Analysis and Evaluation of Visual Information Systems Performance

Michael Grubinger

This thesis is presented in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science and Mathematics
Faculty of Health, Engineering and Science
Victoria University

April 2007

Abstract

This dissertation investigates the system-centred evaluation of visual information retrieval from generic photographic collections.

The development of visual information retrieval systems has long been hindered by the lack of standardised benchmarks. Researchers have proposed numerous systems and techniques, and although different systems clearly have their particular strength, there is a tendency by researchers to use different means of showing retrieval performance to highlight the own algorithm's benefits. For the field of visual information search to advance, however, objective evaluation to identify, compare and validate the strengths and merits of different systems is therefore essential. Benchmarks to carry out such evaluation have recently been developed, and evaluation events have also been organised for several domains. Yet, no efforts have considered the evaluation of retrieval from generic photographic collections (*i.e.* containing everyday real-world photographs akin to those that can frequently be found in private photographic collections as well, *e.g.* pictures of holidays and events).

We therefore first analyse a multitude of variables and factors with respect to the performance and requirements of visual information systems, and we then design and implement the framework and resources necessary to carry out such an evaluation. These resources include: a parametric image collection, representative search requests, relevance assessments and a set of performance measures. In addition, we organise the first evaluation event for retrieval from generic photographic collections and report on its realisation. Finally, we present an analysis and the evaluation of the participating retrieval systems as well as of the evaluation event itself.

Filling this particular gap by making possible a systematic calibration and comparison of system performance for retrieval from generic photographic collections constitutes the main scientific contribution of this research. This dissertation thereby enables a deeper understanding of the complex conditions and constraints associated with visual information identification, the accurate capturing of user requirements, the appropriate specification and complexity of user queries, the execution of searches, and the reliability of performance indicators.

Declaration

“I, Michael Grubinger, declare that the PhD thesis entitled *Analysis and Evaluation of Visual Information Systems Performance* is no more than 100,000 words in length, exclusive of tables, figures, appendices, references and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work”.

Signature Date

List of External Publications

Publications in journals and proceedings of peer reviewed conferences and workshops:

- Michael Grubinger, Clement H. C. Leung. A Benchmark for Performance Calibration in Visual Information Search. In *Proceedings of The 2003 International Conference on Visual Information Systems (VIS'2003)*, pages 414–419, Miami, FL, USA, September 2003. Knowledge Systems Institute.

- Paul Over, Clement H. C. Leung, Horace Ip, Michael Grubinger. Multimedia Retrieval Benchmarks. *Digital Multimedia on Demand, IEEE Multimedia April-June 2004*, pages 80–84, 2004.

- Michael Grubinger, Clement H. C. Leung. Incremental Benchmark Development and Administration. In *Proceedings of The Seventh International Conference of Visual Information Systems (VIS'2004)*, pages 328–333, San Francisco, CA, USA, September 2004. Knowledge Systems Institute.

- Michael Grubinger, Clement H. C. Leung, Paul D. Clough. The IAPR Benchmark for Assessing Retrieval Performance in Cross Language Evaluation Tasks. In *Proceedings of the MUSCLE ImageCLEF Workshop on Image and Video Retrieval Evaluation*, pages 33–50, Vienna, Austria, September 2005.

- Michael Grubinger, Paul D. Clough, Henning Müller, Thomas Deselaers. The IAPR TC-12 Benchmark - A New Evaluation Resource for Visual Information Systems. In *The Proceedings of the International Workshop OntoImage'2006*

Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06, pages 13–23, Genoa, Italy, May 2006.

- Michael Grubinger, Paul D. Clough, Clement H. C. Leung. The IAPR TC-12 Benchmark for Visual Information Search. *IAPR Newsletter April 2006*, Volume 28, Number 2, pages 10–12, 2006.

- Michael Grubinger, Clement H. C. Leung, Paul D. Clough. Linguistic Estimation of Topic Difficulty in Cross-Language Image Retrieval. In *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 558–566, Vienna, Austria, September 2006. Springer.

- Paul D. Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas Lehmann, Jeffery Jensen, William Hersh. The CLEF 2005 Cross-Language Image Retrieval Track. In *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 535–557, Vienna, Austria, September 2006. Springer.

- Paul D. Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, Henning Müller: Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. *Evaluation of Multilingual and Multi-modal Information Retrieval, Lecture Notes in Computer Science (LNCS)*, Alicante, Spain, 2007. Springer - in press.

Publications in working papers and reports:

- Michael Grubinger. Benchmarking for Content Based Visual Information Search. *SCM6102 Research Project Report*, School of Computer Science and Mathematics, Victoria University of Technology, Melbourne, Australia, November 2002.

- Paul D. Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas Lehmann, Jeffrey Jensen, William Hersh. The CLEF 2005 Cross-

Language Image Retrieval Track. In *CLEF Working Notes*, Vienna, Austria, September 2005.

- Michael Grubinger, Clement H. C. Leung, Paul D. Clough. Towards a Topic Complexity Measure for Cross-Language Image Retrieval. In *CLEF Working Notes*, Vienna, Austria, September 2005.

- Paul D. Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, Henning Müller. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *CLEF Working Notes*, Alicante, Spain, September 2006.

Acknowledgements

Many wonderful people have both assisted and supported me during this dissertation, and I offer my thanks to all of them.

First and foremost, I am deeply indebted to my principal supervisor Clement H. C. Leung for his guidance, direction and the countless hours of support he provided during my candidature. I want to thank him for introducing me to the exciting domain of visual information retrieval, showing me the art of writing conference and journal papers, for his advice and feedback on my many proposals, and the regular discussions about my ideas as well as the issues and aspects surrounding my research. His deep knowledge and understanding as well as his contacts within the field of information retrieval have been invaluable for this research. I have benefited immeasurably from his tutelage and expertise.

This dissertation would have not been written without the help of Paul Clough from Sheffield University. Even though he has been extremely busy with his own duties and responsibilities, Paul has given up many hours of his valuable time to help me understand the aspects of text retrieval and the organisation of large-scale evaluation events. I want to thank him for giving me the chance to run my evaluation task at ImageCLEF, the endless Skype conversations, the fruitful discussions and his continuing help and valuable feedback with both publications as well as this dissertation.

I would also like to thank my other ImageCLEF colleagues: Henning Müller for his constructive and detailed feedback on my papers and publications, for establishing the contact with ImageCLEF and also for getting me started with $\LaTeX$; Thomas Deselaers, Allan Hanbury and Mark Sanderson for their fruitful comments on my work; and Carol Peters from CLEF for her support and providing the platform to facilitate the practical application of my research.

My thanks also go to Victoria University for funding this research with an International Postgraduate Research Scholarship, and to Lesley Birch for guiding me during the scholarship application process and her help with all scholarship related issues during the candidature. Further, to the School of Computer Science and Mathematics for providing me with the necessary resources to carry out my research. In particular, I would like to thank the Head of School, Pietro Cerone, and Janet Grady for their immense support and great cooperation; Cameron Giles, Mark Mojic and Danh Ho for their help and technical assistance; and Bai-Ling Zhang, Alasdair McAndrews and Ian Gomm for their valuable feedback, comments and input.

Special thanks to Angela Rojter for correcting my English, improving my academic writing skills and for those long and interesting discussions about grammar in general and the special nuances of Australian English in particular. To my dear friend Petra Siwek for having an independent and critical look at my introduction, and to Eder Kikianty for double-checking and fine-tuning the mathematical elaborations. To Gabriele Sorrentino, Jakub Szajman and Dennis Hammer for helping me out whenever I got stuck with LaTeX. To André Kiwitz and Stefan Richter from *viventura* for supporting this research and providing their valuable images for evaluation purposes.

To my research colleagues Mark Aquilina, Gary Howells, Guandong Xu, Hao Lan Zhang and Murali Krishna Reddy Danda for creating a stimulating and motivating work environment in our student research office. Special thanks go to Klautsi, Rex and Turba for keeping me company during my long and late office hours and making these times a little bit less lonely. I also want to thank my flatmate Chegon Hashim for supporting and nursing an often tired, injured and hungry Abang throughout the entire candidature.

Final thanks to my wonderful friends world-wide for standing by me and continuing that special friendship although they might not have seen me a lot in recent years. I miss the time that I could have otherwise spent with them, and I deeply appreciate their loyalty.

Last but not least, to my beautiful family in Austria, especially my Mum, my Dad and my two gorgeous sisters Andrea and Ingrid for their emotional support, continuous encouragement and everlasting love. Thank you so much for believing in me and always being there for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides a concise introduction to the world of images in the past and the present; it also presents the main motivation for writing this dissertation in the field of visual information retrieval evaluation, lists the scientific contributions achieved within this research and provides an overview of the structure and content of this work.

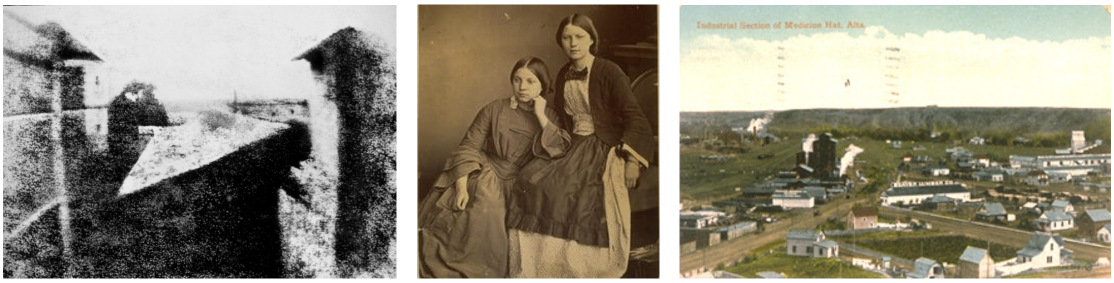
1.1 Visual Information Evolution

Images have always played a vital role in human communication. A long time before the first writing systems evolved in the *Early Bronze Age* a couple of thousand years ago, people had already communicated (or rather expressed themselves) by painting (mainly animal) pictures on the walls of their caves or on rocks. Such cave and rock paintings have been found all over the world, with the oldest known painting at the Grotte Chauvet in France being more than 32,000 years old.

Throughout the past, images kept on fascinating humans: the ancient *Egyptians* and *Greeks* produced many paintings to dignify their gods and also created illustrations inside tombs and sarcophagi as a form of correspondence with the afterlife. The *Romans* went one step further and illustrated scenes of historic events (mainly wars), maps, building plans and portraits to record and convey information; the rise of Christianity in the *Middle Ages* led to the creation of thousands of drawings usually of religious significance, and the *Renaissance* saw a growing importance of paintings and drawings reflecting ideas, art and science. Yet, images were still produced in only rather small quantities, even later throughout the *Early Modern Times*, despite the existence of some image reproduction technologies such as the woodcut (1200s), engraving (1500s) or etching (1600s). The invention of photography in the 19th century, however, would dramatically change this situation.

1.1.1 Traditional Photography

The first remaining photograph was taken by the Frenchman Joseph N. Niépce in 1826 (see Figure 1.1(a)). While certainly being quite intricate and complicated at first (*e.g.* exposing times of several hours!), photographic processes soon improved and photos became more and more widespread until the end of the 19th century: William Fox Talbot invented the positive/negative process in 1841, which allowed the multiple reproduction of photos for the first time and is still widely used in modern photography nowadays; portrait photographers had replaced portrait painters by the 1860s (see Figure 1.1(b)). In 1884, George Eastman patented the photographic (roll) film; four years later, he marketed his *Kodak Number One*, the first camera that was easy to use; and in 1900, he took mass-market photography one step further with the introduction of the *Brownie*, a simple and (the first) inexpensive box camera that introduced the concept of the snapshot.



(a) first photograph (1826) (b) portrait (1865) (c) colour photograph (1910)

Figure 1.1: The evolution of photography.

The early 20th century saw further improvements and refinements which led to photography's increasing popularity. For instance, the first colour photography process, the *Autochrome Lumière*, was marketed in 1907 (Figure 1.1(c)); Ernst Leitz introduced the 35mm format to still photography in 1925 and published the first miniature camera, *Leica*, in the same year; 10 years later in 1935, the first modern colour film, *Kodachrome*, was brought to the market, making colour photography available for the broad public at the same time; and in 1959, the first fully automatic camera, *Optima*, was presented by AGFA. Due to these rapid developments, photography finally became mainstream by mid last century and has witnessed an unparalleled growth in the number, availability and importance of images in all aspects of life ever since.

1.1.2 Digital Imaging

While photography had revolutionised the world of visual information in the 19th century, the computer age and subsequently the digital cameras did the same to the world of photography at the end of the 20th century, providing the technology that allows for the digital capturing, processing, storage and transmission of images.

The start of the era of *digital imaging* is closely connected with Ivan Sutherland, one of the pioneers as far as the involvement of computers in imaging is concerned, who demonstrated the feasibility of computerised creation, manipulation and storage of images as early as 1963 [435]. However, the very high hardware costs hindered the success of digital imaging for a long time, and it took long until the late 1980s that the technology finally took off: when computerised imaging became affordable, fields that had traditionally depended heavily on images for communication (*e.g.* architecture, engineering, medicine) picked up the technology. Soon, photographic libraries, art galleries and museums also began to see the benefits of making their collections available in electronic form and jumped on the bandwagon too, consequently carrying out large digitalisation projects.

This trend continued in the early 1990s and received another massive boost with the introduction of *digital cameras*. In 1991, Kodak presented its *DCS-100*, the first commercially available digital camera; although its price was astronomically high for the normal end-user (around 13,000 USD), it signified the birth of commercial digital photography. Digital cameras soon matured, not least due to the inventions of novel compression formats (JPEG[1] was issued in 1992 and approved in 1994), the integration of liquid crystal display (LCD) in 1995 or reusable storage media such as *CompactFlash* (1996). Although still more expensive than their traditional analogue counterparts, digital cameras soon enjoyed rapidly increasing popularity not only in the commercial but also in the consumer market. This was mainly because they are very easy to handle and no additional costs occur when taking pictures, because photos can directly be displayed on a screen and no longer have to be developed or printed.

As a consequence, digital imaging had started to replace traditional photography at the beginning of the new millennium at a rapid pace and soon prevailed: in 2003, more digital than traditional cameras were sold for the first time. This trend has continued until now, not least due to the integration of digital cameras into other modern devices such as mobile phones or personal digital assistants, and is expected to continue in the future.

Digital imaging, however, is not without criticism either; while traditional photography is an analogue process involving film, optics and photographic paper and thus inherently resists manipulation, digital imaging is purely digital from the very beginning and therefore a highly manipulative medium: pictures can easily be edited, cropped or even combined with illustrations with only a few mouse-clicks. While analogue photography is popularly synonymous with truth ("The camera does not lie!"), digital photography cannot claim the same status; this, for instance, led as far as the majority of courts not accepting digital images as evidence any more due to their manipulative nature.

1.1.3 Digital Images in the Internet Age

In recent years (and especially in the new millennium), the Internet (or more exactly the World Wide Web, short WWW) has produced a massive explosion of images being available throughout the world, allowing their distribution from a local com-

[1] Joint Photographic Experts Group, `http://www.jpeg.org/`

puter or a web server to millions of potential users. While early web pages only contained a few images in addition to the text (if at all), they have now become more and more image centred; there are even web pages that exclusively contain images, with even the text being a part of the images, or the other way round - text being expressed as images.

Many organisations that had contributed to the huge digitalisation trend in the early 90s started to use the new media in order to present their images and products to potential customers throughout the world. Furthermore, photo sharing web-sites such as *FlickR*[2] have also recently been enjoying great popularity in that they facilitate the storage, sharing and management processes of digital images for private home-users as well.

In July 2005, the Internet search engine *Yahoo!*[3] announced that their image index contained more than 1.6 billion entries, with its business rival *Google*[4] issuing a statement the very next day claiming to have indexed more than 2.1 billion images. Both figures certainly have to be handled with care as they might be influenced by market strategies or politics; nevertheless, they both illustrate the sheer amount of images that are available in the WWW.

1.2 Motivation

Despite all the benefits that the process of digitalisation brought along, it did not necessarily make image collections easier to manage. On the contrary, unlike analogue photos, they cannot be archived in boxes or folders and often end up in some computer file directory. While it might certainly be feasible, in the case of small collections, to simply browse through these directories in order to find a desired image, this approach is not applicable any more for slightly larger collections containing several thousand images: more effective techniques are then required to enable successful retrieval afterwards, such as the assignment of descriptive meta-data in the

[2] `http://www.flickr.com/`

[3] `http://www.yahoo.com/`

[4] `http://www.google.com/`

form of keywords or some sort of classification code to each image when it is first added to the collection.

With the ever increasing number of web pages and images available on the WWW, text search engines such as *Google*, *Yahoo!*, *Excite*[5], MSN[6] or *Altavista*[7] soon recognised the need for image retrieval and started to offer searches for multimedia data (such as image, audio or video files). This is usually done using the image filenames or text next to the images on the web page rather than the image features directly, as text retrieval is a very mature field of research quite contrary to image retrieval.

1.2.1 The Visual Information Search Problem

The text-based search approaches executed by the aforementioned internet search engines inescapably raise several questions: What can be done if a web page only contains visual information? Is there a chance to find images although there is no matching word anywhere near it in the surrounding text? What if such text does not refer to the image at all or is wrong? And is there a chance to still retrieve the required image if the text near the image is written in a different language than the query language?

This dissertation is, of course, not the first one to raise these questions; on the contrary, visual information retrieval (VIR) has been one of the most active fields of research in the last 10 to 15 years (see Chapter 2). Especially content-based image retrieval (CBIR) – the idea that an image retrieval request could be satisfied by automatically deriving the information from the images themselves without having to rely on secondary meta-data or associated logical (alphanumeric) image representations – received a lot of attention as CBIR could potentially solve all the questions raised in the last section.

