Content-Based Image Retrieval and the Use of Neural Networks for User Adaptation

Der Technischen Fakultät

der Universität Bielefeld

vorgelegt von

Tanja Katharina Kämpfe

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

Mai 2006

Acknowledgements

This work was done in the Neuroinformatics group, headed by Prof. Dr. Helge Ritter, at the Faculty of Technology, University of Bielefeld. The basis for this work was provided within the BMBF-project *Lernen zur Organisation komplexer Systeme in der Informationsverarbeitung (LOKI)*.

Without the assistance of various people this work would not have come into existence. So, I want to thank Helge Ritter for confidence and fruitful proposals regarding my work, Petra Udelhoven for positive talks and her help in official things, the members of the Neuroinformatics group as well as of the LOKI project for the constructive working environment, Thorsten Uhde, Jens Bories, Axel Saalbach, Till Bovermann and Kai Essig for reviewing this manuscript, Thorsten Twellmann for being a pleasant office colleague, Volker Wendt and Daniel Hänle for developing their diploma theses under my supervision and particularly Tim Nattkemper for various support.

Last but not least, I want to thank my family, Jens Bories, Anne Salich and my friends for having patience with me and supporting me during the last years.

Gedruckt auf alterungsbeständigem Papier °° ISO 9706

Contents

Chapter 1

Introduction

In ancient times the knowledge of a community was concentrated in the mind of the elders and sages. Consequently searching for information meant asking these people. Since those days the world has changed. Today the knowledge and information mankind has collected exceed the mental capacity of any single human mind. Different storage media have been developed: Wall paintings, stone scripts, parchment scripts, books, movies or digital media, to name a few. Today the existing information forms a vast amount of data. Thus the way to get the desired information had to change and therewith the *information retrieval system* altered from an omniscient human mind over a human librarian to an automated system. However, one thing has changed only slightly: The human race generates a visually oriented society. Pictorial information has loomed large in most times and societies.

Thus people are taking pictures – a lot of pictures. Moreover, recent developments regarding digital camera technique boost the human collecting passion. The result is a vast and increasing number of digitally stored images. Therefore getting a desired picture means searching in this enormous and unstructured image set. With the increasing number of images in such a collection the searching for a specific picture becomes more and more difficult and longsome. Thus automated systems to support the search are desired.

Such image retrieval systems should perform in a way, satisfactory for the user. Therefore advanced approaches are necessary for developing systems which perform in a way resembling the human way of retrieving and comparing images. Since this is usually based on the image content, the content itself is the most important feature. Today Content-Based Image Retrieval (CBIR) is established as an important field of research, embracing various research tasks. In this work selected challenges regarding user friendly image retrieval are researched.

Based on the changing technical possibilities and the enormous increase of given images the special challenges of image retrieval are presented in chapter 2. Outstanding tasks regarding image retrieval are reviewed, namely search tasks, similarity searches and the semantic gap. CBIR-systems consist of different components. Interface design, retrieval unit and data storage are analysed regarding their functionalities. Various systems and frameworks are presented, partly developed within this work. These build the basis for elaborated researches of selected image retrieval tasks.

Image retrieval means searching in digital image data. Every image set is different and offers individual qualities and challenges. Thus, in chapter 3 general approaches to describe image sets are reviewed. Grouping images with equal features into subsets, called

categories, is introduced. Image sequences as special subsets are presented. They offer an inherent structure described by time stamps. The image data used in this work (photos, image sequences, webcam images) are presented.

Usually raw image data are little suitable to perform automatic searches. Hence image features based on specific attributes are used to describe the image data. The implemented image features are introduced and analysed regarding the present image data. Therefore the distributions of the image sets in the feature spaces are considered.

Retrieving or organising images can be realised by a number of different approaches. Users usually look at pictures one by one and thus a sequential alignment is desired. A one dimensional Self-Organising Map (1dSOM) is proposed since SOMs are popular for topological preserving mappings. In chapter 4 applications of 1dSOM to align as well as to group images are presented.

Image retrieval research aims at getting automatic approaches. On the other hand, the human user is the most important factor with respect to image retrieval systems. He cannot be replaced or simulated completely. Consequently the systems have to be trained based on user interactions. This will be realised by a *relevance feedback* (chapter 5). General approaches to support the relevance feedback are introduced. The feature *relevant* is put into relation to the feature *interesting*. Similarity models and different methods to achieve user adaptation are presented.

Usually the data spaces representing images do not correspond to the human recognition of images. Thus this data has to be altered to more user adapted representations. Therefore suitable transformations are necessary. The Independent Component Analysis (ICA) computes meaningful directions within a data set. Thus this approach is used for relevance feedback purposes.

ICA is applied to improve image classifications. Image retrieval can be implemented as a classification into relevant and non-relevant images. Such a classifier can be trained based on relevance feedback data. Therefore ICA is inserted as a preprocessing step in a Naive Bayes Classifier. Therewith statistical independent directions are computed to confirm the optimum classification approach. The training of the classifier is based on the relevant data. In doing so the utilisation of the relevance feedback is considered. Moreover, ICA applied on image data is analysed in general.

A number of different image retrieval systems, approaches and components have been developed in the recent years. Their evaluation is miscellanceous since various challenges have to be viewed. For example individual processing steps have to be analysed and entire systems have to be rated regarding their performance. In chapter 6 different ways of CBIR evaluation are reviewed with respect to the presented retrieval systems and approaches.

This work concludes with a summary and propositions for subsequent challenges in chapter 7.

Chapter 2

Information and Image Retrieval

Looking for information, people or objects has always been an important task for human mankind. In the modern world this particularly applies to the retrieval of text and images. Against the background of data storage and camera technique developments, collecting, archiving and retrieving images is reviewed. Specific challenges regarding image retrieval are outlined. Various systems and frameworks focussing different retrieval tasks are presented.

2.1 Historical Developments of Information Storage and Retrieval

Since men started to write down information on any portable media the number of collected data has increased. The spread of knowledge over time and space has become independent from the human author and a human transmitter. Fortified by these developments mankind has turned out to be an information society which requires information retrieval frameworks in numerous situations. This section gives an overview of the historical development of information storage and retrieval with a closer attention to pictorial data in the last paragraph.

