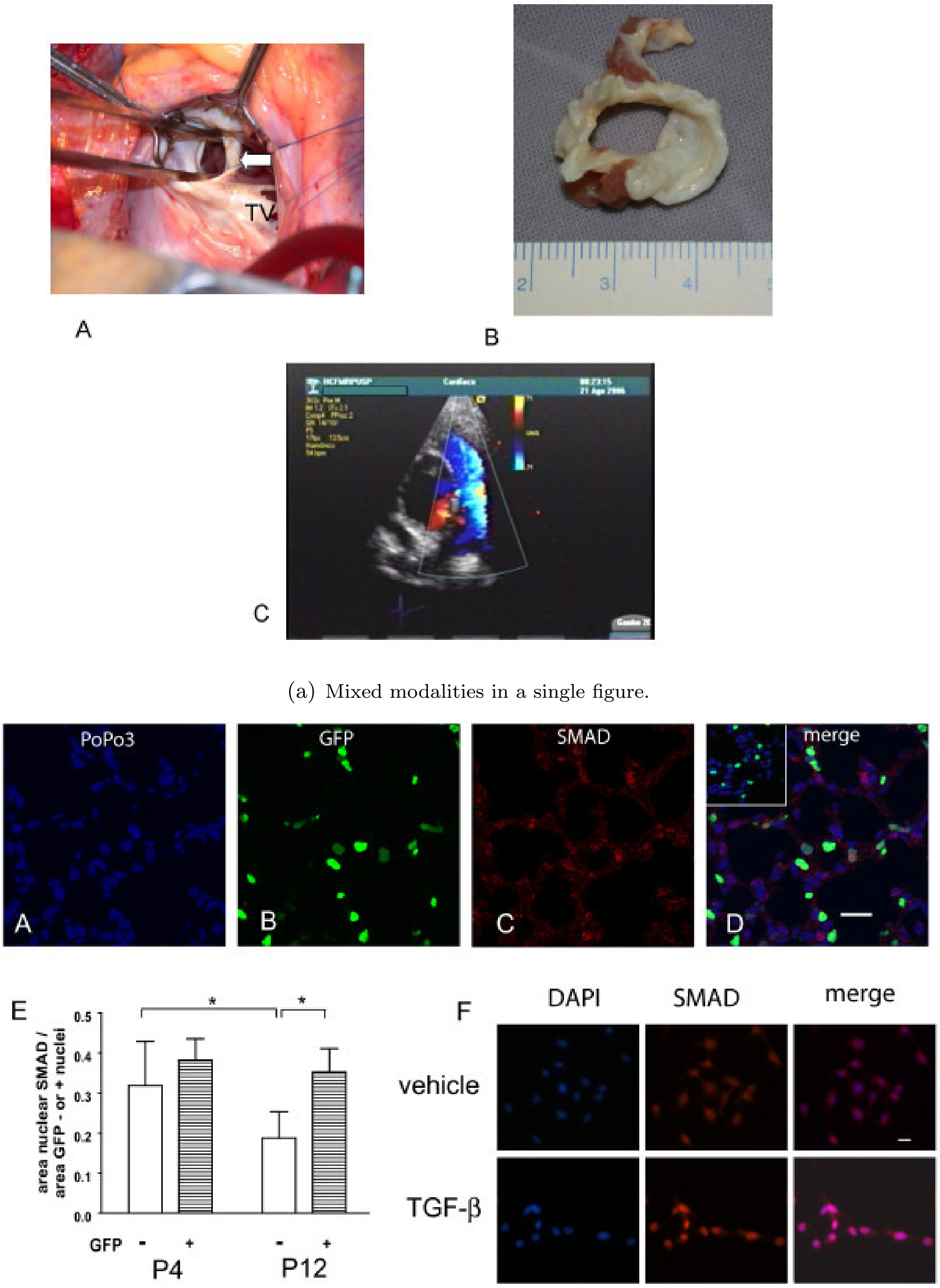
Evaluation The evaluation required the ground truth and the data supplied by the groups in their runs to have a minimum overlap between them. The terminology used in the evaluation is:

- the term *figure*, f, refers to a compound figure as a whole,
- a *subfigure*, f_i, represents a part (or panel) of a figure. The ground truth for the figure f consists of a set of K^f_{GT} subfigures $f_1, \ldots, f_{K^f_{GT}}$;
- the word *candidate*, c_j, refers to the data being evaluated against the ground truth. Separation of figure f consists of a set of K^f_C candidates $c_1, \ldots, c_{K^f_C}$.

A brief summary of the evaluation algorithm for a given figure f is as follows:

- the score $S(f)$ is computed based on the number of correct candidates, $C^f_{correct}$;
- for each subfigure f_i defined in the ground truth the best matching candidate subfigure will be determined. Only one candidate is used in case there are several matches;

[39] For more information about ImageCLEFmed 2015, see `http://www.imageclef.org/2015/medical`.



(a) Mixed modalities in a single figure.

(b) Graphs and microscopy images in a single figure.

Figure 4.12: Examples of compound figures found in the biomedical literature.

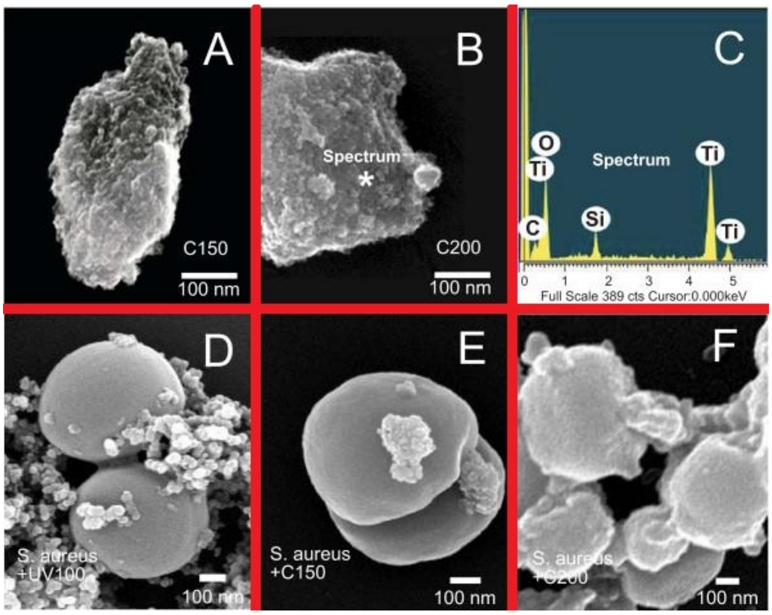


Figure 4.13: Examples of a compound figure separated into subfigures by red lines.

- the main metric used to compare subfigures is the overlap between a candidate subfigure and the ground truth. To be considered a valid match the overlap between a candidate subfigure and a subfigure from the ground truth must correspond to at least 66% of the candidate's size. If the best candidate is an acceptable match, the number of correctly matched figures $C^f_{correct}$ will be incremented. Since only one candidate subfigure can be assigned to each of the subfigures from the ground truth, $|C^f_{correct}| \leq |K^f_{GT}|$;

- the maximum score for the figure is 1 and the normalisation factor used to compute the score will be the maximum between the number of subfigures in the ground truth K^f_{GT} and the number of candidate subfigures K^f_C;

$$S(f) = \frac{|C^f_{correct}|}{max(|K^f_{GT}|, |K^f_C)}|.$$

  Therefore the maximum score is obtained only when the number of candidates K^f_C is equal to the number of subfigures in the ground truth K^f_{GT} and all of them are correctly matched:

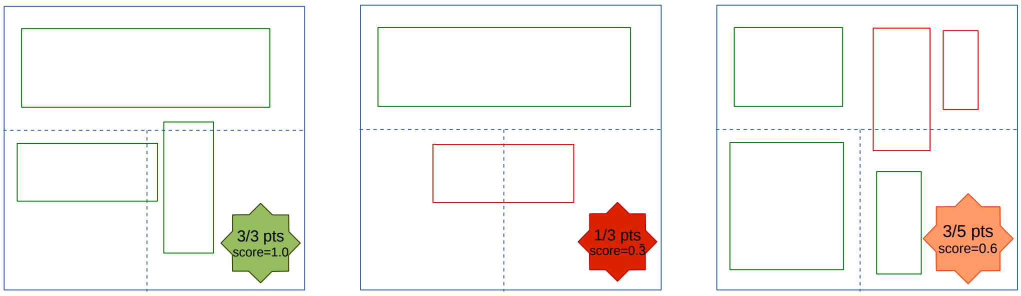
$$|C^f_{correct}| = |K^f_C| = |K^f_{GT}|.$$

Figure 4.14 contains examples showing different candidates being validated against a reference figure (which contains 3 subfigures), along with their scores.

4.3.3 Outcome of the evaluation activities during this thesis

This section reports the outcomes of the ImageCLEFmed, based on the evaluation campaigns organised for the PROMISE visual clinical decision support use cases.

Comparisons between the three years of ImageCLEFmed which were organised during this thesis (2011–2013) are performed based on the material in PROMISE deliverables



(a) Perfect score. All candidates are valid: they are contained above the overlapping threshold in the ground truth.

(b) Not enough candidates. The bottom candidate is not contained to ≥66% in the lower two ground truth subfigures.

(c) Too many candidates. Only the best match in the top area is considered valid.

Figure 4.14: Examples for the separation of a compound figure. Dashed blue lines represent the ground truth, while solid lines represent the candidates. Valid candidates are shown in green and invalid candidates in red.

D6.2 [194] and D6.3 [85]. Appendix B presents the results of the questionnaires sent to the ImageCLEFmed organisers. The main differences between the three years (2011–2013) regarding the the participation and collections are pointed out by the task organisers.

Participation

Appendix B provides a more detailed breakdown on the number of registrations, participations, return participations per task, and submitted runs per task. ImageCLEFmed was attended by people from different academic and industrial institutions. More than 60 groups initially registered each year in the lab showing interest in the benchmarking activities proposed. This proves that ImageCLEFmed has achieved a high visibility. ImageCLEFmed has been one of the most popular tasks able to attract many participants not only from Europe but also from America, Asia and Africa. 34 groups, both researchers and system developers, submitted runs, despite the number of registered participants to the benchmarking activities being much higher than that.

Return participations from the previous year are on average around 40% (30% in 2011 and 47% in 2012 and 40% in 2013), indicating that a large number of researchers rely year after year on the resources created in the context of the ImageCLEFmed evaluation activities. ImageCLEFmed stands out in ImageCLEF with 575 submissions in the 3 years with an average of 192 per year.

In 2013 there were a smaller number of participants and submitted runs that can be due to a change in the evaluation schedule line date of CLEF 2013 and may also be due to the fact that the event was organised outside Europe.

Main advancements

Appendix B presents the main differences between CLEF 2011, 2012 and 2013. Larger collections have been employed each year. In 2011 a new dataset was created and in the following years it has been updated adding new elements. The efforts to make the task more realistic have continued during the three years, not only improving the collections but also modifying query topics and even modifying the hierarchical classification. In 2013, ImageCLEFmed has introduced a new task: compound figure separation.

Main trends and experimental outcomes

Appendix B presents the main trends among the participants approaches, as well as the main experimental outcomes based on the participants' results. The main trend in 2011 was mapping free text onto medical ontologies such as MeSH terms. The MeSH hierarchy was also used and query expansion was often successful. Visual approaches had good early precision. Even if fusion is hard to do multi–modal approaches were often best. In 2012, the main trend was the use of Lucene, concept–based approaches and multiple visual features.

Visual, textual or mixed runs perform differently based on the sub–tasks. It is clear that the expansion of the training set (introduced in ImageCLEFmed 2011) and the use of multiple visual features were successful. The ImageCLEFmed 2012 task was often used to optimise parameters in 2013 as they provided similar data. Finally, visual techniques showed a better performance than multi–modal techniques for the compound figure separation task.

Main problems

Appendix B presents the main problems from the organizational point of view. The main challenge was the low participation compared to the number of registrations. The other main significant problem is the time needed to generate data and also the difficulties identifying appropriate evaluation measures.

Collections

A description of the collections and some statistics are presented in Appendix B. In 2011 the collections were used for the first time and the following years were reused with a greater number of elements. Therefore, the size of the collection and the number of images they contain vary, with the size ranging between 16GB and 18GB and the number of images between 230,000 and 300,000. The collections were multilingual although the tasks were language independent by nature of the images.

Query topics

A further description of the query topics can be found in Appendix B. The number of query topics varies between 10 and 35 in the retrieval tasks while the number of images to classify rates from 1,000 to 2,582. The query topics were provided in English, Spanish, French and German while the corpus is mainly in English.

Ground truth

Appendix B briefly presents the process for the ground truth generation, followed by the task, and also provides estimates of the human effort applied. Ground truth generation is tedious and time–consuming. The task reused ground truth information and extended work over the years. The task employed human assessors, mostly volunteers such as task organisers, students or participants as well as external expert assessors such as physicians.

4.4 Lessons learned

In ten years, the ImageCLEFmed campaign for medical image classification and retrieval has evolved strongly to adapt to current challenges in the domain [119]. Since 2011 the tasks were defined following the described use case (see Chapter 3) making them more realistic. In particular, the case–based task defined is estimated to be closer to clinical routine than the image–based retrieval.

Many systems and techniques have been explored and tested over the years to identify promising techniques and directions. The databases grew from 6,500 to over 300,000 images and now contain a large noise component requiring more complex filtering but being representative of the literature that stores most medical knowledge. Also the tasks increased in complexity from simple visual image retrieval in 2004 to a task consisting of image–based retrieval, case–based retrieval, modality classification and compound figure separation.

ImageCLEF has had an important scholarly impact [236, 237]. More than 200 research groups have worked with the data and many techniques have been compared during its campaigns.

The user surveys and analyses of log files have created insight into the changes in visual information search behaviour and create the basis for system testing. Several clear lessons have been learned over the years:

- a variety of features need to be used for good visual retrieval;
- fusion of visual and text information can improve the results as the two retrieval paradigms are complementary but fusion needs to be done with care as the characteristics of the two are not the same and many poor approaches for fusion actually decrease the text retrieval results;
- mapping of free text to semantics can improve results over using text only;
- using modality information of images can improve performance of image retrieval where one modality is the query objective;
- for the modality classification results the main limiting factor was the training data that did not cover the diversity of the test data; the best techniques all used automatic or manual techniques for the extension of the training data set;
- compound figure separation is an important step to focus search on single figures but keep their context, which is often important.

These lessons learned show the importance of such benchmarks and of systematic evaluation. Research can now be focused on promising techniques and allows concentrating on real research challenges and reproducible approaches, which is clearly not the case when small, private databases are used. Having a forum such as a workshop where participants can compare their experiences with those of other researchers who worked on the same data is another important part. These discussions frequently lead to new, improved research ideas and also collaborations between participants. Research lives off these exchanges and cannot be done alone any more. Sharing work to create resources and evaluation platforms creates an added value for everyone involved and has many advantages in terms of research organization.

4.5 Summary

This chapter describes the *ImageCLEFmed* evaluation framework, starting with an analysis of the scholarly impact of the ImageCLEF campaign. The chapter focuses on the ImageCLEFmed organised during this thesis between 2011 and 2013 but an overview of the history of ImageCLEFmed is given. ImageCLEFmed aims to evaluate multi–modal medical IR systems. To this end, during this thesis two retrieval task are proposed: an image–based and a medical case–based retrieval task. Moreover, modality classification and compound figure separation tasks are constructed because both can help in the retrieval step. To underline the importance of this benchmark with a standardised database, including query topics and relevance judgements, the main outcomes of the evaluation activities during this thesis are provided. These show that based on the same ImageCLEFmed collection multiple visual features combined correctly with text information can improve the performance of the retrieval. Furthermore, approaches including external information, such as MeSH terms or modality information, can also help.

Chapter 5

Case–based Retrieval Techniques

> "El hombre no encontrará nunca soluciones definitivas a sus problemas; cada solución puesta aporta su nuevo lote de problemas propios."
>
> Eduardo García de Enterría

Chapter 3 describes the visual clinical decision support for the medical diagnosis use case. The medical case–based retrieval task is pointed out in the use case validation. Therefore, this thesis focuses on developing a system which addresses the medical case–based retrieval challenge.

ParaDISE [164] is a retrieval engine that has been developed in the context of the Khresmoi project. The main concepts behind its design are scalability, flexibility, expandability and interoperability, allowing it to be used in standalone applications, integrated systems and for research purposes. New components for specific steps and new algorithms for the existing components have been added during this thesis. ParaDISE is programmed in the Java programming language and uses JavaScript Object Notation (JSON) as a data transfer protocol to enable introperability and realistic application development.

This chapter describes the techniques developed to create a medical case–based retrieval system integrated into ParaDISE. Chapter 4 finishes highlighting the main lessons learned thanks to ImageCLEFmed. The developed system integrates most of these lessons. Only the compound figure separation tool is not yet integrated because further work has to be done. In particular four main components are described in detail: a multi–modal baseline with multiple visual descriptors, a data fusion framework, a query–adaptive multi–modal fusion criterion and a modality classification tool for the retrieval step.

5.1 Basic performance

This section details the retrieval tools that are used to create the multi–modal retrieval baseline [161, 80, 83]. Figure 5.1 shows all the basic components of this baseline.

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

The Lucene IR library is used to establish the text retrieval baseline. Lucene was chosen for the experiments because it is fast and easy to install and use. Provided below

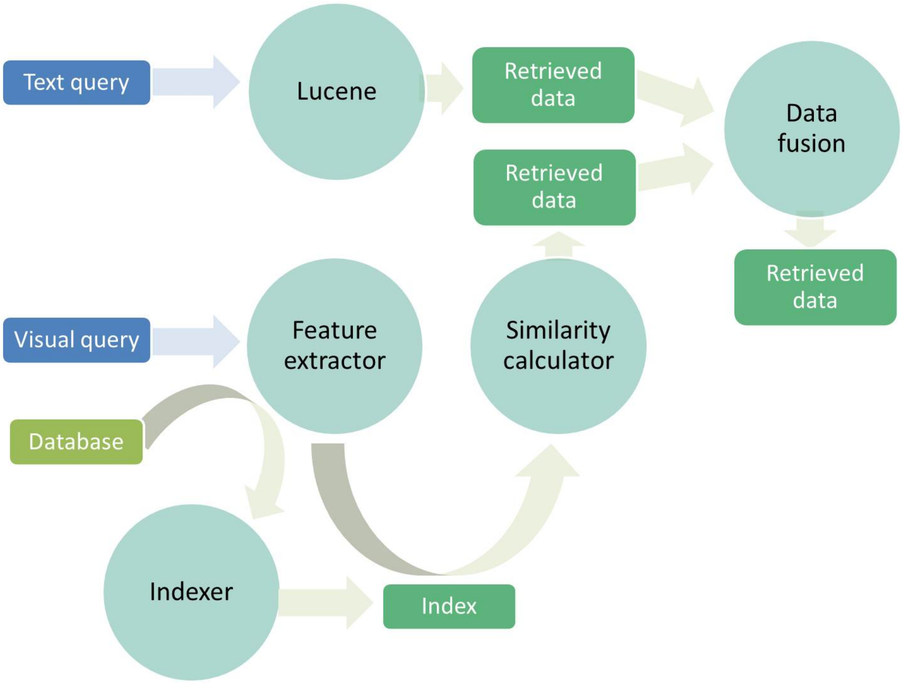


Figure 5.1: Outline of the basic elements of the multi–modal retrieval baseline.

are some details about the way Lucene is used and configured:

- *EnglishAnalyzer* – in Lucene, an analyser is used for tokenization (breaking a stream of text up into words, phrases, symbols or other meaningful elements), stemming (keeping only the root of a word) and stop word removal (excluding common words from the index). The EnglishAnalyzer that is used filters out a list of common English stop words (and, or, is, etc.) and performs stemming based on rules specific to the English language (removing the letter "s" at the end of words, removing common endings like "-ing", "-er", etc.);

- *Multiple boolean operators* – when parsing a text query, Lucene uses a boolean operator for terms separated by a space character (AND, OR). In order to maximise the score of relevant documents, each text query is executed three times: using the OR operator, using the AND operator and finally putting the query into quotes ("...") to perform an exact phrase search. The three result lists are then fused using a reciprocal rank fusion rule [54], in this way boosting the ranking of exact matches;

- *Term frequency–Inverse document frequency (tf/idf) similarity* – several similarity measures are implemented in Lucene. The commonly used tf/idf weighting is applied.

For the visual content of the images, multiple features are used as this was a successfully used technique in ImageCLEFmed (see Chapter 4) [180]. A set of low–level visual descriptors are selected from the descriptors bank of ParaDISE [211, 164] and their combination is explored [83]. The following descriptors are chosen to be investigated:

- *Bag of Visual Words (BoVW) using SIFT (BoVW–SIFT)* [159] – each image is represented by a histogram symbolizing a set of local descriptors represented in visual words from a previously learned vocabulary;

- *BoVW–SIFT with a spatial pyramid matching [154] (BoVW–SPM)* – spatial information is added to the BoVW–SIFT descriptor;
- *BoC* [82] – Each image is represented by a histogram symbolizing the colours from a previously learned vocabulary;
- *BoC with $n \times n$ spatial grid (Grid BoC)* – spatial information is added to the BoC descriptor;
- *Colour and Edge Directivity Descriptor (CEDD)* [42] – colour and texture information is produced by a 144 bin histogram. It only needs a low computational power for its extraction;
- *Fuzzy Colour and Texture Histogram (FCTH)* [43] – this descriptor contains results from the combination of 3 fuzzy systems including colour and texture information in a 192 bin histogram;
- *Fuzzy Colour Histogram (FCH)* [94] – the colour similarity of each pixel's colour associated with all the histogram bins through a fuzzy–set membership function is considered;
- *HSV colour histogram* [228]– the histogram represents the distribution of colours on the HSV (hue–saturation–value) colour space;
- *Colour layout* [134] – this descriptor represents the spatial distribution of the colour of visual signals in a very compact form;
- *Tamura texture* [230] – this descriptor explores an approximation on six visual properties: coarseness, contrast, directionality, line–likeness, regularity and roughness.

For the visual indexing, *histogram intersection* [229] is used for the similarity comparison for each of the descriptors. Histogram intersection has been successfully used as a similarity measure for image retrieval and previous studies have shown that it is robust to many transformations [40, 23, 229].

The indexer creates an Approximate Nearest Neighbour (A–NN) index structure to facilitate fast retrieval. Euclidean Locally Sensitive Hashing (E2LSH) [6] is used as an A–NN indexing method because it deals with a large number of dimensions.

The fusion rule *combMNZ* is selected for the baseline based on previous work due to its good performance on the ImageCLEFmed 2012 tasks [80]. More details on fusion techniques are explained in Section 5.2.

The rest of the chapter describes in detail the main retrieval components studied in this thesis to improve the results obtained with the the baseline approach. Figure 5.2 shows an overview of all the components presented.

5.2 Data fusion

Data fusion is applied in order to achieve more accurate retrieval results than the retrieval results achieved by single sources [87]. Two types of fusion algorithms are considered in this thesis:

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

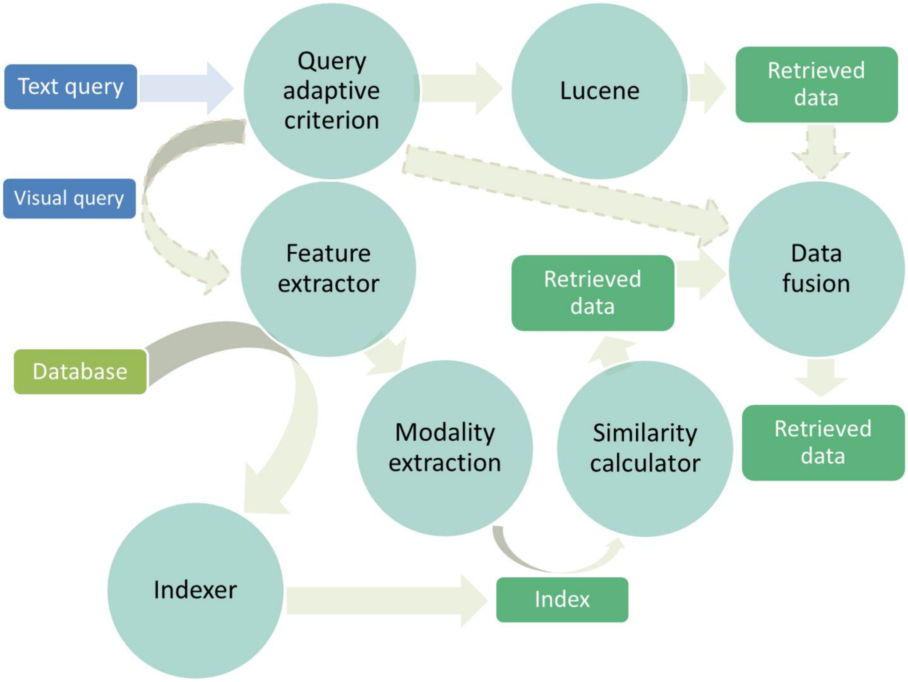


Figure 5.2: Outline of the multi–modal retrieval including a query–adaptive multi–modal fusion criterion and a modality classification filter.

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

To enhance the performance of the medical case–based retrieval task, several fusion strategies are implemented. This section focuses on the description of the fusion strategies.

5.2.1 Visual feature fusion

Several fusion strategies are tested to combine results of each of the query images and of the visual descriptors of the same image to improve visual retrieval. To combine the results/features of multiple query images into a single ranked list, two main fusion strategies are used: early and late fusion [80]. Early fusion integrates unimodal features before making any decision (see Figure 5.3). Unimodal feature vectors are concatenated into one vector using a weighting scheme. Rocchio's algorithm can be applied to merge the vectors of the same feature spaces into a single vector:

$$\vec{q}_m = \alpha\vec{q}_o + \beta\frac{1}{|I_r|}\sum_{j=1}^{|I_r|}\vec{im}(j) - \gamma\frac{1}{|I_{nr}|}\sum_{j=1}^{|I_nr|}\vec{i}(j), \quad j \in \mathbb{N} \tag{5.1}$$

where α, β and $\gamma \in \mathbb{R}$ are weights; $\vec{q}_m \in \mathbb{R}^n$ is the modified query; $\vec{q}_o \in \mathbb{R}^n$ is the original query; $\vec{im}_j$ are the images that belong to I_r, the set of relevant images, or to I_{nr}, the set of non–relevant images. In this scenario there are no non–relevant images and the set of relevant images is the original query. Thus, only the second term of the right part of the equation is used [80].

Late fusion consists of a combination of independent results from various approaches. The ranked lists of retrieval results are fused and not the features (see Figure 5.4). The

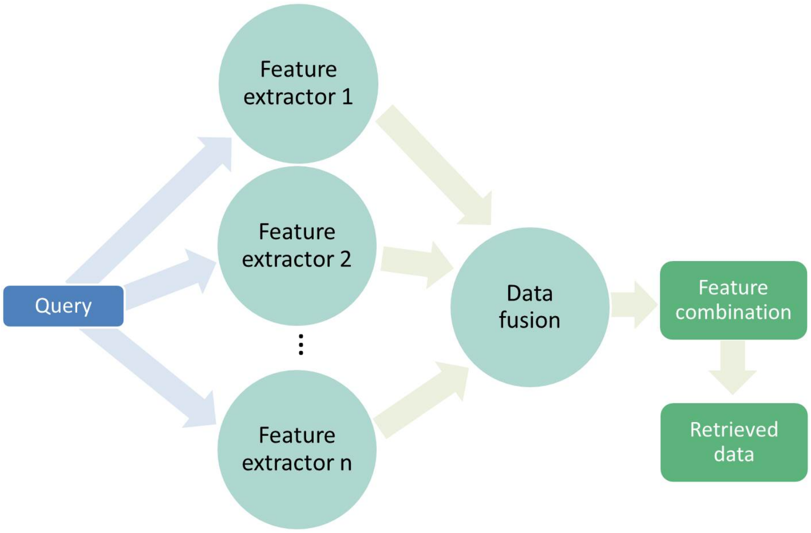


Figure 5.3: General scheme for early fusion.

following fusion rules are used for the experiments because they are commonly used in the biomedical domain:

- combSUM
$$\text{combSUM}(im) = \sum_{j=1}^{N_j} S_j(im) \tag{5.2}$$
with $N_j \in \mathbb{N}$ being the number of descriptors to be combined and $S(im) \in \mathbb{R}$ is the score assigned to image im;
- combMNZ
$$\text{combMNZ}(im) = F(im) * combSUM(im) \tag{5.3}$$
where $F(im) \in \mathbb{N}$ is the frequency of image mi being returned by one input system with a non–zero score;
- combMAX
$$\text{combMAX}(im) = \arg\max_{j=1:N_j}(S_j(im)); \tag{5.4}$$
- combMIN
$$\text{combMIN}(im) = \arg\min_{j=1:N_j}(S_j(im)); \tag{5.5}$$
- RRF
$$\text{RRF}(im) = \sum_{r \in R} \frac{1}{k + r(im)} \tag{5.6}$$
where R is the set of rankings assigned to the images and $k = 60$ for the study [54];
- Borda
$$\text{Borda}(im) = \sum_{r \in R} r(im). \tag{5.7}$$

For further details on the fusion rules see also [61].

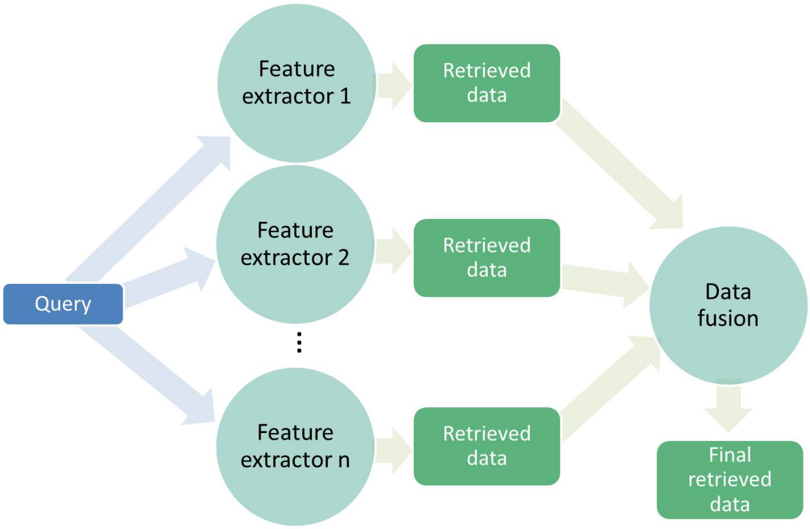


Figure 5.4: General scheme for late fusion.

5.2.2 Multi–modal fusion

There are several ways of combining visual and textual retrieval [89]. In this thesis two approaches are tested: (1) performing both visual and textual retrieval and then combining the results of the two runs; and (2) using textual retrieval as a basis and then reranking results based on visual retrieval.

Combination of visual and textual search

The text and visual retrieval systems are described in Section 5.1. To combine visual and textual ranks, the techniques described in Section 5.2.1 are applied: Borda; combMAX; combMIN; combMNZ; combSUM and RRF. A linear combination of the ranks of the textual and visual runs is also used. Similar to the approach presented by Ramhan et al. [199], the weight of each rank is defined by a function of their performance in terms of MAP:

$$\omega_t = \frac{\mathrm{MAP}(T)}{\mathrm{MAP}(T) + \mathrm{MAP}(V)};\ \omega_v = \frac{\mathrm{MAP}(V)}{\mathrm{MAP}(T) + \mathrm{MAP}(V)} \tag{5.8}$$

where the best MAP scores obtained using text and visual search in ImageCLEFmed 2011 and 2012 are employed. Figure 5.5 shows the fusion process followed in this section.