Unfortunately, despite more than one decade of effort by the CBIR research

[5] `http://www.excite.com/`

[6] `http://www.msn.com/`

[7] `http://www.altavista.com/`

community, this technology still remains in the fledgling stages for most of the image requests, with a real contribution to VIR being limited to information recovery tasks in very specialised fields such as fingerprint and trademark recognition as well as face-based and medical image retrieval. Along with the developing maturity of these endeavours, the realisation of the limitations of general CBIR soon set in:

- the biggest problem is the discrepancy between the semantic categories a user is looking for and the low level features that the present CBIR systems (CBIRS) have to offer (the so called *Semantic Gap*);
- another problem lies in the loss of information that occurs when an image is recorded (*Sensory Gap*);
- moreover, while most real-user information needs are more akin to information discovery, CBIR seems more suitable for information recovery tasks.

These main problems of CBIR (which will be further elucidated in Chapter 2 and are still unsolved to date), and the visual search problem in general, raised more questions: Can these major problems ever be overcome? Improvements to the existing CBIRS seem feasible, but can CBIR solve these problems by itself? Or will some form of combination with text retrieval be inevitable in order to bridge the Semantic Gap? And will CBIR applications ever be suitable for information discovery tasks?

1.2.2 The Need for Image Retrieval Benchmarks

A large number research groups and institutions soon recognised the importance of the visual information search problem, proposing many techniques and building many systems to perform such a search. Although different systems clearly had their particular strengths, there was (and still is) a tendency by researchers to use different sets of data, queries and performance measures to highlight their own system's merits. As a consequence, some degree of bias might exist which makes it difficult to make meaningful comparisons concerning the relative superiority of different algorithms. This lack of objective assessment of retrieval performance (*i.e.* the ability of a system to return relevant images for a specific user need) has hindered the research progress in the field of VIR for some time, because it was quite difficult to distinguish the the promising techniques from the simply glossy ones.

Analogy to the World of Boxing

This entire situation was, to some extent, very similar to the rather unclear situation in current heavyweight boxing, an analogy which I used when I presented this research to the international research community at the *Visual Information Systems Conference* in 2003 [185]. After the opening question "Who is the current world champion in heavyweight boxing?" had raised some eyebrows, no one in the audience could actually come up with a clear answer. This was not a big surprise, as the situation nowadays is not as clear-cut as it was at the times of the great Muhammad Ali more than 30 years ago: now, there are four main boxing federations: WBA[8], WBC[9], IBF[10] and WBO[11] (and several minor ones), each of them claiming to be of utmost significance, accepting different boxers and applying different rules and ranking systems, and each of them promoting different world champions (see Fig-

Jones (WBA) Lewis (WBC) Byrd (IBF) Sanders (WBO)

Figure 1.2: World boxing champions (heavyweight, September 2003).

[8] World Boxing Association: `http://www.wbaonline.com/`

[9] World Boxing Council: `http://www.wbcboxing.com/`

[10] International Boxing Federation: `http://www.ibf-usba-boxing.com/`

[11] World Boxing Organisation: `http://www.wbo-int.com/`

ure 1.2). How will we ever be able to compare their achievements if these boxers actually never (or only rarely) fight against each other? And how should we know who is really the best boxer, and not just the “best-promoted” one? Like with VIR, these questions cannot be answered unless there exists a common basis which allows for an objective evaluation: the boxers actually fighting each other within the same federation, or the image retrieval systems being compared using the same image collections, search queries (topics) and performance measures.

Evaluation in Other Domains

While such common evaluation events have successfully been carried out in the field of text retrieval since 1992 [165], VIR evaluation events followed with a delay of more than ten years: several benchmarks have been created since 2003, and evaluation events have been organised for retrieval tasks from collections containing historic photographs and medical images [62, 63, 64]; the discipline of object recognition saw the evaluation of general object recognition tasks [60, 110] and also more specific ones like the automatic annotation of medical objects [86, 290] and the classification of coin images [323]; evaluation has also been carried out for more specific needs such as copyright protection or detection of text within images [118], and there is an effort to evaluate the usability of interactive retrieval systems that allow relevance feedback [136, 137]; there are also evaluation campaigns for the related fields of video retrieval [214, 215, 328, 401], cross-language information retrieval [102, 339, 340, 341, 342] and multimedia retrieval from structured collections [484, 511].

Evaluation of Retrieval from Generic Photographic Collections

No evaluation efforts, however, had yet been started for ad-hoc[12] retrieval from general collections containing everyday real-world photographs akin to those that can frequently be found in private photographic collections as well, *e.g.* holiday pictures or photos of events (we shall refer to such collections as *generic photographic*

[12] In *ad-hoc* retrieval tasks, only the first response of the system is evaluated, without considering further interaction such as query refinements or relevance feedback.

collections throughout this dissertation).

The lack of existing research for this particular domain is quite astonishing as the evaluation scenario models exactly the operation of the current web-based search engines such as *Google* or *Yahoo!*; and also because there was already a common consent in the late 1990s (see [99]) that the management of private photographic collections and the home-user target group in general would constitute precisely one of the areas where a mass market for VIR technology could develop and hence virtually drive all future development activity, especially in CBIR.

Potential reasons for the lack of evaluation in this area certainly lie in the lack of availability of evaluation resources, as it is very hard to find image collections in this domain which are free of charge and without copyright restrictions, or simply in the fact that, due to the massive potential in the field, several commercial organisations are investigating in the very same field and simply do not want to share information or be compared with their competitors due to strategic plans.

Therefore, the main motivation of the research presented in this dissertation lay in filling this particular gap by providing a benchmark for ad-hoc image retrieval tailored to generic photographic collections and allowing for a systematic calibration and comparison of system performance not possible before.

1.3 Thesis Organisation

This section presents an overview of the structure of this thesis as well as some introductory remarks and its scientific contributions before we go *in medias res.*

1.3.1 Structure of this Dissertation

The overall structure of this dissertation takes the form of eight chapters.

Chapter 1 provides a concise *introduction* to the world of visual information and VIR respectively; it explains the main motivation for the research described in this thesis, introduces its structure and lists its main scientific contributions.

Chapter 2 introduces the general concepts and challenges of *visual information* *retrieval* and, being the first of the two main literature review chapters, provides a number of references for various aspects in this field of research: it first discusses its evolution and exemplifies where and how this technology can (or is likely to be) used; then, information retrieval characteristics including its classification according to user need analyses and the corresponding content-based retrieval technologies are explained; another section is dedicated to concept-based retrieval, concentrating on image semantics and their specifications (*e.g.* MPEG-7); finally, some sample image retrieval systems are explored.

Chapter 3 forms the second part of the literature review and illustrates the development, aims and principles as well as criticism of *visual information retrieval evaluation* (*benchmarks*). An analogy to the closely related and well-matured field of evaluation in text retrieval is found, providing the main framework for benchmarking also in VIR. The most significant part of this chapter comprises the general description of the benchmark components that constitute such a framework, together with the individual analysis of these components in existing evaluation events.

Chapter 4 presents the design of the *document collection* that we created in the frame of this research, including image selection rules according to which the images were specifically selected, strict guidelines to guarantee both consistent and realistic semantic descriptions of the image contents, and the image annotation process itself; it further provides collection statistics and information about the origin of the images and their logical alphanumeric representations, and how the collection, which is available for the research community royalty-free and without hindering copyright restrictions, can be accessed.

Chapter 5 describes the model that we established in order to facilitate the *topic creation process* for image retrieval evaluation events; this comprises the identification of several query dimensions as well as the development of a novel measure for one of these dimensions, namely the retrieval difficulty of such a query.

Chapter 6 illustrates the design, architecture and application of a *parametric* *benchmark administration system* that we developed in order to facilitate and guide the management of the most essential benchmark components as well as to foster a deeper understanding of the complicated processes that underly such evaluation efforts.

Chapter 7 reports on an *evaluation event* for VIR from generic photographic collections, in which we applied all the theoretical work that is described in the previous chapters; this is certainly another major contribution to the research community as it was the first ever evaluation event of its kind. In this chapter, the individual benchmark components are depicted in the light of the event, followed by a description of the techniques used by the participants and a detailed analysis of their results and feedback.

Chapter 8 finally concludes this work by giving a brief summary and critique of the findings and identifying areas for future research.

1.3.2 Remarks

The *bibliography* can certainly not claim to be complete in a fast moving research field such as VIR. However, apart from often-cited and well-established relevant scholarly literature, it contains numerous articles that have influenced this work, be it in a positive or negative way. The section on *external publications* further comprises a complete list of all my publications relevant to this research.

In addition to the main corpus, this thesis also contains several parts that are supposed to make it easier to read: the *glossary* contains all the abbreviations used in the text, as "abbreviationitis" is a disease in VIR which may cause problems for readers who are new to the field; the *notation* allows for consistency within mathematical symbols and equations used in this work; and the *appendix* provides additional research data not directly incorporated in the main content.

Furthermore, this thesis is generally laid out such that already existing work is covered in Chapters 2 and 3, with my own contributions described in Chapters 4 to 7. These chapters, however, also contain some sections where such separation is not as obvious; thus, in order to rule out any misunderstanding throughout this thesis, my own work and scientific contributions will be referred to in the first person (plural), as opposed to already existing work which is expressed using the third person.

1.3.3 Scientific Contributions

The most significant contribution of this research lies in the detection and repair of the lack of evaluation resources for (multilingual) visual information retrieval from generic photographic collections. We have studied and made contributions to

- the *design* and *development* of parametric test *collections*,
- the universality of *image semantics* and the corresponding logical image representations across different languages and world views,
- the matching of *user intentions* and *query specifications* (topics),
- the *complexity* of *queries*,
- the *architecture* and *management* of benchmarks,
- the quantification and analysis of *retrieval performance*, and
- the design of *evaluation events.*

We further show that, with visual information retrieval, it is not just a matter of issuing queries against a database and obtaining results, but rather it requires the analysis of a multitude of variables and factors. The research presented in this dissertation therefore also enables a deeper understanding of the complex conditions and constraints associated with visual information identification, the accurate capturing of user requirements, the correct expression and adequate complexity of user queries, the execution of searches and the reliability of performance indicators.

These contributions make possible a systematic calibration and comparison of system performance for (multilingual) visual information retrieval from generic photographic collections. In particular, significant contributions to the benchmark component of *image collections* include:

- The design for the creation of an image collection of real-world photographs that is specifically developed for image retrieval evaluation [329]; this includes the specification of the collection as well as a framework for image selection and annotation rules [147].

- The implementation of the collection following the aforementioned framework, and the creation of highly qualitative (alphanumeric) image representations in three languages: English, German and Spanish.

- The data collection being available to the research community free of charge and without copyright restrictions that would hinder the redistribution for large-scale evaluation events [146].

One of the key contributions in this research is the introduction of a *parametric benchmark architecture* and the subsequent design and implementation of a benchmark administration system to support this paradigm and to facilitate and guide the use of the individual benchmark components with respect to these parameters. Individual contributions comprise:

- The identification of benchmark collection parameters with respect to the organisation of the images and the management of the corresponding semantic (alphanumeric) image representations.

- The design and implementation of a benchmark administration system [148] to support this parametric benchmark paradigm, facilitating the quick reaction to changes in research directions and the adaptation of the test collection to the specific needs required by different events (*e.g.* expressed by participants' feedback).

- The implementation of an export function to automatically generate subsets of the collection and semantic image representations with respect to predefined parameters; this allows for the subsequent distribution of the subsets without participants having to access the original database.

- The design and development of a topic management system to facilitate the creation, administration, translation and generation of search topics [148].

To improve and facilitate the non-trivial task of *topic creation* for evaluation of visual information retrieval from multilingual collections of real-world photographs, we analyse and advance several of its dimensions:

- The design and implementation of a framework for the topic creation process in order to model a natural, balanced set of representative search queries accurately reflecting real-world user statements of information needs; this includes the identification of several topic dimensions according to which the topics were selected [61].

- The design and validation of a measure for one of these dimensions, topic difficulty [149]; this difficulty measure is a vital dimension for such evaluation events as it allows not only for the control of task difficulty within such an event, but also for the comparison between different events and various data collections.

- The execution of a user need analysis to identify the search behaviour and query patterns for the benchmark collection and to base the topic creation process on realistic and representative topic candidates [61].

Minor contributions made with respect to performance quantification (*e.g. relevance assessments* and *performance indicators*) include:

- The integration of a relevance assessment module into the benchmark administration system.

- The analysis and evaluation of several performance measures, and based on it, the recommendation of a set of performance measures.

Finally, significant achievements in terms of *evaluation events* comprise:

- The organisation of the first large-scale evaluation event for multilingual retrieval from a generic photographic collection.
- The analysis of the retrieval performance of over 150 submitted runs with respect to several submission parameters and topic dimensions.
- The analysis of feedback from participants to evaluate the existing evaluation event and to improve and guide the organisation of future evaluations.

Further literature that was published in the frame of this research includes [144, 145, 150, 234].

Chapter 2

Characteristics and Processes of Visual Information Retrieval

This chapter takes a look into the concepts, characteristics and processes of *visual information retrieval* (VIR) and provides a number of references for various aspects of this field of research. Being the first of the two main literature review chapters, it provides the vital theoretical foundation to understand the functionality of *visual information retrieval systems* (VIRS). The state-of-the-art methods regarding the analysis and evaluation of these systems are then presented in the second part of the literature review in Chapter 3.

First, a general introduction to VIR is presented in Section 2.1, illustrating the evolution, goals and challenges of VIR as well as introducing the main components of a VIRS. Then, Section 2.2 elaborates on the characteristics of visual information queries and provides answers regarding the origin, intentions, modalities and classifications of these expressions of information need in the visual domain.

The core of this chapter can be found in Sections 2.3 and 2.4, which provide a comprehensive overview of the visual and textual features that can be extracted from images. These features are expressed by so called *descriptors* and form the basis for the result generation and presentation process, which is further illustrated in Section 2.5. As already indicated in the introduction, the main focus of this research is on ad-hoc retrieval performance, which is also reflected within this chapter. Thus, other image retrieval aspects such as relevance feedback, interaction speed or usability issues are only peripherally covered in Section 2.6.

Finally, a selection of well-known and often-cited image retrieval systems will be presented in Section 2.7, together with an explanation of their impact on the domain of VIR in general and their influence on the research presented in this dissertation in particular.

2.1 Visual Information Retrieval

This section provides an introduction to visual information retrieval (VIR): it starts with a short glance at the development and evolution of VIR and presents links to some of the most-cited overview articles within the respective field of VIR; then, the general image retrieval process is illustrated and the main components of a VIR system (VIRS) are introduced; finally, some current problems and challenges of VIR are addressed.

2.1.1 Evolution and Overview

The field of VIR first received active research interest in the late 1970s and has been a very active research area ever since, with the driving forces originating from two major research communities: Computer Vision and Database Management. This was not a combined effort, however, as both research fields approached the VIR problem from opposite perspectives, one being content-based and the other one concept-based.

Concept-Based Image Retrieval

The *concept-based image retrieval* approach can be traced back to the late 1970s. It is difficult to reveal when it was exactly started, although a conference on *Database Techniques for Pictorial Applications* held in Florence in 1979 [31] presents an often-cited starting point [249, 287] for VIR in general and concept-based retrieval in particular. These early approaches driven by the *Database Management* community were not based on visual features, but on the idea to first annotate the images by text and to subsequently use *database management systems* (DBMS) for image retrieval based on text [48]. A review of early concept-based retrieval methods can be found in [50]; an overview of image database applications with more than a hundred references from 1983 is presented in [439]; an updated version of research in image database management is provided in [10]; and a survey of the state-of-the-art (1992) image information systems is given in [49]. Image retrieval based on text, however, faces several problems:

- The automatic generation of descriptive texts for a wide spectrum of images is not yet feasible. Thus, a huge amount of labour is required to manually annotate the images, which is obviously an expensive and cumbersome task, especially when the size of collections is large.
- Manually created logical image representations (also called image *annotations*) are by nature subjective, context-sensitive and often incomplete. The subjectivity of human perception and the rich contents in images imply that different people might perceive images in different ways; this combined with the aforementioned imprecision in the annotation process may cause irrecoverable mismatches in subsequent retrieval approaches.

Content-Based Image Retrieval

When these limitations became more and more severe with the emergence of large scale image collections in the early 1990s, *content-based image retrieval* (CBIR) techniques were proposed as a potential solution to the problem. CBIR was first mentioned in 1992 [205] and is based on visual features that are inherent in the images themselves without having to rely on textual representations.

The *Computer Vision* community was hereby the driving force for advances in CBIR, making it one of the most active research areas of the late 1990s and the early 21st century. Some early review articles and surveys on CBIR include [6, 151]; another early review on image databases and pictorial information retrieval can be found in [104].

The amount of scholarly literature on image retrieval techniques had soon increased enormously, especially since 1997. The following articles represent some of the most cited and influential overviews, reviews and surveys of the late 20th century: Smeulders [405] and Eakins [99] both give excellent and very comprehensive reviews of the developments in the field; another critical overview of the developments in the field can be found in [139]. Venters [463] and Eakins [99] provide comprehensive reviews of CBIRS. Rui [368] elaborates on the past, present and future of image retrieval. Gudivada [153] discusses the status quo (1997) of information retrieval in the WWW. Rasmussen [353] provides an overview of image indexing and textual search in image databases, while White [487] illustrates further annotation issues.

More recent work includes the dissertations published by the *VIPER Group* [287, 307], which both provide good resources for image retrieval techniques, especially as regards CBIR, including several hundred citations. Veltkamp [461] presents a very detailed description of the functionality of more than 50 CBIRS. Long [249] introduces some fundamental theories for CBIR as introduction for a book on multimedia retrieval and management [114]. Datta [77] talks about approaches and trends of CBIR after 2000, while Lew [237] discusses major challenges of multimedia information retrieval for the future based on its current state of technology (2006).