2.1.1 Document Collections and Data Storage

The invention of printing by Johann Gutenberg in the 15th century marks a milestone in information storage, duplication and distribution. Data was collected on portable media before but from then on the circulation of discoveries and knowledge around the world has become much easier and the amount of documents containing information has bursted. Consequently the number and dimension of libraries increased in the following centuries. These collections contain predominantly books and therein most of the knowledge is described textually. Indeed further data types offering information have existed at all times. Paintings represent famous persons or important incidents. Maps document geographical knowledge. Numerical data describe population developments as well as mercantile activities. All these different types of information are coded in different data types but mostly stored as paper copies in a library. The number of books and documents reflects the magnitude and importance of these collections. To name an example, the 400 years old Bodleian Library in Oxford [bodleian] is well known and nowadays it holds about 7,000,000 books.

Until the middle of the 20th century this kind of information storage in book libraries had been the state of the art. Then the invention of the computer initiated new techniques to collect information. Henceforth the distribution and duplication of documents have become much more easy, fast and cheaper by switching from hard media like paper books to digital media. Today its development offers data storage by low cost and at the same time easy access. In addition duplication procedures do not cause any information loss.

These developments hold for the data types named above as well as for other data types which require special storage media in the pre-digital era. For example music or sound had to be stored on shellac or vinyl discs and films were available on celluloid bands. Regarding the storage on a digital computer hard disc the data type does not matter. Just the output device to present the information to the user depends on the respective data type.

While many conditions regarding information collections have changed, one attribute is still valid: Their impact is often measured by their size, which means by the number of stored data items. And the modern technologies facilitate recording arbitrary data, e.g. in [Large et al., 2001] is suggested that *more information have been produced between 1970 and 2000 than in the previous 5000 years.*

The resulting information overload is amplified by the increasing usage of the internet since the 1990s. This highly interactive medium is characterised by its broad distribution as well as the lack of any restrictions for publishing. Every user possibly is able to present arbitrary data, text in the same way as pictures, films or sound. The huge variety of different data types available in the internet and particularly the combinations of different data types are subsumed by the term *multimedia data.*

Faced with such an amount of unstructured and varying documents, some questions arose:

- How can I find a specific document?
- How can I detect relevant and reliable informations regarding a desired topic?
- Where is the contents of these documents summarised?

These tasks are subsumed by the term *information retrieval.*

2.1.2 Information Retrieval

People want to utilise different information and data for their own purpose. For example researchers want to upgrade the insights of earlier research activities. Therefore, they often need documents and information other persons had collected. They have to perform an *information retrieval.* Usually this requires an intermediate to bring the searching human and the collected data together. In former times a librarian performed this task and fetched the desired book from the library. Since then the libraries grew and a single person could no longer keep all books in mind. Hence most of the libraries developed specific systems to array their books. Alphabetical orders based on author or title occured as well as systematic or completely individual arrangements which just the local librarians understood. The most successful and persitent ones used card systems and resembled current indexing techniques [Wellisch, 1991].

The basic principle of such indexing systems is to take a set of keys representing the individual book or more general identifying any document of an arbitrary data type. In the

early systems author and title were used as keys and written down on cards added by the physical location of the document. Unfortunately author and title are often not available, may be ambiguous and usually do not represent the content of a document sufficiently. So the main questions according indexing are *What can be a key?* and preceding *What should such keys achieve?*

Usually humans communicate by speech. And they describe the information they are looking for with words. It was self-evident that the keys had to be meaningful words or at least reasonable combinations of letters and numbers. The concept key*word* has been born [Luhn, 1961] [Bowden et al., 1998]. Obviously the assertion of suitable keywords to each document is essential to facilitate the searching for information according to a specific topic. This very important step has to be done a priori and accurately to ensure the retrieval of all but only relevant documents. Unfortunately this mapping is very time-consuming as well as subjective. Indeed the invention of computers offered a lot of approaches to support keyword based information retrieval.

First of all automated systems offer the prospect to manage the keywords. The common index frameworks were implemented directly. Since the early 80's OPACs (Online Public Access Catalogues) [Efthimiadis, 1990] have substituted common card catalogues. In online libraries or internet bookshops like *amazon* [amazon] the title or the author's name constitute the common queries. Further keys are identified by predefined categories like thrillers, horror, nonfiction or science.

Computers perform a lot of virtual arrangements of documents according a priori asserted attributes. Given suitable keywords, the retrieval according these keys is quite easy and a number of very good search algorithms in indexed data sets exist today [Baeza-Yates and Ribeiro-Neto, 1999]. But these systems depend highly on the a priori assignment. On digital stored documents computers can be used to find these attributes. Automated keyword detection is an absorbing field of research. So *How to perform an automated keyword detection?*

In digital text collections the keyword detection may be straightforward: Each word of the text can be used as a keyword and every document containing the desired set of words can be retrieved. But this may result in a bulky useless response. Especially for searching the web this is true, since the internet contains a vast quantity of documents. Consequently the user has to choose his input carefully. Hence most of the common internet search engines like *yahoo* [yahoo] and *google* [google] rank the detected documents according a relevance assumption to help the user.

Unfortunately these relevance rankings are not helpful in any case and the result lists are still very voluminous. To lessen these drawbacks meta search engines – e.g. *metacrawler* [metacrawler] and *searchengineswatch* [searchengineswatch] – have been developed to combine the results of a set of search engines to a more helpful result list.

More advanced approaches to enhance information retrieval in text documents are developed in the research field known as *textmining* [Baeza-Yates and Ribeiro-Neto, 1999]. Known as *the bag of words* [Salton and Buckley, 1988] a term weighting approach to enhance text indexing is established. Other prosperous examples are the clustering of text documents or using Wordnet [Hotho et al., 2003] [Sedding and Kazakov, 2004].

Thus an automated keyword detection in digital text documents is possible and the matching between the input words of the user and the keywords representing the stored documents can be performed straightforward. Indeed users will formulate their query by

words independently from the data type to retrieve. A matching between different data types is postulated and the automated assignment of meaningful keywords to any type of documents is an obvious demand.

In the pre-digital era the kind of data did not matter. The human librarian could assign each kind of document to a keyword, books as well as pictures. But in the online era the digitalisation of information bears new challenges.

For example an image can be represented by the objects it contains. Unfortunately an object detection in images is not generally performed by automated systems. Therewith an enumeration of the contained objects is hard to achieve. See section 2.1.3 for a deeper discussion of the automated indexing of images.