Visual reranking

The reranking method proposed reorders the initial text search results based on the visual descriptors. An initial text search using Lucene returns an ordered set $A = a_1, ..., a_{1,500}$ of the 1,500 articles with the largest scores $S(a)$ assigned to the articles a, thus more than the 1,000 required for the final results list. Instead of accepting these results, the articles' images belonging to A are used to rerank the results. In the visual reranking process the retrieved result list of article A is substituted by a set of the images associated with the retrieved articles. Content–based image retrieval is performed using the topics' query images within this image set using the visual features mentioned above. A sorted list of result images is retrieved and is converted back to the article list preserving

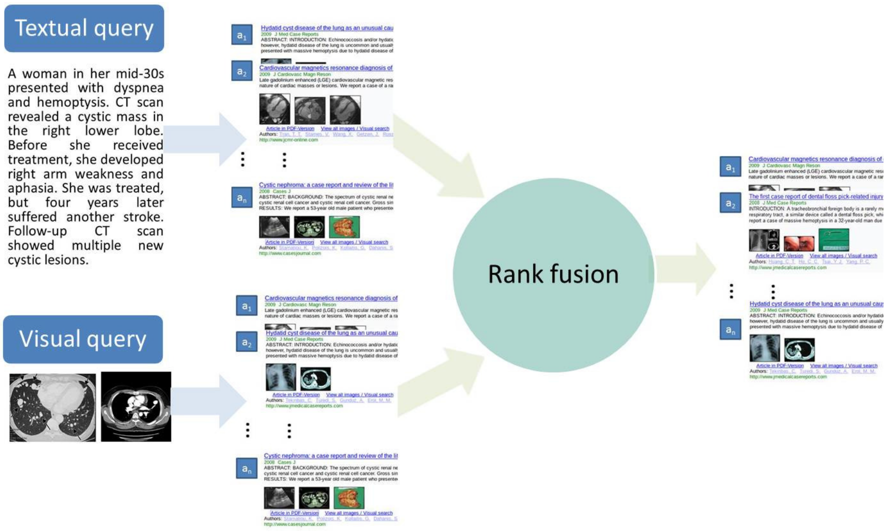

Textual query
A woman in her mid-30s presented with dyspnea and hemoptysis. CT scan revealed a cystic mass in the right lower lobe. Before she received treatment, she developed right arm weakness and aphasia. She was treated, but four years later suffered another stroke. Follow-up CT scan showed multiple new cystic lesions.
Visual query
Rank fusion
a1
a2
an

Figure 5.5: The final rank is obtained by combining both visual and text search.

the order derived by the CBIR. The result is an ordered set $\bar{A}$ where an article $a \in \bar{A}$ if and only if $a \in \bar{A}$ but in a different order. This new order is based on the score values determined by the visual features extracted from the visual information.

The visual reranking process is illustrated in Figure 5.6.

5.3 Query–adaptive multi–modal fusion

In this section, a new method for query–adaptive multi–modal fusion is proposed. The goal is to change the formulation of the retrieval algorithm based on the user query. For this, MeSH terms extracted from the text query are analysed in order to determine the potential use of image queries as a complementary source.

5.3.1 MeSH term extraction

Most of PMC publication records are manually annotated with MeSH terms, which can be retrieved using the Entrez search system API [189, 53]. Each image belonging to a document is represented as a binary histogram which characterises the annotated MeSH terms contained in the document. Each binary histogram is a binary vector–form representation of MeSH term occurrences in the document.

Queries were mapped to MeSH terms by a score–based phrase matching algorithm favouring MeSH terms with words occurring rarely in the document corpus [231]. Matching synonyms were replaced by their primary MeSH terms. Only MeSH terms occurring in the document–MeSH term matrices are considered for query mapping. Hence, textual queries are also represented as a binary histogram of the extracted MeSH terms.

5.3.2 Visual and text synonymy

Collins dictionary [1] defines a "*synonym*" as "*a word that means the same or nearly the same as another word*". Furthermore, Foncubierta–Rodríguez [71] extends the definition

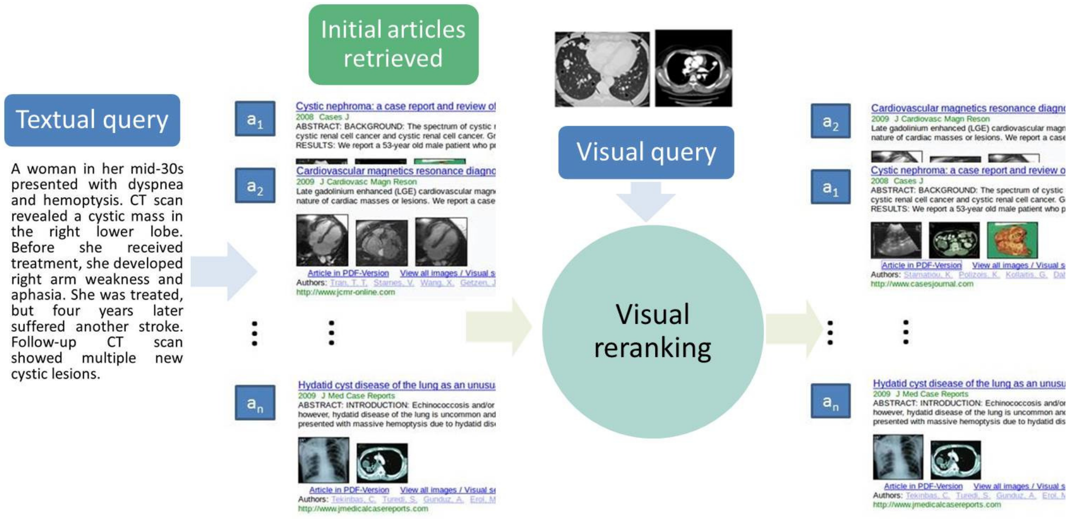

Textual query
A woman in her mid-30s presented with dyspnea and hemoptysis. CT scan revealed a cystic mass in the right lower lobe. Before she received treatment, she developed right arm weakness and aphasia. She was treated, but four years later suffered another stroke. Follow-up CT scan showed multiple new cystic lesions.
Initial articles retrieved
a1
a2
an
Cystic nephroma: a case report and review of
2008 Cases J
Cardiovascular magnetics resonance diagno
2009 J Cardiovasc Magn Reson
Hydatid cyst disease of the lung as an unusu
2009 J Med Case Reports
Article in PDF-Version
View all images / Visual s
Visual query
Visual reranking

Figure 5.6: The proposed visual reranking reorders articles based on images extracted from the initial text search results.

of synonyms to visual words based on criteria derived from Probabilistic Latent Semantic Analysis (PLSA).

Definition 5.1 (Synonyms) *A pair of visual words w_i, w_j can be considered* synonyms *if the following three conditions are met:*

1. *There is at least one visual topic z_k to which both w_i and w_j belong;*
2. *w_i and w_j have a complementary distribution in the collection;*
3. *w_i and w_j have a similar contextual distribution with the rest of the words.*

where a visual topic *z is defined as the representation of a generalised version of the visual appearance modelled by various visual words. It corresponds to an intermediate level between visual words and the complete understanding of visual information. A set of visual topics $Z = \{z_1, \ldots, z_{N_z}\}$can be defined in a way that every visual word can belong to none, one or several visual topics. In this case, visual topics correspond to each of the topics or aspects derived from a PLSA analysis.*

According to this definition of visual synonymy, Foncubierta–Rodríguez [71] defines a synonymy matrix as:

Definition 5.2 (Synonymy visual word space) $\mathbf{S} \in \mathcal{M}_{N_W \times N_W}(\mathbb{R})$ *is a symmetric* synonymy matrix *if:*

$$\mathbf{S} = \begin{pmatrix} 1 & s_{12} & \cdots & s_{1N_W} \\ s_{21} & 1 & \cdots & s_{2N_W} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N_W1} & s_{N_W2} & \cdots & 1 \end{pmatrix} \tag{5.9}$$

where $s_{ij} \in \mathbb{R}$ measures the synonymy of the visual words w_i and $w_j \in W$.

$$s_{ij} = s_{ji} = \begin{cases} 1 \text{ if } i = j \\ \sigma_{ij} \text{ if } w_i, w_j \text{ are synonyms} \\ 0 \text{ otherwise} \end{cases} \tag{5.10}$$

and $\sigma_{ij} \in \mathbb{R}$ *is the synonymy value of the words* w_i *and* w_j*.The* synonymy value *of two words* w_i, w_j *is defined as the maximum significance value for which both words are significant for the same visual topic.*

$$\sigma_{ij} = \sigma_{ij} = \max_{k} \left\{ \min_{i,j} \{p_{i,k}, p_{j,k}\} \right\} \tag{5.11}$$

where $v_{i,j}$ *is the normalised value of the probability* $P(w_i|z_j)$ *obtained from PLSA.*

Medical text can be represented as a histogram of MeSH terms (see Section 5.3.1). Images can also be represented as a histogram of visual features. Descriptors mentioned in Section 5.1 build these histograms. Therefore, it is possible to consider both text and visual features to create a common vocabulary. Definition 5.2 is extended from language modelling techniques, therefore it can also be used for the synonym relation between text and visual information keeping the mathematical sense of synonyms.

The synonymy matrix from a set of MeSH terms and visual descriptors is obtained considering the relative properties of visual words based on their behaviour on training data. For each of the images in the training set, the histogram of MeSH terms and the histogram with the visual features are concatenated. As a result the following symmetric synonymy matrix is obtained:

$$\mathbf{S_{tv}} \;=\; \begin{pmatrix} 1 & t_{12} & \cdots & \cdots & t_{1M} & tv_{1M+1} & \cdots & tv_{1M+N} \\ t_{21} & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & tv_{2M+N} \\ \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \vdots & \vdots \\ t_{M1} & \cdots & \cdots & \cdots & t_{MM} & tv_{MM+1} & \cdots & tv_{MM+N} \\ vt_{M+11} & \cdots & \cdots & \cdots & vt_{M+1M} & v_{M+1M+1} & \cdots & v_{M+1M+N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ vt_{M+N1} & \cdots & \cdots & \cdots & vt_{M+NM} & v_{M+NM+1} & \cdots & 1 \end{pmatrix} \tag{5.12}$$

where t_{ij} is the synonymy value of two MeSH terms, v_{ij} is the synonymy value of two visual features and $tv_{ij} = vt_{jj}$ is the synonymy value of a MeSH term and a visual feature. M is the dimension of the textual histogram (the number or MeSH terms in the set) and N the dimension of the visual histogram.

5.3.3 Query–adaptive fusion criterion

Not all medical case text descriptions need query images to find relevant articles. Often the relevant articles for a query topics do not contain images or contain only general biomedical illustrations (such as statistical figures or graphs). In ImageCLEFmed 2013 [79] best results for medical case–based retrieval were actually achieved by pure text runs. Participants usually decreased their results when using multi–modal approaches. However, this thesis puts forward the hypothesis that visual information can improve the precision of the retrieval.

The basic hypothesis of this work is defined as follows:

Hypothesis 5.1 *If the extracted MeSH terms of a textual query have synonym relations with the visual features, then visual information can improve retrieval.*

Similar to the use of text synonyms, using multi–modal retrieval (text and visual information) only when there is a synonym relation between the text query and the visual features can make the retrieval more consistent because only articles that are really related to the query topics are retrieved [2].

This work focuses on the synonym relation between text and visual features, i.e., on the submatrix of the matrix $\mathbf{S_{tv}}$:

$$\bar{\mathbf{S}} = \mathbf{S_{tv}}(i,j), \quad \forall i \in [M, M+N] \quad \textit{and} \quad \forall j \in [1, M] \tag{5.13}$$

The following criterion is proposed to predict when it is suitable to use visual information in addition to text based on the query:

Definition 5.3 (Query–adaptive fusion criterion) *Let $\vec{q} \in [0,1]^M$ be the binary histogram of MeSH term occurrences in the textual query. If $\exists i | \vec{q}(i) \neq 0$ and $\exists j | \bar{\mathbf{S}}(i,j) \neq 0$ then the textual query is* suitable *to be fused with a visual query.*

For comparative purposes, a second criterion is defined based on the covariance matrix between MeSH terms and visual features. The covariance matrix $\mathbf{V}$ is a measure of the correlation. Given n sets of variables denoted by $\{X_1\}, \ldots, \{X_n\}$, $\mathbf{V}$ is the matrix whose (i,j) entry is the covariance

$$\mathbf{V}(i,j) = cov(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)] \tag{5.14}$$

where μ_i is the expected value of x_i.

To focus on the covariance between text and visual features, the following submatrix of the covariance matrix is selected:

$$\bar{\mathbf{V}} = \mathbf{V}(i,j), \quad \forall i \in [M, M+N] \quad \textit{and} \quad \forall j \in [1, M] \tag{5.15}$$

Replacing the matrix $\bar{\mathbf{S}}$ by $\bar{\mathbf{V}}$, the following criterion is defined:

Definition 5.4 (Covariance–based query–adaptive fusion criterion) *Let $\vec{q} \in [0,1]^M$ be the binary histogram of MeSH term occurrences in the textual query. If $\exists i | \vec{q}(i) \neq 0$ and $\exists j | \bar{\mathbf{V}}(i,j) \neq 0$ then the textual query is* suitable *to be fused with a visual query.*

5.4 Modality classification

This section proposes a modality classification approach based on the ImageCLEFmed 2012 hierarchy presented in Chapter 4. Once the image type information is extracted, the predicted types can be integrated into the search results to generate a final result list. Information on image types can be used in various ways in the retrieval.

5.4.1 Multi–modal classification

The proposed method uses multi–modal information for the representation of the images. The method follows previous work [80, 83] applying a k–NN classifier using weighted voting for the image classification.

Similar to the approach described in Section 5.1, the text representation of the images uses a vector space model with stop word removal, word stemming, tokenization and tf/idf weighting, using the Lucene search engine based on the captions of the images.

A set of low–level visual descriptors are selected from the descriptors bank of ParaDISE [211] and their combination is explored [83]. For visual information extraction, the following descriptors are chosen to be investigated for the modality classification task: CEDD; BoC; BoVW–SIFT; FCTH; FCH; HSV colour histogram; colour layout and Tamura texture [230] (see Section 5.1 for more details on the descriptors selected).

5.4.2 Training set expansion

The number of images per class in the training and test sets of the ImageCLEFmed classification task varies from fewer than ten to several hundred. The distribution of the labelled data among the classes is uneven, making the dataset very suitable for semi–supervised learning.

In this section, a method that uses semi–supervised learning (also referred to as training set expansion in this thesis) to improve the classification accuracy based on the image modalities is proposed. Semi–supervised learning [41] uses a small number of labelled instances and a large amount of unlabelled data for training the classifier. The proposed method uses multi–modal retrieval to expand the training set.

As a result of this semi–automatic learning a larger but "noisy" training set is obtained. This work proposes an iterative procedure to manually correct the expanded set by crowdsourcing. Crowdsourcing allows for dividing the problem into microtasks that can be solved in a short amount of time by users familiar with medical images [91].

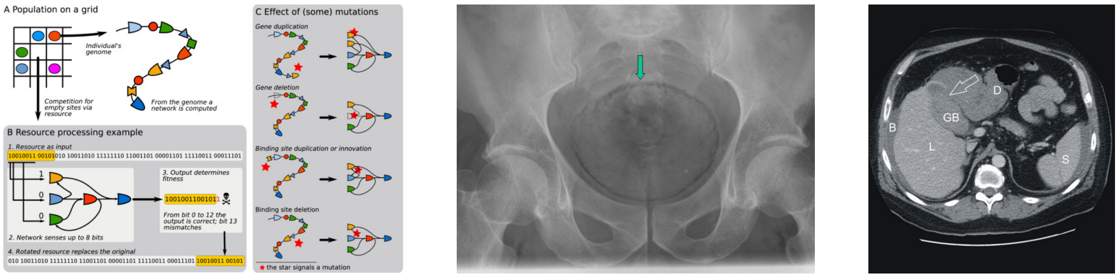
Semi–supervised learning

The training set is denoted as the set of labelled images $\{im_1, \dots, im_{N_l}\} \in I$. Respectively, the corresponding labels are $\{l_1, \dots, l_{N_l}\} \in L$. The set of unlabelled examples is $\{im_{N_l+1} \dots im_{N_l+N_u}\} \in I$. The set that contains all the labelled and unlabelled examples is denoted I. The proposed method labels $N_l \times N_r$ unlabelled examples, where $N_r \in \mathbb{N}$ is a constant, and includes them in the labelled example set. Then, the expanded training set E is used to train the classifier. The labelling of the unlabelled examples is described in the following algorithm:

```
Data: $I; L$
Result: $E \subseteq I; L_E$
$E = \{im_j, j \in [1 \dots N_l]\}$        /* initialise with original training set */
for $im = 1 \dots N_l$ do
    query $im_j$ against I
    retrieve top $N_r$ results $r$
    for $k = 1 \dots N_r$ do
                                          /* do not re-include original examples */
        if $r_j \neq im_m \forall m \in [1 \dots N_l]$ then
            $l_k = y_j$                                /* assign label to result */
            $E = E \cup r_j$                              /* expand training set */
        end
    end
end
remove examples with multiple labels.
```

Algorithm 1: Semi–supervised learning algorithm for training set expansion.

In practice, since it is an automatic classification, images can be retrieve to expand more than one class, resulting on images with multiple labels. After a removing images with multiple labels, the size of the expanded training set is slightly smaller than $N_l + N_l \times N_r$.


A Population on a grid
Individual's genome
Competition for empty sites via resource
From the genome a network is computed
B Resource processing example
1. Resource as input
10010011 00101010 10011010 11111110 11001101 00001101 11110011 00011101
1
0
0
2. Network senses up to 8 bits
3. Output determines fitness
10010011001011
From bit 0 to 12 the output is correct; bit 13 mismatches
4. Rotated resource replaces the original
010 10011010 11111110 11001101 00001101 11110011 00011101 10010011 00101
C Effect of (some) mutations
Gene duplication
Gene deletion
Binding site duplication or innovation
Binding site deletion
★ the star signals a mutation
D
GB
B
L
S

(a) Image automatically classified as "compound".

(b) Image automatically classified as "X–ray".

(c) Image automatically classified as "ultrasound".

Figure 5.7: Images automatically classified. (a) and (b) were correctly classified and (c) not. Crowdsourcing is used to verify their image modality classes.

Crowdsourcing

This work used the CrowdFlower[40] platform to manually correct the automatic training set expansion described in Section 5.4.2. This platform is chosen because its internal interface allows carrying out tasks by a known set of experts, free of charge, to guarantee the precision of the results. More details on the described crowdsourcing task can be found in [76].

The correction task is divided into several steps that are executed in an iterative way:

```
Data: $E, L_E$
Result: $L_E$
$I = \{im_j, j \in [1 \ldots N_l \ldots N_m]\}$
for $j = N_l \ldots N_m$ do
                                                    /* manual verification */
    verify label $l_j$ of image $im_j$
    if $l_j \neq$ "Yes, perfect classification" then
        if $l_j$ is compound image then
            $l_j$ ="COMP";
        else
            relabel $x_i$                            /* manual relabelling */
        end
    end
    if $j = (N_m - N_l)/2$ then              /* automatic reclassification */
        reclassify $im_j, j \in [j \ldots N_l \ldots N_m]\}$ using updated subset
    end
end
```

Algorithm 2: The iterative crowdsourcing algorithm.

Verification The crowdsourcing verification task is set up to verify the automatically given label. Since about 50% of the figures in the biomedical open access literature [44] are compound or multipane images an extra option was added to facilitate the following

[40] CrowdFlower is a crowdsourcing service specialised in microtasking: distributing small, discrete tasks to many online contributors in assembly line fashion (see http://www.crowdflower.com/).

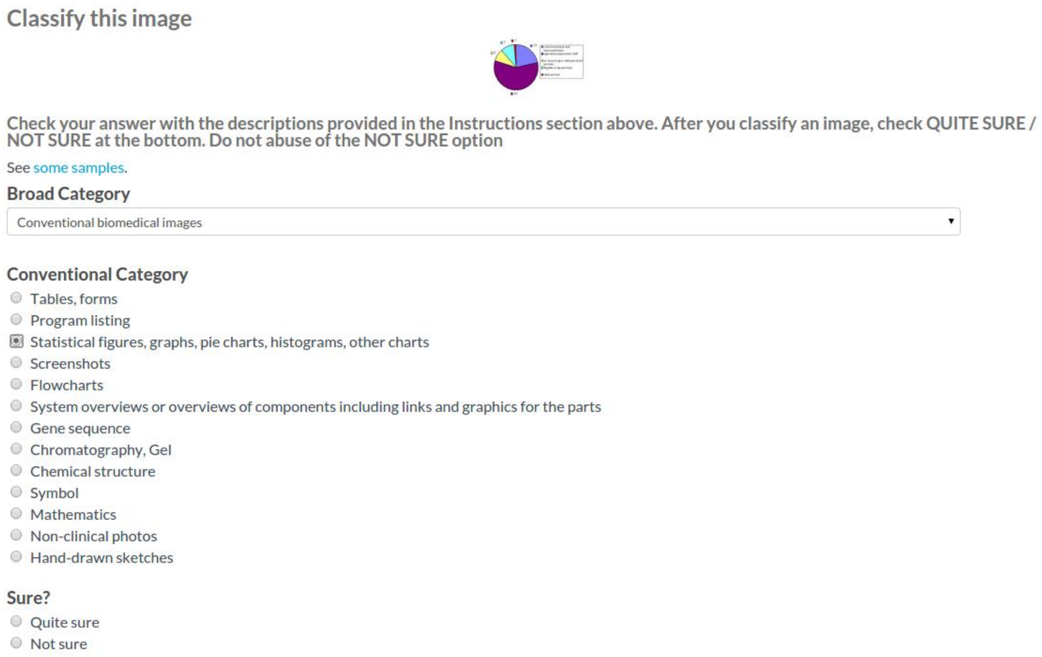
steps.

Therefore, each automatically classified image (see Figure 5.7) is presented with a key question formulated as follow:

- Does the figure correspond to the category?:
  - yes, perfect classification;
  - no, compound image;
  - no, wrong category;
  - not sure.

Relabelling Images that are incorrectly classified automatically or tagged as "not sure" are then manually relabelled in a second crowdsourcing iteration. The images are relabelled into the 31 classes of the hierarchy presented in Chapter 4. The task is therefore presented in a hierarchical structure where a broad class is first asked and then the subclass (see Figure 5.8).

As these images are the ones proven to be difficult to classify in the first iteration, each of the images is classified by two participants. In case of disagreement between the first two answers a third expert labels the image.



Classify this image

Check your answer with the descriptions provided in the Instructions section above. After you classify an image, check QUITE SURE / NOT SURE at the bottom. Do not abuse of the NOT SURE option

See some samples.

Broad Category

Conventional biomedical images ▾

Conventional Category

○ Tables, forms
○ Program listing
● Statistical figures, graphs, pie charts, histograms, other charts
○ Screenshots
○ Flowcharts
○ System overviews or overviews of components including links and graphics for the parts
○ Gene sequence
○ Chromatography, Gel
○ Chemical structure
○ Symbol
○ Mathematics
○ Non-clinical photos
○ Hand-drawn sketches

Sure?

○ Quite sure
○ Not sure

Figure 5.8: Screenshot of the crowdsourcing interface for image modality classification.

5.4.3 Modality filter

Once the image type is extracted, the predicted types can be integrated into the search results to generate a final result list.

The full image dataset is classified using the method presented above. The query images of each query topics are also classified and a set of query modalities is produced.

For each query image a rank list is retrieved and filtered or ranked. Then, the fusion of the various rank lists obtained from each of the query images are combined. Four approaches are tested:

- *Exact* – uses a single modality of the query image for filtering the list of this image;
- *Close* – uses a set of all modalities occurring in the query images of each query topics to filter the list of each image query;
- *Prefix* – is similar to the first but the broadest modality (diagnostic, general, compound) is used instead of the exact modality for boosting the image score in the retrieved set;
- *Diagnostic* – only diagnostic images from the database are retrieved. This approach does not depend on the query.

Images retrieved in the retrieval step are then *filtered* or *reranked*. If images are *filtered*, then images that are classified into one of the query set modalities are discarded. When *reranking* the retrieved images by modality, the images that are classified into one of the query set modalities are placed on top of the other retrieved images.

5.5 Summary

This chapter describes the underlying techniques used to developed a medical case–based retrieval system under ParaDISE. It shows the baseline established that the system is based on. The ParaDISE architecture allows the addition of new components into the system easily. Therefore, this chapter focuses on the development of three components highlighted in the analysis of the ImageCLEFmed outputs.

First, various fusion techniques are presented for both visual and multi–modal descriptors. A query–adaptive multi–modal fusion criterion is then defined to decide the suitability of the use of visual queries to complement text queries. Finally, a modality classification approach is described for its integration in the retrieval. The modality classification method includes a semi–supervised learning algorithm which allows augmenting the training set. Crowdsourcing is proposed to manually correct the image labels.

Chapter 6

Experimental Results

> "Every time man makes a new experiment he always learns more. He cannot learn less."
>
> Richard Buckminster Fuller

The data and evaluation scenario used in this chapter is reused from the ImageCLEFmed benchmark described in Chapter 4. This thesis evaluates the medical case–based retrieval task although the modality classification task is also explored as an intermediate step. For the medical case–based retrieval task, the 1,000 best–ranked articles are retrieved for each query topic in the following experiments. Results are averaged over the total number of queries (26 or 35 for ImageCLEFmed 2012 or 2013) in order to reproduce the exact setup of ImageCLEFmed. The results are computed with the trec_eval software (version 9.0) following the ImageCLEFmed practice (see Chapter 2). The experiments are first carried out on the ImageCLEFmed *2012* collection to optimise the results. Then, they are executed on the ImageCLEFmed *2013* collection to obtain final results. In every experiment, results are compared with the best runs (per type task) submitted by participants to ImageCLEFmed. In addition, each experiment is also compared with other experiments carried out in this thesis.

6.1 Basic performance

This section develops the baseline for the experiments that are carried out during this thesis. See Chapter 5 for more details on the outline selected.

Tables 6.1 and 6.2 show results achieved by the text baseline (*RunT1* and *RunT2*) and the best runs submitted to ImageCLEFmed in 2012 and 2013 respectively. The presented baseline achieved competitive results although not as good as the best one submitted to the competition. The best textual approach used an external corpus for robust and effective expansion term inference [227]. However, this thesis focuses on visual retrieval and the text baseline is only used to test experimental multi–modal approaches.

In CBIR each database needs its corresponding parameter setting for feature extraction. Not all features have the same discriminative power and the performance strongly depends on the extracted features [201]. When combining a set of features some may be irrelevant [21]. Therefore, it is necessary to select well the set of features that the retrieval system will use.

Table 6.1: Results of the approaches of the medical case–based retrieval task when using only text on the ImageCLEFmed 2012 collection.

Run ID	MAP	GMAP	Bpref	P10	P30
Best text ImageCLEF	**0.1690**	**0.0374**	**0.1499**	**0.1885**	**0.1090**
RunT1	0.1670	0.0355	0.1413	0.1731	0.1077

Table 6.2: Results of the approaches of the medical case–based retrieval task when using only text on the ImageCLEFmed 2013 collection.

Run ID	MAP	GMAP	Bpref	P10	P30
Best text ImageCLEF	**0.2429**	**0.1163**	**0.2417**	**0.2657**	**0.1981**
RunT2	0.1791	0.1107	0.1630	0.2143	0.1581

The performance of each of the selected descriptors is studied in Table 6.3. Then, the fusion of the best descriptors is performed to obtain a good feature set (see Table 6.4). Due to time and resource limitations not all the possible combinations are tested and only the descriptors performing well alone are combined. Descriptors with better performance are successively fusioned. If the added descriptor causes worse performance then it is deleted from the list of descriptors used. As presented in Section 5.1, combMNZ is used for a late fusion of the query descriptors and the results of the selected descriptors. Table 6.4 shows that the best performance is obtained with *RunV16* using the following descriptors: BoVW–SPM; Grid BoC; CEDD and Tamura. The selected descriptors are then used as a visual baseline for the following experiments. Table 6.5 shows the results achieved on the ImageCLEFmed 2013 case–based task. The proposed visual baselines, *RunV16* and *RunV18*, perform better than the best visual run submitted to the task in 2012 and 2013, respectively, except at the precision 10 and 30 measures.

6.2 Query topic analysis

Query topics are essential for the IR experiments despite being the most critical element of a collection [225]. Although ImageCLEFmed 2013 query topics were carefully elaborated, differences between the query topics have implications for the performance. Indeed Mandl et al. [160] assess that the variation between query topics is larger than the variations between systems in most of the evaluation activities. In this section, an analysis of the query topics in the ImageCLEFmed 2013 case–based task is provided. The analysis is used to better understand the problem before performing further experiments.

As described in Chapter 4 each of the query topics contains a case description and a few images. The ImageCLEFmed 2013 collection contains 35 query topics. However, ImageCLEFmed reports system effectiveness as an average over the set of query topics. Table 6.6 shows the number of documents judged as relevant in the database for each of the query topics, in total there are only 709 documents judged as relevant for the 35

Table 6.3: Individual performance of selected descriptors on the case–based task of ImageCLEFmed 2012.

Run ID	**Descriptor**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
RunV1	BoVW–SPM	0.0286	0.0009	0.0319	0.0154	0.0077
RunV2	Grid BoC	0.0276	0.0007	0.0399	0.0192	0.0064
RunV3	CEDD	0.0265	0.0004	0.0397	0.0269	0.0103
RunV4	BoC	0.0264	0.0006	0.0329	0.0115	0.0064
RunV5	Colour layout	0.0259	0.0004	0.0399	0.0115	0.0051
RunV6	BoVW–SIFT	0.0226	0.0011	0.0275	0.0192	0.0103
RunV7	Tamura	0.0149	0.0003	0.0384	0.0115	0.0051
RunV8	FCTH	0.0148	0.0003	0.0279	0.0077	0.0026
RunV9	HSV colour histogram	0	0	0	0	0
RunV10	FCH	0	0	0	0	0

Table 6.4: Performance of the combination of several descriptors using combMNZ on the case–based task of ImageCLEFmed 2012. Each of the runs is using descriptors from the runs showed in Table 6.3.

Run ID	**Descriptors**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best visual ImageCLEF	–	**0.0366**	0.0014	0.0347	**0.0269**	**0.0141**
RunV11	RunV1-2	0.0292	0.0010	0.0302	0.0192	0.0103
RunV12	RunV1-3	0.0324	0.0012	0.0330	0.0154	0.0077
RunV13	RunV1-4	0.0296	0.0012	0.0329	0.0154	0.0077
RunV14	RunV1-3,5	0.0282	0.0013	0.0383	0.0192	0.0115
RunV15	RunV1-3,6	0.0307	0.0010	0.0301	0.0231	0.0077
RunV16	**RunV1-3,7**	0.0343	0.0015	**0.0413**	0.0231	0.0115
RunV17	RunV1-3,7-8	0.0329	**0.0018**	0.0370	0.0154	0.0077

Table 6.5: Performance of the visual baseline on the case–based task of ImageCLEFmed 2013 compared with the best visual run submitted to the competition.

Run ID	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best visual ImageCLEF	0.0281	0.0009	0.0335	**0.0429**	**0.0238**
RunV18	**0.0318**	**0.0014**	**0.0629**	0.0343	0.0229

Table 6.6: Number of relevant articles per query topics in the case–based ImageCLEFmed 2013 task.