2.1.2 Retrieval Process and Components

Although the VIR problem was approached from two opposite sides of research, the main functionality of an image retrieval system is in both cases very similar. Figure 2.1 illustrates the main functionality of a VIRS.

The main goal of a VIRS is certainly to "find an image or a set of images that a user is searching for within an image database or an image collection" [287]. The components of such a VIRS and their contributions towards this goals are explained hereinafter.

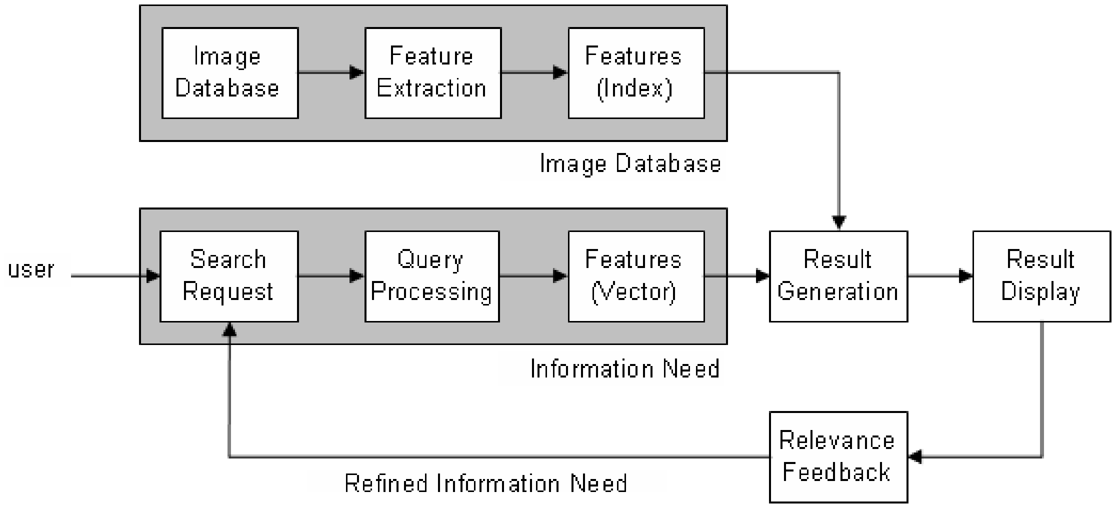


Figure 2.1: Overview of visual information retrieval processes.

Image Database

Most image retrieval systems work with closed image databases that are indexed offline. That is, visual content descriptors are extracted for each image and described in multidimensional feature vectors that can be stored in the so-called *feature index.* Section 2.3 will explain in more detail which visual features can be extracted and how they can be stored in the feature index.

If semantic image representations exist, textual content descriptors can also be extracted and indexed, and several statistics for each of these representations as well as for the entire collection can be precalculated. Section 2.4 will elaborate on textual descriptors and their corresponding statistics.

Information Need and Query Processing

There are several ways users can express their information needs in image retrieval systems (see Section 2.2.1). While these information needs are the same in both content-based (visual) queries and concept-based (text) queries (users want to retrieve relevant images), there are differences in the way the queries are processed.

For content-based queries, the system processes the query image(s) and changes them into the internal representation of feature vectors (see above). This, again, comprises the extraction visual content descriptors of the query images and the storage in multidimensional vectors.

Most concept-based searches also require additional query preprocessing before their elements can be used for retrieval. This includes the use of stemmers in order to avoid vocabulary mismatches caused by morphological variants of the search terms, and query expansion techniques such as the use of thesauri or ontologies.

Result Generation and Presentation

Based on the feature index and statistics provided by the database and the visual and textual descriptors extracted from the query (text and images), the system can then calculate the similarities (or distances) and rank the images in the database accordingly. These rankings of relevant images are carried out on the basis of similarity measures, with many of the different content descriptors requiring a certain type of measure. Both concept-based and content-based similarity measures are further described in Sections 2.5.1 and 2.5.2. Some systems even use a mixture of concept-based and content-based characteristics to improve retrieval results.

The result generation process of both concept-based and content-based retrieval paradigms is, in principle, based on quite different models. The original aim of traditional text retrieval systems was to partition a database into two sets: *relevant* and *non-relevant* documents, even if members of the first set were later ranked by relevance [384]. Yet, such a clear-cut partition of the database is not representative for most current IR approaches any more, and definitely not feasible for for CBIRS in general; on the contrary, their prime aim is to only sort the database in order of the similarity to the query.

Once a system has ranked the images in a database for a specific query, the results are then displayed to the user (see Section 2.5.4).

Relevance Feedback

Many image retrieval systems allow the user to refine the retrieval results by providing information on whether a retrieved image is relevant (positive example) or irrelevant (negative example) to the search request in hand. Based on this feedback, a refined information need is created and the entire image retrieval process is started over again, albeit with slightly different parameters as reflected by the positive and negative examples.

Relevance feedback thus reflects a supervised active learning technique in order to improve the effectiveness of an information system, which is particularly useful due to the incapability of most systems to match users' needs accurately the first time round. Section 2.6.1 provides more information on relevance feedback.

2.1.3 Current Challenges

The two most prominent challenges in image retrieval are illustrated in Figure 2.2 [287]: the *sensory gap* and the *semantic gap*.

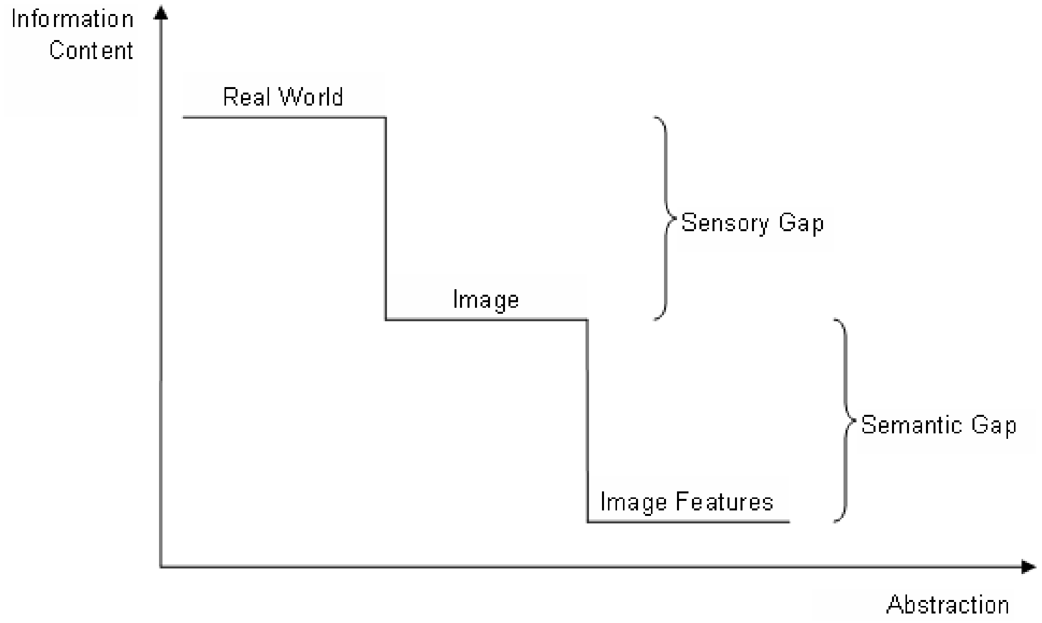


Figure 2.2: Sensory gap and semantic gap.

Sensory Gap

[405] defines the *sensory gap* as the "gap between the object in the real world and the information in a (computational) description derived from a recording of that

scene". In other words, when an image is produced, some information that was present in the real world is automatically lost. This loss of information can be due to missed details because of a too low resolution, partially occluded objects, bad illumination or viewing angles, or any imperfectness of the image capturing device (*e.g.* a camera).

Semantic Gap

The *semantic gap* is defined as the "lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [405]. This reflects the difference between the visual low-level features exhibited by an image and the semantic (objects, relationships, meanings) and abstract (feelings) richness of that image as perceived by a human.

[164] further divides the semantic gap into the gap between the visual descriptors and the object levels (which will be called *lower semantic gap* hereinafter) and the gap between the labelled objects and the full semantics of an image (*upper semantic gap*).

2.2 Visual Information Queries

The first step in the VIR process is the specification of a query in order to satisfy a certain *user information need*. The following questions may arise: Where do such information needs come from? How can such information needs be expressed when using an image retrieval system? And is it possible to classify these information needs?

This section provides an answer to each of these questions: Section 2.2.1 discusses potential image users and the application of VIR in numerous and diverse domains; Section 2.2.2 illustrates several possibilities for the specification of an information need in a VIRS; and Section 2.2.3 presents several classification schemes for search requests.

2.2.1 Identification of User Information Needs

Still images are playing a vital role in all walks of life these days and are extensively used for both recreational use as well as in many aspects of professional life. Across-the-board, images are generally required for a variety of reasons and in relation to past, present and future. For instance, they are used:

- to display *past* data for detailed analysis (*e.g.* radiology images) or record purposes (*e.g.* photographs of an event),
- to illustrate text articles to express *present* emotions or convey any other information that cannot be described in words, or
- to formally record design data (*e.g.* architectural plans) for *future* use.

The high popularity of images automatically creates a need for retrieval mechanisms of images in a repository. Naturally, different professions and user communities exhibit different motivations, attitudes and interaction styles and will hence have specific expectations and requirements for such retrieval mechanisms; [8] provides a good overview of information needs in the visual document domain.

While it is certainly impossible to give a full picture of the uses made of visual information, and a more detailed sociological study of image use and retrieval would be quite out of place in this dissertation, it seems appropriate to provide an introduction to the specific user needs in a selection of representative domains in which both content-based and concept-based retrieval mechanisms are frequently applied. The most relevant for this thesis are VIR in home entertainment and private use as well as in web searching. More comprehensive surveys of potential application areas for VIR in general, and CBIR in particular, can be found in [99, 151].

Home Entertainment and Private Use

The majority of home entertainment deals with images and videos, *e.g.* family photos, holiday snapshots, home videos or recorded scenes from movies or television programs. Digital cameras have resulted in very large collections of private digital photographs, thus computerised systems for the storage of, future browsing in and retrieval from these collections are becoming of increasing significance.

Studies like [363] show that most home users find it easier to manage, organise and search their digital than their non-digital photographs. In general, photographs of the same event (*e.g.* a holiday) are put into one directory. These directories are then named after that particular event, while the photos within in that directory, which are ordered chronologically by default, often remain unnamed. According to the study, three basic types of search requests can be identified for such private photographic collections:

- a set of photos from a particular event (*e.g.* "find pictures from my trip to Spain");
- an individual, remembered photo (*e.g.* "find that photo of my friend Fernando with Marat Safin at the 2007 Australian Open");
- a set of photos taken at different events, but containing a common property like a certain person or a certain activity (*e.g.* "find all photos of my mum", "find all the photos in which I am playing soccer").

For the first two cases, simple browsing based on event and time is usually enough to let users find what they are looking for, as most of them are familiar with their own collections. Yet, this approach is only feasible as long as photos were taken recently and collections are rather small. As collections grow and the photos get older and less familiar, many details (such as names of unfamiliar places) are normally forgotten, and semantic image representations (and concept-based image retrieval) become increasingly significant. This is also true for the third type of search request mentioned above, because without these textual representations one would have to browse the entire collection for such general queries [363].

Web Searching

The need for effective search of text and images on the WWW is an application that cuts across many of the other domains. The rapid development of the WWW in the last ten years has resulted in an indispensable source of information available online these days, making effective concept-based search engines such as *Google* and *Yahoo!* an essential part of daily life. Although the need for image search tools of similar power was already expressed more than ten years ago, for example [189], the state-of-the-art CBIR techniques nowadays are still far from being as well-matured as text-retrieval techniques. Hence, basically all these search engines offer a text-based interface for their image searches (see Section 2.2.2).

The search behaviour for images on the web has been the subject of several investigations, including [138, 192, 422, 423]. Ironically, there is not only a need for software to retrieve images, but also to prevent access to images like the ones that are deemed to be of pornographic nature [99, 121, 420, 421].

Journalism and Advertising

The publishing industry extensively uses images to illustrate books as well as articles in newspapers and magazines. These photographs and pictures gathered over the years are stored in often very large archives that keep on growing day by day. Journalists these days face immense competition and often live by the tyranny of the clock, thus efficient retrieval from these archives can be vital for the success of an article.

Studies like [265, 266, 325] give more insight into the search behaviour of journalists. In most cases, stock shot agencies base their retrieval systems on logical (alphanumeric) image representations, which allow for efficient semantic image retrieval but make these archives expensive to maintain. CBIR is sometimes used to improve retrieval results.

Cultural Heritage and Historical Research

Historians from various disciplines as well as archeologists often rely on visual information sources to support their research activities. This is especially true for the field of arts and archeology, where the access to original works or findings is often restricted because of their geographic distance or ownership regulations, or even impossible due to their physical condition or other circumstances (like theft) - if so, the visual record might be the only evidence available.

Researchers can hence more often than not only refer to surrogates in form of photographs, which are generally collected by museums, art galleries and libraries and have been made available for consultation in digital form in recent years. Comprehensive reviews for the search behaviour in these collections include [125, 272, 400].

Medicine

Due to the increasing reliance of modern medicine on diagnostic techniques, many hospitals have experienced an explosion in the number and importance of medical images such as x-rays, computer tomography (CT), magnetic resonance (MRI) and ultrasound images. Medical images are normally assigned to (textual) case descriptions, which are stored with a person's health record to allow for efficient data recovery in the future.

Studies like [207] indicate that such searches can include the query for a patient's medical history or the investigation of interesting cases in terms of symptoms, diagnoses and treatment aspects; all these requests are predominantly satisfied by the use of concept-based image retrieval. However, there is also an increasing interest in the use of CBIR techniques for the clinical decision-making process, as the visually-based identification of similar past cases might aid diagnosis and further treatment.

Intellectual Property

The protection of intellectual property has long been recognised as one one the prime application areas for CBIR. In trademark image registration, for example, it is crucial to ensure that there is no risk of confusion between existing trademarks and newly introduced ones. CBIR is well-suited for such kind of tasks, especially shape matching algorithms [97, 98].

Other Application Areas of VIR

There are several other areas in which VIR (and in some cases CBIR) has made (or is likely to make) a fruitful contribution. In *crime prevention*, for example, efficient methods for face recognition, automatic fingerprint matching or the detection of pornographic images on the Internet constitutes a vital part of everyday crime solving and prevention processes. In *education and training*, the search for good teaching material to illustrate key points in a lecture or course has become an indispensable part of teachers' and lecturers' preparations. Furthermore, *Architectural and engineering designs* often share a number of common features, thus designers in these professions have to be aware of existing templates or previous designs in case they can be adapted to a problem at hand in order to avoid the time-consuming reinvention of the wheel.

2.2.2 Query Specification

Once a certain information need is identified, the next question arises: How can such an information need be expressed in an image retrieval system?

The following subsections present several query starting points in order to specify what kind of images one wishes to retrieve from a database. The most relevant of these regarding the scope of this research are *concept-based searches* and *query by example.* For a rather formal illustration of the query specification problem, see [405].

Concept-Based Searches

Concept-based image searches allow the user to express an information need by entering text [81, 390]. Such information needs are, in many cases, of a semantic nature; hence, text retrieval searches might represent the most natural way of querying [287].

Several methods of specifying a text query can be identified. *Free-text* searches allow the constraint-free specification of either one or several keywords, search phrases or entire sentences. *Boolean text* searches make it possible to specify own search rules: for example, more significant query terms can be assigned more weight, or non-relevant items for an information need can explicitly be ignored. In *structured text queries*, the text can often be entered in separate fields for exact matches, which can be the case in "advanced" searches, when using databases or structured image representation formats.

Basically all the search engines on the WWW such as *Google* or *Yahoo!* offer a concept-based interface for their image searches: they allow the input of free-text, which is then used in a keyword-based image search whereby the individual query terms are matched with the text nearby images on a web page, their filenames and/or logical alphanumeric representations respectively. Concept-based query interfaces have also frequently been used for most image databases in museums [389], libraries [199], historic photographic collections [66, 355] or image archives used by journalists [266].

While text-based retrieval is a well-matured field of research and can therefore be a very efficient means for concept-based image retrieval as well, such retrieval success is still heavily dependent on good semantic representations of the images.

Query by Example(s)

An image retrieval system employing the *query by example (QBE)* paradigm uses one or more sample images as the starting point for the search of visual information: first, a selection of content-based features (see Section 2.3) is extracted from the sample image; secondly, these features are used as a basis to search for images in the database with similar features (see also Section 2.5.2). The majority of CBIRS like [87, 129, 307, 336, 428, 455] use image examples as their query starting point.

QBE can be classified into *query by external image example* (if the query image is not in the database) and *query by internal image example* (*i.e.* the query image is taken from the database itself). In the latter case, all relationships between the image can theoretically be precalculated [462].

A different classification is used in [287], dividing QBE into methods using user supplied examples and examples proposed by the system. *User supplied examples* can comprise internal images from the database (that the user has to find) or not yet indexed external images that the user possesses. *System supplied examples* are always internal images from the database[1]. Systems that use system supplied examples include [72, 73, 311].

When more than one query image is used, the search is based on the common feature characteristics of the group of query images; this can lead to a better specification of the relevant features and at the same time might remove irrelevant variations in the query. It is further possible to refine these group features by the addition of negative examples (*i.e.* images that are not relevant to a certain information need) [249].

The main benefit of the QBE paradigm is that the image descriptions are implicitly calculated by the system; hence, the user is not required to provide an explicit description of the information need for an image. This search approach is mainly suitable for image recovery needs in which images of the same object (or same set of objects) under different viewing conditions are required.