In general the assignment of a word to a document of an arbitrary datatype is a very hard task and the question for automated keyword assignments is still open. To support the looking for relevant documents according to a specific topic some remarks on keywords are indicated:

- Are keywords impartial?
  To get a universal set of keywords describing a document, these words should be objective. On the other hand every user has his own intention regarding a document. Often this changes even for one user over time. Consequently the keyword detection is a subjective task [Colombo et al., 1999]. An example of such user depending keywords is described in [Weinberg, 1987] as the difference between *aboutness* and *aspects*. While the content of a document can be represented clearly, verbalising ideas and theories is much more difficult.

- To what extent can a limited number of keywords describe the contents of a document?
  The keywords should ensure that the retrieved documents bear relevant information according the user's query. Therefore the keywords must summarise the content adequately. Unfortunately a limited number of keywords cannot subsume every subject of a document. This particularly is true for pictures (see section 2.1.3).

- Is a keyword based information retrieval user-friendly?
  To specify the desired subject keywords are used which must represent the user's need. Since humans are familiar with expressing their intentions with words this may be a convenient approach and is still required by many end-users [Munson and Tsymbalenko, 2001].
  On the other hand the variety of possible search topics is unbounded while the number of provided keywords is limited. It is impossible to represent every user's need. Hence the user has to conform himself to the synopsis (keyword set) of the library. Amongst others, this scares off user who requires indexes on another level of specificity [Weinberg, 1987]. The adaptation of the system to the user would be more user-friendly.

- Which requirements should a set of keywords fulfill?
  Keywords should be meaningful and self-explanatory, identify a specific group and describe the content of the document. Considering the whole set of keywords, every aspect which can be interesting for any user should be covered [Wellisch, 1991].

These challenges are heavy drawbacks of keyword based indexing frameworks for a lot of information retrieval tasks. To avoid these difficulties the keyword detection step may be substituted by a more general feature detection step. The retrieval switches from the exact matching of keywords to the similarity detection of contents. Furthermore an adaptation of the system to the user becomes desireable, since the information need as well as the relevance of a document is user dependent. This leads to modern information retrieval approaches. Here information storage and retrieval take place in one Information Retrieval (IR) system.

Fortified by the increasing independence from an experienced intermediate connecting users and information, the development of user-friendly interfaces becomes more important. Until the early 1980's intermediates had to activate the search engine and interprete the information. Since that time user-friendly interfaces have been developed and users interact with the system without any guidance. In [Large et al., 2001] this is described as *What end-user interfaces have done is create the illusion that searching (often complex) databases is easy.*

Different tasks have to be regarded at this point of an information retrieval framework: The user knows about his particular retrieval task and sometimes the data domain, but he is not familiar with the storage system or the retrieval approach. In a specific scenario this means that the user's intention given by an informal textual description has to be translated to a formal description a computer can act with. A possible approach will be presented according to exemplary retrieval systems in section 2.2.

Today interface design establishes a whole research field, known as human-computer-interaction. Consequently this is an important step in developing information retrieval systems [Eakins et al., 2004]. Electronic forms, hypertext and graphical interfaces are between the document and the user and therewith substitute the human intermediate.

The listed requirements are true for information retrieval in any kind of data. In particular image and multimedia retrieval or searching in the web depends on suitable technical facilities. Furthermore a development from a visual oriented community in the dawn of mankind to a textoriented society promoted by the information transportation media like letterprint back to a visual oriented society today forced by technical inventions like television or visual telephone can be observed. Thus pictorial data become more and more important. Although languages, writing and a large variety of information coding schemes have been developed, humans are still thinking viually. Therefore, visual information retrieval is considered in more detail in the following section.

2.1.3 Visual Information Retrieval

Pictures have been important for the human race at all times and a lot of different techniques and intentions occured during the millennia. The eldest proofs of men-made pictures are wall paintings in caverns which are about 15,000 years old[1]. Since the days of these hand painted images a lot of developments have passed and the variety of different pictorial documents has increased. Especially the invention of photography at the end of the 19th century marks a milestone in image production comparable to the invention of printing.

Today pictorial symbols like signs or trademarks as well as photos and paintings serve a variety of purposes: Restroom labels for female and male, traffic signs, pictures of persons

[1] Altamira cave, Spain

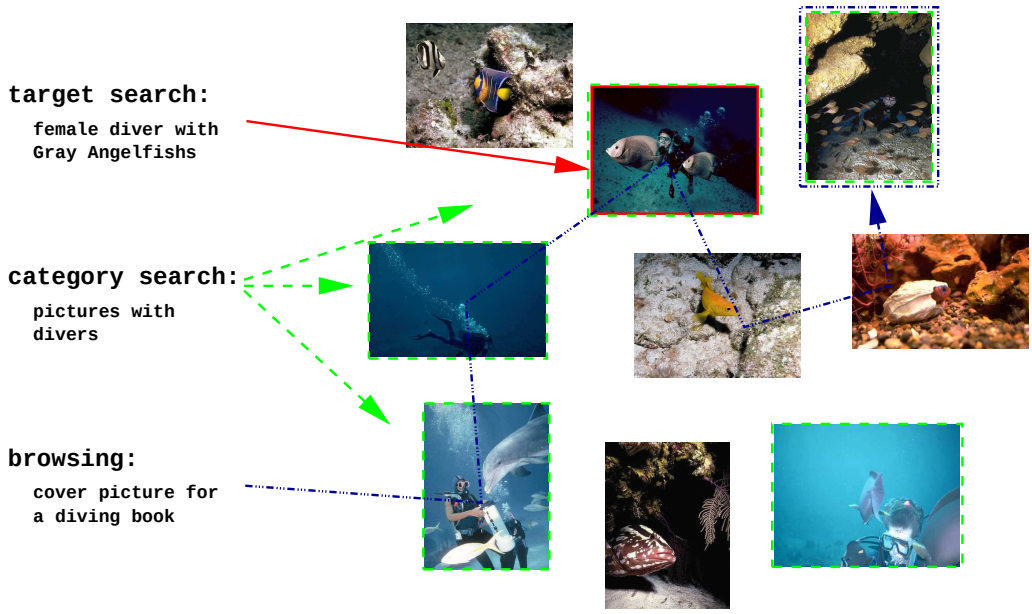


Figure 2.1: Three search tasks are figured in this small image set. The **target search** desires one picture and the **category search** a set of pictures. While **browsing** different images are inspected and choosen based on a quite vague intention. Search tasks like category search are introduce for example in [Newsam et al., 2001].

or situations, illustrated newspaper articles and TV-news, identification of persons by their passport.