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

queries, varying from 1 to 100 relevant articles per query topics, which complicates the task.

ImageCLEFmed query topics were not proposed by the assessors who judge the documents in the pool resulting in fewer documents considered relevant [14]. In addition, most of the submitted runs used only text techniques, only 5 runs were submitted using purely visual techniques. Therefore, most of the documents in the pool were retrieved by systems that used only text techniques and not based on the images belonging to the documents. If a run contains relevant articles that were not judged previously its performance has a negative bias [49]. This thesis focuses on visual techniques, however the articles retrieved using visual techniques might not be judged. In the following the query topics are analysed in detail from a visual point of view.

Figure 6.1 shows the AP per query topic achieved by the run with best MAP submitted on ImageCLEFmed 2013 and by the visual baseline presented in Section 6.1. It is notable that around a third of the query topics got zero AP. Indeed in seven of the query topics both runs got zero AP. Analysing these seven query topics in detail, it is observed that the query images can not retrieve the images belonging to the articles judged as relevant because the images in the relevant articles are visually different to the query images. Figure 6.2 shows one of these query topics where both query images are not visually similar to the images of the articles judged as relevant for that query topics. Therefore, no system will be able to retrieve these articles based only on visual information. In fact, visual information in a multi–modal approach will not contribute to improve the retrieval in these query topics.

In particular many articles contain only graphs, which are not images discriminative for a visual search (see Figure 6.3).

In only two of the seven analysed query topics there is one image in the relevant judged articles that could be visually similar to the query images. However, these images are subfigures of a compound image. One example is shown in Figure 6.4 where each of the two query images are visually similar to subfigures of one image in an article judged as relevant.

Relevance judgements in the medical domain can be cognitively demanding [142]. In this case, the articles were asked to be relevant if they could be useful for a differential diagnosis. After this detailed analysis, it seems that the assessors probably based their decisions mainly on the textual information of the articles and less based on the images that they contain. Therefore, it makes it more difficult to evaluate the system and the

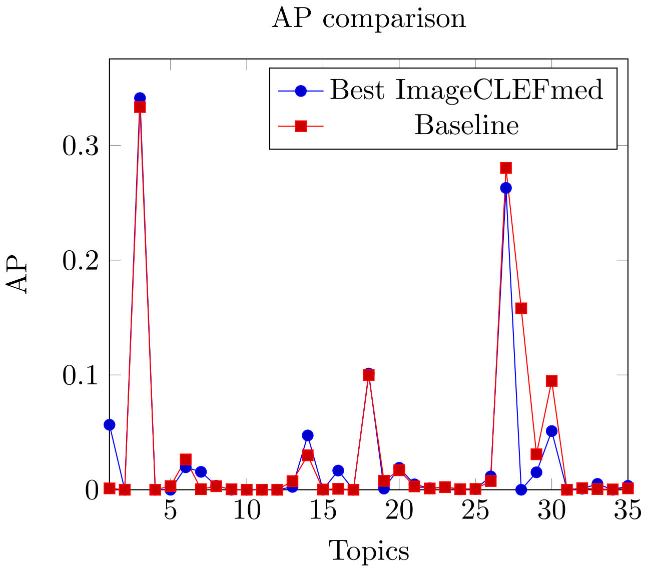
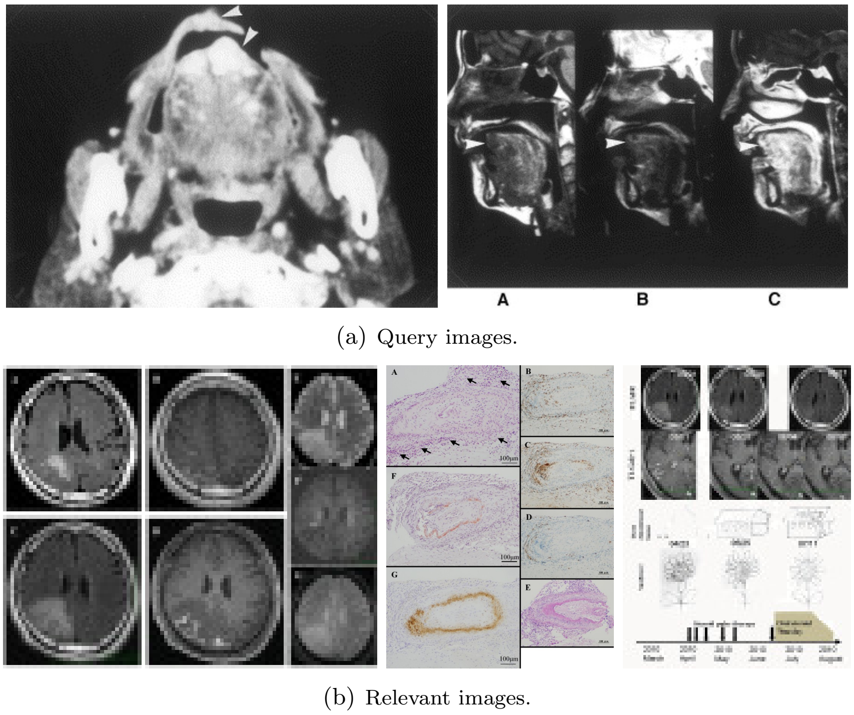


Figure 6.1: AP for individual query topics achieved by the best visual run submitted in ImageCLEFmed 2013.

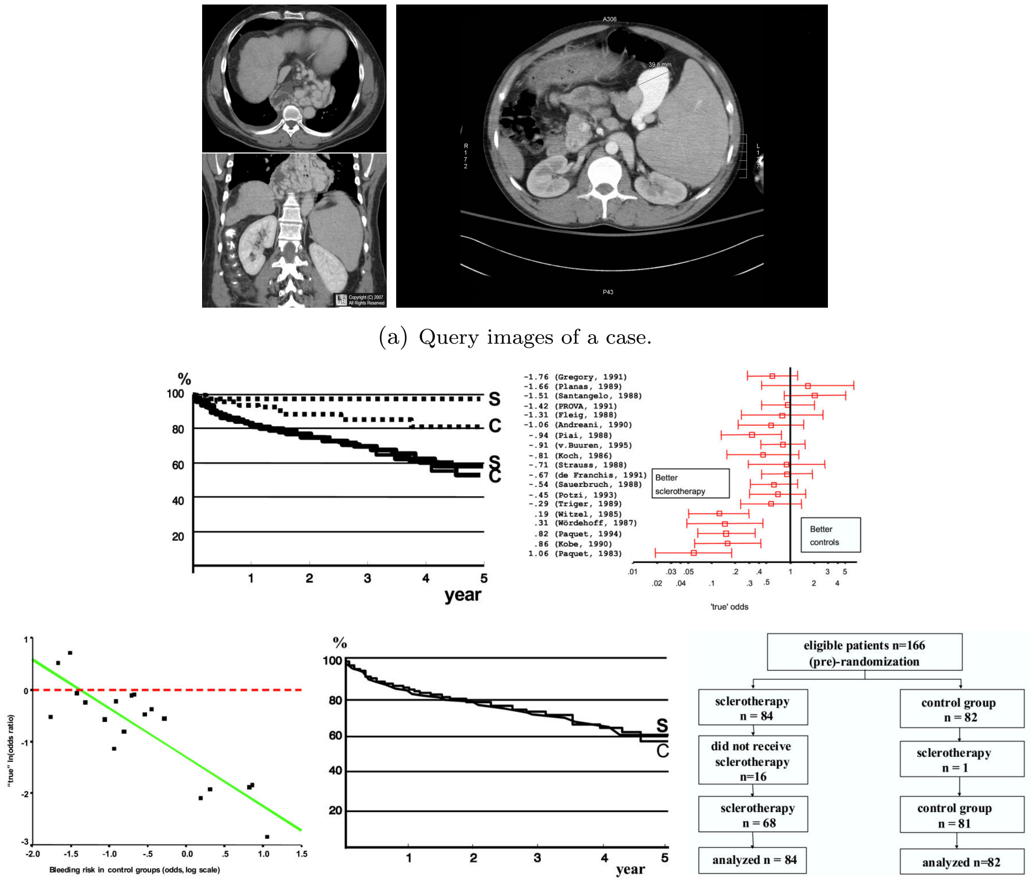


(a) Query images.

(b) Relevant images.

Figure 6.2: Example of (a) query images from a query topic from ImageCLEFmed 2013 and (b) images belonging to the articles judged as relevant for that query topic. It is notable that they are not visually similar.

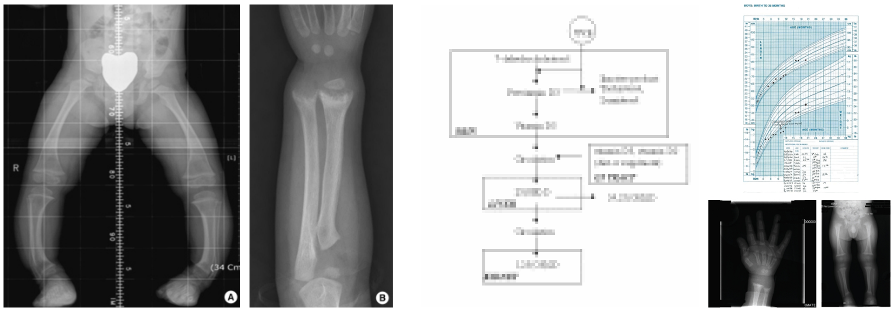
improvements proposed in the following sections, using the ImageCLEFmed 2013 collection. Despite the limitations of the evaluation framework, it provides a good scenario to compare the proposed approaches with the state–of–the–art and with the presented baseline.



(a) Query images of a case.

(b) Images of relevant articles.

Figure 6.3: Example of (a) query images from a query topic from ImageCLEFmed 2013 and (b) images belonging to articles judged as relevant for that query topic. All the images in the article are graphs.



(a) Query images of a case.

(b) Images of relevant articles..

Figure 6.4: Example of (a) query images from a query topic from ImageCLEFmed 2013 and (b) images belonging to the articles judged as relevant for that query topics. There are images in the relevant articles visually similar to the query but as subfigures of a compound figure.

6.3 Comparing fusion techniques

The main objective of this section is to evaluate the effectiveness of the various fusion methods for the medical case–based retrieval task. Fusion is performed in several cases in the retrieval pipelines: to handle multiple query images in CBIR and to combine the various visual and textual features.

Several experiments are conducted for the medical case–based retrieval task:

- Visual fusion (VF);
- Multi–modal Fusion (MF):
  - Modal combination;
  - Reranking.

The visual features are first combined to highlight some distinguishable properties of the images. The first experiment combines the various query and feature fusion techniques when using only visual information. The results of these combinations (VF) are shown in Tables 6.7 and 6.8, on the 2012 and 2013 tasks respectively. For the 2012 collection, best results on visual runs are achieved when the queries are fused with combMAX and the descriptors with Borda (see Section 5.2.1 for more details on the techniques). The experiments indicate that using Borda to fuse the various visual features always outperforms other fusing rules in terms of MAP, bpref and P10. For the analysis, the sets queries of each query topics are combined into one and an overall score is computed. The fusion rules which achieve the best results for the 2012 collection are applied on the 2013 collection, outperforming the best submitted run in terms of MAP, GMAP and bpref.

Since the case–based task has been running, textual approaches have always achieved better results than visual or multi–modal runs. The goal is to enhance the retrieval performance by adding visual information. The results of the second experiment (MF) show the performance of the fusion of textual and visual results (see Table 6.9). To carry out these experiments the best visual approach is used (see Table 6.7) to help in better task accomplishment. The fusion used for RunVF13 is applied because it obtained the best results in terms of MAP, bpref and P10 and good results in terms of GMAP and P30. The fusion rules described in Section 5.2.2 are applied for this experiment. The best result is obtained by RunMF7 using a linear combination of text and visual searches (MAP=0.1806). For the RunMF7, the weights for the linear combination are based on the MAP obtained in 2011 by the best run. Linear combination is one of the simplest and most widely used fusion methods [12]. It improves the fusion more than the other used approaches.

Finally, a reranked approach is carried out (RunMF9). Text retrieval is used to extract a subset of all potential relevant images. In this experiment the retrieval performance is poor, potentially because this visual approach is not optimal for a task where the number of relevant articles is very low.

Table 6.10 presents the results obtained on the ImageCLEFmed 2013 collection when using the fusion rules chosen on the 2012 collection. The presented approach is better than the best mixed (multi–modal) run submitted to ImageCLEFmed.

Table 6.7: Results of the approaches for the medical case–based retrieval task when using various fusion strategies for visual retrieval on the ImageCLEFmed 2012 collection. Several query (QF) and descriptor (DF) fusion techniques are combined in the table.

Run ID	**QF**	**DF.**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best visual ImageCLEF	–	–	0.0366	0.0014	0.0347	0.0269	0.0141
RunVF1	Rocchio	Borda	0.0003	0.0001	0.0065	0	0
RunVF2	Rocchio	combMAX	0.0154	0.0006	0.0175	0.0115	0.0051
RunVF3	Rocchio	combMIN	0.0008	0.0001	0.0130	0	0.0026
RunVF4	Rocchio	combMNZ	0.0211	0.0009	0.0405	0.0192	0.0103
RunVF5	Rocchio	combSUM	0.0226	0.0012	0.0474	0.0231	0.0103
RunVF6	Rocchio	RRF	0.0004	0.0001	0.0067	0	0
RunVF7	Borda	Borda	0.0001	0	0.0057	0	0
RunVF8	Borda	combMAX	0.0005	0	0.0141	0	0
RunVF9	Borda	combMIN	0.0002	0	0.0058	0	0
RunVF10	Borda	combMNZ	0.0058	0.0003	0.0228	0.0077	0.0064
RunVF11	Borda	combSUM	0.0058	0.0003	0.026	0.0077	0.0051
RunVF12	Borda	RRF	0.0002	0	0.0058	0	0
RunVF13	combMAX	Borda	**0.0490**	0.0014	**0.0702**	**0.0308**	0.0141
RunVF14	combMAX	combMAX	0.0381	0.0013	0.0545	0.0269	0.0103
RunVF15	combMAX	combMIN	0.0241	0.0001	0.0505	0.0077	0.0038
RunVF16	combMAX	combMNZ	0.0340	0.0013	0.0410	0.0192	0.0128
RunVF17	combMAX	combSUM	0.0369	**0.0018**	0.0378	0.0231	0.0128
RunVF18	combMAX	RRF	0.0428	0.0013	0.0522	0.0231	0.0167
RunVF19	combMIN	Borda	0.0367	0.0011	0.0647	0.0231	0.0128
RunVF20	combMIN	combMAX	0.0432	0.0011	0.0495	0.0231	0.0128
RunVF21	combMIN	combMIN	0.0138	0	0.0351	0.0038	0.0013
RunVF22	combMIN	combMNZ	0.0285	0.0011	0.0373	0.0154	0.0103
RunVF23	combMIN	combSUM	0.0311	0.0012	0.0376	0.0192	0.0103
RunVF24	combMIN	RRF	0.0337	0.0011	0.0456	0.0231	0.0128
RunVF25	combMNZ	Borda	0.0304	0.0009	0.0402	0.0269	0.0128
RunVF26	combMNZ	combMAX	0.0284	0.0009	0.0326	0.0154	0.0077
RunVF27	combMNZ	combMIN	0.0018	0.0001	0.0226	0	0.0013
RunVF28	combMNZ	combMNZ	0.0343	0.0015	0.0413	0.0231	0.0115
RunVF29	combMNZ	combSUM	0.0339	0.0013	0.0382	0.0231	0.0103
RunVF30	combMNZ	RRF	0.0325	0.0010	0.0401	0.0231	0.0115
RunVF31	combSUM	Borda	0.0406	0.0015	0.0513	**0.0308**	**0.0154**
RunVF32	combSUM	combMAX	0.0314	0.0011	0.0313	0.0192	0.0103
RunVF33	combSUM	combMIN	0.0147	0.0002	0.0313	0.0038	0.0026
RunVF34	combSUM	combMNZ	0.0341	0.0014	0.0393	0.0154	0.0115
RunVF35	combSUM	combSUM	0.0369	**0.0018**	0.0363	0.0192	0.0115
RunVF36	combSUM	RRF	0.0423	0.0015	0.0479	0.0269	0.0128
RunVF37	RRF	Borda	0.0001	0	0.0057	0	0
RunVF38	RRF	combMAX	0.0007	0.0001	0.0091	0	0
RunVF39	RRF	combMIN	0.0002	0	0.0071	0	0
RunVF40	RRF	combMNZ	0.0058	0.0003	0.0228	0.0077	0.0064
RunVF41	RRF	combSUM	0.0021	0.0002	0.0194	0	0.0051
RunVF42	RRF	RRF	0.0002	0	0.0058	0	0

Table 6.8: Results of the approaches for the medical case–based retrieval task when using various fusion strategies for visual retrieval on the ImageCLEFmed 2013 collection. Query (QF) and descriptor (DF) fusion techniques selected after optimization on the ImageCLEFmed 2012 collection are used (RunVF43). The run is compared with the baseline and with the best visual run submitted to the medical case–based retrieval task of ImageCLEFmed 2013.

Run ID	**QF**	**DF.**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best visual ImageCLEF	–	–	0.0281	0.0009	0.0335	**0.0429**	**0.0238**
RunV18	combMNZ	combMNZ	0.0318	**0.0014**	0.0629	0.0343	0.0229
RunVF43	combMAX	Borda	**0.0336**	0.0013	**0.0666**	0.0343	0.0229

Table 6.9: Results of the approaches of the medical case–based retrieval task when using various fusion strategies to combine visual and textual information ("multi–modal fusion") on the ImageCLEFmed 2012 collection.

Run ID	**Multi–modal fusion**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best mix ImageCLEF	–	0.1017	0.0175	0.0857	0.1115	0.0679
RunMF1	Borda	0.0983	0.0206	0.1223	0.0923	0.0538
RunMF2	combMAX	0.0915	0.0213	0.1285	0.0500	0.0564
RunMF3	combMIN	0.0242	0.0014	0.0545	0.0192	0.0128
RunMF4	combMNZ	0.1289	0.0281	0.1481	0.0923	0.0756
RunMF5	combSUM	0.0963	0.0224	0.1379	0.0577	0.0538
RunMF6	RRF	0.1138	0.0214	0.1267	0.1077	0.0538
RunMF7	linearMAP11	**0.1806**	**0.0397**	**0.1578**	0.1808	**0.1064**
RunMF8	linearMAP12	0.1732	0.0390	0.1555	**0.1885**	0.0949
RunMF9	rerank	0.0454	0.0130	0.0842	0.0543	0.0333

Table 6.10: Results of the approaches of the medical case–based retrieval task when using various fusion strategies to combine visual and textual information ("multi–modal fusion") on the ImageCLEFmed 2013 collection. Fusion technique selected after optimization on the ImageCLEFmed 2012 collection is used (RunMF47).

Run ID	**Multi–modal fusion**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best mix ImageCLEF	–	0.1608	0.0779	0.1426	0.1800	0.1257
RunMF10	RunMF7	**0.1889**	**0.1190**	**0.1720**	**0.2257**	**0.1629**

6.4 Query–adaptive multi–modal fusion

In this thesis it is possible to retrieve MeSH terms for 73,584 documents (98.6%) of the ImageCLEFmed dataset and to construct two binary sparse document–MeSH term matrices: one covering all 18,299 MeSH terms referenced by the document corpus and a second matrix covering only 5,583 MeSH terms marked as *major topic* for documents. The two approaches are referred to as *all* and *major* in this thesis.

The synonymy matrix of a set of MeSH terms and each visual descriptor is calculated based on a training set of 5,000 random images from the ImageCLEFmed 2013 database. To study the effect of the latent variable z the synonym matrices are calculated for $N_z = \{50, 100, 200, 300\}$. Minimum significance percentiles $\mathcal{P} = \{0th, 50th, 75th, 99th\}$ are also considered in the study, removing all words with a maximum significance $m_i = \max_j t_{i,j}$ below the given percentile.

The two sets of MeSH terms (*major* and *all*) are analysed. When using the set of *all* MeSH terms, the calculation of the synonymy matrix is restricted to 50,000 synonyms due to computational limitations. All synonyms are calculated when using the *major* set of MeSH terms. The choice of the latent value and the percentile does not affect performance when using *all* the MeSH terms.

The result of the AP per query topic is summarised in Table 6.11. This table shows a comparison between the runs. In general, the *text* approach has a higher AP than the *visual* approach. Fusion of text and visual approaches (*mix*) can improve the AP although for several query topics it is better to use the *text* approach. The query–adaptive criterion presented in Section 5.3.3 allows the automatic selection of the *text* or *mixed* approach for each of the query topics. Table 6.11 shows the AP per query topic for the approaches using *all* and *major* MeSH terms. For the *major* approach, Table 6.11 shows the results for the latent values and percentiles corresponding to the approach with an accuracy of 77.15%. Results are compared with the best mix run submitted to ImageCLEFmed 2013. Table 6.12 shows the accuracy of correct decisions obtained when applying the proposed approach with various parameters and only *major* MeSH terms. These results are not presented for *all* MeSH terms because there is no difference between the parameters, showing the stability of the method. Indeed, using *major* MeSH terms the accuracy of the query–adaptive criterion is always the same except in two cases. Notice that using the text approach was the correct decision in 54.29% of the cases.

Table 6.13 summarises best results achieved with the proposed query–adaptive fusion criterion. This result shows an accuracy of 77.15% when using *major* MeSH terms for most of the parameters values. Accuracy using *all* MeSH terms is lower with 62.86%, probably due to the restriction in the number of synonyms.

For 60% of the query topics, when using *major* MeSH terms CBIR is not used. Using *all* MeSH terms, 23% of the query topics do not apply CBIR. Therefore, the proposed criterion prevents the unnecessary use of visual information making the system more efficient.

One more experiment is carried out to compare the query–adaptive fusion criterion, based on the synonym relation between text and visual features, with the covariance–based query–adaptive fusion criterion. The same $5,000$ images used to calculate the synonymy matrix are used to calculate the covariance matrix. However, using this second criterion on the 35 query topics proposed by ImageCLEFmed in 2013, CBIR is never selected to be used. This means that the covariance between each of the MeSH terms extracted from the 35 text queries and each of the visual features is always zero. In this experiment, the sample size $(5,000)$ is smaller than the dimension which increases the complexity of the

Table 6.11: AP per query topics using various approaches. Correct decisions taken by the proposed approaches are shown in bold type.

#Topic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Best mix ImageCLEF	0.1117	0.0598	0.9167	0.0005	0.2658	0.2066	0.0630	0.0871
Visual	0.0010	0	0.3333	0	0.0034	0.0383	0.0011	0.0033
Text	0.1055	0.0310	0.6789	0.0081	0.4491	0.2207	0.1432	0.0864
Mix	0.1049	0.0306	0.6782	0.0074	0.4492	0.2261	0.1421	0.0799
All	**0.1055**	0.0306	**0.6789**	0.0074	**0.4492**	**0.2261**	**0.1432**	0.0799
Major	**0.1055**	**0.0310**	**0.6789**	**0.0081**	**0.4492**	0.2207	**0.1432**	**0.0864**
#Topic	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Best mix ImageCLEF	0.3548	0.0025	0.0059	0.0119	0.3326	0.1659	0.1380	0.047
Visual	0	0	0	0	0.0058	0.0363	0	0.0022
Text	0.0434	0.0357	0.0038	0.0482	0.2915	0.3044	0.2003	0.0367
Mix	0.0434	0.0357	0.0037	0.0481	0.3049	0.3121	0.1893	0.0344
All	**0.0434**	**0.0357**	**0.0038**	0.0481	**0.3049**	**0.3121**	0.1893	0.0344
Major	**0.0434**	**0.0357**	**0.0038**	**0.0482**	**0.3049**	0.3044	0.1893	0.0344
#Topic	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Best mix ImageCLEF	0.0111	0.1374	0.2097	0.0754	0.0720	0.1985	0.2081	0.0589
Visual	0	0.1000	0.0057	0.0118	0.0010	0.0038	0.0006	0.0033
Text	0.2000	0.0242	0.1896	0.1063	0.1118	0.2419	0.3514	0.1217
Mix	0.2000	0.0669	0.1934	0.1115	0.1098	0.2455	0.3506	0.1228
All	**0.2000**	**0.0669**	**0.1934**	**0.1115**	0.1098	**0.2455**	0.3506	**0.1228**
Major	**0.2000**	**0.0669**	**0.1934**	**0.1115**	**0.1118**	0.2419	**0.3514**	0.1217
#Topic	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
Best mix ImageCLEF	0.0085	0.2202	0.2317	0.0212	0.1325	0.2894	0.5048	0.2435
Visual	0.0005	0.0035	0.2572	0.1699	0.0081	0.0574	0	0.0010
Text	0.0106	0.2780	0.0793	0.0918	0.1642	0.2419	0.5069	0.1381
Mix	0.0102	0.2704	0.3157	0.1278	0.1686	0.2783	0.5063	0.1372
All	**0.0106**	0.2704	**0.3157**	**0.1278**	**0.1686**	**0.2783**	0.5063	0.1372
Major	**0.0106**	**0.2780**	**0.3157**	0.0918	**0.1686**	**0.2783**	0.5063	**0.1381**
#Topic	**33**	**34**	**35**	**Mean**				
Best mix ImageCLEF	0.1045	0.0823	0.0503	0.1608				
Visual	0.1265	0.0002	0.0013	0.0336				
Text	0.2876	0.2820	0.1536	0.1791				
Mix	0.2868	0.2786	0.1404	0.1889				
All	**0.2876**	0.2786	0.1404	0.1890				
Major	**0.2876**	**0.2820**	**0.1536**	0.1885				

dependence [156]. Moreover, the covariance is computed between binary (representing the MeSH term occurrence) and continuous (representing visual features) variables which also complicates the task.

Table 6.12: Accuracy (%) of correct decisions obtained by the proposed approaches when using *major* MeSH terms. The results are shown for several latent values (z) and percentiles $\mathcal{P}$.

$\mathbf{z} \setminus \mathcal{P}$	**0**	**50**	**75**	**99**
50	45.72	77.15	77.15	62.86
100	77.15	77.15	77.15	77.15
200	77.15	77.15	77.15	77.15
300	77.15	77.15	77.15	77.15

Table 6.13: Accuracy (%) of correct decisions obtained by the proposed approaches when using *all* and *major* MeSH terms.

Run	**Accuracy**
Major	77.15
All	62.86

6.5 Modality classification

This section presents the evaluation results for the image modality classification experiments over the ImageCLEFmed 2013 database. It describes the results achieved for the selection of the visual features and the semi–supervised learning approach as well as the crowdsourcing outputs. Finally the image modality classification approach is integrated into the medical case–based retrieval system.

6.5.1 Feature selection

To select the visual descriptors to use for the modality classification task the 2013 ImageCLEFmed training set is divided into two subsets. The accuracy obtained over the training set using each of the visual descriptors is shown in Table 6.14. Then, following the same procedure as for the retrieval task (see Section 6.1), the fusion of the best features is performed to obtain a good feature set (see Table 6.15). Due to time and resource limitations not all the possible combinations are tested and only the features performing well alone are combined in a simple linear way.

6.5.2 Semi–supervised learning and crowdsourcing

In the ImageCLEFmed 2013 modality classification task some of the image categories are represented by only very few annotated examples (see Table 6.16 for the exact numbers). Therefore, the semi–supervised learning algorithm (called Algorithm 1) to expand

Table 6.14: Classification accuracy of each selected feature over the training set.

Feature	CEDD	BoVW	FCTH	BoC
Accuracy (%)	51.75	50.62	49.28	49.25
Feature	FCH	HSV	Colour layout	Tamura
Accuracy (%)	47.50	44.63	44.01	43.26

Table 6.15: Classification accuracy of multiple features combined over the training set.

Feature	**Accuracy**(%)
CEDD+BoVW	55.92
CEDD+BoVW+FCTH	57.15
CEDD+BoVW+FCTH+BoC	58.45
CEDD+BoVW+FCTH+BoC+FCH	**60.16**
CEDD+BoVW+FCTH+BoC+FCH+HSV	58.86
CEDD+BoVW+FCTH+BoC+FCH+HSV+Col. layout	57.42
CEDD+BoVW+FCTH+BoC+FCH+HSV+Col. layout+Tamura	56.88

the training set is applied. All the training images are queried against the full 300,000 images of the ImageCLEFmed 2013 data set and the 10 highest ranked retrieved images of each query are added as training images into the class of the query image. Only the images belonging to the "compound or multipane images" (COMP) class are not queried because this class is well represented.

In [57] an arbitrary k is used for the k–NN classifier, explaining that the choice of k may not be optimal. In this thesis, the classification accuracy for a range of k is computed to investigate the robustness of the method. Table 6.17 shows the results for the experiments using the k–NN classifier over the following training sets:

- *RO* – original training set;
- *RE* – automatically expanded training set;
- *REN* – automatically expanded training set without expanding the compound images.

Using the original training set (*RO*) results achieve 68.93% accuracy when $k = 10$. The accuracy is increased to 72.61% when using $k_r = 10$ for the expansion of the training set (*RE*) and $k = 16$ for the k–NN classifier. Because a large amount of the figures in the dataset are compound images it is proposed to not include the compound image class with the training set expansion [83]. An additional experiment which does not expand this class runs (*REN*) but with worse results than previous experiments.

Half of the expanded training data is first manually corrected using crowdsourcing as explained in Section 5.4.2. The crowdsourcing in this experiment is done with an internal team to limit errors in the classification. A total of eight experts in the medical imaging domain participated in the task. In the past external experts were used but strong quality control is necessary in this case; on the other hand tasks can be finished extremely quickly. In the first step, 50% of the images are verified by crowdsourcing to augment the

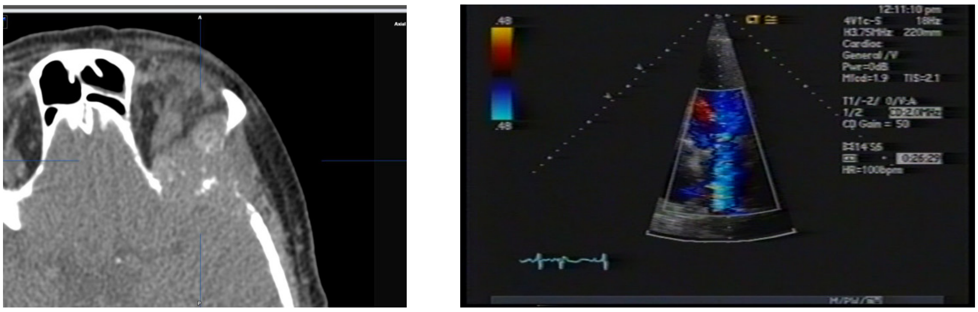
Table 6.16: Distribution in classes of training and test images from the ImageCLEFmed 2013 classification task.