Other Starting Points

The *query by image regions* paradigm is based on the segmentation of all the images in a collection and was used, for example, in [40, 41, 196]. Some CBIRS such as [117, 205, 481] allow the formulation of a visual information need by drawing a *sketch* of an image. Other potential starting points include the *query by spatial relationships* paradigm (*e.g.* "a green round object above a triangular, yellow object"), which is further described in [152, 406] and was used in [40], and the *query by gesture* (QBG) paradigm, which is further illustrated in [210].

[1] There are many systems that offer the user a random set of images to start a query; although these images originate internally from the database, they do not qualify as system supplied examples because the appropriate images that are relevant to one's information need still have to be selected by the user.

2.2.3 Classification Schemes

Although user information needs originate from a vast variety of professions and users and therefore cover very different image collections, many researchers have suggested various classification schemes to generally group the most common types of such search requests.

Enser and McGregor [103, 106] were among the first researchers to attend to such an investigation (1992) and categorised image search requests put to a large picture archive into *unique* and *non-unique* queries. Unique queries are those which can be satisfied by the retrieval of a unique person, object or event (*e.g.* David Beckham, The Titanic), while non-unique ones cannot (soccer players, ships, penalty shootout). Both groups are further subject to refinement in terms of time (a young David Beckham), location (David Beckham at Old Trafford), action (David Beckham missing a penalty), event (David Beckham at the World Cup) or technical specifications. A non-unique query such as "Olympic Games" could also be modified to create "the 1984 Olympic Games in Los Angeles" (unique), refined by location and time.

In 1997, Armitage and Enser [9] used this categorisation scheme on queries from seven picture libraries. They concluded that they found it too broad and adopted a more structured refinement method based on the *Panofsky/Shatford mode/facet matrix* of image analysis. Panofsky [332] had categorised fine art images based on the following three modes:

- *pre-iconography* complies to general image requests.
- *iconography* describes a picture's actual subject matter and complies to specific image requests.
- *iconology* describes a picture's deeper artistic or religious meaning and complies to abstract image requests.

Apart from these general classification attempts, researchers in more specific areas such as journalism [265], art history [169] or medicine [207] also tried to categorise search requests in their fields. However, since these findings are restricted to a very specific domain, they are not generally applicable.

In 2000, Eakins [99] proposed the following three levels of VIR:

- *Level 1:* The lowest level of VIR is based on *primitive features* such as colour, texture, shape, spatial location of image elements, or a combination of these.
- *Level 2:* This level comprises retrieval by *derived attributes* or *semantic content* and corresponds to Panofsky's pre-iconographic level of picture description. Search requests on this level include the retrieval of objects of a given type or class (*e.g.* a car on a street) as well as the retrieval of individual objects or persons (*e.g.* David Beckham, The Titanic).
- *Level 3:* This level comprises retrieval by *abstract attributes* and includes search requests for named events or types of activity, corresponding with iconography (*e.g.* photos of Scottish folk dancing), and search requests for pictures with emotional or symbolic significance, corresponding with iconology (*e.g.* find a photo depicting "grief").

Müller [287] goes one step further and distinguishes five levels of image retrieval: *lower* (retrieval based on global colour or global texture), *low* (retrieval based on local colour, local texture or shape features), *middle* (based on shapes of segmented form and weak segmentation), *high* (based on real objects, persons, places, and strong segmentation), and *higher* (based on invoked feelings and semantics that are very personal to a certain user). Furthermore, Jörgensen [200] presents a conceptual model in the form of a pyramid that distinguishes between not less than 10 levels of image retrieval.

In the context of this dissertation, the distinction of so many different retrieval levels would present a slight overkill. Henceforth, the following rather simple distinction of VIR will be applied within this dissertation:

- *Content-Based Image Retrieval (CBIR)* is based on visual features (such as colour, texture and shape) that can be directly extracted from images without having to rely on meta-data or semantic text representations. CBIR mainly corresponds to *level 1* according to Eakins, to the *low*, *lower* and *middle* levels according to Müller and to the top 4 (syntactic) levels of Jörgensen's pyramid.

- *Concept-Based Image Retrieval*, often referred to as *Text-Based Image Retrieval (TBIR)* or recently also as *Semantic Image Retrieval (SIR)*, is based on alphanumeric features such as logical image representations, which are at this stage of research the only way to retrieve all the other semantic and abstract levels of the aforementioned classification schemata: *levels 2* and *3* according to Eakins, the *high* and *higher* levels according to Müller and the bottom 6 (semantic) levels according to Jörgensen.

The next two sections 2.3 and 2.4 will elaborate on the visual features that form the basis of CBIR and on the textual features that form the basis of TBIR[2] respectively.

2.3 Content-Based Image Descriptors

This section presents background information on visual feature extraction from images, which forms the basis of CBIR: visual features can be directly extracted from the image data itself, without having to rely on the use of concept-based features such as keywords. Concept-based feature extraction will be discussed in Section 2.4.

2.3.1 Descriptor Classification and Requirements

Due to the extensive research on CBIR over the past 15 years, it is impossible to list all the published methods without going beyond the scope of this dissertation. The following sections will therefore provide concise introductions to the main concepts of visual feature extraction, together with some key references. Further, most fundamentals were laid out in the late 20th century, which is also reflected by the literature chosen within this section.

[2] There is no officially recognised abbreviation for Concept-Based Image Retrieval. To avoid confusion with Content-Based Image Retrieval (CBIR), we will hence use TBIR to refer to the Concept-Based Image Retrieval paradigm hereinafter.

Classification of Visual Features

In general, visual features can be classified as *domain-specific* and *general* features. Domain-specific features are application dependent and may include, for instance, human faces or fingerprints; these may involve quite a lot of domain knowledge and are well covered in the Pattern Recognition literature. This section will therefore only concentrate on the second category, the *general features* like colour, texture and shape, which can be used in most applications.

Furthermore, visual content descriptors can be either *global* or *local* descriptors: while global descriptors consider the visual features of the whole image, local descriptors only regard the visual features of objects or regions to describe the image content.

Image Segmentation

In order to compute a local descriptor, the image has to be segmented into parts first. In general, segmentation methods can be classified as follows [405]:

- *Partitions* are the simplest way of segmentation as an image is divided into fixed divisions of the same size and shape, regardless of the data. Although this approach is computationally easy, it does not create regions that are perceptually very meaningful. Nevertheless, partitions have been used for image retrieval, for example, in [428, 483].

- *Sign detection* searches an image for the location of a specific geometric shape with a certain semantic meaning.

- *Region segmentation* divides the image into regions which are internally homogenous according to some criterion (*i.e.* regions with similar properties). This approach, also called *weak segmentation*, has been used by many image retrieval applications such as Blobworld [42].

- *Object segmentation* divides the image into semantically meaningful objects of the real-world. This approach, also called *strong segmentation*, is the most complex form of segmentation and is, at this stage of research, restricted to very limited domains of images [287].

Some pessimistic voices predict that object segmentation (and, as a consequence, *object recognition*) is unlikely to ever succeed at all [249]; and, indeed, the segmentation of images in the context of human perception is far from being solved, despite the publication of some interesting and novel approaches like the segmentation based on the theory of *spectral clustering* [397]. Other recent approaches include the segmentation based on the *mean shift procedure* [67], an expectation-maximisation (EM) based segmentation using a *Gaussian mixture model* [42], the multi-resolution segmentation for images with low depth of field [479], and the segmentation based on a Bayesian framework incorporating the *Markov chain Monte Carlo* technique [448].

Visual Descriptor Requirements

Good visual descriptors should be able to deal with invariances that may be introduced by the imaging process. These invariances can comprise changes in *scale*, *shifting* and *rotation* as well as varying *lighting* conditions, different *viewpoints* or certain *deformations* of the objects in an image.

While humans can still recognise objects and patterns despite such changes, these invariances often mean a loss of information when using computers. Ideally, visual descriptors would be invariant to these changes, but in reality there is a certain tradeoff between the invariance and the discrimination power of a descriptor: the more variance independent a descriptor, the lower is the ability to discriminate between essential differences.

Several aspects of colour invariance are covered in [128]; a survey on shape analysis techniques [248] discusses invariant aspects as well; and [424] provides an overview of invariant pattern recognition methods.

2.3.2 Colour Descriptors

Colour constitutes the most significant feature in an image [433, 453]; it has been used in basically all retrieval systems [42, 43, 69, 117, 126, 129, 327, 428] and has formed an active area of research in image retrieval more than in any other branch of computer vision. The importance of the colour descriptor may be attributed to the superior discriminating potentiality of its three-dimensional domain compared to the single dimensional domain of gray-level images [249, 405].

The colour feature extraction comprises two aspects: the choice of an appropriate *colour space* to quantify the values of the individual pixels of an image, and a *representation model* to describe the distribution of these pixel values for the entire image. This section introduces both aspects and provides examples of the most significant colour spaces and representations. For a more comprehensive discussion on colour-based retrieval, see [385, 394]. A description of the colour descriptors included in MPEG-7 can be found in [261].

Colour Space

Each pixel of an image can be represented as a point in a three-dimensional *colour space*, which is a task that includes the following two challenges: first, the recorded colours of the image might considerably vary with:

- the viewpoint of the camera,
- the position of the illumination,
- the spectrum of the light source,
- the orientation of the surface of an object, and
- the way the light reacts with an object,

and ideally, a colour space should somehow be able to deal with this variability [249]. Second, an appropriate colour space for image retrieval should also guarantee its uniformity; in other words, the mathematical distances between colours should correspond to how viewers perceive that difference, and if a colour pair is equal in distance, they should also be perceived as equal by viewers [270].

RGB. The *RGB colour space* desribes an image in its literal colour properties *red (R)*, *green (G)* and *blue (B)* and is widely used for image display [287, 405]. These three components are also called *additive primaries* because a colour in the RGB space is produced by adding them together when light is emitted (like on a computer screen or on the television). Although the human visual system is also based on red, green and blue receptors, this colour space is only rarely used for retrieval as it is perceptually non-uniform and might only be sensible for the retrieval of images recorded in the absence of variance, as is the case *e.g.* for art paintings [169] that are usually recorded in frontal view under standard conditions, or trademarks [97, 98] for which colour only plays a limited role.

CMY. The *CMY colour space* uses the three colour components cyan (C), magenta (M) and yellow (Y) and is primarily used for printing, where colours are mixed in a subtractive way through light absorption; its three components are therefore also called *subtractive primaries* [249, 287]. Like RGB, the perceptual properties of the CMY colour space do not correlate with human perception, and due to the lack of uniformity it is rarely used for image retrieval.

Opponent Colour Space. The *opponent colour space* is a significant improvement over the RGB and CMY colour spaces, at least as far as image retrieval is concerned [436]. It uses the opponent colour axes (R-G, 2B-R-G, R+G+B) and has the advantage of isolating the brightness information on the third axis. As a consequence, since humans are more sensitive to brightness than to chromatic information, it is possible to down-sample the first two chromaticity axes which are invariant to changes in illumination intensity and shadows.

HSV. The *HSV colour space*, also known as *HSB colour space*, presents a slightly more intuitive way of describing colours and is widely used in computer graphics and image retrieval [40, 42, 327, 428, 483]. It uses the following three components:

- *hue* ($0° \leq H \leq 360°$), sometimes normalised to ($0 \leq H \leq 100$), represents the colour type by quantifying its angle on the colour circle (*e.g.* 0° = red, 120° = green, 240° = blue);
- *saturation* ($0 \leq S \leq 100$) describes the vibrancy (or purity) of the colour: the lower the saturation, the more “grayness” is present and the more faded a colour appears;
- *value* ($0 \leq V \leq 100$) specifies the *brightness* (B) of a colour.

The advantage of this representation lies in the invariancy of the hue component to changes in illumination or camera directions, making it more suitable for object retrieval. Further, the three components (H,S,V) can easily be determined from images stored in the RGB colour space [119]. Sometimes, less importance is given to V [428], or it is even completely omitted [327], due to its sensitivity to lighting conditions.

CIE L*u*v*, L*a*b*. The *CIE L*u*v** and *CIE L*a*b* colour spaces* [357] were developed by the *International Commission on Illumination*[3] with the intention to produce a colour space that is more perceptually linear than other colour spaces; both are hence perceptually uniform and device-independent and thus well suited for retrieval [287]. The three parameters in the model represent:

- the *lightness (L*)* of the colour: $0 \leq L* \leq 100$, where $L* = 0$ yields black and $L* = 100$ indicates white;
- its *position (a*) between magenta and green*: $-128 \leq a* \leq 127$, where negative values indicate green and positive values indicate magenta;

[3] Commission Internationale d’Eclairage, CIE, `http://www.cie.co.at/cie/`

- its *position (b*) between yellow and blue*: $-128 \leq b* \leq 127$, where negative values indicate blue and positive values indicate yellow.

CIE L*u*v* is designed to deal with subtractive colourant mixtures, while CIE L*a*b* is designed to deal with additive colourant mixtures [249]. Both colour spaces can be calculated from the RGB colour space, however the transformation is non-linear and therefore more complex than that of HSV [280]. The precise creation process and more detailed information is illustrated in [357]. Image retrieval systems that employ CIE L*u*v* include [391, 441], while [318], for example, operates with CIE L*a*b*.

Other colour spaces. The *hue-min-max-difference* (HMMD) colour space, in which *hue* has the same meaning as it has with HSV, and *max* and *min* are the maximum and minimum among the R, G and B values respectively, was rigorously tested for inclusion in the MPEG-7 standard, together with RGB, HSV and *YCbCr*, which is a family of colour spaces used in video systems [261]. The *Munsell colour space* is based on the three components hue, value (lightness) and chroma; it is considered to be perceptually uniform and is used by QBIC [117] for its feature representation. A number of invariant colour descriptors is also listed in [128].

Colour Representations

There are several representation models to describe the colour distribution of an image once an appropriate colour space is chosen, *e.g.* colour moments, colour histograms, colour coherence vectors and colour correlograms.

Colour moments are a very simple but compact model that has been used in several retrieval systems including [117, 327]. Its first order (mean), second order (variance) and third order (skewness) colour moments for each of the three colour components (thus only 9 values in total) provide an efficient representation of the colour distribution of images [17, 428, 433]. Since its compactness might imply a lower discrimination power, colour features are often used in a first step to narrow down the search space before other more sophisticated colour features are applied for retrieval [249].

Colour histograms show the proportion of pixels of each colour within an image [436] and have been used in many systems [327, 428]. They are very easy to compute and robustly tolerate the movement of objects within an image and changes in camera viewpoint; they can hence serve as an effective representation of the colour content of an image if its colour pattern is unique in comparison with the rest of the data set. However, the discrimination power of this technique is reciprocally proportional with the number of images in a database: the larger the database, the higher is the likelihood that very different images can have similar colour distributions [333].

As a consequence, simple histogram matching techniques like the *global histogram intersection* [436] do not suffice any more in order to compare the colour features of two images; hence, several improved variations of this technique have been introduced, including *joint histograms* [334], *cumulative colour histograms* [433], region-based colour querying [40] and histograms over multiple scales [126].

But to really solve the problem, the combination of histogram intersection with some sort of spatial information seems necessary [432], as is the case with *colour coherence vectors* (CCV) [333], in *colour correlograms* [178] or in region-based colour descriptors combined with their percentage coverage within these regions [83].

Recent Developments and Alternative Approaches

Some recent developments include *colour descriptor matrices* [453], multi-resolution histograms capturing spatial image information [157] and colour descriptors in wavelet domain [452].

In a slightly alternative approach, [69] attempts to express high-level colour semantics (such as whether a colour is perceived to be warm or cold, or which colours are in harmony or disharmony with each other) to allow retrieval of images evoking a particular mood.

2.3.3 Texture Descriptors

Texture is another fundamental property of images and can be defined as "all that is left after colour and local shape have been considered or in terms of structure and randomness" [405, page 1356]. While image retrieval based on global texture similarity is not considered to be very practical [99] and might just suffice for very specific tasks such as the search for satellite images [238] or images of documents [75], it can be very useful for the distinction between areas with a similar colour (like sea and sky, leaves and grass, *etc.*). Furthermore, the extraction of texture features is normally only meaningful for homogenously textured regions [287], although global texture measures have sometimes been used for general, heterogenous collections of photographs as well [428].

A wide variety of techniques has been developed to characterise texture and to measure texture similarity (between the query and target images). These texture representation methods can be classified into four groups [450]:

- *Statistical methods* such as *co-occurrence matrices* or the *autocorrelation function* describe texture by the statistical distribution of the intensity values within a region of interest.

- *Geometrical methods* characterise the texture by feature primitives and spatial arrangements and tend to be most effective when applied to textures that are very regular. Examples of this category are the *Voronoi tessellation* features or the use of *Laplacian of Gaussian* (LoG) masks.

- *Model-based methods* use mathematical models to describe and synthesise the texture of an image. This includes random field models using the *Markov* (MRF) and *Gibbs* (GRF) *random fields*, *auto-regressive* models as well as *fractal* models.

- *Signal processing methods* analyse the frequency content of an image and use a certain set of filters in order to compute texture features. Examples include *spatial domain filters*, *Fourier domain filters* and multi-resolution filtering techniques such as the *Wavelet* and *Gabor transforms*.

This classification, however, is not without criticism: Long [249], for example, finds it too broad and only makes a distinction between *statistical* and *structural* methods. Further, Zhou argues in his recent dissertation [504] that many methods, as a matter of fact, do fall into several categories. A Markov-Gibbs Random Field (MGRF) model, for instance, would stride over both model-based and statistical categories as it derives a joint probability distribution on statistical image features for texture description. He therefore presents a different categorisation: *descriptive* approaches (*i.e.* statistical and spectral methods) and *generic* approaches (*i.e.* syntactic and probability models).

This section will briefly introduce some of the most influential texture analysis methods that are used for VIR, without attempting to categorise them. For a more comprehensive survey on texture analysis, including an abundance of definitions for texture in general, see [450]. Elaborate comparisons between texture feature descriptors can be found, for example, in [352, 442], and the texture descriptors included in the MPEG-7 standard are described in [261].