The development of digital cameras has effected a substantial progress in collecting visual information since the 1980's [Haslego, 2005]. With the circulation of easy-to-handle equipment the number of people producing images has increased. Museums, archives and scientists produce pictures as well as professional photographers, private persons or governmental organisations. Digital cameras as well as the amount of storage equipment intensify this trend. From now on images can be recorded and archieved with low costs, for example by cameras that are small enough to fit in common mobile telephones or by online connections to automated cameras, called *webcams*. Consequently today an inconceivable amount of miscellaneous pictures exists and is kept in different independent archives.

While still various kinds of pictures like paintings, photographs and films exist today most of them are stored digitally. This motivates automatic management systems to handle the visual information in the image sets like the US NSF Visual Information Management Systems [Jain, 1992].

A usual situation of image retrieval will occur in the following way: A user is looking for the painting *Cafe Terrace on the Place du Forum* by Vincent van Gogh. If he knows the name and the painter this is easy since usually these meta data is stored together with the picture. Or he just knows that it had been painted by van Gogh. In the set of all retrieved van Gogh paintings he will browse until the desired image is found. And maybe he will change his mind and choose another van Gogh painting.

Three types of retrieval tasks occur in this example: Target search, category search and browsing (see figure 2.1):

Target search resembles the searching of an individual and a priori known book in a common library. Regarding image retrieval a specific picture the user has seen before and kept in his mind, is searched. Depending on the kind of image and the information stored together with the picture this retrieval task is easy to solve based on matching suitable metadata.

The aim of a **category search** is a set of images belonging to a somehow defined group (see section 3.1). A category can be defined a priori and labelled by a significant keyword. In this case it may be suitable to assign each picture to the appropriate categories while inserting it into the database. Then the retrieval can easily be performed by database matches of the keywords, e.g. all paintings of van Gogh.

Unfortunately the a priori labelling is a very expensive task and therefore often neglected. Furthermore the searcher may look for a category not defined in advance. In the above example this may be the set of all paintings with a theme located in Arles, France.

Browsing an image set means scanning through a set of pictures, sometimes without a well defined target. The imagination of the desired image may arise or change in the searchers mind while scanning the images stored in the collection. Comparable with the browsing in a shoe carton of private photos browsing in digital data can be performed without any guidelines. Nevertheless retrieval systems may give some assistance based on the documents the user recently has browsed through. Usually he has to rate the seen pictures and the system can adapt to the user. On the other hand this restrains the search space and the user may not find the most suitable image.

Up to now no outstanding difference regarding retrieval tasks between pictorial data and other types of data is obvious. Consequently it may be appropriate to perform visual information retrieval in a similar way as textual information retrieval by presenting the content by keywords and use an indexing framework. If it would be possible to extract a textual description of the image content automatically, common text retrieval can be used for content-based image retrieval tasks [Laaksonen et al., 2001].

Unfortunately at this point a difficulty arises: In text documents the words carry the semantic and the keyword detection can be implemented as a filter process on these small entities of the document. In difference to that the pixels of images do not provide any semantic description. Since current image segmentation approaches do not correspond to a reliable object recognition, automatically detected image segments are not suitable to represent image semantics. Thus an automatic assertion of pictures to text is not possible. Two interdependent questions illustrates the arising challenge: *What are keywords of images?* and *How to generate keywords of images?*

A workaround to manage the lack of keywords is performed in successful and popular internet search enginges, e.g. yahoo [yahoo] or google [google]. These systems offer image retrieval based on the image content such that the user has to announce a query by a keyword. The system then searchs for web-sites containing that word close to the presentation of a picture. Therewith the image search is just a text retrieval enhanced by the search for any picture identified by the data structure but not by the visual content of the picture. In [Munson and Tsymbalenko, 2001] this approach is stated as more user friendly than content based retrieval approaches.

In [Yee et al., 2003] such image search engines for the Internet are slightly compared based on number of results per page and existence of links. A slight review is presented in [TASI]. To show the impact of a probabilistic model in image gathering from the web, in [Yanai and Barnard, 2005] Google image search is used as a benchmark.

A plain approach to record the image content is still a human based labelling. But this is not practicable due to a lot of facts: The number of stored pictures exceeds the labelling capacities of men. A lot of pictures are taken automatically and inserted in image collections without any human observation. Since men regard images on different levels the labelling is very unreliable [Eakins, 2002]. The lowest level bases on primitive features like the predominant colour. More complex is an inferential view with logical descriptions and well-defined objects. And on the most advanced level just abstract attributes are used, e.g. spirits, impressions or feelings are desired. Each level causes different labellings and has to be kept in mind. At least regarding images in the internet the language and the cultural background of the different people complicates the human based labelling [Colombo et al., 1999]. Furthermore humans usually rate images or their similarity just on a transient view and a difference between linguistic and visual interpretations of images is observed [Enser, 1995].

Consequently different user based ratings of the same image in different treatments conflict with the required non-ambiguous description. Further problems will remain after an automated keyword detection: As in the famous saying *A picture tells more than 1,000 words* is subsumed a limited number of keywords cannot describe an image content completely [Smeulders et al., 2000].

Classical computer vision deals with a related problem, the demand *tell me, what's on this image*, and provides a large set of different more or less suitable approaches. Most of them base on code vectors or code vector histograms and are subsumed under the term *image features*. These are computed automatically and are called *low-level image features*, to distinguish them from high-level, semantic covering features. Low-level image features represent the computable image content, e.g. the predominant colour of a picture or a region within the picture. Usually such a feature is a vector of real numbers and therewith conflicts with database matches. A similarity search [Pecenović et al., 1998] [Eidenberger and Breiteneder, 2002] or classification task should be performed instead.

A data driven approach to get suitable keys for indexing images may be the detection of *representative blocks* [Zhu et al., 2000]. Based on vector quantisation image fragments are assigned to salient picture clippings, known as codebook elements. Based on such a codebook common text retrieval techniques can be used for image retrieval. Indeed textual queries are not possible in this approach.

To keep the user-friendliness of textual descriptors as well as the computability of low-level features in [Pecenović et al., 1998] these features are combined. Motivated from text retrieval they use latent semantic indexing (LSI) and singular value decomposition (SVD). Therewith an indexing can be performed independetly from the user.