Modality		**# of training images**	**# of test images**
Compound or multipane images		1,105	1,014
Diagnostic images	**Radiology**		
	Ultrasound	60	85
	Magnetic Resonance	97	90
	Computerized Tomography	113	186
	X-Ray, 2D Radiography	70	344
	Angiography	54	18
	PET	16	3
	Combined modalities in one image	22	1
	Visible light photography		
	Dermatology, skin	79	28
	Endoscopy	64	20
	Other organs	70	112
	Printed signals, waves		
	Electroencephalography	21	9
	Electrocardiography	29	96
	Electromyography	18	1
	Microscopy		
	Light microscopy	91	121
	Electron microscopy	51	20
	Transmission microscopy	46	20
	Fluorescence microscopy	33	33
	3D reconstructions	46	26
Generic biomedical illustrations			
Tables and forms		65	29
Program listing		28	22
Statistical figures, graphs, charts		102	101
Screenshots		91	20
Flowcharts		94	20
System overviews		89	16
Gene sequence		68	21
Chromatography, Gel		55	30
Chemical structure		62	19
Mathematics, formulas		20	5
Non–clinical photos		96	37
Hand–drawn sketches		46	54

Table 6.17: Accuracy (%) obtained applying the k–NN classifier for the ImageCLEFmed 2013 modality classification using the original training set as well as the two versions of expanded training sets.

k	**RO**	**RE**	**REN**	k	**RO**	**RE**	**REN**
2	65.05	62.19	43.67	14	67.49	72.49	56.92
3	68.19	67.11	49.13	15	67.34	72.26	57.07
4	68.62	68.38	50.14	16	66.99	72.61	57.50
5	68.70	69.55	51.41	17	66.99	72.26	57.96
6	68.81	69.70	52.54	18	66.68	72.41	58.04
7	68.85	70.63	53.20	19	66.37	72.38	57.57
8	68.50	71.10	54.20	20	66.56	71.83	57.81
9	68.26	71.68	54.39	21	66.21	72.49	57.61
10	68.93	71.52	55.60	22	66.18	**72.61**	57.88
11	68.62	71.76	55.52	23	65.79	72.57	57.69
12	68.00	72.26	56.49	24	65.71	72.14	57.77
13	67.76	71.87	56.88	25	65.71	72.22	57.77

training set and automatically classify the remaining images. Thanks to the first question in the platform 21% of the images are relabelled as compound figures during the same crowdsourcing job. Almost 20% of the images then have to be relabelled manually (see Figure 6.6). For the *relabelling* task, the correctly labelled images and the images labelled as compound are added to the initial training set. The new training set containing more than 9, 000 images is used to automatically relabel the non–labelled images (images tagged as "Wrong category" or "Not sure"). A total of 1, 600 images are automatically relabelled and then verified via crowdsourcing. 16% of the previously wrongly labelled images are now well labelled after the automatic relabelling. Some images that are correctly relabelled can be seen in Figure 6.5.



(a) Image reclassified as "CT".

(b) Image reclassified as "US".

Figure 6.5: Images correctly reclassified after the training set expansion verification.

New labels of this first iteration of the correction procedure (see algorithm 2) are then updated in the expanded data. This procedure leads to the creation of two more training sets:

- *REH* – automatically expanded training set with half of the expanded images manually relabelled;

Table 6.18: Accuracy (%) obtained applying the k–NN classifier for the ImageCLEFmed 2013 modality classification using the automatically expanded training set as well as the two versions of expanded training sets where half of the expanded images are manually relabelled.

k	**RE**	**REH**	**RENH**	k	**RE**	**REH**	**RENH**
2	62.19	65.88	53.51	14	72.49	70.59	71.76
3	67.11	69.82	60.44	15	72.26	70.40	72.57
4	68.38	69.78	61.76	16	72.61	70.28	72.34
5	69.55	70.40	64.20	17	72.26	70.44	72.65
6	69.70	70.40	65.44	18	72.41	70.28	72.65
7	70.63	70.36	67.07	19	72.38	70.48	72.84
8	71.10	70.86	67.96	20	71.83	70.21	73.03
9	71.68	70.55	68.19	21	72.49	69.90	**73.34**
10	71.52	70.75	69.35	22	72.61	69.82	73.30
11	71.76	70.63	70.44	23	72.57	69.74	73.19
12	72.26	71.02	70.71	24	72.14	65.25	73.23
13	71.87	70.40	70.90	25	72.22	64.97	72.76

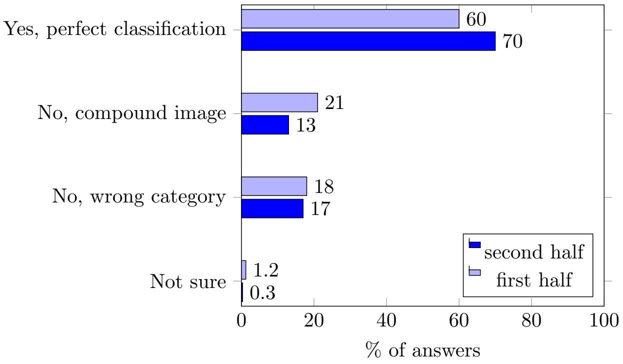


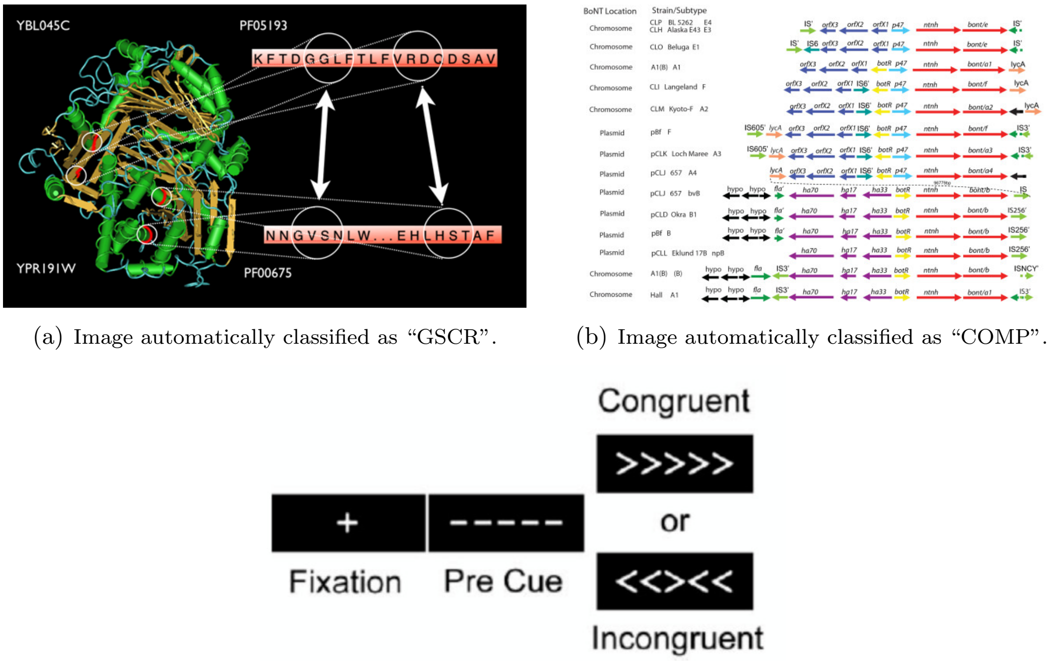
Figure 6.6: Each bar represents the distribution of each of the answers in the two verification crowdsourcing task.

- *RENH* – automatically expanded training set without expanding the compound images. Half of the expanded images are manually relabelled.

Therefore, two more experiments are run using the new labels (*REH* and *RENH*) increasing the accuracy to 73.34% with $k = 21$. This result is obtained without including the images tagged as "No, compound images" during the crowdsourcing iteration (*RENH*) in the training data. Table 6.18 shows the results for the experiments using the k–NN classifier over these training sets.

To crowdsource the label correction of the images of the second half of the expanded training set, images are previously automatically reclassified using the *RENH* training set and a k–NN classifier. $k = 21$ is chosen because it generates a higher accuracy in the experiments.

Hence, the number of correctly labelled images increases and fewer images have to be relabelled during the crowdsourced relabelling.



(a) Image automatically classified as "GSCR".

(b) Image automatically classified as "COMP".

(c) Image automatically classified as "GTAB".

Figure 6.7: Images incorrectly classified automatically but that are also difficult to classify manually.

Figure 6.6 shows the distribution of each of the answers in the two verification tasks carried out during the iterative process. Since the second half of the set is reclassified using the updated training set after the first half is relabelled, the second half of the set is better labelled.

In general, the relabelling task is more difficult for the experts than the verification task. More knowledge about the modality classes is necessary and indeed the classes are not always easy to identify. Figure 6.7 shows some examples of images incorrectly labelled, which experts also found difficult to classify. Often full size versions of the images are necessary to clearly determine the modality.

Finally, a training set with 19,905 images containing correct labels is obtained. Therefore, three more training sets are created:

- *RET* – automatically expanded training set of all of the expanded manually relabelled images;
- *RENT* – automatically expanded training set without expanding the compound images. All of the expanded images are manually relabelled.
- *REW* – automatically expanded training set without the images labelled as "compound" or "correct". All of the expanded images are manually relabelled.

Table 6.19 presents the results of all the used training set as well as the average and the standard deviation (SD) over the k's. Using the training set *RET* with all the correct labels and $k = 3$ the accuracy obtained is 69.16%. Two last experiments are carried out. One is done without including the images tagged as "No, compound images" during the crowdsourcing iteration (*RENT*) into the training data. Considering the hypothesis

Table 6.19: Accuracy (%) obtained applying the k–NN classifier for the ImageCLEFmed 2013 modality classification using various training sets. Average (Avg) and standard deviation (SD) over the k's is also showed.

k	**RO**	**RE**	**REN**	**REH**	**RENH**	**RET**	**RENT**	**REW**
2	65.05	62.19	43.67	65.88	53.51	68.19	70.67	66.37
3	68.19	67.11	49.13	69.82	60.44	69.16	73.5	69.86
4	68.62	68.38	50.14	69.78	61.76	68.66	73.27	70.13
5	68.70	69.55	51.41	70.40	64.20	67.34	73.30	70.44
6	68.81	69.70	52.54	70.40	65.44	67.53	**73.89**	70.98
7	68.85	70.63	53.20	70.36	67.07	67.45	73.42	70.71
8	68.50	71.10	54.20	70.86	67.96	66.83	72.68	71.83
9	68.26	71.68	54.39	70.55	68.19	66.80	73.27	71.25
10	68.93	71.52	55.60	70.75	69.35	66.52	73.23	71.33
11	68.62	71.76	55.52	70.63	70.44	66.49	72.84	71.41
12	68.00	72.26	56.49	71.02	70.71	66.83	72.80	71.68
13	67.76	71.87	56.88	70.40	70.90	65.83	72.80	71.33
14	67.49	72.49	56.92	70.59	71.76	65.40	72.57	71.17
15	67.34	72.26	57.07	70.40	72.57	65.13	72.10	71.02
16	66.99	72.61	57.50	70.28	72.34	64.94	72.14	70.36
17	66.99	72.26	57.96	70.44	72.65	64.20	71.83	70.59
18	66.68	72.41	58.04	70.28	72.65	64.20	71.91	70.55
19	66.37	72.38	57.57	70.48	72.84	64.24	71.37	70.24
20	66.56	71.83	57.81	70.21	73.03	64.01	71.52	70.52
21	66.21	72.49	57.61	69.90	73.34	63.62	71.41	70.52
22	66.18	72.61	57.88	69.82	73.30	63.27	71.10	70.24
23	65.79	72.57	57.69	69.74	73.19	62.92	71.02	70.24
24	65.71	72.14	57.77	65.25	73.23	62.81	71.29	69.97
25	65.71	72.22	57.77	64.97	72.76	62.46	71.33	69.55
Avg	67.35	71.08	55.20	69.72	69.32	65.62	**72.30**	70.51
SD	1.21	2.37	3.61	1.72	5.06	1.96	**0.94**	1.06

that the correctly classified and compound images from the expanded dataset do not add information, another experiment is done adding only the images tagged as “No, wrong category” and “Not sure” during the crowdsourcing iteration (REW) to the original training set. Best results are obtained using the $RENT$ training set and $k = 6$ achieving 73.89% accuracy. The use of this training set also achieved best accuracy on the average over the k values. Indeed the standard deviation is also the lowest (0.94) showing that the results obtained with most of the k values are very close to the average (see Table 6.19). Using k–NN, it can be seen by the average accuracy that the variant without the added compound images ($RENT$) of the proposed multi–modal approach achieves the best results. The results obtained by the full expanded dataset (RET) are poorer than the ones obtained without the extra compound images ($RENT$). This can be explained by the fact that the compound figures belong to the best–represented class in the test set and did not need to be expanded. It also demonstrates that the application of semi–supervised learning is not trivial and the algorithms used should take into account the data distribution across classes.

6.5.3 Modality classification for retrieval

Two techniques to integrate the image modality into the medical case–based retrieval are investigated in this thesis: *filtering* and *reranking* (see Section 5.4.3). To carry out these experiments all the images are classified using the best classification approach presented above. Relatively low standard deviations of the algorithm suggest that the k–NN algorithm is stable across k choices (see Table 6.19). Therefore, the $RENT$ training set and $k = 6$ are used. Experiments are first carried out on the ImageCLEFmed 2012 collection to choose the best method before applying it to the ImageCLEFmed 2013 collection. Results are also compared with the run without any image modality modification presented in Section 6.3 and with the best result achieved in ImageCLEFmed.

Table 6.20 shows the performance of the four approaches of modality filtering and reranking applied to the visual retrieval step. Using both methods, filtering and reranking, “close” and “prefix” approaches obtain slightly better results that the approach without using any modality integration. Therefore, these approaches are further explored for multi–modal retrieval.

However, modality reranking is not helpful and filtering is even detrimental to text retrieval (see Table 6.21). This may occur because text techniques retrieve numerous documents which contain no relevant images or even no images. Section 6.2 analyses in detail the relevant images of the ImageCLEFmed query topics.

Therefore, multi–modal retrieval is carried out integrating the modality only into the visual runs using the approaches “close” and “prefix”. The filtered or reranked visual ranks are then combined with the text rank. Table 6.22 presents the results compared with the approach before applying any modality integration. Similar results are obtained when applying a “close” approach or not modality modification. The benefit of applying an image modality filter is the reduction of the search space. For this reason, the “close” and “prefix” filtering approaches are also investigated for the ImageCLEFmed 2013 collection. Table 6.23 shows that both filtering approaches achieve slightly better results than without applying any modality filtering. Moreover, both approaches outperform the best multi–modal ImageCLEFmed results. More important is the reduction of the image search space obtained in the CBIR step.

Table 6.20: Results of the visual approaches when using various modality integration strategies on the ImageCLEFmed 2012 medical case–based retrieval task.

	Run ID	**Method**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
	Best visual ImageCLEF	–	0.0366	0.0014	0.0347	0.0269	0.0141
	RunVF13	–	0.0490	0.0014	0.0702	**0.0308**	0.0141
Filter	RunVC1	exact	0.0434	0.0009	0.0579	**0.0308**	0.0128
	RunVC2	close	**0.0493**	**0.0015**	0.0691	**0.0308**	**0.0154**
	RunVC3	prefix	0.0490	0.0014	**0.0707**	0.0269	**0.0154**
	RunVC4	diagnostic	0.0394	0.0014	0.0573	0.0231	0.0128
Rerank	RunVC5	exact	0.0435	0.0008	0.0590	0.0308	0.0115
	RunVC6	close	**0.0493**	0.0013	0.0692	**0.0346**	**0.0141**
	RunVC7	prefix	0.0490	0.0014	**0.0714**	0.0308	**0.0141**
	RunVC8	diagnostic	0.0402	**0.0017**	0.0568	0.0269	0.0115

Table 6.21: Results of the textual approaches when using various modality integration strategies on the ImageCLEFmed 2012 medical case–based retrieval task.

	Run ID	**Method**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
	Best text ImageCLEF	–	**0.1690**	**0.0374**	**0.1499**	**0.1885**	**0.1090**
	Baseline (T1)	–	0.1670	0.0355	0.1413	0.1731	0.1077
Filter	RunTC1	exact	0.0795	0.0071	0.0877	0.1154	0.0705
	RunTC2	close	0.1562	0.0290	0.1333	0.1654	0.1064
	RunTC3	prefix	0.0834	0.0089	0.0907	0.1154	0.0705
	RunTC4	diagnostic	0.1662	0.0322	0.1450	0.1731	0.1051
Rerank	RunTC5	exact	0.1670	0.0355	0.1413	0.1731	0.1077
	RunTC6	close	0.1670	0.0355	0.1413	0.1731	0.1077
	RunTC7	prefix	0.1670	0.0355	0.1413	0.1731	0.1077
	RunTC8	diagnostic	0.1670	0.0355	0.1413	0.1731	0.1077

Table 6.22: Results of the multi–modal approaches when using various modality integration strategies on the ImageCLEFmed 2012 medical case–based retrieval task.

	Run ID	**Method**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
	Best mix ImageCLEF	–	0.1017	0.0175	0.0857	0.1115	0.0679
	Baseline (MF7)	–	0.1806	0.0397	0.1578	0.1808	0.1064
Fil.	RunMC1	close	**0.1808**	**0.0400**	**0.1598**	0.1808	**0.1077**
	RunMC2	prefix	0.1803	0.0394	0.1572	**0.1846**	0.1064
Rer.	RunMC3	close	0.1807	0.0398	0.1583	0.1808	**0.1077**
	RunMC4	prefix	0.1784	0.0391	0.1534	0.1808	0.1064

Table 6.23: Results of the multi–modal approaches when using various modality filtering strategies on the ImageCLEFmed 2013 medical case–based retrieval task.

Run ID	**Method**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best mix ImageCLEF	–	0.1608	0.0779	0.1426	0.1800	0.1257
Baseline (MF10)	–	0.1889	0.1190	0.1720	0.2257	0.1629
RunMC5	close	0.1892	0.1191	0.1720	**0.2286**	0.1629
RunMC6	prefix	**0.1904**	**0.1208**	**0.1732**	0.2257	**0.1638**

6.6 Final approach

This section presents the results achieved in a final approach. The medical case–based retrieval method is executed by combining the techniques studied in this chapter. The combination of the following procedures is applied:

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

Optimal choice of parameters was decided in the previous sections.

Table 6.24 shows how the proposed final approaches, with "close" or "prefix" filters, outperform the best multi–modal approach submitted to ImageCLEFmed. Indeed, results show that filtering by a broad modality in the hierarchy ("prefix" filter) improves the retrieval performance. Moreover, the combination of the techniques suggested in this thesis has enabled a more efficient search, limiting the use of CBIR only to suitable cases, and reducing the search space thanks to the modality filtering.

Table 6.24: Results of the *final* multi–modal approaches when using "close" and "prefix" modality filtering strategies on the ImageCLEFmed 2013 medical case–based retrieval task.

Run ID	**Method**	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best mix ImageCLEF	–	0.1608	0.0779	0.1426	0.1800	0.1257
RunF1	close	0.1892	0.1191	0.1720	**0.2286**	0.1629
RunF2	prefix	**0.1904**	**0.1208**	**0.1732**	0.2257	**0.1638**

6.7 Result discussion

Section 6.1 presents a basic retrieval approach which is used as a baseline for comparison. The proposed visual approach achieves better results than the best visual run submitted to the ImageCLEFmed case–based task in 2012 and 2013, except at the precision 10 and 30 measures. On the other hand, for text retrieval, better approaches were submitted to ImageCLEFmed.

In Section 6.2, it is observed that visual retrieval is not optimal for many query topics of the medical case–based retrieval task of 2013. The retrieval performance can be enhanced more effectively when there is a sufficient number of relevant articles, which is not the case for all the presented query topics.

In Section 6.3, several combination strategies are applied. Experimental results demonstrate the impact of the types of fusion rules used on the retrieval performance. Visual fusion achieves best results when the queries are fused with combMAX and the descriptors with Borda. These results are consistent with previous studies [19, 104], which state that combination using similarity scores is better than using only ranks when the sources are commensurable. This is the case of the fusion of the queries where the proposed approach outperforms when using combMAX, a score–based technique. Furthermore, Belkin et al. [19] also state that a combination method based on ranked outputs is better when the sources have incompatible scores. Hence, Borda, a rank–based method, obtains best results when fusing several descriptors. Despite the low performance of the visual search, the effectiveness of the multi–modal approaches is improving the text search for the first time in the medical case–based retrieval task. Results also outperformed the best multi–modal runs submitted to ImageCLEFmed 2013 by a weighted linear combination of visual and text retrieval. It demonstrates the effectiveness of the proposed multi–modal framework.

A major challenge is the low performance of the visual retrieval approach for the medical case–based retrieval task. To overcome this, a query–adaptive fusion criterion for the use of multi–modal techniques in medical case–based retrieval is presented in Section 6.4. The textual information of MeSH terms are integrated with the visual descriptors creating a matrix of synonym relations between both kinds of features (text and visual). The synonym matrix is then used to decide if a text query is suitable for a multi–modal approach or if text alone would lead to best results. Experimental results indicate that it is indeed effective, showing that correct decisions are taken in 77.15% of the cases. The results are also very stable regarding parameter choices. On the other hand, the use of a covariance matrix in place of the synonymy matrix is inappropriate for the query–adaptive fusion criterion on the ImageCLEFmed collection. This can be due to the nature of the data because the covariance is computed between binary (representing the MeSH term occurrence) and continuous (representing visual features) variables. In addition, the estimation of the high–dimensional covariance matrix is challenging due to the large number of parameter to be estimated and it affects the results [156, 13, 138]. The current work opens an area of research on multi–modal decision for medical case–based retrieval. Moreover, by facilitating decision–making the criterion avoids the unnecessary use of CBIR. It makes the retrieval system more efficient.

Furthermore, image classification is applied to enhance the quality of the retrieval system. Modality classification is important in medical image retrieval systems, both for overall retrieval quality and because it is a funtionality requested by users. Section 6.5 describes a method for improving medical image classification and therefore medical case–based retrieval. The method uses multi–modal retrieval to exploit unlabelled data for semi–supervised learning to create an expanded training set. Crowdsourcing is used to manually correct the assigned labels and, therefore, to improve the quality of an automatic modality classification task. Increasing the size of the training set is shown to improve the quality of the automatic classification. Manual correction of such a noisy training set also significantly improves performance, demonstrating the effectiveness of the presented method. For the ImageCLEFmed classification task, better accuracy scores than the ones reported in this thesis were obtained by two participants in ImageCLEFmed 2013 [79] (81.68% and 78.04%). However, these multi–modal approaches either used much more sophisticated classifiers such as SVM or multiple classifiers [139, 174] compared to k–NN on more complex visual features. Moreover, a medical image modality filter is also presented in this chapter to filter out non–relevant images, which has the possibility to remove some noise from the results. Image modality filtering improves the performance of a simple visual retrieval and, therefore, a multi–modal retrieval. For each query image, descriptors are extracted and compared with the image descriptors stored in the database. Therefore, image filtering reduces the search space focusing the search only on the query modalities.

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

6.8 Summary

This chapter relates the different factors that led to performance improvement of a medical case–based retrieval system. After the presentation of a baseline, ImageCLEFmed query topics are analysed to better understand the task.

All the techniques presented in Chapter 5 are implemented for the ImageCLEFmed database and studied in this chapter. These techniques are then combined to define a final approach. The final approach outperforms the best multimodal approach submitted to ImageCLEFmed. Moreover, the query–adaptive criterion and the modality filtering examined contribute to improve the effectiveness of the retrieval system.

A detailed reading of the results is done at the end. It shows what can be said about the system performance by taking a close look at the performance measures.

Chapter 7

Shangri–La: a Web–based Interface for Medical Case–based Retrieval

> “Success is getting what you want. Happiness is enjoying what you get.”
>
> Ralph Waldo Emerson

Chapter 3 defines a use case to support clinical decisions. Furthermore, Chapter 5 proposes an approach for medical case–based retrieval and Chapter 6 evaluates the results obtained from the use of the proposed approach on the ImageCLEFmed collection. The techniques described are integrated into the ParaDISE system.

This chapter presents a novel web–based retrieval interface, called *Shangri–La*, based on the ParaDISE system and the multi–modal techniques proposed in this thesis. The goal of the interface is to provide a front–end with which the user can interact and control the underlying medical case–based retrieval system [25]. The web–based interface seises the opportunities and challenges given by the Internet to easily share the developed system. Shangri–La provides multi–modal retrieval functionalities that allow the user to find relevant articles querying the system with a text case description and/or visual examples. Currently, the ImageCLEFmed 2013 dataset (see Chapter 4) is accessible from Shangri–La and supported by the proposed system. However, it can easily be extended to other datasets.

This chapter provides a detailed overview of the interface features. A complete version of the proposed interface is available at the following address `http://shangrila.khresmoi.eu/`.

7.1 System architecture

Shangri–La is developed as a client–side only application, based entirely on HyperText Markup Language 5 (HTML5) and JavaScript.

The interface accesses several web services that use a REpresentational State Transfer (REST) style architecture [211]. The used web services are described below (see Figure 7.1):

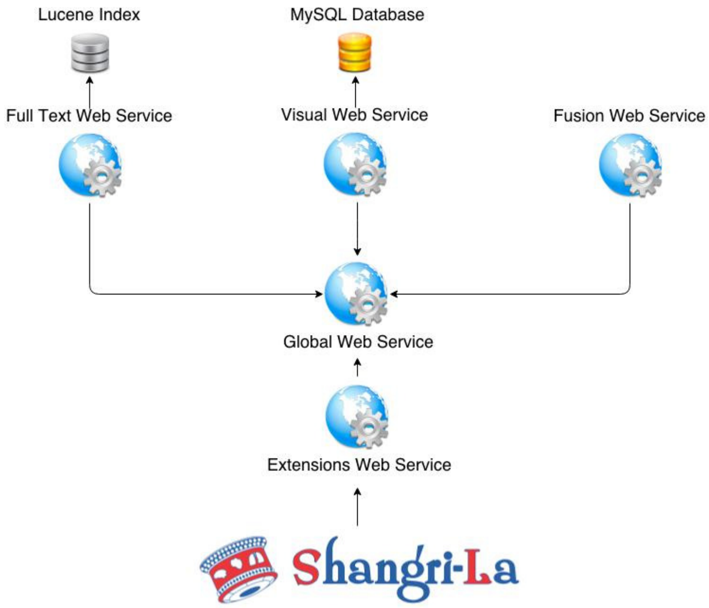


Figure 7.1: Medical case–based retrieval system service layer architecture.

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

For more details on the techniques used by the web service see Chapter 5.

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

7.2 Interface functionality

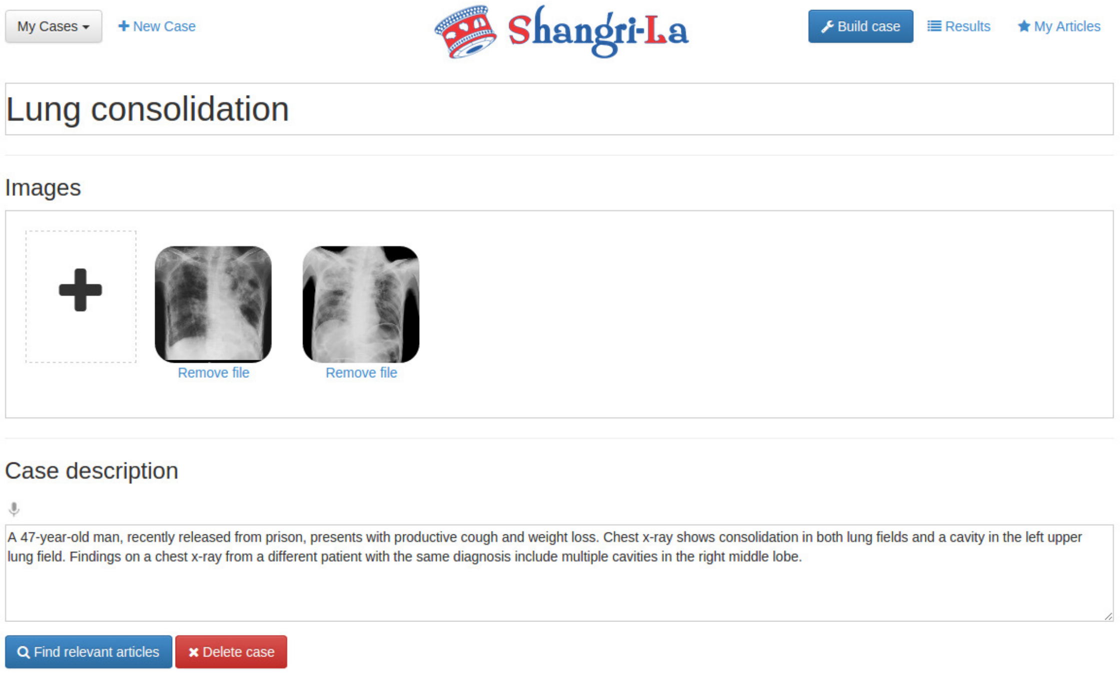
Shangri–La enables people to interact with the medical case–based retrieval system with the least amount of user effort. The interface hides the complexity of the system implementation, giving users a simple site to collect the desired information. To keep the interaction clear and concise, Shangri–La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

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

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

7.2.1 Build Case page

The goal of the *Build Case* page is to simply and easily capture the user's information need. The medical case–based retrieval system was developed in this thesis to support inputs including a text case description and image examples. Query images can be uploaded from a storage device using a file browser dialogue selection method. In addition, drag and drop facilitates query image uploading. Both options are available to satisfy user needs. A text area is used to enter text which can contain multiple lines of textual information for long case descriptions. Shangri–La also supports real–time speech recognition which transcribes a spoken query into text using the Google Chrome Speech API. However, in the current version many phrases or words are not recognised. An example of a build case query is shown in Figure 7.2.



Lung consolidation

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

Figure 7.2: Screenshot of the *Build Case* page from the Shangri–La interface. This page shows an example of a query.

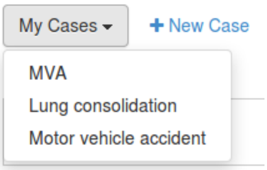


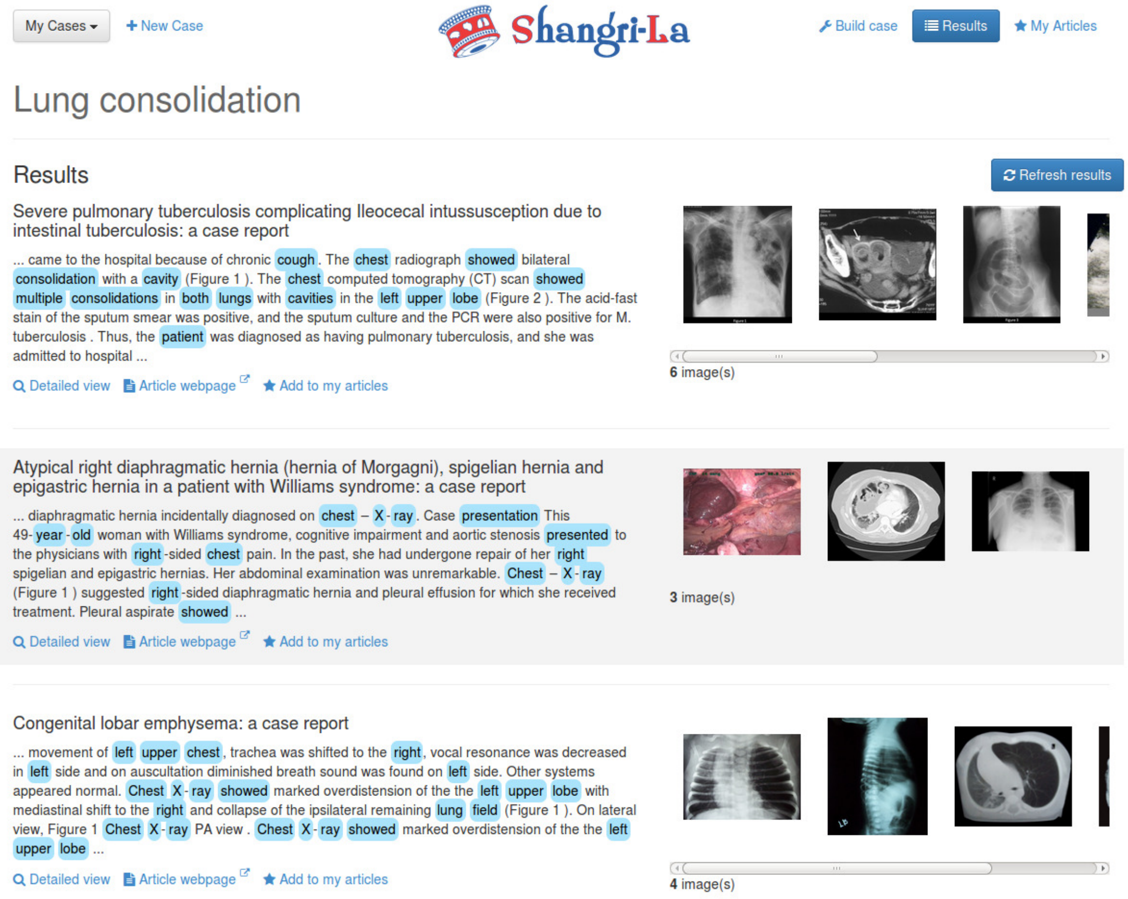
Figure 7.3: Screenshot of the drop–down menu from the Shangri–La interface. It allows users to navigate between their cases.