Co-occurrence Matrices

Co-occurrence matrices are used to quantify the information about the neighbourhoods of pixels in certain distances and directions and present a rather simple approach to describe the texture in an image [140, 287]. Since these matrices can be numerous and very large, several features such as entropy, contrast, symmetry and homogeneity are usually extracted to provide a more compact representation. Examples of the use of co-occurrence matrices include web image retrieval [327] and the analysis of high-resolution computed tomography lung images [398].

MRF, SAR and Fractal Models

Generalised versions of the *Markov Random Field (MRF)* models [133, 216] have been very successful in texture modeling for the past decades. These models are based on the assumption that the intensity of each pixel in the image depends on the intensities of only the neighbouring pixels. In the context of image retrieval, sliding masks for localisation are used for the computation [242].

The *Simultaneous Auto-Regressive (SAR)* model is an instance of the MRF models, only using fewer parameters, and pixel intensities are taken as random variables [249]; texture is considered as the outcome of a deterministic dynamic system subject to state and observation noise [263, 405, 440].

Improvements of the SAR model include the *rotation-invariant SAR (RISAR)* model (because the SAR model itself is very vulnerable to rotation) and the *multi-resolution SAR (MRSAR)* model to describe textures of different granularity and allow multi-scale texture analysis [263, 349, 457].

The use of *fractals* [260, 335] to describe the texture of images presents another model-based approach. Fractals are a set of self-similar functions in the so-called fractal dimension and show some correlation with the perceived roughness of image texture, and have been used for image retrieval, for example, in [16, 203].

Wavelet Transformation

The wavelet transformation [78, 259, 430] provides a multi-resolution approach to texture analysis and representation [52, 222, 408]; its computation involves recursive filtering and sub-sampling, which is introduced in [430] and further explained in [431]. Although there exists an abundance of different wavelet filters, the choice of a particular one of them is not critical for texture analysis [133].

Despite having their roots in approximation theory and signal processing, wavelets have successfully been applied to many problems in computer graphics such as image compression and image editing. Even though wavelets are very sensitive to noise, they are often used to describe texture for image retrieval [326, 327, 481].

Gabor Filter Transformation

The Gabor filter transformation is regularly used in image retrieval to describe the global structure or texture of images [254, 383, 428]. The algorithm's use of the Gabor filter in order to extract texture features, which is further explained in [220, 482], is akin to the orientation and frequency-selective processes in the primary visual cortex [79, 287].

According to a comparison carried out in [254], approaches based on Gabor filters yield better retrieval results than the aforementioned wavelet or MRSAR based techniques. Consequently, many similar approaches soon followed, like [494], and texture descriptors based on Gabor filters with different directions and various scales were also included in the MPEG-7 standard [261].

Tamura and Wold Features

So far, the majority of the above-named texture features are not suited for retrieval applications in which the user wants to use a verbal description of the image. This lack of texture descriptions in terms of perceptual properties led to the development of the Tamura and Wold features, which were both designed in accordance with psychological studies on the human perception of texture [249].

The Tamura features include single-valued measures like *coarseness*, *contrast*, *directionality*, *linelikeness*, *regularity* and *roughness*, which are all further described in [438]. Directionality, contrast and courseness were used in some early retrieval systems like *QBIC* [318] and *Photobook* [336], while regularity, coarseness and directionality were accepted as MPEG-7 standard [261].

The Wold components comprise *harmony*, *evanescence* and *indeterminism* [123, 245, 429] and correspond to periodicity, directionality and randomness respectively: periodic textures have a strong harmonic component, highly directional textures have a strong evanescent component, and less structured textures tend to have stronger indeterministic components [249].

Recent Approaches and Alternative Ideas

Some recently published work includes texture features modelled on the marginal distribution of wavelet coefficients using generalised Gaussian distribution [91], rotation-invariant texture retrieval with gaussianised steerable pyramids [451], and texture analysis using level-crossing statistics [382] or generic Markov-Gibbs image models [504] respectively.

Further, Ma and Manjunath [255] developed a texture thesaurus to assist texture retrieval in images on the basis of similarity to automatically-derived codewords representing essential classes of texture within a collection.

2.3.4 Shape Descriptors

Research in the field of human image understanding, such as [29], has reported that natural objects are primarily recognised by their shape; the retrieval based on shape features might therefore be the most obvious content-based approach and has consequently been used in many CBIRS [130, 318, 336].

Ideally, such shape features are invariant to rotation, translation and scaling and are extracted once an image has already been segmented into regions or objects. Unfortunately, robust and accurate image segmentation still presents an intricate problem [249], which limits the use of shape features to special applications in which these regions and objects are already defined and therefore easy to segment. Examples include mono-object image collections like databases for car, fish and trademark retrieval [57, 97, 98, 188, 282, 330].

Shape representations can generally be divided into two categories: *boundary-based* and *region-based* methods.

- *Boundary-based* methods, as their name indicates, are based on the outer boundary of the shape and include finite element models, scale space and Fourier-based shape descriptors.

- *Region-based* methods, in contrast, use the entire shape region for the calculation of shape descriptors such as statistical moments.

For example, if a query was based on the second shape in the first row of Figure 2.3 (taken from [32]), region-based methods would return the shapes in the first row, whereas boundary-based methods would return the shapes from the second column. This section will briefly introduce the basics of some of the shape features that have commonly been used in image retrieval applications. More comprehensive surveys and overviews for shape descriptors can be found in [248, 273], and [459] discusses similarity measures and algorithms for shape matching. The shape features that are accepted in MPEG-7 can be found in [32]. For an overview and a performance comparison of the state-of-the-art shape similarity measures, see [460].

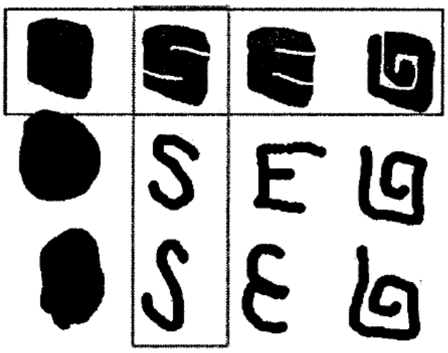


Figure 2.3: Example of boundary-based and region-based similarity.

Region-Based Methods

Classical region-based shape representations make use of statistical moments. In general, the geometric (*i.e.* algebraic) moment $m_{p,q}$ of order $(p+q)$ of an object $O \subseteq \mathbb{R}^2$ in an image is given by

$$m_{p,q} = \int_{(x,y)\in O} x^p y^q \, dx \, dy \tag{2.1}$$

where $p, q = 1, 2, ...\infty$. The infinite set of moments as defined in (2.1) uniquely determines the shape of an object, however not all of these variations contain all the desired invariance properties as well.

Algebraic Moment Invariants. Based on these geometric moments, a set of seven invariant region-based moments was defined by Hu in 1962 [177]; these moments are invariant with respect to scaling (size), shifting (position) and rotation (orientation), hence the name *moment invariants* ([287] prefers to call them *invariant moments*). This original set of moment invariants is still used regularly, especially in trademark retrieval [57, 188, 223]. Improved versions based on these seven moments include [204, 231, 497].

Orthogonal Moment Invariants. *Zernike*, *Pseudo-Zernike* and *Legendre* moments present examples of orthogonal moment invariants, an enhancement of the algebraic invariants in that they also allow images to be recovered from these moments. Teh [442] described, examined and compared these orthogonal moments with various other moments including algebraic, *complex* and *rotational* moments; he concluded that Zernike and Pseudo-Zernike moments outperform the rest regarding the capability of shape representation. This result was later also confirmed for the case of trademark retrieval [208].

The Angular Radial Transform. The only region-based descriptor that is included in the MPEG-7 standard is the *angular radial transform* (ART) [32]. It is defined on a unit disk in polar coordinates and takes into account all the pixels that describe an object, which makes it quite robust to noise. Its feature vector is formed by a number of normalised coefficients, and the shape similarity measure is simply the *Manhattan* (L_1) *distance* between two such vectors [460].

Other region-based descriptors. Other descriptors include the grid descriptor [251], the Hausdorff distance on region [180] and the image edge orientation histogram [187]. Naturally, most region-based methods can also be applied to contour shapes. See [460] for a short description and comparison of these descriptors.

Global object features. Single valued measures characterising the entire object, such as area, circularity, eccentricity, compactness, major axis orientation, Euler number, concavity tree, shape numbers, and algebraic moments, can also be used for shape description [348, 459]. The first three of these, for instance, were used in QBIC [318]. Definitions for most of these measures can be found in [97, 98, 135].

Boundary-Based Methods

Fourier Descriptors. The main idea of Fourier-based shape descriptors is to describe the shape of an object in the frequency domain by determining the *Fourier transform* of its boundary [206, 337]. General shape properties are characterised by the lower frequency coefficients, whereas shape details are reflected by the higher frequency coefficients. While translation invariance is obtained by the boundary representation, rotation invariance can be achieved by discarding the phase components of the complex coefficients and scale invariance via dividing the amplitude of the coefficients by the amplitude of the DC component. Fourier-based shape descriptors have been employed in CBIR systems like MARS [326] and are frequently used in trademark retrieval [223]. Recent improvements include the exploitation of the amplitude and phase of Fourier descriptors using dynamic time warping [21].

Curvature Scale Space. Another boundary-based approach to shape retrieval makes use of the *curvature scale space* (CSS) whereby an image is described by the zero-crossings of its curvature, *i.e.* the points in which the curvature of the object boundary changes from convex to concave (and vice versa). This process is repeated at several scales, with the resulting image being low-pass filtered (and thus reducing the number of zero-crossings and producing a smoother contour) before each iteration, until the entire contour is finally convex. The maxima in CSS are then used together with its eccentricity, circularity and aspect ratio at various scales to represent the object [282] or to compare it with other objects [460]. CSS has successfully been used for image retrieval, for instance [354], and is also included in the MPEG-7 standard [32]. Recent work based on the CSS includes [190, 211].

Shape Contexts. Shape matching and object recognition using *shape contexts* [24] are recently developed boundary-based methods in which a shape representation for each contour point of the boundary is built by using statistics (*i.e.* quantised angular and distance intervals) of other contour points "seen" by the point in consideration. A two-dimensional histogram is then used to represent the obtained view of a single point. In order to compare two contours, the correspondence of contour points is established that minimises the distances of the corresponding matrices [460]. Two algorithms for efficient shape matching based on shape contexts are presented in [284].

Other boundary-based methods. Similar approaches based on the curvature of the boundary include the *turning angle function*, the *tangent space representation*, the *convex parts correspondence*, the *contour edge histogram*, *string matching*, *chain code nonlinear elastic matching*, and *Delaunay triangulation angles*. Short descriptions and links to further literature for most of these descriptors can be found in [460]. Examples of more recent work are using the *deformation effort* [393], the *distance set correspondence* [143] or the *contour to centroid triangulation* [12] as shape descriptors.

Recent Methods and Alternative Ideas

Mehtre [273] showed, in a comparison study, that combined boundary-based and region-based methods outperform the individual boundary-based or region-based techniques. *Skeleton-based* methods, for example, use both boundary information and region information and include the *medial axis*, *smoothed local symmetries* and the *processing inferring symmetric axis* [449].

Similar to the texture thesaurus, Hove [175] introduces a thesaurus for shapes and objects and integrates it into an image retrieval system called VORTEX. Other recent work not named above includes the introduction of a novel skeleton polygonal shape (*the linear axis*), a new similarity measure for partial shape matching [449], and the supervised learning of edges and object boundaries [92].

2.3.5 Other Content-Based Descriptors

Other content-based features than can be used for CBIR include *local invariants* and *spatial information.*

Local Invariants

Features based on local invariants include *salient regions* and *interest (corner) points* and have extensively been used in image retrieval. Interest points are scale and affine invariant and can deal with significant transformations and illumination changes; such points can be found where there is, for instance, a corner in an image or another salient feature. *Salient* points are those which survive the longest when an image is gradually blurred [77, 287].

Since these interest points provide a compact representation of significant image regions, yielding good discrimination power and efficient indexing, they have shown to be very effective features for image retrieval [276, 386]. Likewise, wavelet-based salient points have also been for image retrieval [443], and colour interest points are discussed in [141]. A comparative evaluation of a range of interest point detectors can be found in [275].

Spatial Information

Spatial constraints comprise the information of the position of pictorial data within an image and have been an essential aspect of geographical information systems for many years [56, 366]. Similar techniques to access data by spatial locations have also been applied for retrieval in general image collections. These include the *spatial location* of (and the *spatial relationships* between) regions or objects, and spatial relations between *points of interest.*

Spatial indexing is seldom meaningful on its own, but can be quite effective in combination with other features such as colour [411, 432]; spatial constraints can be used to distinguish between regions or objects with similar colour or texture properties. For example, regions of the sea and the blue sky might exhibit very similar colour histograms, but their spatial locations in an image are different (under normal circumstances).

The *2D strings* [51] and variations of it [227, 228] are the most widely used representation of spatial relationships. Other examples are the spatial quad-tree [377] and symbolic images [152]. An overview of possible spatial relationships in image retrieval can be found in [101]. Systems using spatial information for retrieval include [196, 386, 409, 410].

Since reliable segmentation of objects or regions in natural images is often not feasible, searching images based on spatial relations remains a tricky research question [249]. One of the attempts to tackle this problem is the *radon transform*, which exploits the spatial distribution of visual features without requiring a sophisticated segmentation beforehand [155, 478].

2.3.6 Index Organisation

A vital issue not mentioned so far is the actual storage and organisation of these visual content descriptors. Most systems extract a certain subsection of the aforementioned feature descriptors for all the images in the database and store them as a vector in the so called *feature index*. Such feature vectors, however, tend to exhibit a very high dimensionality (normally of the order of 10^2), which makes efficient CBIR unfeasible for large-sized image collections from a computational point of view. Hence, a *dimension reduction* is often carried out before setting up an efficient indexing scheme.

Dimension Reduction

Feature vectors can be considerably reduced in dimension without significant degradation in retrieval quality [112, 317, 488]. Consequently, dimension reduction techniques such as the *principal component analysis* (PCA) and the *Karhunen-Loeve transform* (KLT) have frequently been used for image retrieval.

The main idea of the PCA is the linear mapping of the input data to a coordinate space such that the axes are aligned to reflect the maximum variations in the data.

This technique was used, for example, in QBIC where the feature vector was reduced from 20 to only two or three dimensions [117, 318]. The KLT technique is able to locate the most significant sub-space and has been used for dimension reduction in [46, 112]. Further approaches used in information retrieval include techniques based on neural networks [45] or on clustering [376]. Other strategies to reduce the feature space are described in [412].

Indexing Techniques

Early image retrieval systems used simple files in a directory or entries in databases to store the extracted visual descriptors of an image. For example, QBIC makes use of DB2 [117] and *VIPER* uses mySQL [253] as an alternative to inverted files; both options, however, performed very poorly from a computational perspective as most file systems only used linear search within directories and most databases only allowed for efficient operations on fixed size elements. Other approaches to use database systems to access image features include [81, 129].

Although most modern databases are now offering higher performance table searches and integrated modules to index images, like Oracle interMedia [271], researchers [100, 236] turned to similarity-based storage techniques which allow the use of tree-based indices to achieve logarithmic performance [237]. Existing techniques which allow the efficient similarity searching include the R-tree [156] and its improved versions, the R+ tree [395] and R* tree [23]. Further techniques are Linear Quad-Trees [462], TV trees [243], SS+ trees [218], k-d trees [100], priority k-d trees [488], K-d-B trees [360] and grid files [320]. An overview of several tree structures and their properties can be found in [405]. For a comparison with respect to index generation time, search time and error probabilities, see [487]. Other good reviews and comparisons include [317, 488].

Most of these multidimensional techniques exhibit a reasonable performance for a small number of dimensions and are not scalable for dimensions higher than 20 [111]. Another problem is that most of these approaches assume that the Euclidean distance forms the base for feature comparison, which is not necessarily true for many image retrieval applications [249]. In fact, the human perception of a certain visual content may not effectively be simulated by the Euclidean distance, yet it might not be even metric at all [368]. One attempt to solve these problems [502] is the use of hierarchical indexing schemes based on *self-organisation maps* (SOM). Other approaches include the use of incremental clustering techniques for dynamic information retrieval as proposed in [55] and further improved in [195, 458].

Some recent literature reports on improvements to k-d trees [392] and the effective use of vector quantisation to guarantee efficient search in large image databases [499]. In a rather alternative approach, *inverted files* that have proven to be very useful for text retrieval [490] are shown to be efficient also for image retrieval when the feature space is only very sparsely populated [426, 428].

2.4 Concept-based Image Descriptors

The last section presented the state-of-the-art techniques for automatic visual feature extraction. However, most genuine users of image collections formulate their search queries at the opposite site of the semantic gap, namely in terms of semantic retrieval requests using text (see Section 2.2). As a consequence, a heavy dependency on the *concept-based image retrieval paradigm* continues to be exhibited in the commercial use of picture collections [105, 107, 164]. The process of the concept-based retrieval paradigm can be described as follows:

1. First, the textual expression of the query is preprocessed using stemmers and stop word lists and might then be mediated by a thesaurus and/or classification schemes (*e.g.* ontologies) in order to couch the query in terms of a controlled (authorised) vocabulary (see Section 2.5.1).

2. Then, the textual representations associated with each image in the collection are matched against this (modified) query expression. Section 2.5.1 describes the essential matching techniques and similarity measures for concept-based image retrieval.

3. Finally, if the textual representation of an image matches the query expression (sufficiently), the image associated with that textual representation is added to the set of relevant images and presented to the user for consideration.