Regarding the content detection of images one very important point becomes obvious: There is a wide gap between the human interpretation of images and the current computational possibilities to deal with pictorial data. This difference is called the *semantic gap* [Eidenberger and Breiteneder, 2002]. Furthermore the human based measurement of

visual similarity differs from mathematical distance metrics applied on low-level features.

According to [Eidenberger, 2004] a semantic enrichment of low-level features can uncover higher-level similarities between the query and the database canditates and narrow the semantic gap. But since both, an overall semantic description and the human intention cannot be generated in advance the user has to *teach* the system. Computationally this interactive image understanding means, that the user has to rate the systems performance, particularly the retrieval results. Based on these ratings the system's algorithm should be able to adapt to the user's need. This kind of user influence is called *relevance feedback*.

Although a lot of different approaches to improve the user rating exist, the user adaptation of the retrieval system is not solved satisfactorily. For example the relevance feedback requires very fast retrieval performance, since little can be computed in advance and stored in the database. Context dependencies are further restrictions complicating the distribution and evaluation of these algorithms.

Corresponding to the high relevance of feature detection and relevance feedback, these tasks are discussed more detailed in section 3.3 and chapter 5 respectively. As a conclusion may remain the remark of Laaksonen et al. [2001]: *Since the task of image retrieval is to find pictures a user would regard interesting, the user himself is an inseparable part of the query process.* Consequently the human subjectivity has to be strongly respected, quite harder than in other computer vision tasks.

2.2 Content-Based Image Retrieval Systems

Today CBIR-systems perform the role of an omniscient expert regarding a specified image collection. Whereas the enquiring part is still the user, modern CBIR-systems are automated computer programs and therefore based on mathematical algorithms. They may be welcome in a lot of application fields, e.g. crime prevention, photo-journalism, fashion design, trademark registration, medical diagnostic or education [Simon and Verstegen, 2004]. Furthermore a lot of people will use CBIR-systems for their image collection like private users, news agencies, scientists or scholars. Image retrieval can be performed on different kinds of pictures, namely dynamic scenes, image sequences, single images, subimages or image regions corresponding real world objects. These images are summed up by the term *pictorial entity* or the set of *pictorial entities*.

On a very base level each CBIR-system consists of quite a small number of units (see figure 2.2): A **user-interface** for query formulation, result presentation and – if performed – user rating input, a **retrieval unit** which rates the pictorial entities according to the user's query and possibly performs a kind of adaptation as well as a **data repository** keeping the image set.

In the **user interface** the subjective and discontinuous human image interpretation has to be matched to the defined algorithmic descriptions of an automated system. There arise a lot of different challenges subsumed in table 2.1. And two directions of data-transfer emerge at this point. The first direction goes from the user into the system for query formulation or ratings in a relevance feedback framework. The other way round the system has to present the retrieval results in a suitable manner. The latter task can be handled in a quite obvious way, as the system easily can assume the results to the desired

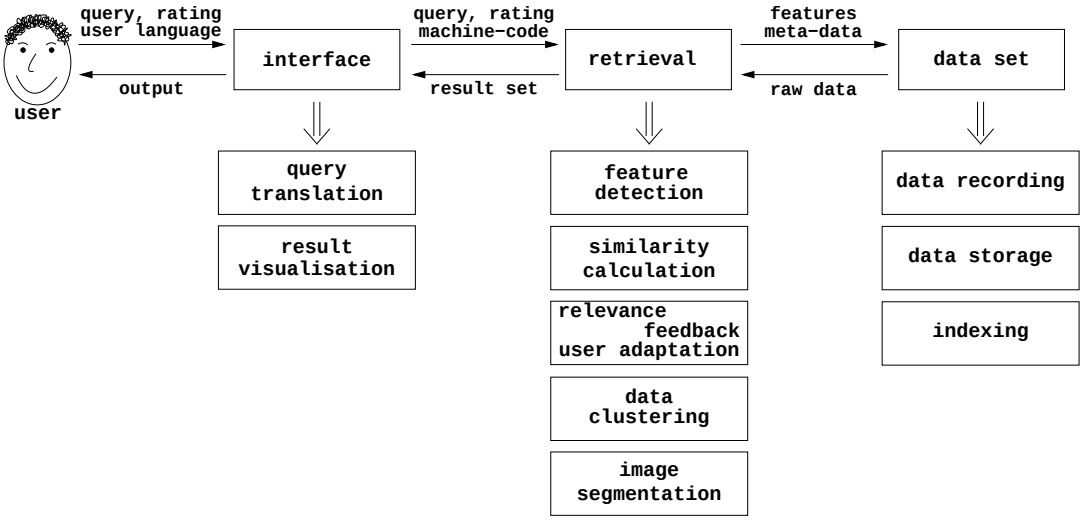


Figure 2.2: General overview of CBIR-systems: Basically each CBIR-system consist of three modules, which may be compound of further, partially optional components.

output mode, a set of images. The interface design is a more complex task. Especially the request for user ratings has to be figured out.

The **retrieval unit** constitutes the centre of the CBIR framework. A variety of partly essential, partly optional components are subsumed in it. The most important one is the retrieval step. One obvious realisation provides the alignment of the pictorial entities based on the similarity between each pictorial entity and the query. But even this self-evident implementation causes a number of new challenges. For example the similarity can be computed on the raw pixel values or on different image features. Although frequently realised as the inverse of a distance measure, the term *similarity* is not well-defined in a computer vision framework, since the various distance measures do not represent the human distance perception. First of all the human similarity judgement does not satisfy Euclidean metrics, which are preferable for automated mathematical analyses.

An optional but often very suitable component within the retrieval unit is the realisation of relevance feedback. In general the user ratings of preceding retrieval results have to be transformed into parameters affecting the retrieval step. Apparently the implementation depends on the implementation of the user rating as well as the retrieval step. Similarity and relevance feedback are presented more detailed in chapter 5.

Further steps within the retrieval unit may be feature detection, an adaptable grouping step (clustering), a combination step of different meta data (any textual description, the context of a picture or medical diagnoses) or an unrestricted number of components emanating from the developer's phantasy.