In addition, users can navigate through the cases they already built thanks to a drop–down menu (see Figure 7.3). It allows users to find their cases with ease.

7.2.2 Results page

The goal of the *Results* page is to provide all of the information needed to support the medical case–based retrieval task. A users' study [67] shows that users prefer retrieved results which display, accompanying the title, lines in the document which fulfil the search condition and not the first lines of the document. Therefore, Shangri–La displays the resulting articles of the search in a ranked list, with basic information containing: the title, relevant lines which fulfil the search criteria taken from the body of the article, and images (if available). In addition, terms contained in the text query are highlighted. An example of an outcome displayed in the *Results* page is shown in Figure 7.5.

Furthermore, the interface provides a link to the corresponding article as well as a bookmark option (see Section 7.2.3). A detailed view of the article is also possible without going to the original source. It includes its title, abstract and images contained in the article. Moreover, the user can click on the images contained in a retrieved article to see a larger view.



Lung consolidation

Results

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

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

Detailed view Article webpage Add to my articles

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

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

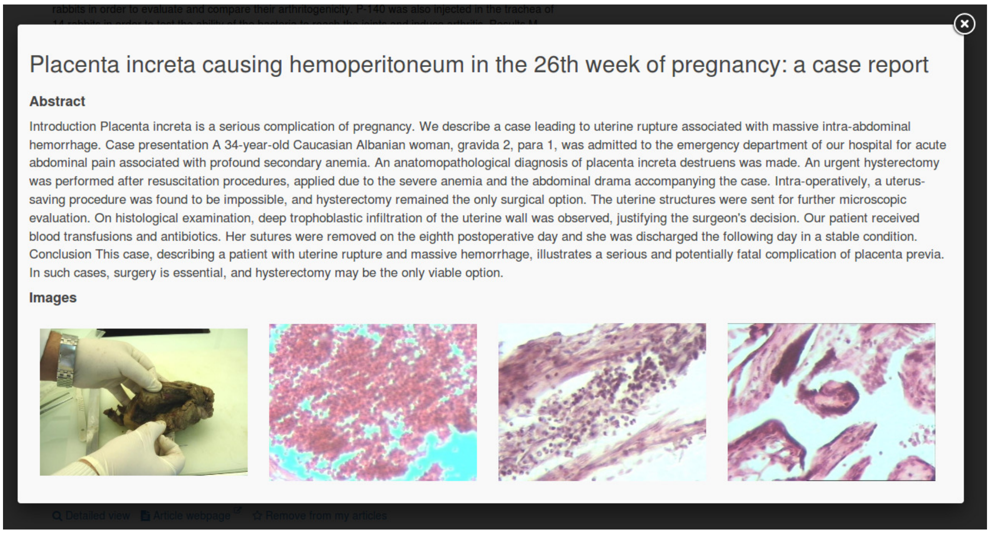
Detailed view Article webpage Add to my articles

Congenital lobar emphysema: a case report

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

Detailed view Article webpage Add to my articles

Figure 7.4: Screenshot of the *Results* page from the Shangri–La interface. This page shows a ranked list of articles resulting from a search query.



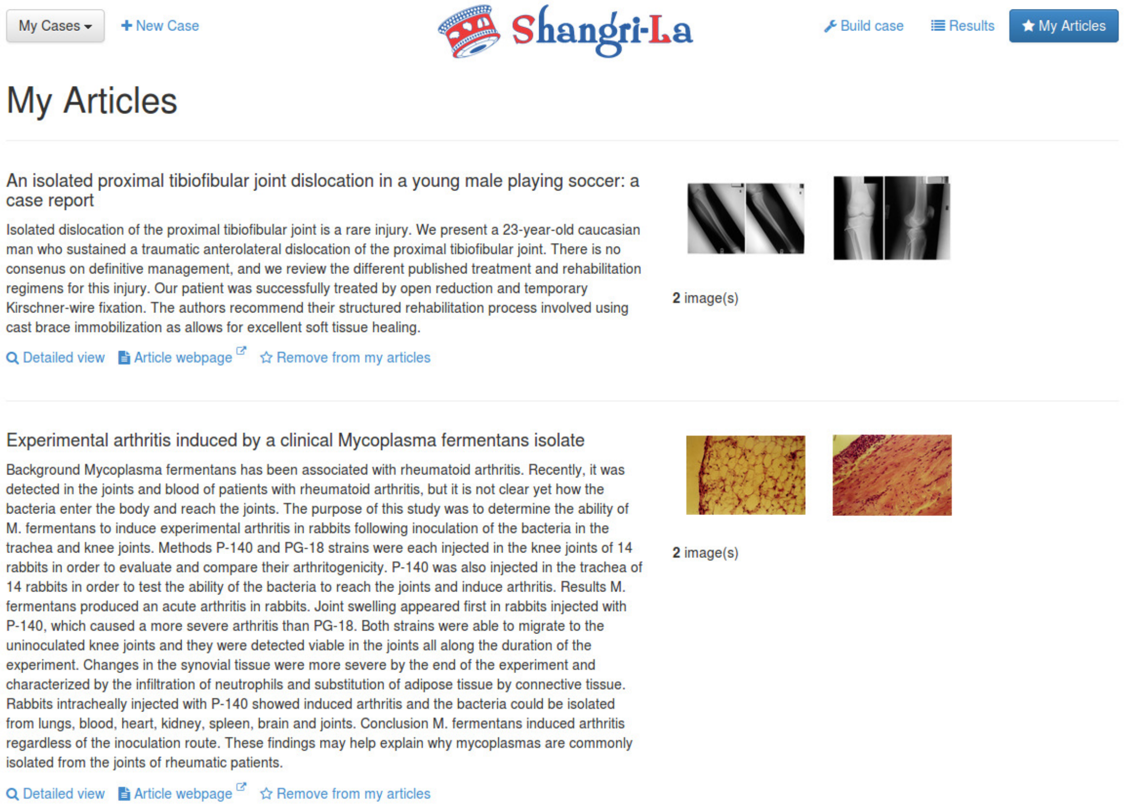
Placenta increta causing hemoperitoneum in the 26th week of pregnancy: a case report

Abstract

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

Images

Figure 7.5: Screenshot of the *detailed view* of an article from the Shangri–La interface. This page shows the title, abstract and images from a selected article.



My Articles

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

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

Detailed view Article webpage ☆ Remove from my articles

2 image(s)

Experimental arthritis induced by a clinical Mycoplasma fermentans isolate

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

Detailed view Article webpage ☆ Remove from my articles

2 image(s)

Figure 7.6: Screenshot of the *My Articles* page from the Shangri–La interface. This page shows a selection of articles that the user added to easily revisit.

7.2.3 My Articles page

Bookmarks, also referred to as favourites or hotlist, are a common tool to facilitate revisitation when the relevant (previously visited) result is somewhere in the results list, even if is not among the first results. Indeed, a recent survey found that 97% of the users are aware of the bookmarks function and 85% regularly save web pages using this method [214].

Shangri–La allows the user to bookmark articles from the results list of a case search. The *My Articles* page displays the favourite articles that are added by the user. An overview of the selected articles is shown in the same format as in the *Results* page (see Section 7.2.2). The user can interact with the article selection from this page. The page allows checking the overview information display, visiting the article or even deleting articles from the list. Figure 7.6 shows a view of the actual *My articles* page for a selection of articles.

7.3 Summary

This chapter describes a multi–modal search interface for biomedical articles based on medical cases. To facilitate the interaction between a user and the medical case–based retrieval system developed in this thesis, a web–based interface, called Shangri–La, is presented. The design of this web–based interface takes into account the main features of a retrieval system described in the use case (see Chapter 3). This chapter describes the architecture of the medical case–based retrieval system and the interface's functionalities.

Chapter 8

Conclusions and outlook

"It is the combination of reasonable talent and the ability to keep going in the face of defeat that leads to success."

Daniel Goleman

This thesis has investigated a medical retrieval system use case associated with an evaluation task. The use case is developed and validated based on users' needs as well as its role in the retrieval application. An evaluation framework focuses on medical tasks, ImageCLEFmed, was developed based also on the use case analysis. Furthermore, a medical case–based retrieval system is proposed and evaluated. Finally, this system is integrated into a web–based interface, Shangri–La.

In this chapter, the objectives of the work in this thesis are revisited, the contributions and limitations are summarised and promising directions for future research are discussed.

8.1 Objectives revisited

As stated in Chapter 1, the problem addressed in this thesis is *medical visual information retrieval* and its *system evaluation*.

A use case is described and analysed (see Chapter 3). It consists of finding medical cases/images similar to the one under observation for supporting a clinician's decision–making during medical diagnosis. The developed use case is validated through an iterative process. First, the initial specifications of the use case are described. Then, validation is carried out to validate and improve the realism, accuracy and coverage of the use case.

Based on the validated use case the evaluation resources for the ImageCLEFmed benchmark 2011–2013 are developed (see Chapter 4). Thanks to the organization of this evaluation campaign in the context of this thesis, research in medical retrieval was encouraged.

In particular, taking into account the opinions of interviewees, this thesis focuses on the medical case–base retrieval task. The goal is to retrieve articles from the biomedical literature that potentially help clinicians in the process of diagnosing a case. This task is considered closer to the clinical routine than a classical image–based retrieval task. However, it is also a much more complicated problem. In this work, the main aspect is to bring the visual information available in the medical cases into a retrieval system. Due to the nature of the data collected in the ImageCLEFmed campaign, it is really difficult to evaluate and to show the improvements that visual information brings to solving the problem. However, ImageCLEFmed has brought up several clear lessons learned over the campaigns. These lessons encourage work on the following technical aspects of a medical case–based retrieval system: the use of multiple visual features as well as a special focus on information fusion; studying the semantic relation between visual and text information and how to best use it; and the use of modality information of images including extension of the training data set (see Chapters 5 and 6).

Finally, Shangri–La, a web–based retrieval interface, is implemented to integrate the multi–modal medical case–based retrieval approach proposed in this thesis (see Chapter 7).

8.2 Contributions

The main scientific contributions of this thesis are summarised in this section. The introduction (Chapter 1) already mentions these achievements and the different chapters of this thesis also detail them.

The "visual clinical decision support" use case is defined and validated. It was used to define an evaluation framework for medical retrieval systems, the ImageCLEFmed.

To better define the evaluation framework during this thesis, a detailed analysis of the the scholarly impact of ImageCLEF in the years previous to this thesis was carried out. Since 2011, ImageCLEFmed was organised in the context of this thesis. Freely available databases and ground truth were generated following a preparation process also detailed in this thesis. Evaluation of participant systems and comparison of techniques was carried out to draw conclusions and extract lessons learned.

The medical case–based retrieval task is studied in more detail in this thesis because it is refereed as the most realistic problem by the surveyed people. This task aims to retrieve articles from the biomedical literature that might help in the diagnosis of a given case which includes images. As a result, a medical case–based retrieval system is implemented. This thesis focuses on the integration of the visual information into the retrieval system. To achieve this goal different fusion strategies are implemented providing evidence that multi–modal medical case–based retrieval systems can obtain good performance using appropriate fusion. However, query topics have big differences in performance results even though ImageCLEFmed query topics are carefully created. To overcome this problem, a query–adaptive multi–modal fusion criterion is created to change the formulation of the retrieval algorithm based on the user query. The proposed method integrates the textual information of MeSH terms with visual descriptors creating a matrix of synonym relations between both kinds of features (text and visual). The synonym matrix is then used to decide if a text query is suitable for a multi–modal approach or if text alone would lead to best results.

Furthermore, image modality classification is also used to extract relevant information from the images to filter/rerank results lists. This thesis presents a biomedical image modality classification approach. A semi–supervised learning technique is applied to expand the uneven ImageCLEFmed training set; and to manually correct the assigned labels a crowdsourcing platform is used. Consequently, a larger and more accurate image collection for evaluating image modality classification is created.

The performance of the experiments is assessed on the very challenging collection of the medical case–based retrieval task of ImageCLEFmed 2013. Experimental results indicate that the approach is indeed effective despite the low performance. Better results than the best ImageCLEFmed submitted multi–modal approach are achieved. The search space is reduced by using CBIR only when suitable in addition to a modality filter. Moreover, the methods applied can be reproduced easily by other researchers and may serve for further investigation on medical case–based retrieval.

Finally, the web–based retrieval interface Shangri–La makes the proposed case–based retrieval system accessible from a web browser. It simplifies the search for relevant information to the clinicians when studying their cases. Furthermore, it can allow them to obtain additional information in less time benefiting their clinical workflow.

8.3 Limitations

Despite the mentioned contributions in Section 8.2 this chapter also identifies the limitations of the work done in this thesis.

The database used for ImageCLEFmed may not be optimal for the medical case–based retrieval task. It contains many generic biomedical illustrations, such as statistical figures or non clinical photos, and compound images. This has clear implications for the systems evaluation. The creation or reuse of a database containing medical cases would be more appropriate. However, the ImageCLEFmed collection needs to be open–access and to allow its redistribution to the participants. In addition, the goal was to use a large scale collection for retrieval. Therefore, PMC was chosen because it fulfils these requirements and no database was found containing medical cases that was large enough and redistributable.

The medical case–based retrieval approach proposed in this thesis has limitations. The query–adaptive multi–modal fusion is constrained to the MeSH terms extracted from the queries and manualy annotated MeSH terms from the articles. It does not use further text information. Moreover, only matrices of synonyms and covariances are explored and no other relations between text and visual information. Another limitation is that the proposed approach does not deal with compound images, which are between 40% and 50% of the collection. Finally, the modality classification approach is limited to the k–NN classifier being dependent to the k parameter although the results are stable under various k elections.

8.4 Perspectives

This section identifies potential directions for future work for research initiated by this thesis. Three perspectives are proposed here: evaluation, technical aspects and system integration.

8.4.1 Evaluation

After the detailed analysis of the query topics of the medical case–based task, further work should be done to correctly evaluate visual systems. Average effectiveness scores hide a big variation [33]. Future evaluation campaigns may consider different weights for "difficult" or "easy" query topics. Mizzaro [169] already proposed to reward high effectiveness on difficult query topics more than high effectiveness on easy query topics, and to penalise low effectiveness on easy query topics more than low effectiveness on difficult query topics. ImageCLEFmed currently uses GMAP to emphasise "difficult" query topics improvements in their performance.

Furthermore, a large number of the figures in PMC are compound figures (images consisting of several subfigures). When data from articles is made available digitally, often the compound images are not separated but made available in a single block. IR systems for images should be capable of distinguishing the parts of compound figures that are relevant to a given query. A major step for making the content of the compound figures accessible is the detection of compound figures and then their separation into subfigures that can subsequently be classified into modalities and made available for research.

ImageCLEFmed proposes four types of tasks in 2015:

- *Compound figure detection* – compound figure identification is a required first step to make available compound images from the literature. Therefore, the goal of this task is to identify whether a figure is a compound figure or not. The task makes training data available containing compound and non compound figures from the biomedical literature.
- *Multi–label classification* – characterization of compound figures is difficult, as they may contain subfigures from various image modalities. This task aims to label each compound figure with each of the modalities (of the 30 classes of a hierarchy shown in Figure 8.1) of the subfigures contained without knowing where the separation lines are.
- *Figure separation* – this task was first introduced in 2013. The task makes available training data with separation labels of the figures, and then a test data set where the labels are made available after the submission of the results. In 2015, a larger number of compound figures are distributed.
- *Subfigure classification* – similar to the modality classification task organised in 2011–2013 this task aims to classify images into the 30 classes of the hierarchy shown in Figure 8.1. The images are the subfigures extracted from the compound figures distributed for the figure separation task.

To carry out ImageCLEFmed 2015, new data is developed for correct evaluation. Crowdsourcing is used for this aim.

8.4.2 Medical case–based retrieval techniques

Medical case–based retrieval is a challenging problem. Therefore, there is a huge number of possible directions for future work. The most promising ones related to the work started in this thesis are presented here.

Semantic relations between images and text could be further studied. MeSH terms are formalised in a hierarchical structure, representing increasing levels of specificity [188]. Including hierarchical relationships between MeSH terms in the presented synonym approach could be a very powerful tool when searching for specific query topics. Future work also includes a study of synonym relation between visual descriptors and terms of UMLS. In addition, other relations between text and visual information can be study such as the correlation.

In future evolutions of query–adaptive multi–modal fusion, visual query reweighting based on synonym relations between text and visual features will be performed. Usually, an optimal data description does not exist, as the suitable data representation is strongly user–query dependent. This method can also be explored for automatic visual descriptor selection based on the user-query.

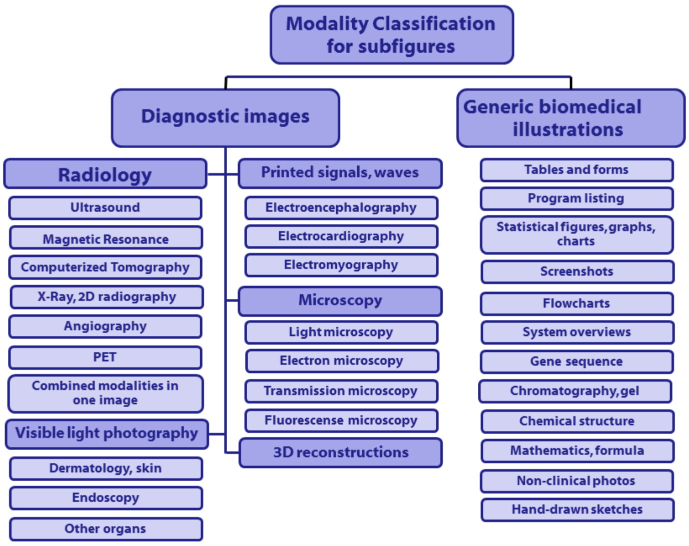


Figure 8.1: The image class hierarchy that was developed in 2015 for document images occurring in the biomedical open access literature.

This thesis demonstrated the effect of semi–supervised learning on classification performance. More importantly, other supervised learning techniques can benefit from the proposed method. Future developments could evaluate the training set expansion using other classifiers, such as SVM or Random Forests and even using more complex visual features. Evaluation with multi–kernel classifiers could also better demonstrate the added value of multi–modal semi–supervised learning. Crowdsourcing is used in this thesis to improve the quality of automatic image modality classification. The iterative nature of the shown crowdsourcing process will be continued to progressively generate a large and discriminative training set so all images of PMC can be automatically classified and then made accessible for retrieval tasks.

Finally, since the visual retrieval performance is rather low, future work will concentrate more on better ways to increase the visual performance such as compound figure separation methods. Current retrieval systems could greatly benefit from the use of multi–label classification approaches [258] by: defining models that can use the dependencies between the extracted images; and defining models that can express the importance of a label in a compound image. In addition, compound images are naturally redundant sources of information with natural dependencies occurring between the different regions of the image. Two research topics may improve the task of multi–label classification of compound images: compressive sensing and deep learning. Compressive sensing models [28] can make use of the natural redundancy and sparsity of images to define compressed features carrying more information than with standard feature extraction approaches. Deep learning models [69] can define an automatic and highly meaningful feature extraction that, adding to the compressive sensing models may also improve the multi–label classification task.

8.4.3 System integration

The Shangri–La interface brings to the clinicians a medical case–based retrieval interface. Plans for the future include user tests to fully determine its effectiveness and overall usability. User tests will show how real users actually interact with the system. The analysis will ensure that the features described and validated in the use case correspond to real cases. Furthermore, it will help to address users needs and expectations.

Speech recognition is already integrated in Shangri–La although there is room for improvement of the quality. Indeed, CLEF eHealth Lab is organizing a clinical speech recognition task in 2015. This task is related to converting verbal nursing handover to written free–text records. Research on this area will minimise word–detection errors by addressing the correctness of the speech recognition engine itself and the post–processing methods for the recognised text.

Other alternative human–computer interaction tools can be integrated into the interface. Motion sensors, such as the Leap Motion controller [248], allow gesture recognition enabling interaction with the system without touching a mouse or a keyboard. It would facilitate its use in a sterile clinical setting, such as an operating room. Leap Motion controller was already integrated into a medical CBIR interface also based in ParaDISE [251]. In addition, Google Glass, a wearable device which is also able to interact with online services, was connected to the same CBIR system [250]. An application to use the proposed medical case–based retrieval system with Google Glass can also be developed. Thanks to its built–in capabilities, Google Glass would allow clinicians to keep contact with the patient while obtaining additional information.

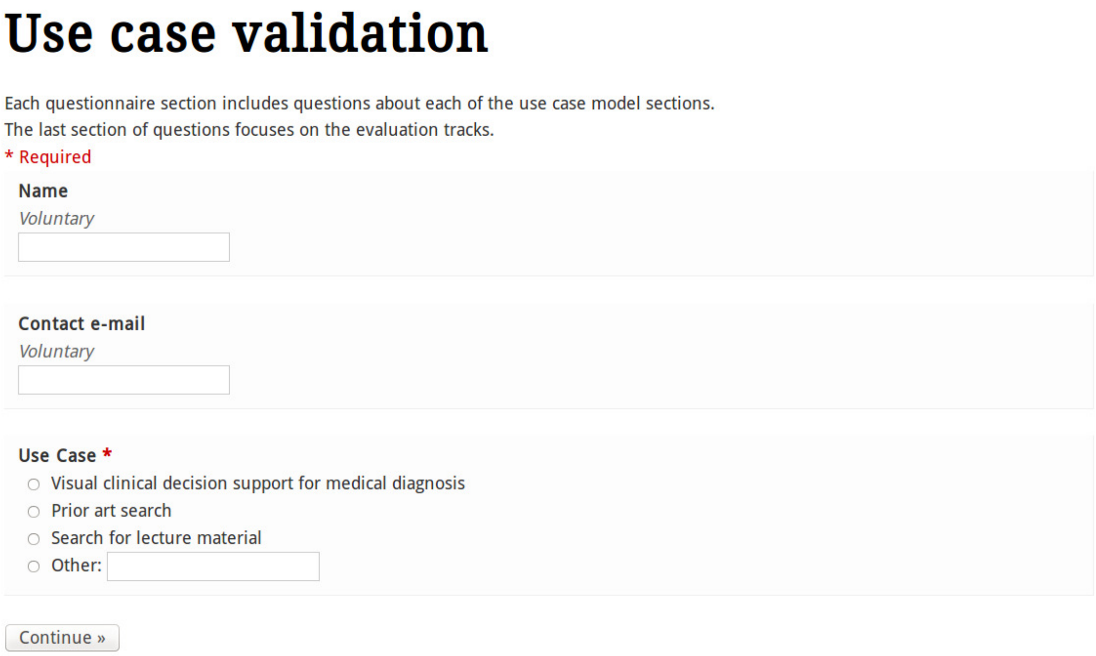
Appendix A

Medical Use Case Validation 2012

Some of the results from the questionnaire included in this appendix are presented in Chapter 3. The questionnaires are used to carry out the use case validation. In this appendix, the HTML form is first presented (see Section A.1). Afterwards, Section A.2 shows the anonymised results of the use case validation. For each question, the answers from the four experts interviewed are distinguished. In this appendix no analysis on the significance of the answers are reported. All importante responses are analised in Chapter 3.

A.1 Survey Questions

The following form is used to validate the visual clinical decision support use case.



Use case validation

Each questionnaire section includes questions about each of the use case model sections.
The last section of questions focuses on the evaluation tracks.

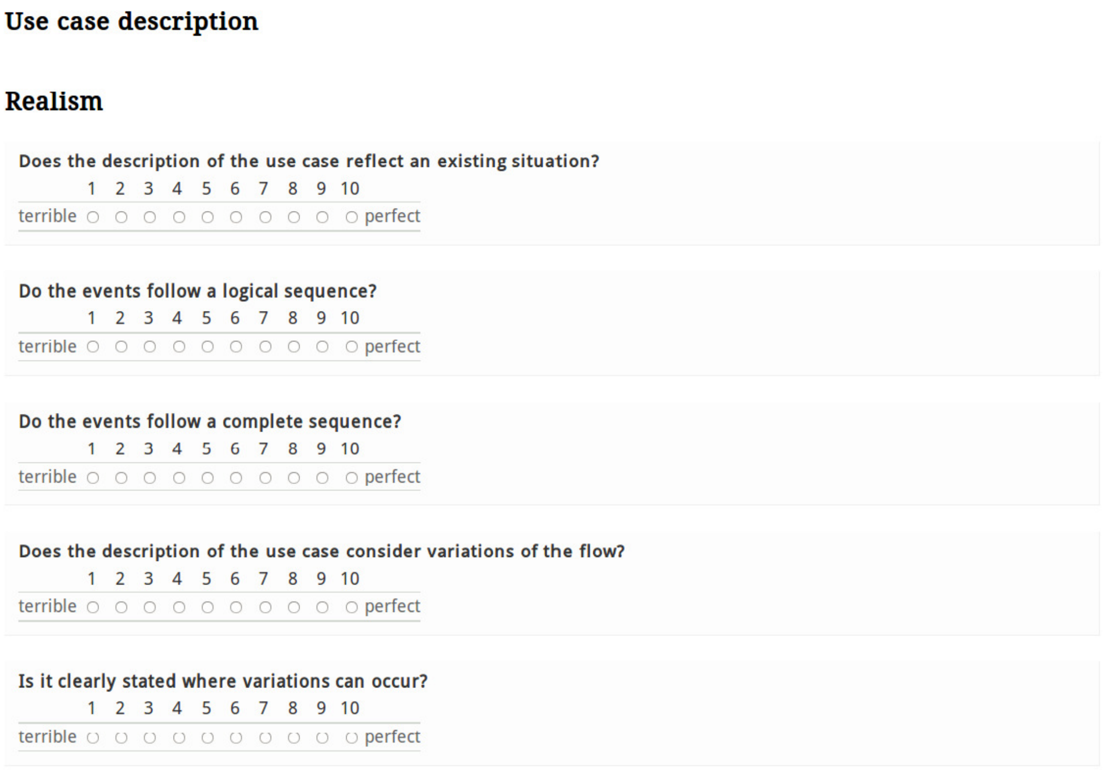
* Required

Name
Voluntary

Contact e-mail
Voluntary

Use Case *

○ Visual clinical decision support for medical diagnosis
○ Prior art search
○ Search for lecture material
○ Other:



Use case description

Realism

Does the description of the use case reflect an existing situation?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Do the events follow a logical sequence?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Do the events follow a complete sequence?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Does the description of the use case consider variations of the flow?

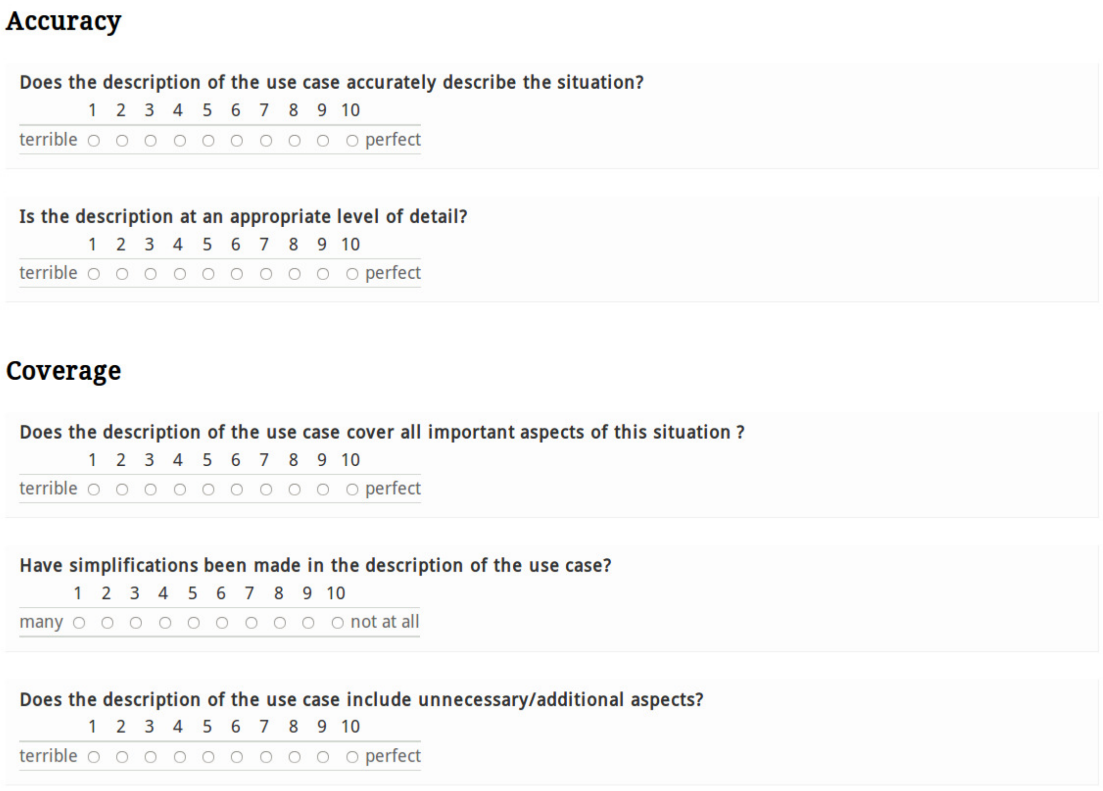
1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Is it clearly stated where variations can occur?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect



Accuracy

Does the description of the use case accurately describe the situation?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Is the description at an appropriate level of detail?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Coverage

Does the description of the use case cover all important aspects of this situation ?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Have simplifications been made in the description of the use case?