The concept-based image retrieval paradigm thereby translates the task into a simple text-matching operation, totally ignoring the fact that the original information need is given in the visual domain. Hence, an effective linguistic representation of the semantic content of an image is obviously a prerequisite for successful retrieval.

This section provides an overview of several approaches to assign textual representations to images. Some classification attempts for semantic image representations are described in [82, 183]. Related issues such as textual query preprocessing and concept-based similarity measures can further be found in Section 2.5.1.

As for further literature, a simple introduction to text retrieval approaches is presented in [358]. More information on information (text) retrieval can be found in Rijsbergen's seminal compendium [454], which is still well-worth reading although it was published as early as 1979, while a very recent (2007) and comprehensive introduction on information retrieval focussing on mathematic foundations as well as algorithms is given in [262].

2.4.1 Content-Independent Meta-Data

Content-independent meta-data provides information related to an image without describing it directly. Examples include the name of the photographer, the date, time and cost of production or technical specifications such as the camera or lens used. This type of information can normally not be extracted from a photograph or film, and although such extraction could be interesting for text search purposes, it is universally not considered useful for the evaluation of VIR systems [161, 234].

2.4.2 Keyword Representations

The terms in a textual search request that are actually used to find relevant documents are commonly referred to as *keywords*. Thus, the *keyword representation* of an image is based on the anticipation of these query terms a user is likely to enter in order to retrieve that particular image. These representations can either correspond to a complete image or only be associated with a certain region within such an image [161].

The main goal of keyword representations is to capture the essential entities and relationships in visual materials. This, however, raises the question of how general or specific these keywords should be, and how much background data should be included about these entities and relationships. In principal, keywords could be used to describe images throughout all conceptual levels. While it seems obvious to list, for instance, the objects illustrated in an image (*e.g.* soccer players, ball, goal, referee), keywords can also be used for the description of the general setting (*e.g.* World Cup Final) or abstract concepts such as associated feelings and emotions (*e.g.* glory, triumph) [200, 437].

Keyword representations are generally classified into three types: they can either be arbitrarily chosen to describe an image (*uncontrolled vocabulary*), be based on a set of standard terms without any hierarchy (*controlled vocabulary*), or use a controlled vocabulary that is also hierarchically structured (*ontology*).

Uncontrolled Vocabulary

Keyword representations using arbitrarily chosen terms tend to be the least expensive of the three types, and surprisingly it has been shown that retrieval based on a uncontrolled vocabulary is not necessarily inferior to that based on its controlled counterpart [376, 414]. Further advantages include the possibility of searches for novel topics that might not yet be included in controlled vocabularies, the ability to perform very specific searches if the exact subject can be well-defined, and the fact that more results can often be provided in a shorter time span, because there is no need to browse through controlled vocabulary subject headings [13].

In contrast, the main drawback of using an uncontrolled vocabulary can be found in its inability to deal with the inherent ambiguity of natural language, which can lead to very low precision scores as many irrelevant items are being retrieved (false positives). Moreover, one might have to spend more time on preparing a search strategy in order to incorporate all the terms that cover a certain topic, which might also involve a certain level of specialist knowledge [13].

The *FlickR* photo archive (see Section 3.2.7) presents one famous example of an image database with keyword representations based on uncontrolled vocabularies.

Controlled Vocabulary

Keyword representations based on a controlled vocabulary only use terms from a list of words and phrases that were carefully preselected by trained professionals with expertise in the subject area. Retrieval based on such preselected terms, which ideally exhibit unambiguous, non-redundant definitions, allows for greater focus and levels of relevance, greater precision and lower recall searches and fewer missed citations because of terminology problems or spelling variations.

The use of controlled vocabulary in retrieval, however, may also give rise to an under-representation of new, unusual or very specific topics within such an alphanumeric image description. Controlled vocabulary is more expensive to implement than its uncontrolled counterpart, and it has to be updated regularly due to the constantly evolving nature of human knowledge. Further, users need to be made aware that a controlled vocabulary is available, and they will also have to learn how to use it when performing a search [13].

Ontologies

In computer science, the term ontology has come to be used to refer to a hierarchical data structure that contains all the relevant entities as well as their relationships and rules within a certain domain. As a consequence, if such a hierarchical structure is added to a list of keywords (controlled vocabulary) that are used to describe images within a collection, a domain-specific ontology is created for that collection. Since ontologies in information retrieval contexts are based on hypernym relations (*e.g.* a cat *is an* animal), they automatically produce a *taxonomy* within that collection as well [161].

Apart from the aforementioned benefits associated with controlled vocabularies, the use of ontology-based keyword representations further allows for the possibility of hierarchical searching. The aim of using ontologies to describe multimedia resources is to provide well-structured information to improve the accuracy of retrieval. Ontologies provide a top-down approach to bridging the *upper semantic gap* (the gap between the symbolic labels of the single objects of an image and the full semantic information conveyed by that image). As a consequence, ontologies have been shown to be crucial for the Semantic Web [161], which has further given rise to the development of several languages for their formalisation; examples include the *Web Ontology Language* (OWL) and *Resource Description Framework* (RDF). The major disadvantage of this approach is that the development of ontologies is a highly complex process – even in very limited domains as indicated by [183, 388].

Although most of the literature about ontologies originates from the field of text retrieval and management, there has recently been an increasing interest in the use of ontologies also with multimedia collections. Some early work by Schreiber [388], for example, describes the use of ontologies as a tool for the semantic representation of (and search for) images, an approach which was later improved [174] and extended by the integration of spatial information [173]. Several other approaches to semantically annotating multimedia data using ontologies include [2, 95, 176, 387].

Probably one of the most famous ontologies within the information retrieval community is *WordNet*[4], an online lexical reference system which organises nouns, verbs and adjectives into synonym sets (synsets); these synsets are arranged in a hierarchy, with each of them representing one underlying lexical concept [113]. One example of the effective use of *WordNet* in VIR is [509], whereby an ontology of portrayable objects is constructed by pruning the *WordNet tree*. Other applications using *WordNet* for ontology-based keyword representations include [19, 191].

[4] `http://wordnet.princeton.edu/`

Limitations

An extensive study by Tam and Leung [437] has shown that keyword searches on free-text and keyword based representations suffer from low precision and low recall respectively, especially for retrieval from very large databases. These poor retrieval results can be attributed to the lack of syntax in keyword search and representations. In particular, some of the most serious limitations of keyword representations for annotation and retrieval are that they cannot:

- associate modifiers (*i.e.* adjectives, adverbs) with an entity (noun) or action (verb). For instance, the search for "tall, Austrian referee booking blond, Dutch footballer" could also retrieve "Dutch referee booking tall, blond, Austrian footballer";
- detect relationships between search words, and therefore the search for "woman riding bicycle" could also return an image semantically represented by "woman reading book on bench; boy riding bicycle on path behind bench";
- give more weight to semantically more significant query terms. People are usually more interested in entities than modifiers: for instance, someone looking for a "red car" would probably prefer an image of a green car over an image of a tomato.

One approach [153] in order to overcome these problems and to enhance retrieval results is to structure the query and representation terms by assigning syntactic class indicators such as nouns, verbs or adjectives - which is the underlying idea of *Structured Representations.*

2.4.3 Structured Representations

The aim of structured representations is to reduce the parsing problems of *natural language processing* (NLP) and to remove the limitations of keyword annotation and retrieval. The main idea is to enable searches for entities (subjects, objects) and relationships (verbs) in association with terms that either modify them (adjectives) and/or further explain the setting of the image (adjuncts).

Ternary Fact Model

One of the first approaches (1995) to express the semantic content of an image in a structured format was the *ternary fact model* (TFM) which comprised an underlying visual entity-relationship representation, a rule-based conceptional hierarchy and other features to support semi-automatic annotation and to enhance retrieval performance. In particular, TFM used pairs (binary facts) and triplets (ternary facts) as the main building blocks of its description [235] and formed the basis for the development of a concept-based query system [434].

The specification of ternary facts, however, can yield some problems as TFM cannot explicitly determine the roles played by each of the facts. For example, "Michael Jordan throwing the ball to Scottie Pippen" could also be interpreted as "Scottie Pippen throwing the ball to Michael Jordan" or even as "Michael Jordan throwing Scottie Pippen to the ball". This dependency on word order and the use of prepositions automatically limits the utility of TFM in other languages as well [437].

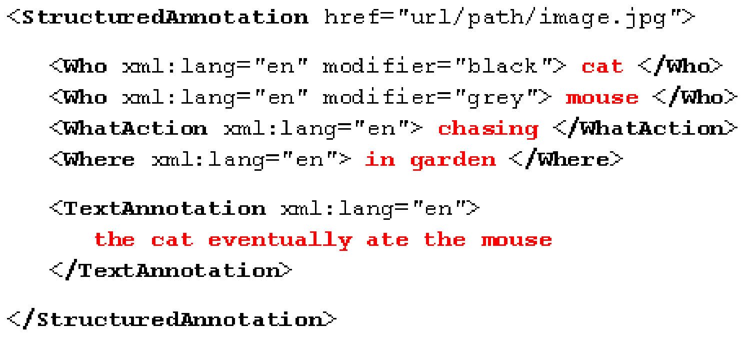
MPEG-7 Description Scheme

The Moving Picture Expert Group (MPEG[5]) also soon recognised the need for a (manually produced) semantic image representation and included a structured *Description Scheme* (DS) within the MPEG-7 standard (or, more formally, *Multimedia Content Description Interface*). The main objective of MPEG-7 was to provide a comprehensive set of tools to describe multimedia content such that users can search, browse and retrieve that content more efficiently and effectively. These tools include Description Definition Languages (DDL), Description Schemes (DS) and Descriptors (D). The definition and examples of the use of these tools are provided in [374]. Overview articles include [80, 268], while publications such as

[5] `http://www.mpeg.org/`

[32, 261] concentrate on specific aspects of MPEG-7. Further information on MPEG in general and on its other standards can be found in [13, 287].

The MPEG-7 DS allows a simple structured description of the entities, actions, places, times and reasons that are represented in audiovisual materials. The logical image representation is hereby treated as a data type of XML (Extensible Markup Language) and comprises several tags to structure image descriptions (*Who*, *When*, *Where*, *Why*, *WhatObject*, *WhatAction*, and free-text representations) as well as modifiers to further describe each of them. For example, an image showing "a black cat chasing a grey mouse in a garden" would be represented as indicated in Figure 2.4.



```
<StructuredAnnotation href="url/path/image.jpg">

   <Who xml:lang="en" modifier="black"> cat </Who>
   <Who xml:lang="en" modifier="grey"> mouse </Who>
   <WhatAction xml:lang="en"> chasing </WhatAction>
   <Where xml:lang="en"> in garden </Where>

   <TextAnnotation xml:lang="en">
      the cat eventually ate the mouse
   </TextAnnotation>

</StructuredAnnotation>
```

Figure 2.4: Structured image representation according to the MPEG-7 DS.

This structure, however, again gives rise to similar problems as with the TFM, because there are cases in which it cannot distinguish between main sentence elements (such as subjects and objects) and therefore provides a certain ambiguity. In the aforementioned example, both cat and mouse are encapsulated by the *who* tag, leaving no indication of who is actually chasing whom!

Improved Structured Representations

Tam and Leung [437] also pointed out that the "MPEG–7 DS might not have enough structure to provide guidance to annotators" and also criticised the unclear use of the *Why* tag as well as the remaining question whether fields should contain single words (linked to controlled vocabulary or ontologies) or complete noun-phrases.

Hence, as an enhancement of the Structured DS, Tam tried to link the entities and relationships to appropriate items in relevant databases and ontologies. However, the derived database design, which included several redundancies, was neither successfully validated nor adopted as a standard.

One approach to overcome the discrepancy between the grammatical elements of a free-text description and their corresponding representation as tags in a structured format was published in [144] and later reused in a similar way in, for example, [13]. An improved design to typify a structured representation in a relational database was published in [148]. Several publications such as [179, 447] have expressed the need to move the MPEG-7 description of multimedia information closer to ontology languages such as RDF or OWL.

2.4.4 Free-Text Representations

Unlike the previously mentioned keyword and ontology-based representations, free-text representations are neither based on a predefined structure nor restricted to a certain vocabulary or ontology; in other words, an image can be described by any combination of words and/or sentences.

Due to the lack of restrictions and regulations, free-text representations offer a very easy annotation method and are hence often used to describe images. This is especially true for the area of private photographic collections, where it is generally hard to motivate users to annotate images at all, and further rules and regulations would hardly increase these motivation levels [219, 363].

However, retrieval based on free-text descriptions alone can be quite difficult as they require accurate NLP to match the terms and syntactic structure of the query terms and the free-text representations. The parsing of truly unrestricted natural language is certainly not a trivial task and often requires further human intervention to improve poor automatic NLP results [364, 437].

As a consequence, rather than being used as stand-alone descriptions, free-text representations are often used in addition to the other types of representations in order to depict image concepts that cannot be adequately described by the use of keywords, do not fit into provided representations structures or are not covered by a provided ontology: "There is no way a domain ontology can be complete – it will not include everything a user might want to say about a photograph" [388].

2.4.5 Automatic Annotation

Regardless of their particular benefits and limitations exhibited by all aforementioned types of logical image representations, they all have one serious disadvantage in common: the manual annotation of image collections is a very tedious, time-consuming and quite expensive process, which in the case of very large image databases can also be an impractical and unfeasible undertaking [13, 105, 139]. Although there are a couple of very innovative approaches for the manual annotation of images, like [464], there is a clear need for the development of systems to identify symbolic labels that can be automatically be used to describe an entire image (or parts of it).

Automatic annotation can be very beneficial to reduce the amount of manual effort required to annotate (large) image collections and, at the same time, represents a "bottom-up" attempt to bridge the *lower semantic gap* (the gap between visual descriptors and symbolic labels of objects) by learning which combination of visual descriptors corresponds to which object and what the labels of that object should be [164].

Mori et al. [285] were probably the first to experiment with automatic annotation and tried to apply a co-occurrence model to keywords and low-level features to rectangular image regions. In general, the state-of-the-art techniques in automatic annotation can be divided into two categories: segmentation and scene oriented approaches.

Segmentation Based Automatic Annotation

In segmentation based automatic annotation, an image is first segmented into regions (also called "blobs") before the actual annotation algorithm is executed. For example, a method which uses a machine translation model to translate between keyword representations and a discrete vocabulary of blobs is described in [96]. Although this method is now outdated and has been outperformed by many other approaches [193], it is often cited within the literature because the data-set proposed within that publication has become a popular benchmark for image annotation systems [164].

More recent and effective methods are based on an extension to the *latent Dirichlet allocation* [18] or on the use of cross-media relevance models [193], continuous-space relevance models [226] and inference network approaches to link regions with their alphanumeric representations [274]. Models that use rectangular regions rather than blobs include [115, 194, 283].

Scene Oriented Automatic Annotation

The second type of automatic annotation techniques uses the global information of the entire image and therefore takes a more scene-oriented approach. For example, vector space representations created from local descriptors of salient regions within an image can be used for automatic annotation by propagating semantics from similar images [163]. Further, [498] showed that the use of simple global features (together with non-parametric density estimation and the technique of "kernel smoothing") can produce results comparable to those of [226, 274].

One problem of most of these approaches, scene oriented and segmentation based, is that they explicitly apply a certain number of textual descriptions to an image (so called "*hard* annotations"). This may lead to the creation of similar but wrong labels, which can cause problems in subsequent retrieval [164, 193].

Alternative Approaches

Some recently published dissertations [13, 162] provide a few alternative as well as innovative approaches to reduce the manual annotation effort. For example, Hare [162] proposes the creation of a simple semantic space of documents (images) and terms (keywords) using a linear algebraic technique. The main idea is that similar documents and terms would share similar positions within this semantic space, which implicitly generates semantic image representations in a "soft" manner (as opposed to the aforementioned "*hard* annotations").

Another novel method that aims to reduce manual description efforts uses implicit concept-based image indexing [13]. Although not fully automatic, this technique uses fuzzy expert systems, the categorisation of image components and the relative importance of these components for each of the levels of an ontology, which can subsequently be used to generate or predict other semantic concepts within an image.

Finally, the overview papers of recent evaluation events such as [60, 86, 110, 323] present an updated review of the state-of-the-art technology in object recognition and automatic annotation, and also provide links to many recently developed systems.

2.5 Result Generation and Presentation

The last two sections discussed various descriptors for visual and textual features of an image and/or its logical semantic representation. Once these features are extracted from a certain query (text and images), the search results can be calculated by the system by comparing these descriptors. The result of this comparison is hereby not only a single image, but rather a list of images ranked by their similarity with the query (again, approaches that are not relevant for ad-hoc image retrieval will not not be further discussed in this section, for instance those specific to image classification tasks, *etc.*). Three different approaches for result generation can be identified: *concept-based*, *content-based* and combined approaches.

2.5.1 Concept-Based Result Generation

The result generation process of TBIR is based on a pure text-matching operation in which the textual expression of the user request (*i.e.* word lists, phrases, sentences, extended text) is matched against the textual representation associated with each image within a collection. Most TBIR systems (TBIRS) preprocess the query request before term weighting and scoring algorithms are applied in order to improve retrieval results.

Query Request Processing

Query request processing comprises several techniques to enhance the subsequent retrieval performance, such as tokenisation, stop word lists, stemmers and query expansion methods including the use of ontologies or thesauri [262].

Tokenisation. The first step of request processing is *tokenisation*, whereby a query sentence is split into several terms (tokens), while certain other characters (such as punctuation marks) are discarded.

Stop Word Lists. Then, *stop words* – extremely common and semantically non-selective words such as articles, conjunctions or prepositions – are sometimes removed from a query expression using a *stop word list.* There is no common rule to how long or complete such a stop word list should be: some applications use quite large stop word lists (200 - 300 terms), others very small ones (7 – 12 terms), while most Web search engines refrain from using stop word lists at all.