Technical requirements and retrieval functions determine the design of the **data repository**. The kind of the stored data can be very different, depending on the systems functionality. If all computer vision steps are processed during each retrieval step, in the repository just the raw image data have to be stored. Unfortunately a lot of reasonable computation steps are very time consuming. Therefore, they are performed in advance and their results are stored in the database.

A possible preprocessing step may be an automatic image partitioning. In this case the image segments are stored equitable to the entire images. The term pictorial entity has been introduced to subsume the different image types.

Another preprocessing step is the feature detection. Usually the corresponding algorithms are executed a priori and the feature vectors are stored for each pictorial entity. In most systems precomputed data are not changed during CBIR runs. Indeed user may enhance the image segmentation or add new pictures which should be integrated into the collection. The real implementation depends on the whole system and should be regarded in conjunction with the example systems.

	user	**system**
image features	unspecified, subjective	specified by algorithms, computational, objective
distances	not metric, nonlinear	usually metric
processing speed	fast	slow
number of treatable image	quite small	in fact unbound, but time dependent
reliability	changeable	repeatable
exactness	low	as high as possible

Table 2.1: Each CBIR-system can be considered from at least two points of view: (1) the users view and (2) the systems view.

Table 2.2 lists a number of different CBIR-systems along with some important attributes. In the following sections some example systems are presented based on these processing units as well as the initial motivation and some application possibilities.

2.2.1 PicSOM

Using the very powerful neural approach of Self-organising Maps (SOM) [Kohonen, 1997] Laaksonen et. all. have developed the framework PicSOM (Picture Self-Organising Maps) [Laaksonen et al., 2000], [Laaksonen et al., 2001]. Based on tree-structured SOMs (TR-SOMs) content-based image retrieval tasks are investigated.

Self-Organising Maps (SOMs) are neural networks which are widely used for different applications as well as analysed and enhanced theoretically [Kaski et al., 1998], [Oja et al., 2002]. Since the early 1990's SOMs are a well known approach to visualise data structures. On the two-dimensional grid of a classical SOM multi-dimensional data can be presented by conserving the topological relations. This is used in the well known document exploration tool WEBSOM [Kohonen et al., 2000] for text retrieval in the world wide web. Since the inherent structure of pictures differs from the structure of text, a new system has been developed for searching in a large picture collection.

The PicSOM framework uses SOMs to arrange images on a set of maps. The trained maps are used to find regions of the data space which may contain interesting images. Therewith a special approach to perform relevance feedback has been developed. Pictures are very complex and different features may be suitable to present the images. Hence a set of SOMs is used whereas each one is applied in another feature space. Furthermore

system	focus of the system
QBIC [IBM]	well known commercial image retrieval system by IBM
Photobook and FourEyes [Minka]	research project at the MIT, image segmentation and image models to perform keyword based retrieval
picHunter [Cox et al., 2000]	bayesian networks to group the images, enhanced evaluation by psychophysical experiments
MARS [Huang]	framework to investigate a number of approaches, mainly relevance feedback, region based retrieval and multimedia retrieval
Viper [Marchand-Maillet] and GIFT [Müller, 2002]	communication protocol MRML (Multimedia Retrieval Markup Language), extended evaluation by the benchathlon [benchathlon], open source version for common users
VisualSeek[Anastassiou, 2005] and WebSeek [Chang]	spatial features for region based retrieval
blobworld [Carson et al., 2002]	image segmentation for region based retrieval
NETRA [Ma and Manjunath, 1999]	multimedia retrieval, region based retrieval
SPIRIT and ARTISAN [Graham and Eakins, 1998] [Hussain and Eakins, 2004]	retrieval of trademark images
Visual Retrieval Ware [convera]	commercial retrieval ware with an upgrading for visual retrieval, semantics and indexing
PROMETHEUS [Verstegen, 2003]	image retrieval system for art history and archaeology
SemView [Wang et al., 2003]	semantic retrievel, distributed search in a set of databases
CAIRO [Geisler et al., 2001]	parallel programming and cluster architectures for image retrieval
CIRES [Iqbal and Aggarwal, 2002a]	specialies structure feature for retrieving manmade objects, grouping the data set by multi-class classification

Table 2.2: A selection of image retrieval systems.

a tree-structured SOM (TR-SOM) is used to get an acceptable calculation speed as well as a gradual search (beginning on the top level the search can be improved by diving into deeper SOM-layers). Comparable with the WEBSOM tool PicSOM supports a target search as well as the exploration of a collection or browsing through it.

On the image set a number of MPEG-7 features [Manjunath et al., 2000], [Koskela et al., 2001b] is calculated and stored in a file system to get the corresponding feature vectors of each individual picture. Based on each feature space TR-SOMs are trained individually for each feature space. This is done a priori, since the training of SOMs needs some time and cannot be done online. Furthermore an image segmentation is implemented [Sjoberg et al., 2003] to enhance the system.

A search session can be started in two ways: A query by pictorial example or a browsing through the data set based on topological neighbourhoods are possible. Adapted from a variable number of query images a set of probably similar images is presented. For this the regions of SOMs where the query images are located are marked and pictures of these regions are presented in the next step. To enhance the results the user can rank the presented images as relevant. All presented and not rated pictures are treated as non-

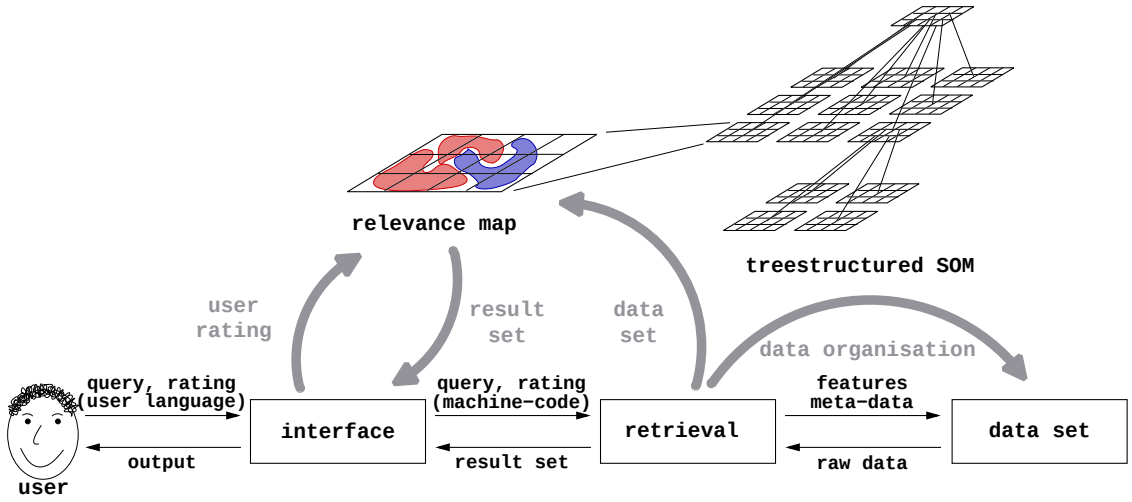


Figure 2.3: The CBIR-system PicSOM: The main attributes of this framework are the tree-structured SOM to organise the stored images and the relevance maps on the several SOMs.

relevant. Based on these labelling the associated map regions are marked as relevant. A low-pass filter on the map extends the points of relevant objects to regions containing probably relevant objects [Koskela et al., 2002].