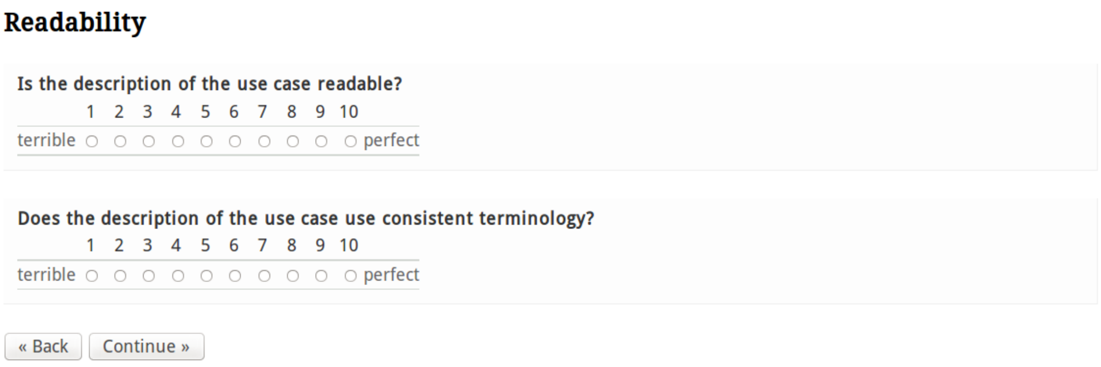
1 2 3 4 5 6 7 8 9 10

many ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ not at all

Does the description of the use case include unnecessary/additional aspects?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect



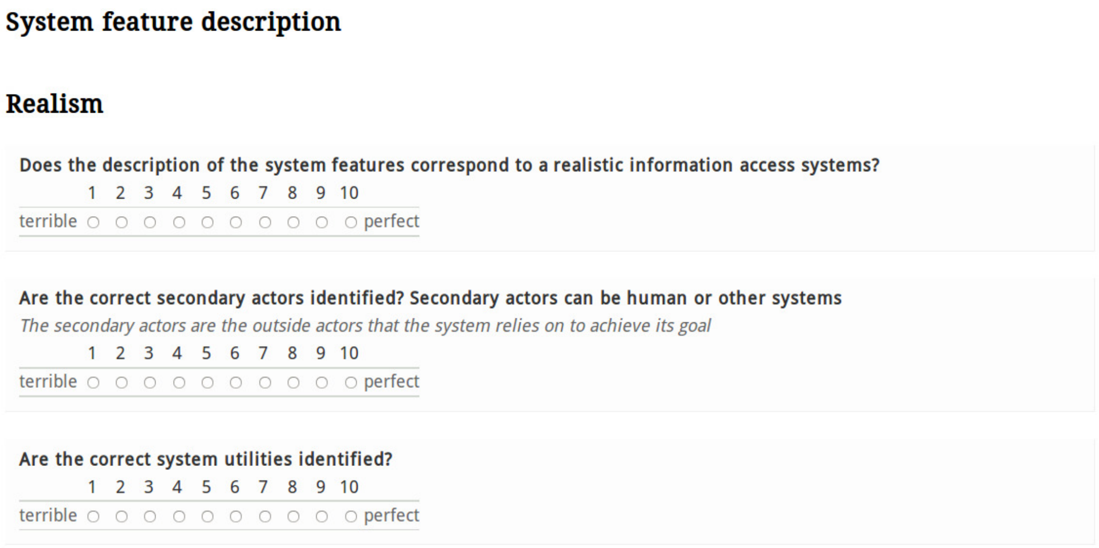
Readability

Is the description of the use case readable?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Does the description of the use case use consistent terminology?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect



System feature description

Realism

Does the description of the system features correspond to a realistic information access systems?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

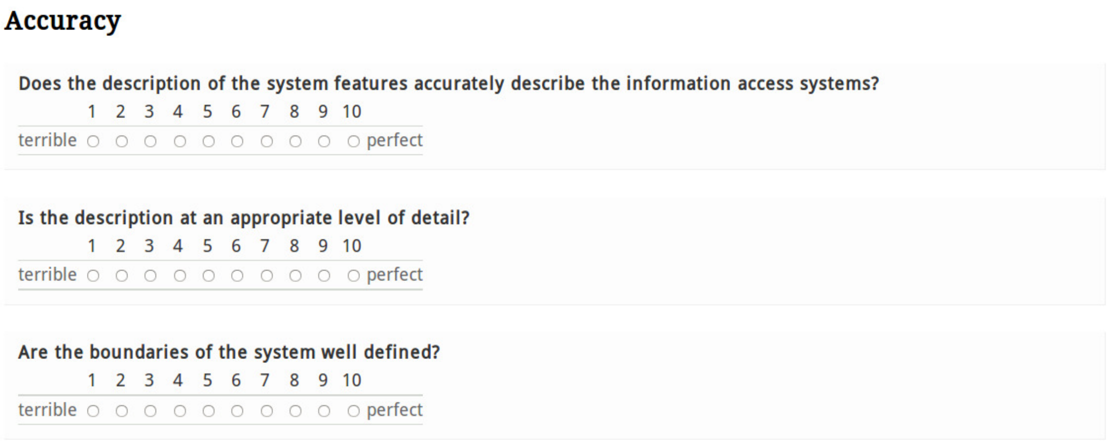
Are the correct secondary actors identified? Secondary actors can be human or other systems

The secondary actors are the outside actors that the system relies on to achieve its goal

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Are the correct system utilities identified?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect



Accuracy

Does the description of the system features accurately describe the information access systems?

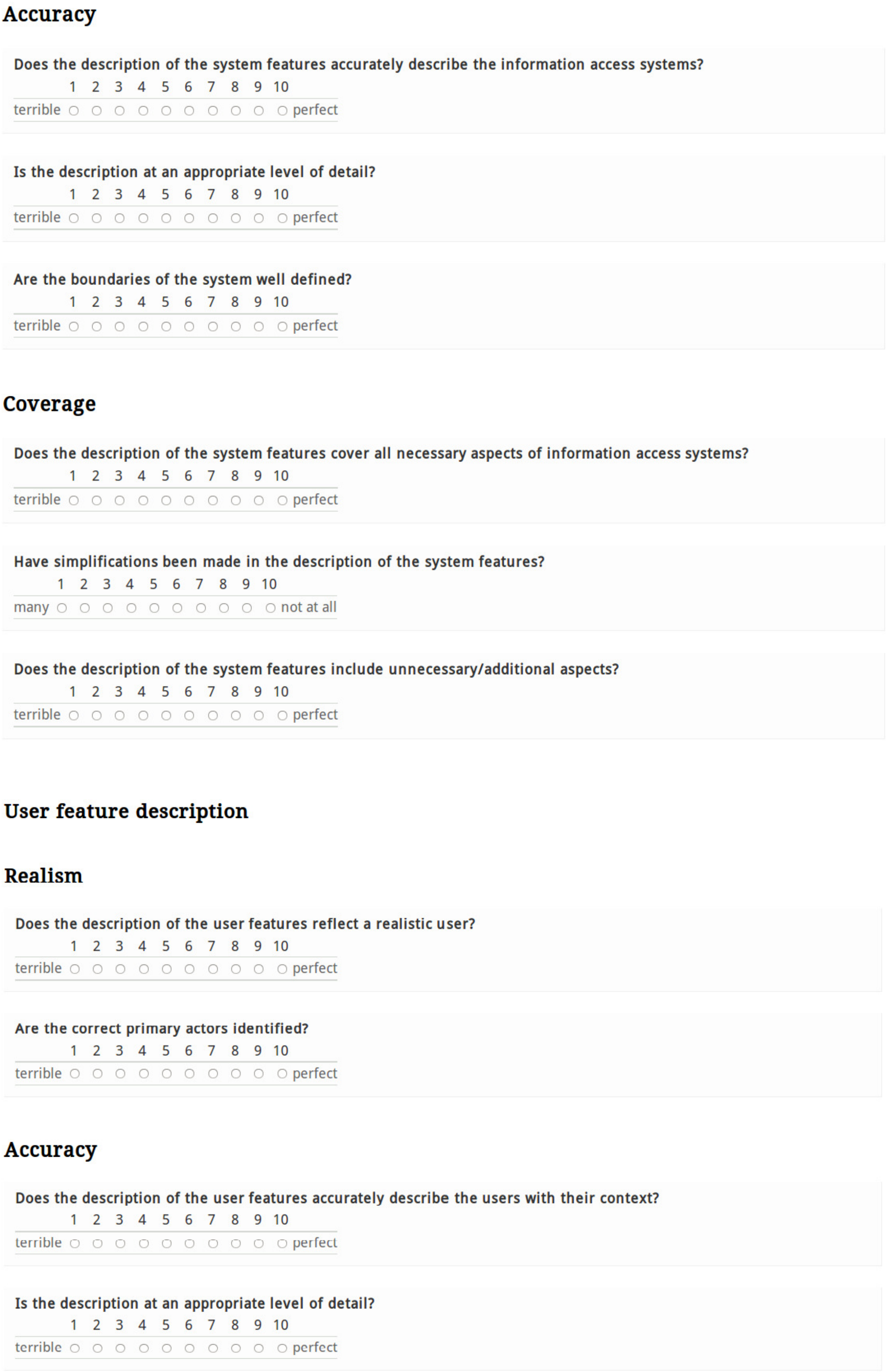
	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Is the description at an appropriate level of detail?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Are the boundaries of the system well defined?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect



Accuracy

Does the description of the system features accurately describe the information access systems?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Is the description at an appropriate level of detail?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Are the boundaries of the system well defined?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Coverage

Does the description of the system features cover all necessary aspects of information access systems?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Have simplifications been made in the description of the system features?

	1	2	3	4	5	6	7	8	9	10	
many	○	○	○	○	○	○	○	○	○	○	not at all

Does the description of the system features include unnecessary/additional aspects?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

User feature description

Realism

Does the description of the user features reflect a realistic user?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Are the correct primary actors identified?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

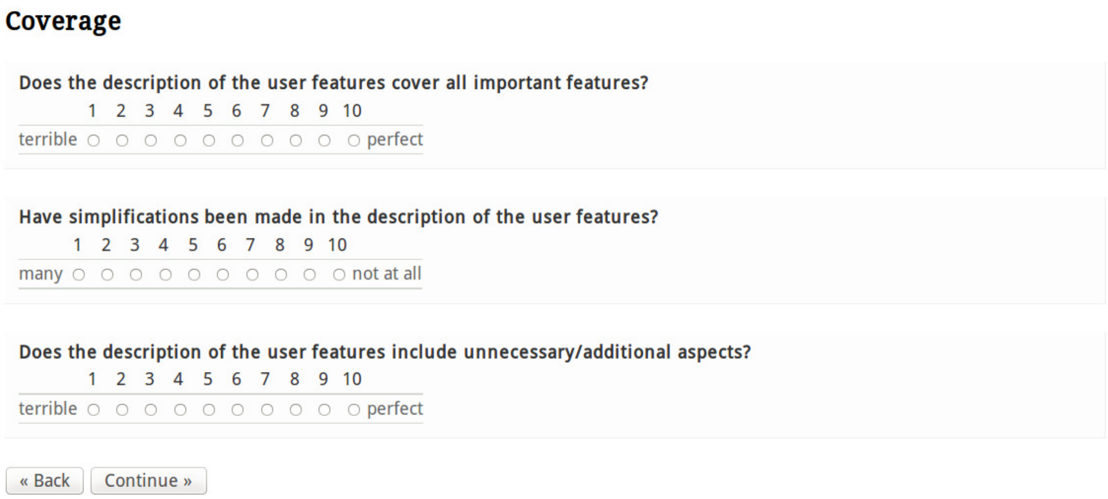
Accuracy

Does the description of the user features accurately describe the users with their context?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Is the description at an appropriate level of detail?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect



Coverage

Does the description of the user features cover all important features?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Have simplifications been made in the description of the user features?

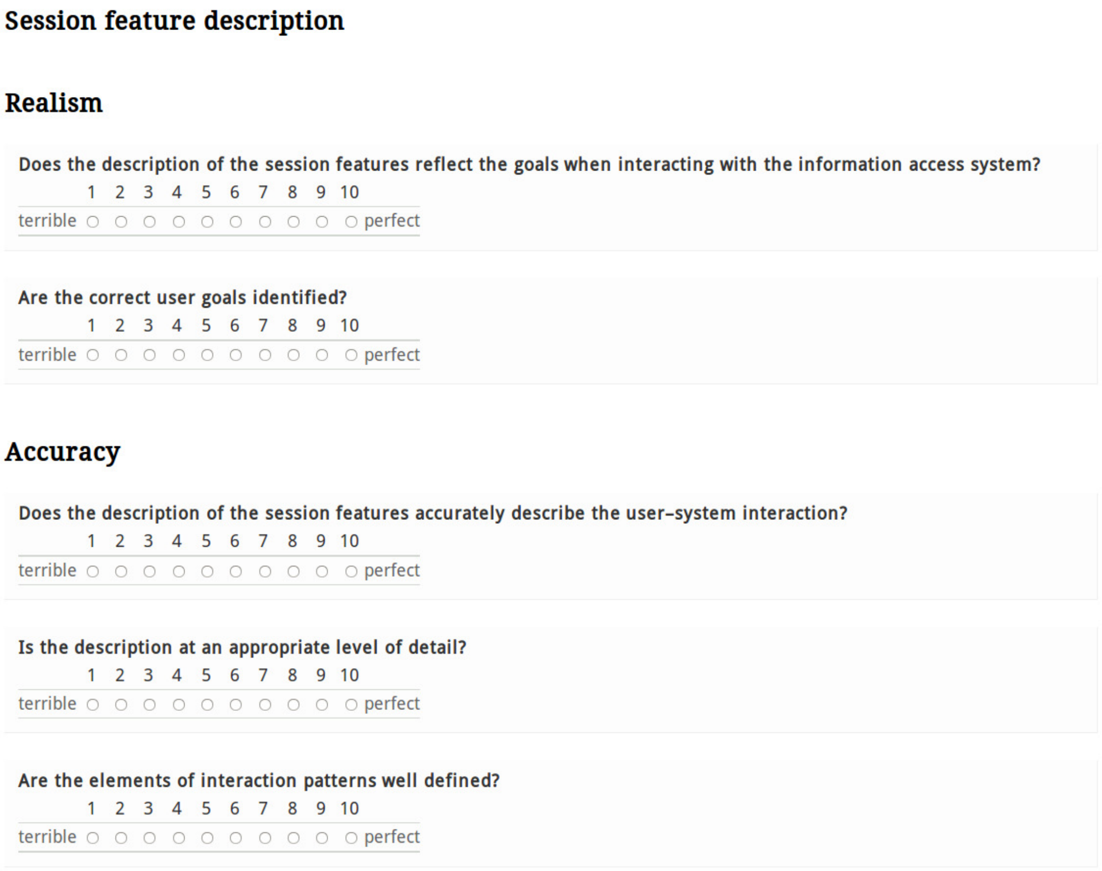
1 2 3 4 5 6 7 8 9 10

many ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ not at all

Does the description of the user features include unnecessary/additional aspects?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect



Session feature description

Realism

Does the description of the session features reflect the goals when interacting with the information access system?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Are the correct user goals identified?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Accuracy

Does the description of the session features accurately describe the user–system interaction?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Is the description at an appropriate level of detail?

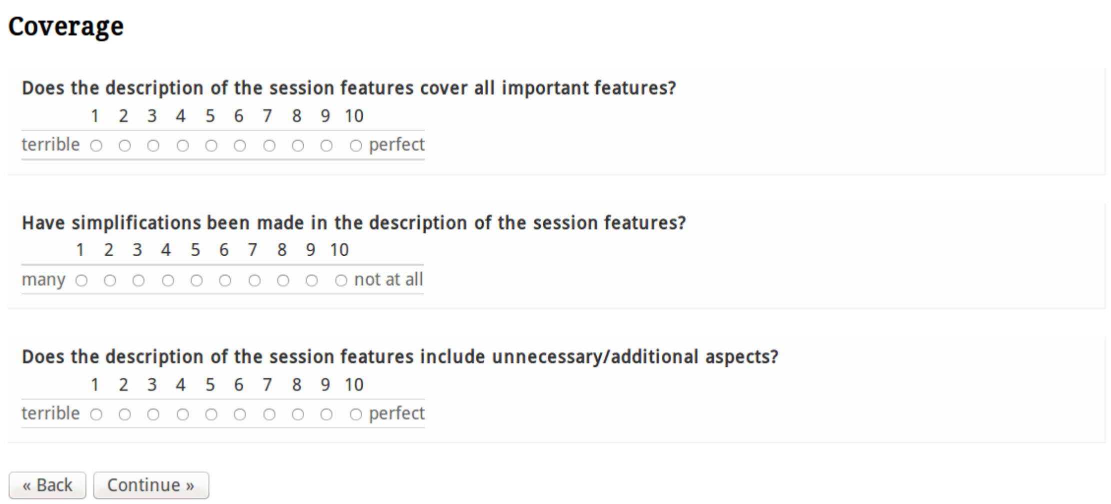
1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect

Are the elements of interaction patterns well defined?

1 2 3 4 5 6 7 8 9 10

terrible ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ perfect



Coverage

Does the description of the session features cover all important features?

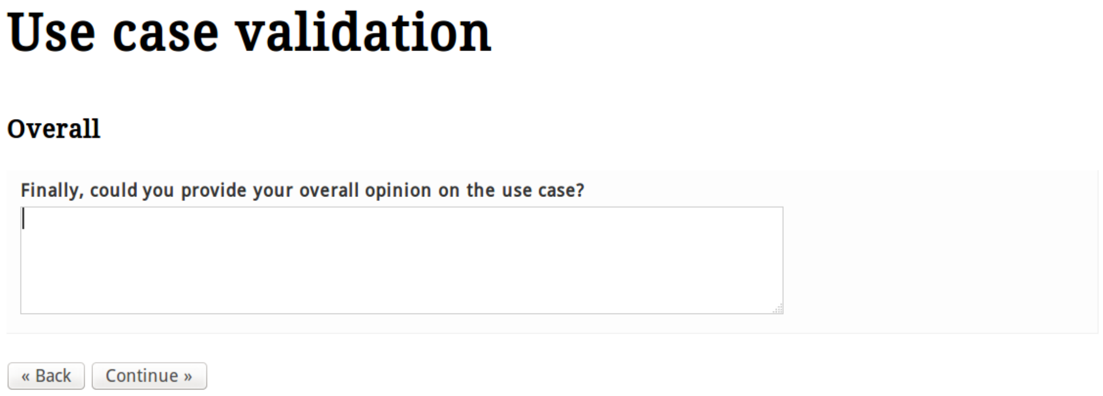
	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Have simplifications been made in the description of the session features?

	1	2	3	4	5	6	7	8	9	10	
many	○	○	○	○	○	○	○	○	○	○	not at all

Does the description of the session features include unnecessary/additional aspects?

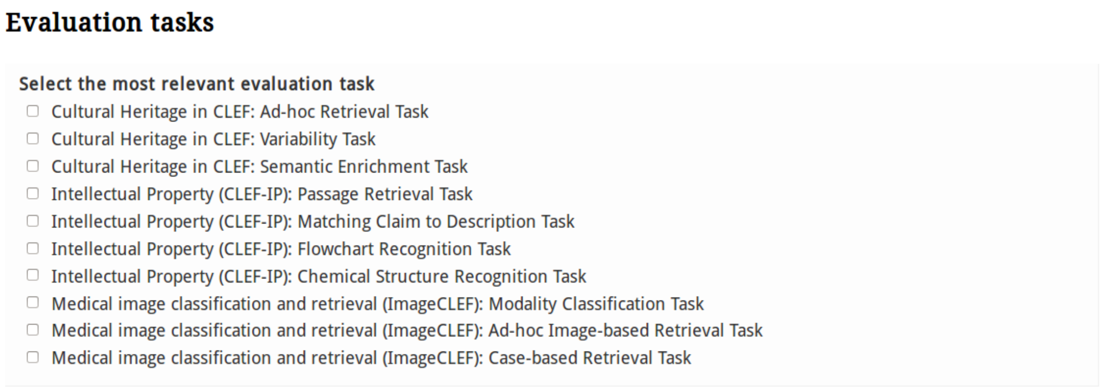
	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect



Use case validation

Overall

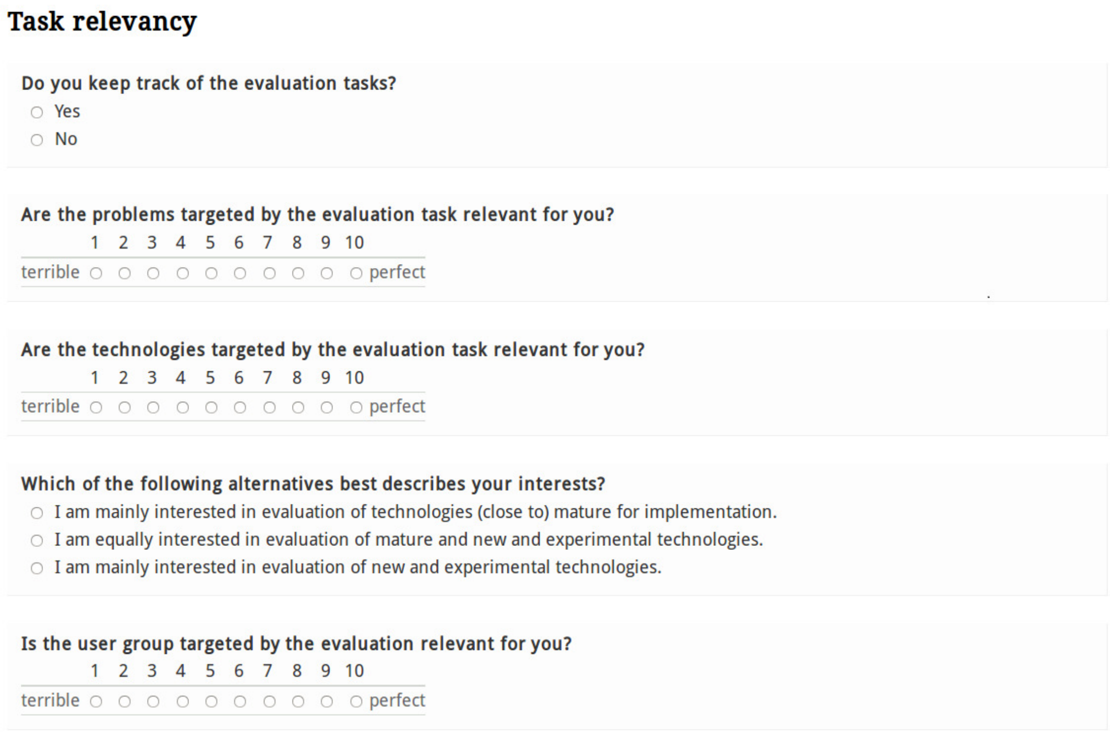
Finally, could you provide your overall opinion on the use case?



Evaluation tasks

Select the most relevant evaluation task

- □ Cultural Heritage in CLEF: Ad-hoc Retrieval Task
- □ Cultural Heritage in CLEF: Variability Task
- □ Cultural Heritage in CLEF: Semantic Enrichment Task
- □ Intellectual Property (CLEF-IP): Passage Retrieval Task
- □ Intellectual Property (CLEF-IP): Matching Claim to Description Task
- □ Intellectual Property (CLEF-IP): Flowchart Recognition Task
- □ Intellectual Property (CLEF-IP): Chemical Structure Recognition Task
- □ Medical image classification and retrieval (ImageCLEF): Modality Classification Task
- □ Medical image classification and retrieval (ImageCLEF): Ad-hoc Image-based Retrieval Task
- □ Medical image classification and retrieval (ImageCLEF): Case-based Retrieval Task



Task relevancy

Do you keep track of the evaluation tasks?

○ Yes
○ No

Are the problems targeted by the evaluation task relevant for you?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Are the technologies targeted by the evaluation task relevant for you?

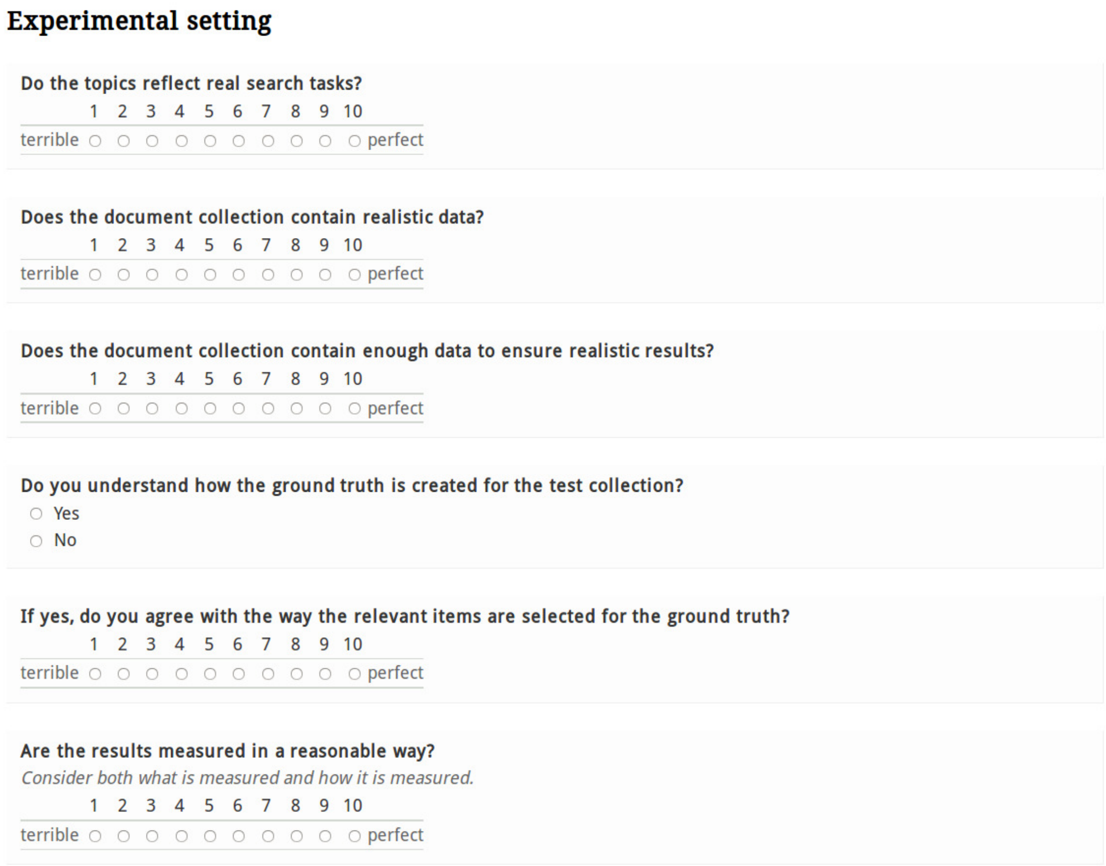
	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Which of the following alternatives best describes your interests?

○ I am mainly interested in evaluation of technologies (close to) mature for implementation.
○ I am equally interested in evaluation of mature and new and experimental technologies.
○ I am mainly interested in evaluation of new and experimental technologies.

Is the user group targeted by the evaluation relevant for you?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect



Experimental setting

Do the topics reflect real search tasks?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Does the document collection contain realistic data?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Does the document collection contain enough data to ensure realistic results?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Do you understand how the ground truth is created for the test collection?

○ Yes
○ No

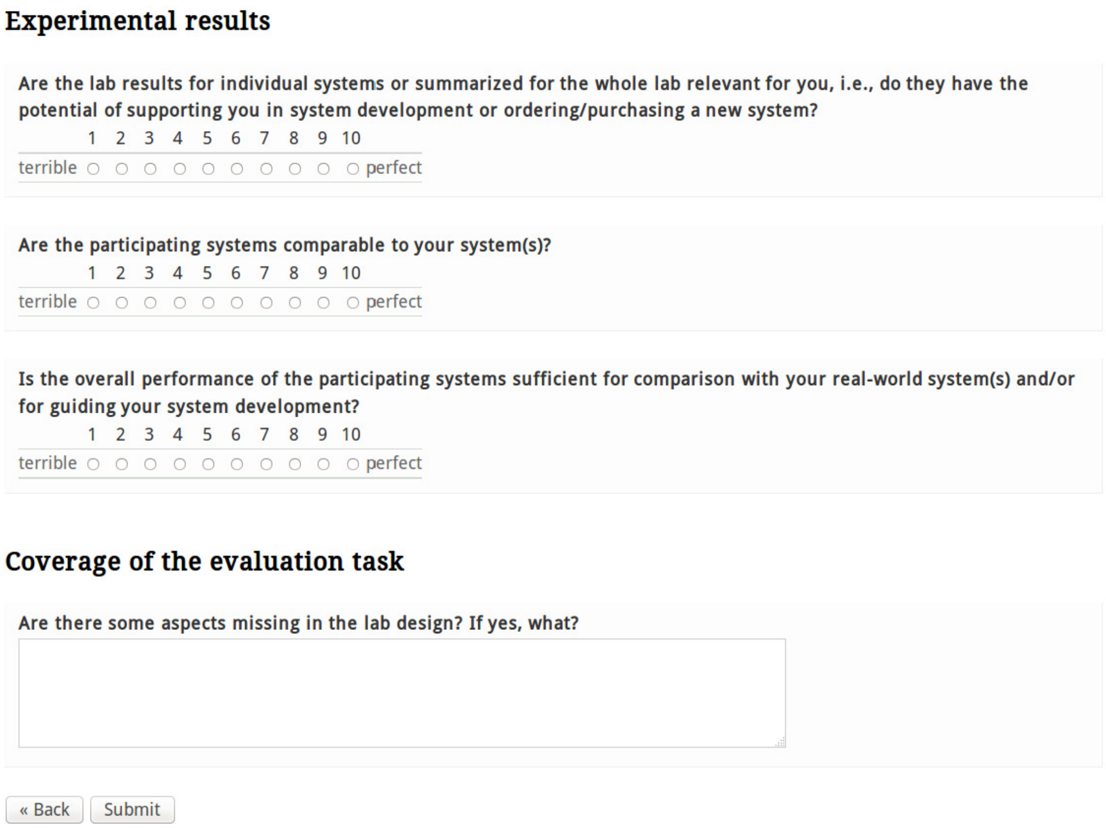
If yes, do you agree with the way the relevant items are selected for the ground truth?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Are the results measured in a reasonable way?

Consider both what is measured and how it is measured.

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect



Experimental results

Are the lab results for individual systems or summarized for the whole lab relevant for you, i.e., do they have the potential of supporting you in system development or ordering/purchasing a new system?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Are the participating systems comparable to your system(s)?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Is the overall performance of the participating systems sufficient for comparison with your real-world system(s) and/or for guiding your system development?

	1	2	3	4	5	6	7	8	9	10	
terrible	○	○	○	○	○	○	○	○	○	○	perfect

Coverage of the evaluation task

Are there some aspects missing in the lab design? If yes, what?

A.2 Survey Responses

Answers to the use case validation questionnaire can be found below.

Use Case Description

Realism

- Does the description of the use case reflect an existing situation?
  - *Subject 1*: 7;
  - *Subject 2*: 8;
  - *Subject 3*: 9;
  - *Subject 4*: 8.
- Do the events follow a logical sequence?
  - *Subject 1*: 7;
  - *Subject 2*: 8;
  - *Subject 3*: 9;
  - *Subject 4*: 8.

- Do the events follow a complete sequence?
  - *Subject 1*: 5;
  - *Subject 2*: 9;
  - *Subject 3*: 7;
  - *Subject 4*: 8.
- Does the description of the use case consider variations of the flow?
  - *Subject 1*: 6;
  - *Subject 2*: 3;
  - *Subject 3*: 4;
  - *Subject 4*: 7.
- Is it clearly stated where variations can occur?
  - *Subject 1*: 7;
  - *Subject 2*: 3;
  - *Subject 3*: 5;
  - *Subject 4*: 7.

Accuracy

- Does the description of the use case accurately describe the situation?
  - *Subject 1*: 7;
  - *Subject 2*: 9;
  - *Subject 3*: 7;
  - *Subject 4*: 8.
- Is the description at an appropriate level of detail?
  - *Subject 1*: 8;
  - *Subject 2*: 9;
  - *Subject 3*: 7;
  - *Subject 4*: 8.