Stemmers. The goal of stemmers is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form (stems, roots) to avoid subsequently missed matches due to trivial word variations (morphological variants). This includes, for example, the transformation of plural to singular forms (cat, cats ⇒ cat) or of conjugated verb forms to its infinitive (go, goes, going, gone ⇒ go). The most common algorithm for stemming English is the *Porter* stemmer [346]. Examples of other popular stemmers are the *Lovins* [250] and *Paice* [331] stemmers.

Lemmatisers. While stemming simply chops off the ends of words in the hope of arriving at the base form most of the time, lemmatisation provides a more sophisticated approach by using dictionaries and morphological analysis to achieve the same goal. For example, stemmers often have problems with irregular verbs (consider: am, are, is, was, were ⇒ be), while most lemmatisers would easily arrive at the infinitive of that verb (be).

Thesauri and Ontologies. A *thesaurus* is a list of synonyms a search engine can use to find matches for particular terms that do not directly appear in the logical image representations. Query expansion using thesauri can therefore be an effective method to further enhance retrieval results. *Ontologies* can also be used for query expansion to further retrieve images that have been annotated by using *hypernyms* of the requested query term.

Term Weighting

The principal aim of traditional TBIR systems is to partition a database into two sets: *relevant* and *non-relevant* documents. Only the relevant documents are then displayed to the user, which are then ranked by relevance [384]. In its simplest statistical approach, the calculation of this ranking is based on three different sources of weighting data.

Inverse Document Frequency. The *inverse document frequency idf(i)* – also referred to as *collection frequency weight cwf(i)* – for term t_i is based on the observation that terms that occur in only a few documents contain more information than the ones that occur in many, and is defined as

$$\mathrm{idf}(i) = \log D - \log r \tag{2.2}$$

where $D = |\mathcal{D}|$ denotes the total number of documents of a collection $\mathcal{D}$, and r the number of documents term t_i occurs in.

Term Frequency. The *term frequency tf(i,j)* quantifies the number of occurrences of term t_i in document $D_j \in \mathcal{D}$. The underlying assumption here is that the more often a term occurs in a document, the more likely it is to be important for that document. In terms of TBIR, term frequencies can often be neglected because each keyword is, under normal circumstances, only applied once to describe an object in an image. Thus, $tf(i,j) = 1$ for most of the terms t_i in document D_j.

Document Length. The length $dl(j)$ of a text document $D_j \in \mathcal{D}$ is the third input to weighting: a term that occurs as often in a long as in a short text document is likely to be more valuable for the latter. The *normalised document length ndl(j)* is normalised by the length of an average document in the collection $\mathcal{D}$:

$$\mathrm{ndl}(j) = \frac{dl(j)}{\frac{1}{D}\sum_{j=1}^{D} dl(j)} \tag{2.3}$$

where $D = |\mathcal{D}|$ is the total number of documents.

Matching and Ranking

The three types of weights are not very representative for ranking if they are used individually and, therefore, need to be combined to give a meaningful matching score for a particular document against a certain request. An immense number of variations for possible combinations have been proposed in the past, but rather than listing all of them, only a simple and a more sophisticated combination approach are presented hereinafter.

TF-IDF Weighting. The *tf-idf weight* (term frequency - inverse document frequency) is a weight often used in information retrieval and is defined as follows:

$$\text{tf-idf}(i,j) = \mathrm{tf}(i,j) * \mathrm{idf}(i) \tag{2.4}$$

where *tf(i,j)* denotes the term frequency of term *t(i)* in document D_j and *idf(i)* is defined as in (2.2). This approach is probably the most used weighting scheme in text retrieval [262] and has even been used as a distance measure for CBIR as well [428, 483].

Okapi BM25 Weighting. The Okapi BM25 function provides an example of a more sophisticated weighting combination that has proved very effective in trials during the TREC programme [359] and is defined as

$$\mathrm{BM25}(i,j) = \frac{\mathrm{idf}(i) * \mathrm{tf}(i,j) * (K_1 + 1)}{K_1 * ((1-b) + (b * \mathrm{ndl}(j))) + \mathrm{tf}(i,j)} \tag{2.5}$$

where K_1 and b are tuning constants.

Other Approaches

Apart from the *probabilistic* term matching approaches introduced above, there also exist two other major models to information retrieval: *vector-space models* and *language models*.

Vector-Space Models. In the vector-space model, n-dimensional vectors are used to represent both the document and the query (whereby n denotes the number of distinct terms observed in the document collection). The cosine measure, which computes the similarity between a query and a document as the cosine of the angle between their vectors, is one of the best-known techniques under the vector-space model. One example is the use of *pivoted cosine document length normalisation* [399].

Language Models. Language model approaches involve the estimation of the likelihood that both the query and the document could have been generated by the same language model, which is a probability distribution aiming to capture the statistical regularities of natural language use. One example of a language model based retrieval is the *query likelihood approach with Dirichlet smoothing* [501].

Further Reading. Within the scope of this thesis, it is only feasible to cover a small proportion of the wide range of text retrieval mechanisms and algorithms. A concise overview of other probabilistic, vector-space and language model approaches is provided in [152]; more detailed descriptions of these techniques can be found, for example, in [262, 417, 418].

2.5.2 Content-Based Result Generation

Unlike with concept-based result generation, it is not possible for content-based retrieval methods to divide the images in the database into relevant and irrelevant sets as response to a specific query. Instead, the prime aim of content-based result generation is to rank the images in a database in order of the similarity to the query image(s). The similarity of two images is thereby calculated by comparing the features extracted from the query image (*e.g.* a feature vector) with the features of the images in the database (*i.e.* in the feature index).

A number of similarity functions based on empirical estimates of the distribution of features exists to quantify such comparison. These so-called *distance* or *similarity measures* strongly depend on the feature space; the choice for a particular measure will most certainly affect the retrieval performance significantly.

This section will cover some similarity measures commonly used in CBIR. Let $D(I, J)$ denote the distance measure between two images I and J, $f_i(I)$ the number of pixels in bin i of the image I, and B the total number of bins in the histogram.

Histogram Intersection

The *histogram intersection* was one of the first measures that was introduced to quantify the difference between histograms in the context of VIR. It was first described by [436] and is defined as follows:

$$D(I, J) = \frac{\sum_{i=1}^{B} \min(f_i(I), f_i(J))}{\sum_{i=1}^{B} \min(f_i(J))} \tag{2.6}$$

Although histograms (and histogram intersection) are mostly used to compare colour features, they could theoretically also be used for texture or shape properties as long as each dimension of the image feature vector is independent and of equal significance. Further, research has shown that histogram intersection is fairly insensitive to changes in image resolution, histogram size, viewing point, depth or inclusion [249, 405].

Histogram intersection has been used in many early CBIRS, including [187, 326, 428], and is often blurred using a low-pass filter because it is very sensitive to small shifts [391], or a histogram cross correlation is performed, *e.g.* [129].

Minkowski-Form Distance

The *Minkowski-Form distance* (MFD) is the most widely used metric for image retrieval and, like the simple histogram intersection, treats all the bins of the histogram independently. This distance is defined as:

$$D(I, J) = \left(\sum_i |f_i(I) - f_i(J)|^p \right)^{\frac{1}{p}} \tag{2.7}$$

When $p = 1$, then $D(I, J)$ corresponds to the *Manhattan (L_1) distance*, when $p = 2$, then $D(I, J)$ complies with the *Euclidean (L_2) distance*, and when $p = \infty$, then $D(I, J)$ is L_∞ respectively.

Examples of the use of L_1 include [318]; L_2 was used in [369] to compute the similarity between texture features, in Blobworld [43] for texture and shape comparison, and was further employed in other systems such as [187, 318]; and L_∞ was used in [476] to compute the similarity between texture features. Sometimes, different forms of the Minkowski distance are used for different features. For example, Netra [256, 257] makes use of L_2 for colour and shape and L_1 for texture.

Quadratic Form Distance

The independent treatment of histogram bins, as employed by simple histogram intersections or the Minkowski form distances, ignores the fact that certain pairs of bins can correspond to features which are perceptually more similar than other pairs, which might lead to many false negatives [433]. One answer to this problem is the *quadratic form distance* (QFD), which is defined as follows:

$$D(I, J) = \sqrt{(F_I - F_J)^T A (F_I - F_J)} \tag{2.8}$$

where F_I and F_J are vectors that list all entries in $f_i(I)$ and $f_j(J)$, $A = [a_{ij}]$ denotes a similarity matrix, a_{ij} the similarity between i and j, and T the transpose of the matrix.

Since the QFD considers the cross similarity of histogram bins (*e.g.* colours), it can yield perceptually better results than the distance measures named above [249]. As a consequence, many retrieval systems such as [158, 318] have used the QFD as a similarity measure, especially for colour features.

Mahalanobis Distance

When all dimensions in an image feature vector are dependent on each other and exhibit different levels of importance, then the *Mahalanobis distance* is a good choice as a similarity measure. For example, this measure seems appropriate to quantify the similarity of features describing salient points [40]. It is defined as

$$D(I,J) = \sqrt{(F_I - F_J)^T C^{-1} (F_I - F_J)} \tag{2.9}$$

where C is the covariance matrix of the feature vectors.

Hausdorff Distance

One of the most studied similarity measures for shape features is the *Hausdorff distance* [180, 449, 459], which is defined as

$$H(I,J) = \max \left\{ \vec{h}(I,J), \vec{h}(J,I) \right\} \tag{2.10}$$

where $\vec{h}(I,J)$ denotes the *directed Hausdorff distance* from I to J, which is the lowest upper bound (supremum) over all points in I of the distances to J,

$$\vec{h}(I,J) = \sup_{i \in I} \inf_{j \in J} D(i,j), \tag{2.11}$$

with $D(i,j)$ the underlying distance such as the Euclidean distance L_2. Unfortunately, the Hausdorff distance is very sensitive to noise as a single outlier can influence the distance value. A review of other shape similarity measures can be found in [459].

Kullback-Leibler (KL) Divergence

The *Kullback-Leiber divergence (KLD)* is a measure mainly used for texture similarity with the following definition:

$$D(I,J) = \sum_i f_i(I) \log \frac{f_i(I)}{f_i(J)} \tag{2.12}$$

It describes the level of compactness in which one feature distribution can be coded if the other distribution is used as a "codebook" [249].

Other Similarity Measures

Other similarity measures include, for example, the *Jeffrey Divergence* or the *Bhattacharyya*, *Maximum Likelihood* and *Fréchet distances*. An interesting alternative was examined by utilising text retrieval features such as term frequency and collection frequency also in image retrieval systems such as [428, 483]. Excellent overviews of similarity measures in general and for shape matching in particular can be found in [405] and in [459] respectively.

2.5.3 Combined Result Generation

Various studies have shown that the combination of content-based and concept-based approaches can lead to better results than using both approaches separately [62, 63, 105, 390, 483].

There are several possibilities to achieve such a combination of TBIR and CBIR results. For example, [27, 172] show that retrieval results can be improved if the results based on TBIR are reordered using the results of CBIR. Other approaches, such as [3, 54, 197] that experimented with merging visual and textual runs, show similar improvements.

2.5.4 Result Display

Once a system has ranked the images in a database for a specific query, the results are then displayed to the user. Most retrieval systems display the thumbnail versions of the images separated over several result pages. The number of thumbnails shown on the first page varies, although 20 seems to have become a golden standard as it is used by most online search engines such as *Google* or *Yahoo!* as well.

In image retrieval research, the presentation of search results has not been of primary concern for a long time [287], with more formal research being published only in recent years. A user study on various ways of arranging images for browsing purposes [362] reports that text and image based result arrangements exhibit their own merits and demerits. Innovative arrangements for retrieval results of personal images are explored in [181]. Other recent work dealing with efficient ways for browsing and visualisation of search results include [240, 246, 380, 446, 495].

2.6 Related Issues

Although they are not directly within the scope of this research, there are related areas of VIR that deserve a brief introduction: relevance feedback and user interaction.

2.6.1 Relevance Feedback

Relevance Feedback (RF) is a query modification technique that is used to improve the effectiveness of information systems by capturing the user's precise needs through iterative feedback and query refinement [77, 249]. Although originally developed for text retrieval [361, 375], the ability to refine searches in response to user indications of relevance is particularly useful for image retrieval [252, 301, 370, 371, 375, 405, 428, 462, 493], especially because of the incapability of most systems to match user needs accurately the first time round (ad-hoc retrieval).

Users can normally judge the relevance of a set of images displayed on a screen within seconds. Hence, the main idea of RF is to present the user with a list of candidate images, combined with the request to mark the ranked images returned by the system as either relevant (positive examples) or irrelevant (negative examples) to a query. This feedback results in a refined information need, in which the relevant and irrelevant images are reflected by modifications concerning the feature space, semantic space, parameter space or classification space [237]. In a sense, RF therefore enables the iterative establishment of a link between high-level concepts and low-level features.

Since the main focus of this research lies in the evaluation of ad-hoc image retrieval, RF is not elaborated on further. A comprehensive review can be found in [505]. In [301], several strategies for positive and negative RF are compared, and [287] describes various possibilities for relevance feedback. [77, 237] summarise some recent concepts and provide further links to existing literature.

2.6.2 User Interaction

The success of efficient interaction between a user and a computer system or application often depends on such a system's *usability* [319, 347], and VIR systems are no exception hereby. In fact, usability issues in image retrieval applications do not only concern the user interface, but also affect several other system components and can be very crucial for their individual performance in particular as well as the entire system's performance in general [287]. These usability issues in VIR include:

- *query specification:* the flexible formation and modification of search requests (see also Section 2.2.2);

- *result presentation:* the clearly arranged presentation of the results in a user interface (see also Section 2.5.4);

- *relevance feedback:* the system should provide effective relevance feedback mechanisms (see also Section 2.6.1);

- *interaction speed:* the ability of a system to react instantaneously, which can be improved through, *e.g.* search pruning [305, 425];

- *learning aptitude:* a system should be able to learn a certain user's behaviour, for example, by analysis of user log files similar to the market basket analysis in the data mining literature [4, 5], or by learning feature weights [300, 306];

- *general usability:* the system should be easy to learn, efficient to use, and easy to remember; it should also have a low error rate and be pleasant to use [319].

This thesis is mainly concerned with the evaluation of retrieval performance for ad-hoc image retrieval, thus these aforementioned usability issues will not be discussed in further detail hereinafter. The fourth chapter of Müller's dissertation [287] presents a very comprehensive overview of user interaction in VIR and is well-worth reading. The Cross-Language Evaluation Forum (CLEF) has an evaluation campaign for interactive mechanisms in information retrieval (iCLEF), which has further comparisons and literature on usability issues [136, 137]. Another interesting article on user interaction can be found in [296].

2.6.3 Current Retrieval Problems

As indicated in the introduction, the most significant challenge for VIR is to bridge the semantic gap. CBIR and TBIR are hereby trying to approach this challenge from opposite sides: while CBIR is pursuing a "bottom-up" strategy (*e.g.* object recognition, automatic annotation) to bridge the lower semantic gap, TBIR attempts to attack the problem with a "top-down" approach (*e.g.* ontologies, structured alphanumeric representations) to bridge the upper semantic gap. Both directions encounter several limitations, which are further described hereinafter.

Limitations of Concept-Based Retrieval Methods

The main limitations of concept-based retrieval methods include:

- Logical image representations are highly subjective as different levels of prior knowledge and experience can influence the interpretation and understanding of an image [264].

- Images can mean different things to different people, or mean different things to the same person under different circumstances or at different times [396].

- Some features such as special textures or complex shapes are difficult to describe and cannot be clearly represented using text.

- Logical image representations may sometimes be incomplete. Although there is no limit to how semantically rich one could make such a representation, it is possible that some image features may not be mentioned (especially in highly complex images).

- Query by keyword (QBK) can produce low results due to its lack of syntax. Ontologies or structured representations, on the other hand, can be quite complex to create and be very cumbersome when generating search queries.

- Automatic annotation approaches are still far from reaching the quality of manual annotation. The manual annotation of visual resources, on the other hand, presents a very tedious, time-consuming and expensive process.

Limitations of Content-Based Retrieval Methods

Examples of the limitations of content-based retrieval methods comprise:

- The performance of traditional CBIRS is far from the users' expectations in real-world applications due to the semantic gap [63, 247, 281].

- While the sensory gap does not influence TBIR, different recording parameters such as illumination source and orientation, deformation and rotation can heavily affect the performance of CBIR.

- The QBE search paradigm is not practicable as most search requests are more akin to image discovery than image recovery; therefore, most queries in real-world applications are concept-based.

- The query specification for content-based image discovery (such as selecting colours, shapes, or texture patterns, or drawing a sketch) are very cumbersome in comparison to the ease of specifying a query for TBIR.

- TBIR is computationally much less complex than CBIR and only takes a small fraction of the latter's processing time to complete a search.

2.7 Image Retrieval Systems

An immense number of image retrieval systems has been developed ever since the first applications emerged in the early 1980s [48]. Several overviews on image database systems, image retrieval or multimedia information systems have been published in the literature, such as [82, 151, 295, 439]. An almost exhaustive overview describing (mainly CBIR) systems and their corresponding techniques in that phase was compiled in [99, 463]. Further, Veltkamp [461] provides information about features, types of query specification, matching techniques (similarity measures), index organisation, relevance feedback and result presentation of more than 50 retrieval systems (2002). More recent surveys include [295], in which techniques and systems used for medical image retrieval are reviewed.

Rather than attempting to list as many systems as possible, this section will only present a selection of well-known and often-cited image retrieval systems and explain their impact on the domain of VIR or their influence on the research presented in this thesis respectively. In particular, in accordance with the scope of the thesis, their query paradigms, feature extraction and similarity measures will be elaborated on, while related issues like usability or relevance feedback are not further discussed.