Therewith a local relevance feedback is performed since just the pictures in the neighbourhood of already presented pictures are rated. *White spots* may remain on the maps. At the same time the filter mask acts as a window function and searching in the respective map regions becomes more detailed in subsequent steps. This local relevance feedback does not influence the images but only the relevance labelling of the maps. Thus the nonlinearity of image similarities is respected. In order to assist browsing, map regions with relevant labelled pictures are coloured based on user ratings (see figure 2.3).

To preserve the experiences of past search sessions a longterm learning is implemented [Koskela and Laaksonen, 2003]. For this purpose the set of relevant labelled pictures according to each query is stored. On these sets latent semantic indexing (LSI) is performed as an inter-query learning step to get a user-interaction feature [Koskela, 2003].

The prevailing quality of PicSOM is the unranked result set based on a relevance value calculated in the feature spaces. Most of the other CBIR-systems calculate a result list based on a distance value related to any kind of example. Based on the hierarchical relevance labelling approach PicSOM can detect pictures with quite different visual attributes. For example photos as well as sketches can be retrieved in the same session.

The evaluation of the PicSOM system is quite good. The individual features are evaluated [Koskela et al., 2001a] as well as a comparison with other systems [Rummukainen et al., 2003]. Recently the system has been evaluated in the TRECVID competitions [Koskela et al., 2005].

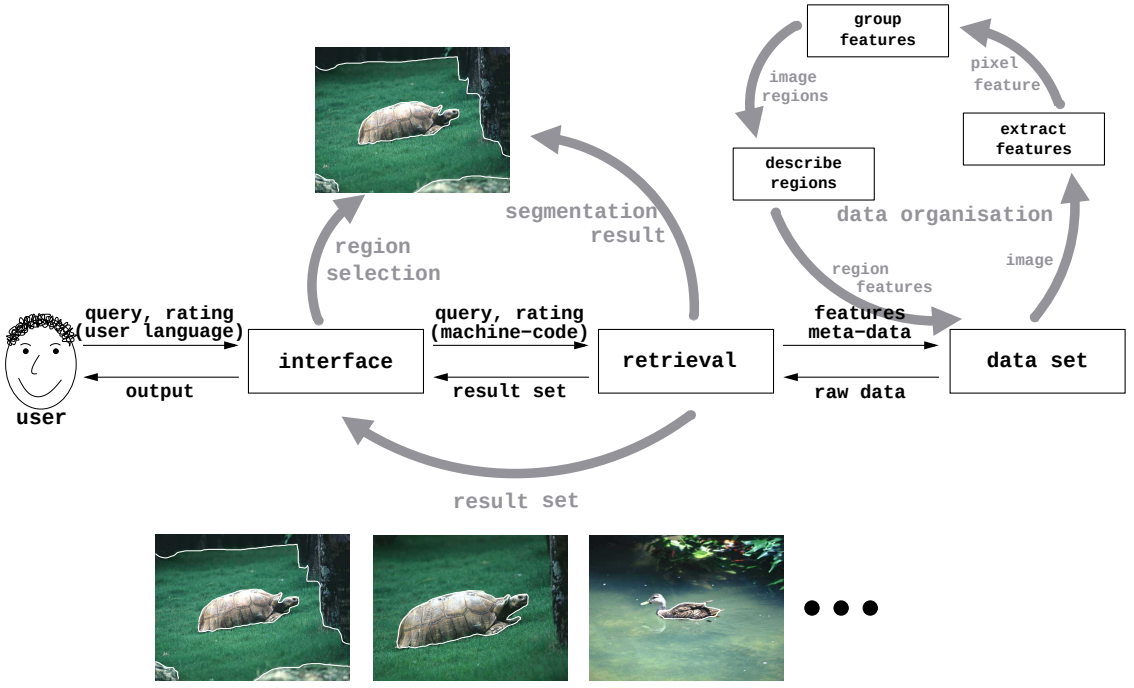


Figure 2.4: The CBIR-framework blobworld: The focus of the system is the image segmentation. Thus image retrieval is based on local attributes and an object retrieval is approximated.

2.2.2 blobworld

Searching for images showing specific objects is an often required retrieval task. To support this an image segmentation is necessary to cut out the objects. *Blobworld* [Carson, 2004] [Carson et al., 2002] is an image retrieval framework particularly addressed to this challenge. Although object retrieval is not the main task, the assumption that each image is a combination of different meaningful regions resembles this.

The image segmentation is performed in a preprocessing step. Based on texture, colour and position the pixels of each image are grouped to clusters. To represent the texture contrast, polarity and anisotropy in the neighbourhood of each pixels are computed. The colour is described by the values in the L*a*b-space[2]. These features are combined to one vector and added by the (x, y)-coordinates of the pixels.

Based on this an Expectation Maximisation approach is used to estimate the parameters of a Mixture of Gaussians model. Then the pixels are grouped to connected clusters. The common texture and colour attributes are stored in the data collection to represent the different regions [Carson et al., 2002].

The segmentation results are presented in the interface. Thus the user can select the image region that satisfies his intention. The retrieval is implemented as a similarity search based on the local features. In the result list the images are presented whereas the most similar region is highlighted. Therewith the user can understand why the images are retrieved.

[2]L is the lightness, a is the *redness/greeness* and b is the *yellowness/blueness*.