Coverage

- Does the description of the use case cover all important aspects of this situation?
  - *Subject 1*: 7;
  - *Subject 2*: 7;
  - *Subject 3*: 8;
  - *Subject 4*: 8.
- Have simplifications been made in the description of the use case?
  - *Subject 1*: 6;
  - *Subject 2*: 7;
  - *Subject 3*: 8;
  - *Subject 4*: 8.
- Does the description of the use case include unnecessary/additional aspects?
  - *Subject 1*: No answer;
  - *Subject 2*: 6;
  - *Subject 3*: 10;
  - *Subject 4*: 9.

Readability

- Is the description of the use case readable?
  - *Subject 1*: 9;
  - *Subject 2*: 9;
  - *Subject 3*: 10;
  - *Subject 4*: 10.
- Does the description of the use case use consistent terminology?
  - *Subject 1*: 9;
  - *Subject 2*: 9;
  - *Subject 3*: 8;
  - *Subject 4*: 9.

System Feature Description

Realism

- Does the description of the system features correspond to a realistic information access systems?
  - *Subject 1*: 9;
  - *Subject 2*: 9;
  - *Subject 3*: 9;
  - *Subject 4*: 9.
- Are the correct secondary actors identified? Secondary actors can be human or other systems
  - *Subject 1*: 6;
  - *Subject 2*: 7;
  - *Subject 3*: 8;
  - *Subject 4*: 8.
- Are the correct system utilities identified?
  - *Subject 1*: 7;
  - *Subject 2*: 8;
  - *Subject 3*: 8;
  - *Subject 4*: 9.

Accuracy

- Does the description of the system features accurately describe the information access systems?
  - *Subject 1*: 9;
  - *Subject 2*: 7;
  - *Subject 3*: 7;
  - *Subject 4*: 9.
- Is the description at an appropriate level of detail?
  - *Subject 1*: 8;
  - *Subject 2*: 8;
  - *Subject 3*: 6;
  - *Subject 4*: 9.

- Are the boundaries of the system well defined?
  - *Subject 1*: 6;
  - *Subject 2*: 10;
  - *Subject 3*: 4;
  - *Subject 4*: 9.

Coverage

- Does the description of the system features cover all necessary aspects of information access systems?
  - *Subject 1*: 7;
  - *Subject 2*: 7;
  - *Subject 3*: 4;
  - *Subject 4*: 9.
- Have simplifications been made in the description of the system features?
  - *Subject 1*: 8;
  - *Subject 2*: 7;
  - *Subject 3*: 6;
  - *Subject 4*: 8.
- Does the description of the system features include unnecessary/additional aspects?
  - *Subject 1*: No answer;
  - *Subject 2*: 8;
  - *Subject 3*: 8;
  - *Subject 4*: 9.

User Feature Description

Realism

- Does the description of the user features reflect a realistic user?
  - *Subject 1*: 6;
  - *Subject 2*: 10;
  - *Subject 3*: 8;
  - *Subject 4*: 9.
- Are the correct primary actors identified?
  - *Subject 1*: 7;
  - *Subject 2*: 10;
  - *Subject 3*: 10;
  - *Subject 4*: 9.

Accuracy

- Does the description of the user features accurately describe the users with their context?
  - *Subject 1*: 7;
  - *Subject 2*: 9;
  - *Subject 3*: 7;
  - *Subject 4*: 9.
- Is the description at an appropriate level of detail?
  - *Subject 1*: 8;
  - *Subject 2*: 9;
  - *Subject 3*: 9;
  - *Subject 4*: 9.

Coverage

- Does the description of the user features cover all important features?
  - *Subject 1*: 6;
  - *Subject 2*: 9;
  - *Subject 3*: 9;
  - *Subject 4*: 8.
- Have simplifications been made in the description of the user features?
  - *Subject 1*: 6;
  - *Subject 2*: 4;
  - *Subject 3*: 10;
  - *Subject 4*: 9.
- Does the description of the user features include unnecessary/additional aspects?
  - *Subject 1*: No answer;
  - *Subject 2*: 8;
  - *Subject 3*: 9;
  - *Subject 4*: 9.

Session Feature Description

Realism

- Does the description of the session features reflect the goals when interacting with the information access system?
  - *Subject 1*: 9;
  - *Subject 2*: 9;
  - *Subject 3*: 10;
  - *Subject 4*: 9.
- Are the correct user goals identified?
  - *Subject 1*: 8;
  - *Subject 2*: 10;
  - *Subject 3*: 9;
  - *Subject 4*: 9.

Accuracy

- Does the description of the session features accurately describe the user–system interaction?
  - *Subject 1*: 8;
  - *Subject 2*: 4;
  - *Subject 3*: 8;
  - *Subject 4*: 9.
- Is the description at an appropriate level of detail?
  - *Subject 1*: 7;
  - *Subject 2*: 4;
  - *Subject 3*: 9;
  - *Subject 4*: 9.
- Are the elements of interaction patterns well defined?
  - *Subject 1*: 7;
  - *Subject 2*: 6;
  - *Subject 3*: 9;
  - *Subject 4*: 9.

Coverage

- Does the description of the session features cover all important features?
  - *Subject 1*: 7;
  - *Subject 2*: 4;
  - *Subject 3*: 5;
  - *Subject 4*: 9.
- Have simplifications been made in the description of the session features?
  - *Subject 1*: 7;
  - *Subject 2*: 4;
  - *Subject 3*: 9;
  - *Subject 4*: 9.
- Does the description of the session features include unnecessary/additional aspects?
  - *Subject 1*: No answer;
  - *Subject 2*: 8;
  - *Subject 3*: 9;
  - *Subject 4*: 10.

Overall

- Finally, could you provide your overall opinion on the use case?
  - *Subject 1*: It has interesting applications that could be expanded e.g. for educational purposes in Medicine schools. It may also be appropriate to include the administrative internal affairs of a hospital with the handling of a patient e.g. referrals and legal implications;
  - *Subject 2*: Quite complete description of the use case, but further information is required on the user-system interaction;
  - *Subject 3*: It is generally very good. There are very few points that are missing in each section.;
  - *Subject 4*: Very clear and easy to understand. Also, the practical examples that are given for certain terms (such as multi–modal information), as well as the illustration in the session feature section really help to ground things in reality.

Evaluation Task

- Select the most relevant evaluation task
  - *Subject 1*: Medical image classification and retrieval (ImageCLEF): Case-based Retrieval Task;
  - *Subject 2*: Medical image classification and retrieval (ImageCLEF): Case-based Retrieval Task;
  - *Subject 3*: Medical image classification and retrieval (ImageCLEF): Case-based Retrieval Task;
  - *Subject 4*: Medical image classification and retrieval (ImageCLEF): Case-based Retrieval Task.

Task Relevancy

- Do you keep track of the evaluation tasks?
  - *Subject 1*: No;
  - *Subject 2*: No;
  - *Subject 3*: Yes;
  - *Subject 4*: No.
- Are the problems targeted by the evaluation task relevant for you?
  - *Subject 1*: 10;
  - *Subject 2*: 10;
  - *Subject 3*: 9;
  - *Subject 4*: 8.
- Are the technologies targeted by the evaluation task relevant for you?
  - *Subject 1*: 10;
  - *Subject 2*: 10;
  - *Subject 3*: 10;
  - *Subject 4*: 9.
- Which of the following alternatives best describes your interests?
  - *Subject 1*: I am equally interested in evaluation of mature and new and experimental technologies;
  - *Subject 2*: I am equally interested in evaluation of mature and new and experimental technologies;
  - *Subject 3*: I am equally interested in evaluation of mature and new and experimental technologies;
  - *Subject 4*: I am mainly interested in evaluation of new and experimental technologies.

- Is the user group targeted by the evaluation relevant for you?
  - *Subject 1*: 10;
  - *Subject 2*: 10;
  - *Subject 3*: 10;
  - *Subject 4*: 8.

Experimental settings

- Do the query topics reflect real search tasks?
  - *Subject 1*: 6;
  - *Subject 2*: 10;
  - *Subject 3*: 8;
  - *Subject 4*: 8.
- Does the document collection contain realistic data?
  - *Subject 1*: 9;
  - *Subject 2*: 10;
  - *Subject 3*: 10;
  - *Subject 4*: 8.
- Does the document collection contain enough data to ensure realistic results?
  - *Subject 1*: 8;
  - *Subject 2*: 8;
  - *Subject 3*: 5;
  - *Subject 4*: 7.
- Do you understand how the ground truth is created for the test collection?
  - *Subject 1*: No;
  - *Subject 2*: Yes;
  - *Subject 3*: Yes;
  - *Subject 4*: Yes.
- If yes, do you agree with the way the relevant items are selected for the ground truth?
  - *Subject 1*: No answer;
  - *Subject 2*: 7;
  - *Subject 3*: 9;
  - *Subject 4*: 8.

- Are the results measured in a reasonable way?
  - *Subject 1*: 6;
  - *Subject 2*: 7;
  - *Subject 3*: 5;
  - *Subject 4*: 8.

Experimental Results

- Are the lab results for individual systems or summarised for the whole lab relevant for you, i.e., do they have the potential of supporting you in system development or ordering/purchasing a new system?
  - *Subject 1*: 9;
  - *Subject 2*: 9;
  - *Subject 3*: 8;
  - *Subject 4*: 8.
- Are the participating systems comparable to your system(s)?
  - *Subject 1*: 8;
  - *Subject 2*: 7;
  - *Subject 3*: 9;
  - *Subject 4*: 8.
- Is the overall performance of the participating systems sufficient for comparison with your real-world system(s) and/or for guiding your system development?
  - *Subject 1*: 8;
  - *Subject 2*: 6;
  - *Subject 3*: 5;
  - *Subject 4*: 8.

Coverage of the Evaluation Task

- Are there some aspects missing in the lab design? If yes, what?
  - *Subject 1*: Measurement of clinical accuracy in the search. Follow up of the cases for system optimization;
  - *Subject 2*: No answer;
  - *Subject 3*: There is no measurement of query times and index sizes. This makes them difficult to evaluate for realistic scenarios;
  - *Subject 4*: No answer.

Appendix B

ImageCLEFmed Questionnaires

In Chapter 4, some of the results from the questionnaire were related to the results from the ImageCLEFmed lab. The questionnaires were sent to the ImageCLEF organisers. In this appendix, most results from the questionnaire filled by ImageCLEFmed organisers are presented. For each variable, the results from years 2011 to 2013 are distinguished. In this appendix no analysis on the significance of the answers are reported. All important responses are reported in Chapter 4.

Participation in the ImageCLEFmed 2011–2013

- Number of years the task is part of CLEF:
  - *2011*: 8;
  - *2012*: 9;
  - *2013*: 10.
- Registrations:
  - *2011*: 60;
  - *2012*: 85;
  - *2013*: 63.
- Participations:
  - *2011*: 17;
  - *2012*: 17;
  - *2013*: 10.
- Return participations:
  - *2011*: 5;
  - *2012*: 8;
  - *2013*: 4.

- Submissions allowed per participant:
  - *2011*: 10;
  - *2012*: 10 per subtask;
  - *2013*: 10 per subtask.
- Total submissions:
  - *2011*: 207;
  - *2012*: 202;
  - *2013*: 166.
- Submissions system:
  - *2011*: ImageCLEF;
  - *2012*: ImageCLEF;
  - *2013*: ImageCLEF.

Main outcomes of the ImageCLEFmed 2011–2013

- Task type:
  - *2011*: Retrieval and classification;
  - *2012*: Retrieval and classification;
  - *2013*: Retrieval, classification and annotation.
- Main differences/advances from previous year:
  - *2011*: Larger, totally different dataset;
  - *2012*: Larger dataset; improved hierarchy in the classification task;
  - *2012*: Larger number of retrieval query topics; more compound images in the classification task; compound figure separation new task.
- Main problems:
  - *2011*: Many groups do not submit runs;
  - *2012*: Many groups do not submit runs;
  - *2013*: Slightly less participants.

Main trends in the approaches employed by the participants of the ImageCLEFmed 2011–2013 and the main experimental outcomes

- Main trends (among the participants' approaches):
  - *2011*: Query expansion was often successful; mapping to MeSH terms; using the MeSH Hierarchy;
  - *2012*: Lucene; concept–based approaches; used of multiple visual features;
  - *2013*: ImageCLEFmed 2012 database was used to optimise parameters.

- Main experimental outcomes (among the participants' approaches):
  - *2011*: Multi–modal approaches are often best; visual has good early precision; fusion is hard to do;
  - *2012*: Visual, textual or mixed runs perform differently based on the subtasks; same or similar descriptors differ on results; expansion of the training set and the used of multiple visual features were successful;
  - *2013*: Visual techniques perform better for compound figure separation.

Collections used in the ImageCLEFmed 2011–2013

- Collection:
  - *2011*: PMC;
  - *2012*: PMC;
  - *2013*: PMC.
- Number of documents:
  - *2011*: 230,000 images;
  - *2012*: Over 300,000 images of 75,000 articles;
  - *2013*: Over 300,000 images of 75,000 articles.
- Size:
  - *2011*: 16 GB;
  - *2012*: 18 GB;
  - *2013*: 18 GB.
- Languages:
  - *2011*: Mainly English;
  - *2012*: Mainly English;
  - *2013*: Mainly English.
- Collection created for the lab:
  - *2011*: Yes;
  - *2012*: Yes;
  - *2013*: Yes.
- Number of years collection used in lab:
  - *2011*: 1;
  - *2012*: 1;
  - *2013*: 2.

- Parts of the collection used in previous years of the lab:
  - *2011*: None;
  - *2012*: Yes;
  - *2013*: Yes.

Topics used in the tasks of ImageCLEFmed 2011–2013

- What constitutes a topic for this task?
  - Classification:
    - *2011*: An image from the medical literature;
    - *2012*: An image from the medical literature;
    - *2013*: An image from the medical literature.
  - Retrieval:
    - *2011*:A multimedia query that consists of a textual part, the query title in three languages, and a visual part, one or several example images;
    - *2012*: An information need in four languages and images;
    - *2013*: An information need in four languages and images.
  - Separation:
    - *2011*: No applicable;
    - *2012*: No applicable;
    - *2013*: A compound image.
- Topics
  - Classification:
    - *2011*: 1,000 images 18 classes;
    - *2012*: 1,000 images 31 classes;
    - *2013*: 2,582 images 31 classes.
  - Retrieval:
    - *2011*: 30;
    - *2012*: 30 image–based query topics and 10 case–based query topics;
    - *2013*: 35 image–based query topics and 35 case–based query topics.
  - Separation:
    - *2011*: No applicable;
    - *2012*: No applicable;
    - *2013*: 1,429.
- Languages
  - Classification:
    - *2011*: Mainly English;
    - *2012*: Mainly English;
    - *2013*: Mainly English.

  - Retrieval:
    - *2011*: English, French, German;
    - *2012*: English, French, German, Spanish;
    - *2013*: English, French, German, Spanish.
  - Separation:
    - *2011*: No applicable;
    - *2012*: No applicable;
    - *2013*: Mainly English.

Ground truth generation for the tasks in ImageCLEFmed 2011–2013

- How many documents were assessed?:
  - *2011*: Retrieval: $\sim 30,000$ (pooling: top 50); Classification: 2,000 images;
  - *2012*: Retrieval: $\sim 30,000$ (pooling: top 50); Classification: 2,000 images;
  - *2013*: Retrieval: $\sim 30,000$ (pooling: top 50); Classification: 3,753 images.
- How many assessors were employed?:
  - *2011*: ~ 15;
  - *2012*: Classification: 18; Retrieval: 11;
  - *2013*: Classification: ~ 10; Retrieval: ~ 15.
- Who were the assessors?:
  - *2011*: Medical doctors in a medical information program in Portland OR, USA;
  - *2012*: Classification: Researchers in the medical imaging; Retrieval: Medical doctors in a medical information program in Portland OR, USA ;
  - *2013*: Classification: Researchers in the medical imaging; Retrieval: Medical doctors in a medical information program in Portland OR, USA.
- How much time did the assessors spend?:
  - *2011*: ~ 250 hours;
  - *2012*: Classification: 96 hours; Retrieval: 235 hours;
  - *2013*: Classification: ~ 180 hours; Retrieval: ~ 250 hours.

Nomenclature

α, β, γ	Factors for Rocchio weighting
$\mathcal{P}$	Percentile
μ_i	Expected value of x_i
ω	Factors for linear combination weighting
σ_{ij}	Synonymy value of the words w_i and w_j
$\vec{im}$	Image represented as a vector
$\vec{q}$	Query represented as a vector
A	Set of articles
a	Article
c_j	Candidate
$C^f_{correct}$	Set of correct candidates
E	Expanded dataset
$E(X)$	Expectation of variable X
f	Figure
f_i	Subfigure
I	Set of images
i, j, k, l, m, n	Counters for diverse purposes
im	Image
K^f_C	Set of candidates for the figure f
K^f_{GT}	Ground truth for the figure f
L	Set of labels
$MAP(T)$	MAP score obtained using text retrieval techniques
$MAP(V)$	MAP score obtained using visual retrieval techniques

N, M	Number of elements in a set
R	Set of rankings
r	Rank belonging to the set R
$S(x)$	Score assigned to element x
t_{ij}	Synonymy value of two MeSH terms
tv_{ij}	Synonymy value of a MeSH term and a visual feature
V	Covariance matrix
v_{ij}	Synonymy value of two visual features
W	Set of words
w	Visual word
Z	Set of visual topics
z	Visual topic

Glossary

AJAX Asynchronous JavaScript And XML

AMIA American Medical Informatics Association

A–NN Approximate Nearest Neighbour

AP Average Precision

BoC Bag of Colours

BoVW Bag of Visual Words

CBIR Content–Based Image Retrieval

CEDD Colour and Edge Directivity Descriptor

CLEF Conference and Labs of the Evaluation Forum

CORI Clinical Outcomes Research Institute

CT Computer Tomography

DFR Divergence from Randomness

E2LSH Euclidean Locally Sensitive Hashing

FCH Fuzzy Colour Histogram

FCTH Fuzzy Colour and Texture Histogram

FIRE Forum for Information Retrieval Evaluation

FP7 European Seventh Framework Programme

GMAP Geometric Mean Average Precison

GP Genetic Programming

HON Health On the Net

HTML5 HyperText Markup Language 5

iCLEF Interactive CLEF

ICPR International Conference for Pattern Recognition

ImageCLEF Cross–Language Retrieval in Image Collections

INEX INitiative for the Evaluation of XML retrieval

IR Information Retrieval

IRMA Image Retrieval in Medical Applications

JSON JavaScript Object Notation

k–**NN** *k*–Nearest Neighbours

LBP Local Binary Patterns

LSA Latent Semantic Analysis

MAP Mean Average Precision

MeSH Medical Subject Headings

MIR Mallinckrodt Institute of Radiology

MRI Magnetic Resonance Imaging

NCBI National Center for Biotechnology Information

NL Natural Language

NLM National Library of Medicine

NoE Network of Excellence

NTCIR NII Testbeds and Community for Information access Research

OHSU Oregon Health and Science University

ParaDISE Parallel Distributed Image Search Engine

PEIR Pathology Education Instructional Resource

PET Positron Emission Tomography

PLSA Probabilistic Latent Semantic Analysis

PMC PubMed Central

Pn Precision n

PoP Publish or Perish

PROMISE Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

REST REpresentational State Transfer

RRF Reciprocal rank fusion

RSNA Radiological Society of North America

SIFT Scale Invariant Feature Transform

SVM Support Vector Machine

tf/idf Term frequency–Inverse document frequency

TREC Text REtrieval Conference

TRECVid TREC Video Retrieval Evaluation

UML Unified Modelling Language

UMLS Unified Medical Language System

Visceral Visual Concept Extraction Challenge in Radiology

List of Figures

List of Tables

Bibliography

[1] Collins: English dictionary. `http://www.collinsdictionary.com/`. Accessed: 2014-12-14.

[2] WHSL medical subject headings for PubMed searching: Medical subject headings (MeSH). `http://libguides.wits.ac.za/whsl-mesh`. Accessed: 2014-12-14.

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

[4] A. Aamodt and E. Plaza. Case–based reasoning: Foundational issues, methodological variations, and systems approaches. *AIC*, 7(1):39–59, 1994.

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

[6] A. Andony and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *47th Annual IEEE Symposium on Foundations of Computer Science, 2006. FOCS'06*, pages 459–468, 2006.

[7] M. Angelini, N. Ferro, B. Larsen, H. Müller, G. Santucci, G. Silvello, and T. Tsikrika. Measuring and analyzing the scholarly impact of experimental evaluation initiatives. In *Italian Research Conference on Digital Libraries*, 2014.

[8] A. Arampatzis, K. Zagoris, and S. A. Chatzichristofis. Fusion vs. two–stage for multimodal retrieval. In *Advances in Information Retrieval*, pages 759–762. Springer, 2011.

[9] J. Arias, J. Martinez-Gomez, J. A. Gamez, A. García Seco de Herrera, and H. Müller. On the application of discrete bayesian networks for modality classification of medical images. *Computer Vision and Image Understanding*, Submitted.

[10] N. Aswani, T. Beckers, E. Birngruber, C. Boyer, A. Burner, J. BystroN, K. Choukri, S. Cruchet, H. Cunningham, J. Dedek, L. Dolamic, R. Donner, S. Dungs, I. Eggel, A. Foncubierta-Rodríguez, N. Fuhr, A. Funk, A. García Seco de Herrera, A. Gaudinat, G. Georgiev, J. Gobeill, L. Goeuriot, P. Gómez, M. Greenwood, M. Gschwandtner, A. Hanbury, J. Hajic, J. Hlaváčová, M. Holzer, G. Jones, B. Jordan, M. Jordan, K. Kaderk, F. Kainberger, L. Kelly, S. Mriewel, M. Kritz, G. Langs, N. Lawson, D. Markonis, I. Martinez, V. Momtchev, A. Masselot, H. Mazo, H. Müller, P. Pecina, K. Pentchev, D. Peychev, N. Pletneva, D. Pottecherc, A. Roberts, P. Ruch,

M. Samwald, P. Schneller, V. Stefanov, M. A. Tinte, Z. Uresová, A. Vargas, and D. Vishnyakova. Khresmoi – multimodal multilingual medical information search. In *Proceedings of the 24th International Conference of the European Federation for Medical Informatics*, 2012.

[11] N. Aswani, T. Beckers, E. Birngruber, C. Boyer, A. Burner, J. Bystro, K. Choukri, S. Cruchet, H. Cunningham, J. Ddek, L. Dolamic, R. Donner, S. Dungs, I. Eggel, A. Foncubierta-Rodríguez, N. Fuhr, A. Funk, A. García Seco de Herrera, A. Gaudinat, G. Georgiev, J. Gobeill, L. Goeuriot, P. Gómez, M. Greenwood, M. Gschwandtner, A. Hanbury, J. Hajič, J. Hlaváčová, M. Holzer, G. Jones, B. Jordan, M. Jordan, K. Kaderk, F. Kainberger, L. Kelly, S. Kriewel, M. Kritz, G. Langs, N. Lawson, D. Markonis, I. Martinez, V. Momtchev, A. Masselot, H. Mazo, H. Müller, J. a. Palotti, P. Pecina, K. Pentchev, D. Peychev, N. Pletneva, D. Pottecherc, A. Roberts, P. Ruch, A. Sachs, M. Samwald, P. Schneller, V. Stefanov, M. A. Tinte, Z. Urešová, A. Vargas, and D. Vishnyakova. Khresmoi – a multilingual semantic search of medical text and images. In *MedInfo 2013*, 2013.

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

[13] J. Bai and S. Shi. Estimating high dimensional covariance matrices and its applications. *Annals of Economics and Finance*, 12(2):199–215, 2011.

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

[15] J. Bar-Ilan. Which h–index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2):257–271, 2008.

[16] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[17] S. Bedrick, S. Radhouani, and J. Kalpathy-Cramer. *Improving Early Precision in the ImageCLEF Medical Retrieval Task*, volume 32 of *The Information Retrieval Series*, pages 397–413. Springer Berlin Heidelberg, 2010.

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

[19] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3):431–448, 1995.

[20] S. T. Bhairnallykar and V. B. Gaikwad. Content based medical image retrieval with SVM classification and relevance feedback. *International Journal of Applied Information Systems*, 2013.

[21] A. Bhalerao and C. Reyes-Aldasoro. Volumetric texture description and discriminant feature selection for MRI. In R. Moreno-Díaz and F. Pichler, editors, *Computer Aided Systems Theory - EUROCAST 2003*, volume 2809 of *Lecture Notes in Computer Science (LNCS)*, pages 573–584. Springer Berlin/Heidelberg, 2003.

[22] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134. ACM, 2003.

[23] S. Boughorbel, J.-P. Tarel, and N. Boujemaa. Generalized histogram intersection kernel for image recognition. In *IEEE International Conference on Image Processing*, volume 3, pages III–161. IEEE, 2005.

[24] H. C. Bow, J. R. Dattilo, A. M. Jonas, and C. U. Lehmann. A crowdsourcing model for creating preclinical medical education study tools. *Academic Medicine*, 88(6):766–770, 2013.

[25] G. Brajnik, S. Mizzaro, and C. Tasso. Evaluating user interfaces to information retrieval systems: A case study on user support. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 128–136. ACM, 1996.

[26] M. Braschler and C. Peters. The CLEF campaigns: Evaluation of cross–language information retrieval systems. *CEPIS UPGRADE III*, 3:78–81, 2002.

[27] A. Broder. A taxonomy of web search. In *ACM SIGIR forum*, volume 36, pages 3–10. ACM, 2002.

[28] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

[29] C. Buckley. *The SMART Project at TREC*, pages 301–320. Springer, 2005.

[30] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information retrieval*, 10(6):491–508, 2007.

[31] C. Buckley, G. Salton, and J. Allan. The SMART information retrieval project. In *Proceedings of the workshop on Human Language Technology*, pages 392–392. Association for Computational Linguistics, 1993.

[32] C. Buckley, A. Singhal, M. Mitra, and G. Salton. New retrieval approaches using SMART: TREC 4. In *Proceedings of the Fourth Text REtrieval Conference (TREC–4)*, pages 25–48, 1995.

[33] C. Buckley and E. M. Voorhees. *Retrieval System Evaluation*, pages 53–75. Springer, 2005.

[34] G. J. Burghouts and J.-M. Geusebroek. Performance evaluation of local colour invariants. *Compututer Vision and Image Understanding*, 113(1):48–62, 2009.

[35] J. Cancela Gónzalez, V. Ferro Llanos, M. d. P. García Jorge, A. García Seco de Herrera, J. M. Gómez Cordero, and I. Ibarz Gabardós. Modelado de sistemas de telemedicina. Technical report, Technical University of Madrid, 2009.

[36] J. Cano, J.-C. Pérez-Cortes, J. Arlandis, and R. Llobet. Training set expansion in handwritten character recognition. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 548–556. Springer, 2002.

[37] L. Cao, Y.-C. Chang, N. Codella, and M. Merler. IBM T.J. Watson research center, multimedia analytics: Modality classification and case–based retrieval task of ImageCLEF2012. In *Working Notes of CLEF 2012*, 2012.

[38] Y. Cao, Y. Li, H. Müller, C. E. Kahn Jr., and E. Munson. Multi–modal medical image retrieval. In *SPIE Medical Imaging*, 2011.

[39] A. Castellanos, J. Benavent, X. Benavent, and A. García-Serrano. Using visual concept features in a multimodal retrieval system for the medical collection at ImageCLEF2012. In *Working Notes of CLEF 2012*, 2012.

[40] R. Chakravarti and X. Meng. A study of color histogram based image retrieval. In *Sixth International Conference on Information Technology:New Generations ITNG*, pages 1323–1328, 2009.

[41] O. Chapelle, B. Schölkopf, A. Zien, et al. *Semi–Supervised Learning*, volume 2. MIT press Cambridge, 2006.

[42] S. A. Chatzichristofis and Y. S. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Lecture notes in Computer Sciences*, volume 5008, pages 312–322, 2008.

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

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

[45] C. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. Technical report, ASLIB Cranfield Research Project, Cranfield, 1966.

[46] P. Clough, H. Müller, T. Deselaers, M. Grubinger, T. M. Lehmann, J. Jensen, and W. Hersh. The CLEF 2005 cross–language image retrieval track. In *Cross Language Evaluation Forum (CLEF 2005)*, Lecture Notes in Computer Science (LNCS), pages 535–557. Springer, September 2006.

[47] P. Clough, H. Müller, and M. Sanderson. The CLEF 2004 cross–language image retrieval track. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 597–613, Bath, UK, 2005. Springer.

[48] P. Clough, H. Müller, and M. Sanderson. *Seven Years of Image Retrieval Evaluation*, pages 3–18. Springer, 2010.

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

[50] P. Clough, M. Sanderson, and H. Müller. The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In *The Challenge of Image and Video Retrieval (CIVR 2004)*, volume 3115 of *Lecture Notes in Computer Science (LNCS)*, pages 243–251. Springer, jul 2004.

[51] A. Cockburn. *Agile Software Development*. Addison–Wesley, 2002.

[52] J. Collins and K. Okada. A comparative study of similarity measures for content–based medical image retrieval. In *Working Notes of CLEF 2012*, 2012.

[53] N. R. Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 41(D1):D8–D20, 2013.

[54] G. V. Cormack, C. L. A. Clarke, and S. Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, New York, NY, USA, 2009. ACM.

[55] C. E. Crangle, A. Zbyslaw, J. M. Cherry, and E. L. Hong. Concept extraction and synonymy management for biomedical information retrieval. In *The thirteenth Text REtrieval Conference (TREC 2004)*, 2004.

[56] G. Csurka, S. Clinchant, and G. Jacquet. Medical image modality classification and retrieval. In *9th International Workshop on Content–Based Multimedia Indexing*, pages 193–198. IEEE, 2011.

[57] G. Csurka, S. Clinchant, and G. Jacquet. XRCE's participation at medical image modality classification and ad–hoc retrieval task of ImageCLEFmed 2011. In *Working Notes of CLEF 2011*, 2011.

[58] R. Datta, W. Ge, J. Li, and J. Z. Wang. Toward bridging the annotation–retrieval gap in image search. *Advances in Multimedia Computing*, 14(3):24–35, 2007.

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

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

[61] A. Depeursinge and H. Müller. Fusion techniques for combining textual and visual information retrieval. In H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors, *ImageCLEF*, volume 32 of *The Springer International Series On Information Retrieval*, pages 95–114. Springer Berlin Heidelberg, 2010.

[62] A. Deshpande, Z. Ives, and V. Raman. Adaptive query processing: Why, how, when, what next? In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 1426–1427. VLDB Endowment, 2007.

[63] E. Di Buccio, M. Dussin, N. Ferro, I. Masiero, and G. Silvello. PROMISE Participative Research labOratory for Multimedia and Multilingual Information Systems

Evaluation. In M. Agosti, F. Esposito, C. Meghini, and N. Orio, editors, *Digital Libraries and Archives*, volume 249 of *Communications in Computer and Information Science*, pages 140–143. Springer Berlin Heidelberg, 2011.

[64] M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña López. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine*, 39(4):396–403, 2009.

[65] Y. Dong, S. Gao, K. Tao, J. Liu, and H. Wang. Performance evaluation of early and late fusion methods for generic semantics indexing. *Pattern Analysis and Applications*, 17(1):37–50, 2014.

[66] K. Dramé, F. Mougin, and G. Diallo. Query expansion using external resources for improving information retrieval in the biomedical domain. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

[67] O. Drori. Improving display of search results in information retrieval systems–users' study. Technical report, Center for Research in Computer Science of the Leibniz, 2000.