2.7.1 Early Image Retrieval Systems

Most of the early image retrieval systems were based on database management systems that provided the architecture to support a concept-based retrieval approach. These systems such as [48], which is often cited as one of the first image retrieval applications, seem to be more relevant for computer historians and are only of limited significance with respect to the scope of this thesis; they are thus not further discussed. Reviews and surveys in the early 1980s include [50, 439], with updated versions such as [10, 49] available at the beginning of the 1990s.

2.7.2 Commercial Image Retrieval Systems

The beginning of the 1990s saw the emergence of the first commercial CBIR systems. The most popular of these are *QBIC* and *Virage*, which are further described below. Other commercial CBIR systems include *RetrievalWare* by Excalibur Technology[6], *ImageFinder* by Attrasoft[7] and *IMatch* by MWLabs[8].

QBIC

Query By Image Content (QBIC) [117, 318] is a very well-known, commercial CBIRS developed by the *IBM Almaden Research Centre*[9]; it is by far the most cited system in the image retrieval literature, where it is commonly regarded as the first application (1993) that really performed content-based retrieval depending on a number of features that can be selected by the user [287, 295]. QBIC also offers some functionality for video retrieval, which is further explained in [117].

QBIC extracts colour *features* for individual objects or the entire image in several colour spaces (including RGB, CIE L*a*b* and Munsell) and represents them in a 256-dimensional colour histogram. The texture features include modified versions of the Tamura features coarseness, contrast and directionality. The shape features comprise shape area, circularity, eccentricity, major axis orientation, and a set of algebraic moment invariants. Queries can be specified by providing sample images (QBE), user-constructed sketches, and/or certain colour and texture patterns which can be selected from a sampler. QBIC also allows for *textual* representations of images, which can then be used for querying as well. As one of the first systems to apply multidimensional *indexing* to enhance the speed performance of the system, QBIC uses R* trees to index colour, texture and shape features (for the latter ones, KLT is performed first). QBIC's *similarity measures* include weighted *Euclidean distance* (L2) for texture and shape comparison and QFD for the comparison of colour histograms.

[6] `http://www.excaltech.com/`

[7] `http://attrasoft.com/image.htm`

[8] `http://www.mwlabs.de/`

[9] `http://wwwqbic.almaden.ibm.com/`

Virage

Another well-known commercial system for image [14] and video [159] retrieval is offered by *Virage*[10]. It is one of the most successful products, with companies such as CNN or NBC among their customers; as a consequence, information about the retrieval techniques are not always made public. Virage also offers an extensible framework in the form of an *application programming interface* (API) for the development of client retrieval applications.

Apart from simple features such as global and local colour, texture and shapes, various domain specific primitives (*i.e.* feature types, computation, indexing and corresponding similarity measures) can be defined when developing an application. The graphical user interface (GUI) provided for the development of a query interface includes facilities for image queries (QBE), keywords for concept-based retrieval, the support for several image file formats and even queries by sketch. Both similarity measures and indexing schemata have to be provided in the aforementioned primitives by the developers themselves.

2.7.3 Academic Image Retrieval Systems

Since commercial systems can be expensive to acquire (and often specific information relating to their underlying algorithms and techniques is withheld due to commercial purposes), research has benefited more from freely available systems provided by academia. Some of the most influential well-known systems are that presented hereinafter are *PicHunter*, *PhotoBook*, *Blobworld*, and *MARS*.

Photobook

Photobook [336] was developed by *MIT Media Laboratory*[11] and represents one of the first academic prototypes for VIR. Its implementation includes retrieval mechanisms for two-dimensional shapes, texture images and face recognition; the technology

[10] http://www.virage.com/products/vir-irw.html

[11] http://www.media.mit.edu/

for the latter one was also used by *Viisage Technology*[12] in their *FaceID package* employed in several US police departments.

Although Photobook also supports colour features, the main concentration is on texture and shape. The texture features are computed as the sum of the three orthogonal Wold components: periodicity, directionality and randomness. The shape description is based on the extraction of the boundary, which is then described by corners and curvature points. Queries are created by selecting still images from grid (QBE) or by entering an annotation filter (TBIR). Shape similarity is primarily calculated using the deformation effort; other similarity measures include Euclidean, Mahalanobis, vector space angle, histogram, Fourier peak and wavelet tree distances.

PicHunter

PicHunter [70, 71, 72, 73] is another example of a freely available image retrieval system and was developed by the *NEC Research Institute* (which is now a part of *NEC Laboratories America*[13] after a merger in 2002). Its functionality is thereby based on the assumption that a user is looking for an exact image in the database and therefore presents one of the first applications for target testing searches.

The content descriptors are mainly based on hidden alphanumeric representations as well as colour features in the HSV and RGB spaces that are represented as colour histograms and correlograms (both HSV) as well as CCV (RGB). The queries are specified using QBE, and the similarity between the individual features (*i.e.* colour vectors) is calculated using the L_1 distance.

MARS

MARS [326, 370] stands for *Multimedia Archival and Retrieval System(s)* and describes a series of systems first developed by the *Department of Computer Science* at the *University of Illinois at Urbana-Champaign*[14] and further improved at the

[12] `http://www.viisage.com/`

[13] `http://www.nec-labs.com/`

[14] `http://www-db.ics.uci.edu/pages/research/mars.shtml`

Department of Information and Computer Science at the *University of California, Irvine*[15].

In MARS, colour is represented in a two-dimensional histogram over the HS coordinates of the HSV colour space (the V component is neglected because it can be influenced by lighting conditions); the texture features coarseness and directionality are also stored in histograms, while the contrast of the texture is stored in a scalar; the boundary of the shape is described using Fourier descriptors (FD).

Queries allow boolean operators and can comprise any combination of the low level features colour, texture, and shape (that can be chosen from a palette) as well as textual descriptions (as keywords can be integrated as well). Histogram intersection is used to compare colour histograms, the weighted sum of the Euclidean (L_2) distance for texture similarity, and the weighted sum of the standard deviations of the magnitude and phase angles of the FD coefficients for shape similarity.

Blobworld

Blobworld [41] was developed by the *Computer Science Division* of the *University of California at Berkeley*[16] and was one of the first retrieval systems to use image regions for the query process. Several updated versions with significant changes were published over time [42, 43] until the research project finally ended in 2004.

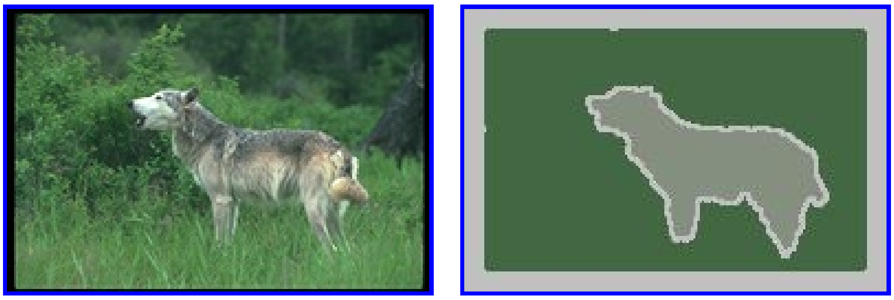


Figure 2.5: Blobworld: a real and segmented image of a wolf.

Before the feature extraction process is started, the image is first segmented into regions. The first versions [40, 41, 43] used 6 features for segmentation, a colour

[15] `http://www.ics.uci.edu/`

[16] `http://elib.cs.berkeley.edu/blobworld/`

histogram based on the HSV space to store the colours and ellipses to symbolise the regions, whereas the latter versions such as described in [42] made use of 8 features for segmentation, the CIE L*a*b* space and the real boundaries of the regions as illustrated in Figure 2.5 respectively. For each region, mean contrast and anisotropy are used as texture features, while approximate area, eccentricity and orientation quantify the corresponding shapes.

The query interface allows the user to select a category (to limit the search space) and the regions (blobs) of an initial image. The importance of the selected blob as well as the importance of the colour, texture, location and shape within that blob can be indicated and form the basis for retrieval. The colour histograms are hereby matched using QFD, and the Euclidean (L_2) distance quantifies the similarity of the texture descriptors and of the centroids. R* trees are used for indexing purposes.

Other Academic Systems

Other academic systems that are alluded in the context of this thesis include the following:

PicToSeek [129, 130] defines colour and shape invariants as features in content-based queries to guarantee invariancy of camera viewpoint, illumination conditions as well as the geometry of the objects.

NeTra [257] is another example of a system that uses image segmentation. First, an image is divided into regions of homogenous colour, and then colour, texture, shape and spatial location are extracted from those regions.

VisualSEEk [410, 411] employs a similar approach and also decomposes each image into regions of dominant colours. Again, feature properties and spatial relations are retained for each region.

ASSERT is described in [398] and is specifically targeted towards retrieval of high-resolution computed tomography images of the lung.

2.7.4 Internet Image Search Systems

Near the turn of the 21st century, the basic concept of similarity search was also transferred to several internet search engines, such as *WebSEEk*, *WebSeer*, *WebMARS* and *ImageRover*, which are further illustrated below. This section also presents a description of the nowadays extremely popular online systems such as *Google*, *Yahoo!*, and *PicSearch* as well as a few examples of meta-search engines.

WebSeer

WebSeer [124] was developed by the *Department of Computer Science* at the *University of Chicago*[17] and is often cited as one of the first web-based image retrieval applications, for example in [237], because the system collects the target images from the WWW before any calculation is started.

First, very simple colour features are extracted from an image collected from the WWW and stored in a simple colour histogram using the RGB colour space in order to separate photographs from drawings. Then, keywords are extracted from textual information provided on the web pages, including the image's filename and logical representation as well as associated hyperlinks, alternate texts and HTML (Hypertext Markup Language) titles. The search requests are subsequently based on these keywords, and the user has further options to specify image dimensions, file sizes or whether the required image should be a photograph or a drawing.

WebSEEk

WebSEEk [406] was developed by the *Image and Advanced Television Laboratory at Columbia University*[18] and is basically the online version of VisualSEEk. Like WebSeer, the target images are first collected from the WWW by an autonomous Web robot.

The retrieval algorithm relies on text and colour based queries, whereby the colour is stored in the HSV space in a normalised histogram using 166 bins and

[17] `http://www.cs.uchicago.edu/`

[18] `http://www.ctr.columbia.edu/WebSEEk/`

the QFD as the similarity measure. Spatial relations and texture matching features are also supported. The query process is initiated by selecting a subject from the available catalogue or by entering a topic (keywords). The colour features can further be used from the second search iteration on.

ImageRover

Another often cited early WWW image search system is *ImageRover* [391, 441], which was developed at the *Computer Science Department of Boston University*[19].

Image retrieval is based on textual and visual statistics that are combined in one single index. *Latent semantic indexing* (LSI) is hereby used to capture the textual statistics, while colour and texture orientation histograms are employed to store the visual features of the entire image and of five subregions. The colour histograms are computed in the CIE L*u*v* space and use 64 bins, whereas the texture direction distribution is based on steerable pyramids. Indexing is done by an optimised KD tree. The query process starts with keywords, and the visual features are further used to refine the queries through relevance feedback.

WebMARS

WebMARS [327] presents another system worth mentioning as it is the web-based extension of MARS and was created to manage images and text in HTML documents.

The underlying technique is based on the creation of a multimedia object model, with its contained objects forming the basis for the execution of the queries. The visual (and textual) features as well as the similarity measures comply with the ones used in MARS: HSV colour space, simple histogram intersection and colour moments are mainly used. The combination of textual and visual features is thereby shown to produce better results than any of the two techniques used alone.

[19] `http://www.cs.bu.edu/groups/ivc/ImageRover/`

Recent Popular Internet Search Engines

The aforementioned web-based search engines can be considered as precursors for the popular Internet search engines of the present. The three most used search engines for web-based image retrieval worldwide are provided by *Google*, *Yahoo* and *PicSearch*, whereby all of them are exclusively performing concept-based retrieval.

PicSearch[20] was the first of the three big providers to launch its image retrieval engine in 2001. Although perhaps not as famous as the other two competitors, its success is certainly not inferior, in comparison, as *PicSearch* provides (or has provided) their service as underlying technology for other leading Internet properties such as *MSN Search* (now *Windows Live Search*[21]), *Ask.com*[22] and *Lycos*[23].

Google Image Search[24] might be the most popular image search engine for the WWW. It was launched in December 2001 and, like PicSearch, relies on purely concept-based retrieval including image filenames, link texts pointing to the images and texts adjacent to the image.

Yahoo! Search[25] had ironically used Google as a mirror until it finally launched its own image search engine in 2004 (based on the technology of its acquisitions such as *Inktomi*[26], *AlltheWeb*[27] and *Altavista*[28]) and is now Google's fiercest rival.

Meta-Search Engines

Meta-search engines are search engines that send user requests to several other search engines and/or databases and subsequently combine the individual results from each of them.

The first meta-search engine for images might have been *MetaSeek* [25], which used the combination of several early image search engines such as *QBIC*, *Vi-*

[20] http://www.picsearch.com/

[21] http://search.live.com/

[22] http://www.ask.com/

[23] http://www.lycos.com/

[24] http://images.google.com/

[25] http://www.yahoo.com/

[26] http://www.inktome.com/

[27] http://www.alltheweb.com/

[28] http://www.altavista.com/

rage, *WebSEEk* and *VisualSEEk*. Currently popular meta-search engines include *MetaCrawler*[29], *DogPile*[30] and *WebCrawler*[31], which use the top search results of other aforementioned popular search engines such as *Google*, *Yahoo!*, *Windows Live Search*, *Ask.com*, *About.com*[32], and *LookSmart*[33].

2.7.5 Recent and Influential Systems

The first decade of the 21st century has witnessed an exploding number of systems being developed and presented (see [295, 461] for recent surveys). The two most significant and influential systems in the context of this research are GIFT and FIRE.

GIFT

GIFT [291, 307] stands for *GNU Image Finding Tool* and is an open source CBIRS based on *VIPER* [428], the result of a research effort by the *Vision Group*[34] at the *Centre Universitaire d'Informatique*[35] (Computer Science Department) of the *University of Geneva*, Switzerland. A demonstration of the current version of GIFT can be found at the *VIPER web page*[36], and the latest version of the program can be downloaded from the GNU web page[37] free of charge under the GNU General Public License (GPL); the source code is also available from GNU's *SourceForge* clone *Savannah*[38].

As far as feature extraction is concerned, GIFT uses a palette of 166 colours in HSV colour space and represents these colours in a global colour histogram (whereby bins containing zero pixels are discarded) as well as in square blocks ranging from 16x16 to 128x128 pixels after the image has been normalised to 256x256 pixels).

[29] `http://www.metacrawler.com/`
[30] `http://www.dogpile.com/`
[31] `http://www.webcrawler.com/`
[32] `http://www.about.com/`
[33] `http://search.looksmart.com/`
[34] `http://vision.unige.ch/`
[35] `http://cui.unige.ch/LeCUI.html`
[36] `http://viper.unige.ch/demo/php/demo.php`
[37] `ftp://ftp.gnu.org/gnu/gift`
[38] `http://savannah.gnu.org/projects/gift/`

Further, Gabor filters are used for local and global texture features, and inverted files for indexation. A simple histogram intersection is used as a distance measure for the colour histograms, while quite surprisingly, concept-based similarity measures such as simple *td/idf weighting* are used to rank the images. A number of articles have been published on GIFT's retrieval techniques and the underlying architecture, including [287, 291, 307, 425, 426, 427, 428].

This system has been of importance for this project as it was used at *ImageCLEF* to provide a CBIR baseline run for participants who wanted to explore combined (*i.e.* concept-based and content-based) retrieval approaches and did not have the know-how or time to further investigate CBIR techniques [62, 63].

FIRE

FIRE [85, 87] stands for *Flexible Image Retrieval Engine* and was developed by the *Human Language Technology and Pattern Recognition Group* of the *RWTH Aachen University*[39]. Like with GIFT, an online demonstration is available[40] and the application and its source code can be downloaded[41] under a GNU GPL. FIRE was used for a visual pre-analysis of the search topics at *ImageCLEFphoto 2006* and helped to classify these topics according to their "visuality" (see also Chapter 7).

FIRE provides concept-based and/or content-based retrieval of images. Its text retrieval engine implements a variant of the *Smart-2* retrieval metric, which itself is based on the *tf-idf* metric. As for the visual features, colour histograms are used to describe the colour and the Tamura features for texture respectively. The default distance for both is the JD, although FIRE also allows for the selection of up to 40 different similarity measures including the L_1 and L_2 distances as well as histogram intersection.

[39] `http://www-i6.informatik.rwth-aachen.de/`

[40] `http://www-i6.informatik.rwth-aachen.de/~deselaers/cgi_bin/fire.cgi`

[41] `http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html`

2.8 Summary

This chapter introduced and explored the main concepts and challenges of VIR.

First, a general introduction was given in which the evolution, goals and challenges of VIR were addressed and the main components of an image retrieval system were explained. The characteristics of visual information queries were examined, offering answers to questions including why there is a need for VIR, who its current and potential users are, what exactly these users look for, and how these user needs can be classified and expressed.

Then, a comprehensive overview of the visual and textual descriptors that can be extracted from images and their logical alphanumeric representations was given, and the major similarity measures were presented. These descriptors and measures form the basis for the result generation and presentation process, which were subsequently covered as well. One section touched on related issues such as relevance feedback, usability aspects and the current problems within this field of research: the most predominant being the semantic gap, which researchers have unsuccessfully tried to bridge from either side: bottom-up (CBIR) approaches as well top-down (TBIR) approaches.

Finally, a number of well-known and often-cited image retrieval systems were presented and their impact on the domain of VIR in general, as well as their influence on the research presented in this thesis in particular, was explained.

The first of the two main literature review chapters provided the fundamental theoretic foundations needed to understand the functionality of VIR systems; this is an essential prerequisite for the comprehension of the next chapter, which deals with the state-of-the-art methods regarding the analysis and evaluation of these systems.