[68] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision–ECCV 2002*, pages 97–112. Springer, 2002.

[69] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[70] F. A. Faria, R. T. Calumby, and R. d. S. Torres. RECOD at ImageCLEF 2011: Medical modality classification using genetic programming. In *Working Notes of CLEF 2011*, 2011.

[71] A. Foncubierta-Rodríguez. *Description and Retrieval of Medical Visual Information based on Language Modelling.* PhD thesis, University of Geneva, 2014.

[72] A. Foncubierta-Rodríguez, A. García Seco de Herrera, and H. Müller. Medical image retrieval using bag of meaningful visual words: Unsupervised visual vocabulary pruning with PLSA. In *ACM Multimedia Workshop on Multimedia Information Indexing and Retrieval for Healthcare*, MIIRH '13, pages 75–82. ACM, 2013.

[73] A. Foncubierta-Rodríguez, A. García Seco de Herrera, and H. Müller. *Meaningful Bags of Words for Medical Image Classification and Retrieval.* Springer, 2014.

[74] A. Foncubierta-Rodríguez and H. Müller. Ground Truth Generation in Medical Imaging: A Crowdsourcing Based Iterative Approach. In *Workshop on Crowdsourcing for Multimedia, ACM Multimedia*, oct 2012.

[75] N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas. INEX: Initiative for the evaluation of XML retrieval. In *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, volume 2006, pages 1–9, 2002.

[76] A. García Seco de Herrera, A. Foncubierta-Rodríguez, D. Markonis, R. Schaer, and H. Müller. Crowdsourcing for Medical Image Classification. In *Annual Congress SGMI 2014*, 2014.

[77] A. García Seco de Herrera, A. Foncubierta-Rodríguez, and H. Müller. Medical case–based retrieval: Integrating query MeSH terms for query–adaptive multi–modal fusion. In *SPIE Medical Imaging*. International Society for Optics and Photonics, 2015.

[78] A. García Seco de Herrera, A. Foncubierta-Rodríguez, E. Schiavi, and H. Müller. 2D-based 3D volume retrieval using singular value decomposition of detected regions. In *Medical Computer Vision. Large Data in Medical Imaging*, Lecture Notes in Computer Science, pages 185–195. MICCAI workshop, Springer International Publishing, 2014.

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

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

[81] A. García Seco de Herrera, D. Markonis, R. Joyseeree, , R. Schaer, A. Foncubierta-Rodríguez, and H. Müller. Using semi–supervised learning for image modality classification. In *Proceedings of the Multimodal Retrieval in the Medical Domain (MRMD) Workshop*, Lecture Notes in Computer Science. Springer, 2015.

[82] A. García Seco de Herrera, D. Markonis, and H. Müller. Bag of colors for biomedical document image classification. In H. Greenspan and H. Müller, editors, *Medical Content–based Retrieval for Clinical Decision Support*, MCBR–CDS 2012, pages 110–121. Lecture Notes in Computer Sciences (LNCS), October 2013.

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

[84] A. García Seco de Herrera and H. Müller. *Fusion Techniques in Biomedical Information Retrieval*, pages 209–228. Springer, 2014.

[85] A. García Seco de Herrera, H. Müller, M. Gäde, F. Piori, and T. Tsikrika. Report on the outcomes of the third year evaluation activities. Deliverable D6.3 of the PROMISE project, University of Applied Sciences Western Switzerland (HES–SO), 2013.

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

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

[88] Y. Gkoufas, A. Morou, and T. Kalamboukis. IPL at ImageCLEF 2011 medical retrieval task. In *Working Notes of CLEF 2011*, 2011.

[89] J. Glasgow and I. Jurisica. Integration of case–based and image–based reasoning. In *AAAI Workshop on Case–Based Reasoning Integrations*, pages 67–74, Menlo Park, California, 1998. AAAI Press.

[90] B. Godin and C. Doré. Measuring the impacts of science: Beyond the economic dimension. *History and Sociology of S&T Statistics*, 2004.

[91] B. M. Good and A. I. Su. Crowdsourcing for bioinformatics. *Bioinformatics*, 16(29):1925–1933, 2013.

[92] K. Gottlieb and G. Marino. *Diagnostic Endosonography: A Case–based Approach.* Springer Berlin Heidelberg, 2014.

[93] M. Gu, A. Aamodt, and X. Tong. Component retrieval using conversational case–based reasoning. In *Intelligent information processing II*, pages 259–271. Springer-Verlag, London, UK, 2005.

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

[95] A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R. Smith. The holy grail of multimedia information retrieval: So close or yet so far away? In *Proceedings of the IEEE*, volume 96, pages 541–547, 2008.

[96] D. K. Harman. *The TREC Test Collections*, pages 21–52. Springer, 2005.

[97] A.-W. Harzing. Citation analysis across disciplines: The impact of different data sources and citation metrics. `http://www.harzing.com/data_metrics_comparison.htm`, 2010. Accessed: 2014-10-27.

[98] W. Hersh, C. Buckley, T. Leone, and D. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR94*, pages 192–201. Springer, 1994.

[99] W. Hersh, J. Jensen, H. Müller, P. Gorman, and P. Ruch. A qualitative task analysis for developing an image retrieval test collection. In *ImageCLEF/MUSCLE Workshop on Image Retrieval Evaluation*, pages 11–16, Vienna, Austria, 2005.

[100] W. Hersh, H. Müller, J. Kalpathy-Cramer, E. Kim, and X. Zhou. The consolidated ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging*, 22(6):648–655, 2009.

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

[102] J. E. Hirsch. An index to quantify an individuals scientific research output. *Proceedings of the National Academy of Sciences (PNAS)*, 102(46):16569–16572, 2005.

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

[104] D. F. Hsu and I. Taksa. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, 8(3):449–480, jan 2005.

[105] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Reranking methods for visual search. *Advances in Multimedia Computing*, 14(3):14–22, 2007.

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

[107] H. Hwang Kyung, H. Lee, and D. Choi. Medical image retrieval: Past and present. *Health Information Research*, 18(1):3–9, 2012.

[108] *Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010), Instanbul, Turkey.* IEEE Computer Society, 2010.

[109] N. C. Ide, R. F. Loane, and D. Demner-Fushman. Essie: A concept–based search engine for structured biomedical text. *Journal of the American Medical Informatics Association*, 14(3):253–263, 2007.

[110] I. Jacobson. Object–oriented development in an industrial environment. *ACM Special Interest Group on Programming Languages (SIGPLAN) Notices*, 22(12):183–191, 1987.

[111] I. Jacobson, M. Christerson, P. Jonsson, and G. Övergaard. *Object Oriented Software Engineering: A Use Case Driven Approach.* Addison–Wesley Professional, 1992.

[112] P. Jacsó. Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3):297–309, 2006.

[113] P. Jacsó. The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, 32(3):437–452, 2008.

[114] A. K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, 1996.

[115] A. Järvelin, R. Berendsen, G. Eriksson, P. Hansen, K. Friberg Heppin, J. Karlgren, V. Petras, M. Gäde, M. Lupu, F. Piroi, A. García Seco de Herrera, S. Rietberger, and M. Braschler. Use case inventory and final specification of the evaluation tasks. Deliverable D2.4 of the PROMISE project, University of Gothennburg, 2013.

[116] A. Järvelin, G. Eriksson, P. Hansen, T. Tsikrika, A. García Seco de Herrera, M. Lupu, M. Gäde, V. Petras, S. Rietberger, M. Braschler, and R. Berendsen. Revided specification of evaluation tasks. Deliverable D2.2 of the PROMISE project, University of Amsterdam, 2012.

[117] O. A. Jiménez del Toro, O. Goksel, B. Menze, H. Müller, G. Langs, M.-A. Weber, I. Eggel, K. Gruenberg, M. Holzer, G. Kotsios-Kontokotsios, M. Krenn, R. Schaer, A. A. Taha, M. Winterstein, and A. Hanbury. VISCERAL – VISual Concept Extraction challenge in RAdioLogy: ISBI 2014 challenge organization. In O. Goksel,

editor, *Proceedings of the VISCERAL Challenge at ISBI*, number 1194 in CEUR Workshop Proceedings, pages 6–15, Beijing, China, May 2014.

[118] C. E. Kahn Jr. and C. Thao. Goldminer: A radiology image search engine. *American Journal of Roentgenology*, 188(6):1475–1478, 2007.

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

[120] J. Kalpathy-Cramer and W. Hersh. Multimodal medical image retrieval: Image categorization to improve search precision. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR'10, pages 165–174, New York, NY, USA, 2010. ACM.

[121] J. Kalpathy-Cramer and H. Müller. *Systematic Evaluations and Ground Truth*, pages 497–520. Springer, 2011.

[122] J. Kalpathy-Cramer, H. Müller, S. Bedrick, I. Eggel, A. García Seco de Herrera, and T. Tsikrika. The CLEF 2011 medical image retrieval and classification tasks. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, September 2011.

[123] J. Kalpathy-Cramera and W. Hersh. Automatic image modality based classification and annotation to improve medical image retrieval. *Studies in Health Technology and Informatics*, 129:1334–1338, 2007.

[124] N. Kando. Overview of the second NTCIR workshop. In *Proceedings of the 2nd NTCIR Conference on Evaluation of Information Access Technologies*, 2001.

[125] N. Kando. Overview of the third NTCIR workshop. In *Proceedings of the 3rd NTCIR Conference on Evaluation of Information Access Technologies*, 2003.

[126] N. Kando. Overview of the fourth NTCIR workshop. In *Proceedings of the 4th NTCIR Conference on Evaluation of Information Access Technologies*, 2004.

[127] N. Kando. Overview of the fifth NTCIR workshop. In *Proceedings of the 5th NTCIR Conference on Evaluation of Information Access Technologies*, 2005.

[128] N. Kando. Overview of the sixth NTCIR workshop. In *Proceedings of the 6th NTCIR Conference on Evaluation of Information Access Technologies*, 2007.

[129] N. Kando. Overview of the seventh NTCIR workshop. In *Proceedings of the 7th NTCIR Conference on Evaluation of Information Access Technologies*, 2008.

[130] N. Kando. Overview of the eighth NTCIR workshop. In *Proceedings of the 8th NTCIR Conference on Evaluation of Information Access Technologies*, 2010.

[131] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the 1st NTCIR Conference on Evaluation of Information Access Technologies*, 1999.

[132] N. Kando, D. Oard, T. Kato, and M. Sanderson. Preface from NTCIR-10 general chairs. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, 2013.

[133] J. Karlgren, G. Eriksson, M. Frieseke, M. Gäde, P. Hansen, A. Järvelin, M. Lupu, H. Müller, V. Petras, and J. Stiller. Initial specification of the evaluation tasks. Deliverable D2.1 of the PROMISE project, Swedish Institute of Computer Science (SICS), 2011.

[134] E. Kasutani and A. Yamada. The MPEG–7 color layout descriptor: A compact image feature description for high–speed image/video segment retrieval. In *Proceedings of the International Conference on Image Processing*, ICIP'2001, pages 674–677, 2001.

[135] L. Kennedy, S.-F. Chang, and A. Natsev. Query–adaptive fusion for multimodal search. *Proceedings of the IEEE*, 96(4):567–588, 2008.

[136] A. Kent, M. M. Berry, F. U. Luehrs, and J. W. Perry. Machine literature searching viii. operational criteria for designing information retrieval systems. *American documentation*, 6(2):93–101, 1955.

[137] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.

[138] S.-H. Kim. *The Effects of High Dimensional Covariance Matrix Estimation on Asset Pricing and Generalized Least Squares*. PhD thesis, Georgia Institute of Technology, 2010.

[139] I. Kitanovski, I. Dimitrovski, and S. Loskovska. FCSE at medical tasks of ImageCLEF 2013. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

[140] J. Kludas, E. Bruno, and S. Marchand-Maillet. Information fusion in multimedia information retrieval. In *Proceedings of 5th International Workshop on Adaptive Multimedia Retrieval (AMR)*, volume 4918, pages 147–159, Paris, France, June 2008. ACM.

[141] J. Kludas, E. Bruno, and S. Marchand-Maillet. Can feature information interaction help for information fusion in multimedia problems? *Multimedia Tools and Applications*, 42(1):57–71, March 2009.

[142] B. Koopman and G. Zuccon. Why assessing relevance in medical IR is demanding. In *SIGIR 2014, Medical Information Retrieval (MedIR) Workshop*, 2014.

[143] M. Kreuzthaler, M. Bloice, K.-M. Simonic, and A. Holzinger. *Navigating through Very Large Sets of Medical Records: An Information Retrieval Evaluation Architecture for Non–standardized Text.*, volume 7058 of *Lecture Notes in Computer Science*, pages 455–470. Springer, 2011.

[144] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi. Pooling for a large–scale test collection: An analysis of the search results from the first NTCIR workshop. *Information Retrieval*, 5(1):41–59, 2002.

[145] C. Kurtz, C. F. Beaulieu, S. Napel, and D. L. Rubin. A hierarchical knowledge–based approach for retrieving similar medical images described with semantic annotations. *Journal of biomedical informatics*, 2014.

[146] C. Kurtz and D. L. Rubin. Utilisation de relations ontologiques pour la comparaison d'images décrites par des annotations sémantiques. In *Conférence Francophone sur l'Extraction et la Gestion de Connaissance*, January 2014.

[147] M. Kwiatkowska and S. Atkins. Case representation and retrieval in the diagnosis and treatment of obstructive sleep apnea: A semiofuzzy approach. In *Proceedings European Case Based Reasoning Conference*, ECCBR'04, 2004.

[148] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content–based image retrieval on the world wide web. In *IEEE Workshop on Content–Based Access of Image and Video Libraries*, pages 24–28, Washington, DC, USA, June 1998. IEEE Computer Society.

[149] C. Lacoste, J.-P. Chevallet, J.-H. Lim, D. T. H. Le, W. Xiong, D. Racoceanu, R. Teodorescu, and N. Vuillenemot. Inter–media concept–based medical image indexing and retrieval with UMLS at IPAL. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *CLEF*, volume 4730 of *Lecture Notes in Computer Science (LNCS)*, pages 694–701. Springer, September 2007.

[150] C. Lacoste, J.-H. Lim, J.-P. Chevallet, and D. T. H. Le. Medical–image retrieval based on knowledge–assisted text and image indexing. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(7):889–900, 2007.

[151] A. Lakdashti and M. S. Moin. A new content–based image retrieval approach based on pattern orientation histogram. In A. Gagalowicz and W. Philips, editors, *MIRAGE*, volume 4418 of *Lecture Notes in Computer Science*, pages 587–595. Springer, 2007.

[152] M. Larson, E. Newman, and G. Jones. Overview of VideoCLEF 2008: Automatic generation of topic–based feeds for dual language audio–visual content. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. Jones, M. Kurimo, T. Mandl, A. Peñas, and V. Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706 of *Lecture Notes in Computer Science*, pages 906–917. Springer Berlin/Heidelberg, 2009.

[153] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of the Seventeenth Annual Conference on Neuronal Information Processing Systems*, volume 16, pages 553–560, 2003.

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

[155] W.-Y. Lee, P.-T. Wu, and W. Hsu. Adaptive learning for multimodal fusion in video search. In *Advances in Multimedia Information Processing-PCM 2009*, pages 659–670. Springer, 2009.

[156] J. Li and S. X. Chen. Two sample tests for high–dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940, 2012.

[157] Y. Li, N. Shi, and H. D.Frank. Fusion analysis of information retrieval models on biomedical collections. In *Proceedings of the 14th International Conference on Information Fusion*. IEEE Computer Society, 2011.

[158] E. D. Liddy. Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science and Technology*, 24(4):14–16, 1998.

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

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

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

[162] D. Markonis, A. García Seco de Herrera, I. Eggel, and H. Müller. Multi–scale visual words for hierarchical medical image categorization. In *SPIE Medical Imaging 2012: Advanced PACS–based Imaging Informatics and Therapeutic Applications*, volume 8319, pages 83190F–11, February 2012.

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

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

[165] J. L. Martínez-Fernández, J. V. Román, A. M. García-Serrano, and J. C. González-Cristóbal. Combining textual and visual features for image retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 680–691. Springer, 2006.

[166] M. McCandless, E. Hatcher, and O. Gospodnetic. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA, 2010.

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

[168] D. Mitry, T. Peto, S. Hayat, J. E. Morgan, K.-T. Khaw, and P. J. Foster. Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of images in the epic norfolk cohort on behalf of the UK biobank eye and vision consortium. *PLOS ONE*, 8(8), 2013.

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

[170] S. Montani and R. Bellazzi. Supporting decisions in medical applications: the knowledge management perspective. *International Journal of Medical Informatics*, 68:79–90, 2002.

[171] J. G. Moreno, J. C. Caicedo, and F. A. González. Bioingenium at ImageCLEFmed 2010: A latent semantic approach. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[172] A. Mourão and F. Martins. NovaMedSearch: a multimodal search engine for medical case–based retrieval. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR'13, pages 223–224, 2013.

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

[174] A. Mourão, F. Martins, and J. a. Magalhães. NovaSearch on medical ImageCLEF 2013. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

[175] H. Müller, C. Boyer, A. Gaudinat, W. Hersh, and A. Geissbuhler. Analyzing web log files of the Health On the Net HONmedia search engine to define typical image search tasks for image retrieval evaluation. In *MedInfo 2007*, volume 12 of *IOS press, Studies in Health Technology and Informatics*, pages 1319–1323, Brisbane, Australia, 2007.

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

[177] H. Müller, T. Deselaers, T. M. Deserno, P. Clough, E. Kim, and W. Hersh. Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross–Language Evaluation Forum (CLEF 2006)*, volume 4730 of *Lecture Notes in Computer Science (LNCS)*, pages 595–608, Alicante, Spain, 2007. Springer.

[178] H. Müller, T. Deselaers, T. M. Deserno, J. Kalpathy-Cramer, E. Kim, and W. Hersh. Overview of the ImageCLEFmed 2007 medical retrieval and medical annotation tasks. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007)*, volume 5152 of *Lecture Notes in Computer Science (LNCS)*, pages 472–491. Springer, 2008.

[179] H. Müller, C. Despont-Gros, W. Hersh, J. Jensen, C. Lovis, and A. Geissbuhler. Health care professionals' image use and search behaviour. In *Proceedings of the Medical Informatics Europe Conference (MIE 2006)*, IOS Press, Studies in Health Technology and Informatics, pages 24–32, Maastricht, The Netherlands, aug 2006.

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

[181] H. Müller, A. Geissbuhler, S. Marchand-Maillet, and P. Clough. Benchmarking image retrieval applications. In *Proceedings of the Conference on Visual Information Systems (VISUAL 2004)*, pages 334–337, San Francisco, CA, USA, 2005.

[182] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, and S. Antani. Creating a classification of image types in the medical literature for visual categorization. In *SPIE Medical Imaging*, 2012.

[183] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, C. E. Kahn Jr., and W. Hersh. Overview of the CLEF 2009 medical image retrieval track. In *Proceedings of the 10th International Conference on Cross–language Evaluation Forum: Multimedia Experiments*, CLEF'09, pages 72–84, Berlin, Heidelberg, 2010. Springer–Verlag.

[184] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, J. Reisetter, C. E. Kahn Jr., and W. Hersh. Overview of the CLEF 2010 medical image retrieval track. In *Working Notes of CLEF 2010 (Cross Language Evaluation Forum)*, September 2010.

[185] H. Müller, J. Kalpathy-Cramer, C. E. Kahn Jr., W. Hatt, S. Bedrick, and W. Hersh. Overview of the ImageCLEFmed 2008 medical image retrieval task. In C. Peters, D. Giampiccolo, N. Ferro, V. Petras, J. Gonzalo, A. Peñas, T. Deselaers, T. Mandl, G. Jones, and M. Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, volume 5706 of *Lecture Notes in Computer Science (LNCS)*, pages 500–510, Aarhus, Denmark, September 2009.

[186] H. Müller, X. Zhou, A. Depeursinge, M. Pitkanen, J. Iavindrasana, and A. Geissbuhler. Medical visual information retrieval: State of the art and challenges ahead. In *2007 IEEE International Conference on Multimedia and Expo*, pages 683–686. IEEE, 2007.

[187] S. Navarro, R. Muñoz, and F. Llopis. Diversity promotion: Is reordering top–ranked documents sufficient? In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 120–123. Springer, 2010.

[188] S. J. Nelson, W. D. Johnston, and B. L. Humphreys. *Relationships in Medical Subject Headings (MeSH)*, pages 171–184. Springer, 2001.

[189] J. Ostell. The Entrez search and retrieval system. Technical report, National Center for Biotechnology Information (US), 2014.

[190] G. Övergaard and K. Palmkvist. Use cases: Patterns and blueprints, 2004.

[191] A. P. Pentland, R. W. Picard, and S. Scarloff. Photobook: Tools for content–based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, June 1996.

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

[193] A. Philip, B. Afolabi, A. Oluwaranti, and O. Oluwatolani. Development of an image retrieval model for biomedical image databases. In C. Jao, editor, *Efficient Decision Support Systems — Practice and Challenges in Biomedical Related Domain*, pages 311–329. Intech, 2011.

[194] F. Piori, V. Petras, M. Gäde, B. Larsen, T. Tsikrika, A. García Seco de Herrera, and H. Müller. Report on the outcomes of the second year evaluation activities. Deliverable D6.2 of the PROMISE project, University of Applied Sciences Western Switzerland (HES–SO), 2012.

[195] G. Quellec, M. Lamard, L. Bekri, G. Cazuguel, C. Roux, and B. Cochener. Medical case retrieval from a committee of decision trees. *IEEE Transactions on Information Technology in Biomedicine*, 14(5):1227–1235, 2010.

[196] S. Radhouani, J. Kalpathy-Cramer, S. Bedrick, B. Bakke, and W. Hersh. Using media fusion and domain dimensions to improve precision in medical image retrieval. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 223–230. Springer, 2010.

[197] S. Radhouani, J. Kalpathy-Cramer, S. Bedrick, and W. Hersh. Medical image retrieval, a user study. Technical report, Medical Inforamtics and Outcome Research, OHSU, Portland, OR, USA, June 2009.

[198] E. Rahm and A. Thor. Citation analysis of database publications. *SIGMOD Record*, 34:48–53, December 2005.

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

[200] B. L. Ranard, Y. P. Ha, Z. F. Meisel, D. A. Asch, S. S. Hill, L. B. Becker, A. K. Seymour, and R. M. Merchant. Crowdsourcing–harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, 29(1):187–203, 2014.

[201] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi. A simultaneous feature adaptation and feature selection method for content–based image retrieval systems. *Knowledge–Based Systems*, 39:85–94, 2013.

[202] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross–modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia*, pages 251–260. ACM, 2010.

[203] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, pages 13–19. ACM, 2004.

[204] A. Rosset, H. Müller, M. Martins, N. Dfouni, J.-P. Vallée, and O. Ratib. Casimage project — a digital teaching files authoring environment. *Journal of Thoracic Imaging*, 19(2):1–6, 2004.

[205] B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic impact assessment of NISTs text retrieval conference (TREC) program. Technical report project number 0211875, National Institute of Standars and technology, 2010.

[206] T. Sakai. NTCIREVAL: A generic toolkit for information access evaluation. In *The Forum on Information Technology*, volume 2, pages 23–30, 2011.

[207] T. Sakai and H. Joho. Overview of NTCIR–9. In *Proceedings of the 9th NTCIR Conference on Evaluation of Information Access Technologies*, 2011.

[208] G. Salton. The state of retrieval rystem evaluation. *Information Processing & Management*, 28(4):441–449, 1992.

[209] M. Sanderson and M. Braschler. Best practices for test collection creation and information retrieval system evaluation. Technical report, TrebleCLEF Consortium, 2009.

[210] T. Saracevic. Evaluation of evaluation in information retrieval. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 138–146. ACM, 1995.

[211] R. Schaer, D. Markonis, and H. Müller. Architecture and applications of the Parallel Distributed Image Search Engine (ParaDISE). In *FoRESEE 2014, 1st International Workshop on Future Search Engines at INFORMATIK 2014*, 2014.

[212] S. Selvarajah and S. R. Kodituwakku. Analysis and comparison of texture features for content based image retrieval. *International Journal of Latest Trends in Computing*, 2:108–113, 2011.

[213] L. G. Shapiro, I. Atmosukarto, H. Cho, H. J. Lin, S. Ruiz-Correa, and J. Yuen. Similarity-based retrieval for biomedical applications. In *Case–Based Reasoning on Images and Signals*, volume 73 of *Studies in Computational Intelligence*, pages 355–387. Springer, 2008.

[214] S.-T. Shen and S. D. Prior. My favorites (bookmarks) schema: One solution to online information storage and retrieval. In *Proceedings of the 2013 International Conference on Information Systems and Design of Communication*, pages 33–40. ACM, 2013.

[215] Z. Shi, B. Gu, F. Popowich, and A. Sarkar. Synonym–based query expansion and boosting–based re–ranking: A two–phase approach for genomic information retrieval. In *The Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[216] M. Simpson, M. M. Rahman, S. Phadnis, E. Apostolova, D. Demmer-Fushman, S. Antani, and G. Thoma. Text– and content–based approaches to image modality classification and retrieval for the ImageCLEF 2011 medical retrieval track. In *Working Notes of CLEF 2011*, 2011.

[217] M. S. Simpson and D. Demner-Fushman. Biomedical text mining: A survey of recent progress. In *Mining Text Data*, pages 465–517. Springer, 2012.

[218] M. S. Simpson, D. Demner-Fushman, S. K. Antani, and G. R. Thoma. Multimodal biomedical image indexing and retrieval using descriptive text and global feature mapping. *Information Retrieval*, 17(3):229–264, 2014.

[219] M. S. Simpson, D. You, M. M. Rahman, D. Demmer-Fushman, S. Antani, and G. Thoma. ITI's participation in the ImageCLEF 2012 medical retrieval and classification tasks. In *Working Notes of CLEF 2012*, 2012.

[220] M. S. Simpson, D. You, M. M. Rahman, D. Demner-Fushman, S. Antani, and G. Thoma. ITI's participation in the 2013 medical track of ImageCLEF. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

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

[222] C. L. Smith and P. B. Kantor. User adaptation: Good results from poor systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 147–154. ACM, 2008.

[223] W. Song, D. Zhang, and J. Luo. BUAA AUDR at ImageCLEF 2012 medical retrieval task. In *Working Notes of CLEF 2012*, 2012.

[224] E. Sormunen. Liberal relevance criteria of TREC: Counting on negligible documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–330. ACM, 2002.

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

[226] S. Stathopoulos, I. Lourentzou, A. Kyriakopoulou, and T. Kalamboukis. IPL at CLEF 2013 medical retrieval task. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

[227] C. Sungbin, J. Lee, and J. Cho. SNUMedinfo at ImageCLEF 2013: Medical retrieval task. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

[228] S. Sural, G. Qian, and S. Pramanik. Segmentation and histogram generation using the HSV color space for image retrieval. In *Proceedings of the International Conference on Image Processing*, ICIP'2002, pages 589–592, 2002.

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

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

[231] M. Taschwer. Textual methods for medical case retrieval. Technical report, Institute of Information Technology (ITEC), Alpen-Adria-Universität Klagenfurt, Austria, May 2014.

[232] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee. The scholarly impact of TRECVid (2003–2009). *Journal of the American Society for Information Science and Technology*, 62(4):613–627, 2011.

[233] P. Tirilly, K. Lu, X. Mu, T. Zhao, and Y. Cao. On modality classification and its use in text–based image retrieval in medical databases. In *9th International Workshop on Content–Based Multimedia Indexing*, 2011.

[234] C.-F. Tsai. Bag–of–words representation in image annotation: A review. *International Scholarly Research Notices*, 2012, 2012.

[235] S. Tsevas and D. K. Iakovidis. Fusion of multimodal temporal clinical data for the retrieval of similar patient cases. In *10th International Workshop on Biomedical Engineering*, pages 1–4. IEEE, 2011.

[236] T. Tsikrika, A. García Seco de Herrera, and H. Müller. Assessing the scholarly impact of ImageCLEF. In *CLEF 2011*, Springer Lecture Notes in Computer Science (LNCS), pages 95–106, sep 2011.

[237] T. Tsikrika, B. Larsen, H. Müller, S. Endrullis, and E. Rahm. The scholarly impact of CLEF (2000–2009). In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 1–12. Springer, 2013.

[238] T. Tsikrika, H. Müller, and C. E. Kahn Jr. Log analysis to understand medical professionals' image searching behaviour. In *Proceedings of the 24th European Medical Informatics Conference*, MIE'2012, 2012.

[239] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–18. ACM, 2006.

[240] D. Ünay, Z. Çataltepe, and S. Aksoy, editors. *Proceedings of the 20th International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos, ICPR Contest Reports.* Springer, 2010.

[241] E. Uwimana and M. E. Ruiz. Integrating an automatic classification method into the medical image retrieval process. In *AMIA Annual Symposium Procreedings*, pages 747–751, 2008.

[242] K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. The university of Amsterdam's concept detection system at ImageCLEF 2009. *Lecture Notes in Computer Science*, 6242:261–268, 2010.

[243] S. Viswa. Efficient retrieval of images for search engine by visual similarity and re ranking. *International Journal of Advanced Computer Research*, 3(2):2277–7970, 2013.

[244] E. M. Voorhees and L. P. Buckland, editors. *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), November 2006.

[245] E. M. Voorhees and D. K. Harman. *The Text Retrieval Conference*, pages 3–19. Springer, 2005.

[246] E. M. Voorhees and W. Hersh. Overview of the TREC 2012 medical records track. In *The Twenty–first Text REtrieval Conference Proceedings TREC*, 2012.

[247] J. Z. Wang. Region–based retrieval of biomedical images. In *Proceedings of the ACM Multimedia Conference*, pages 511–512, nov 2000.

[248] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler. Analysis of the accuracy and robustness of the leap motion controller. *Sensors*, 13(5):6380–6393, 2013.

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

[250] A. Widmer, R. Schaer, D. Markonis, and H. Müller. Facilitating medical information search using google glass connected to a content-based medical image retrieval system. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2014.

[251] A. Widmer, R. Schaer, D. Markonis, and H. Müller. Gesture interaction for content–based medical image retrieval. In *ICMR*, 2014.

[252] H. Wu, K. Sun, X. Deng, Y. Zhang, and B. Che. UESTC at ImageCLEF 2012 medical tasks. In *Working Notes of CLEF 2012*, 2012.

[253] S. Wu. Linear combination of component results in information retrieval. *Data & Knowledge Engineering*, 71(1):114–126, 2012.

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

[255] K. Y. Yip and M. Gerstein. Training set expansion: An approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics*, 25(2):243–250, 2009.

[256] B. Yu, M. Willis, P. Sun, and J. Wang. Crowdsourcing participatory evaluation of medical pictograms using Amazon mechanical turk. *Journal of Medical Internet Research*, 15(6), 2013.

[257] D. Zhang and G. Lu. Review of shape representation and description techniques. *Pattern Recognition*, 37(1):1–19, 2004.

[258] M.-L. Zhang and Z.-H. Zhou. A review on multi–label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 99, 2013.

[259] X. Zhou, A. Depeursinge, and H. Müller. Information fusion for combining visual and textual image retrieval. In *20th IEEE International Conference on Pattern Recognition (ICPR)*, pages 1590–1593, aug 2010.

[260] X. Zhou, M. Han, Y. Song, and Q. Li. Fast filtering techniques in medical image classification and retrieval. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

Index