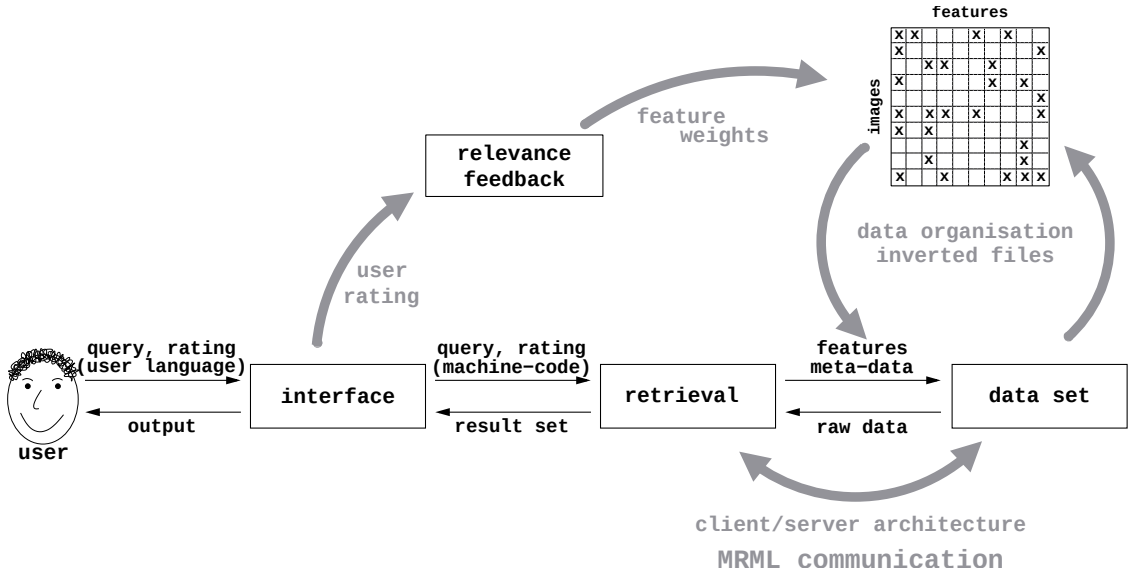


Figure 2.5: The CBIR-system GIFT/Viper: A client/server architecture is proposed to be most flexible for image retrieval systems. Motivated by text retrieval research *inverted files* are used to perform the retrieval.

The image segmentations are evaluated visually. Therefore developers as well as the user can inspect the segmentation results. Indeed this qualitative evaluation is quite incomplete. More detailed experiments to analyse the user satisfaction are not documented.

To show suitability of a segmentation for image retrieval tasks the performance is compared with retrieval results based on global image features. Therefore precision-recall-diagrams are presented.

2.2.3 GIFT/Viper

Based on the image and multimedia retrieval research of the University of Geneva the CBIR-system *Viper* [Müller, 2002] [Squire et al., 1999] is published as GIFT (Gnu Image Finding Tool) in the GNU Project [GIFT]. While developing Viper common approaches of text retrieval are applied to images. Furthermore a client/server architecture is proposed as suitable for image retrieval. To establish this the communication protocol MRML (Multimedia Retrieval Markup Language) [MRML] is developed. Therewith the evaluation of image retrieval systems should be forced to be comparable.

Based on text retrieval approaches *inverted files* have been used to perform the retrieval [Müller et al., 1999]. Therefore the existence or absence of numerous features is detected for each image. Textual features are used in the same way as visual features. Thus each image has $O(10^3)$ features. Images offering the same features as the query are retrieved as relevant.

A relevance feedback approach is implemented to enhance the retrieval results [Müller et al., 2000a]. Basically the set of relevant labelled images is enlarged and therewith the selection and weighting of suitable image features. Therefore the frequencies of the individual features in the relevant labelled image set is measured. The assumption is that features frequent in one image (category) are suitable to detect this. On the other hand

features which are frequent in the entire data set are less suitable to distinguish between different images (categories).

To get as much rating as possible and perform the relevance feedback a demo version of Viper has been presented in the Internet. The user interaction has been stored in log files and used for relevance feedback.

The GIFT version offers the same features. The user specific image collection is indexed in a preprocessing step and the inverted files are computed. Then the user can search in his image collection. The client/server architecture and the XML-based communication protocol MRML offers the possibility to enhance the system with further modules.

Recognising the importance of image retrieval evaluations the developers tried to establish the communication protocol used in Viper as an image retrieval evaluation standard [Müller et al., 2001b] [Müller et al., 2001a]. Therewith the *benchathlon* [benchathlon] has been initiated. This should offer comparing evaluation strategies to rate CBIR-systems in relation to other approaches. Unfortunately this has not been accepted and the benchathlon dropped off (see section 6.4.3).

2.2.4 INDI

The CBIR-system INDI (**I**ntelligent **N**avigation in **D**igital **I**mage Databases) has been developed within the LOKI[3]-project [Kämpfe et al., 2002]. The main intentions were to create a framework for developing, analysing and testing CBIR relevant approaches, mainly adaptable approaches for human-computer-interaction in an image retrieval situation.

Without a capable retrieval unit the development of a user-friendly and multi-modal interface is improper. Consequently the retrieval unit offers a lot of modifications to analyse. The performed search task is a similarity search suitable for a target search as well as a category search. Although the system can handle arbitrary pictures, most of the analyses are based on the *artexplosion*-photo collection (see section 3.2.1).

Since user interaction is the main focus of the LOKI-project, the user interface consists of different modules. In especially it performs a multi-modal communication via touchscreen or gesture recognition as well as speech input [Bauckhage et al., 2003], [Käster et al., 2003] (see figure 2.6).

The main input data is independent from the input path. An initial query image has to be determined, either by choosing one arbitrary picture of a random subset of the database or by presenting a new picture. Since the system performs a relevance feedback, in further retrieval steps the pictures have to be rated by the user. Five rate levels (very good, good, medium, bad, very bad) are determined and can be chosen in each communication situation. Furthermore the user has to give the system calls like *search* and *new search*.

A list of similar images to a query is presented as the result. Some intra-system parameters can be displayed for evaluation or analysing purposes.

As pictorial entities entire images can be used as well as any kind of subimages resulting from an arbitrary segmentation step. The initial version cuts off subimages by a rough grid while later on a segmentation algorithm based on salient points has been implemented. For each pictorial entity a set of N_f low-level feature vectors is computed a priori and stored in the database.

[3]Lernen zur Organisation komplexer Systeme in der Informationsverarbeitung

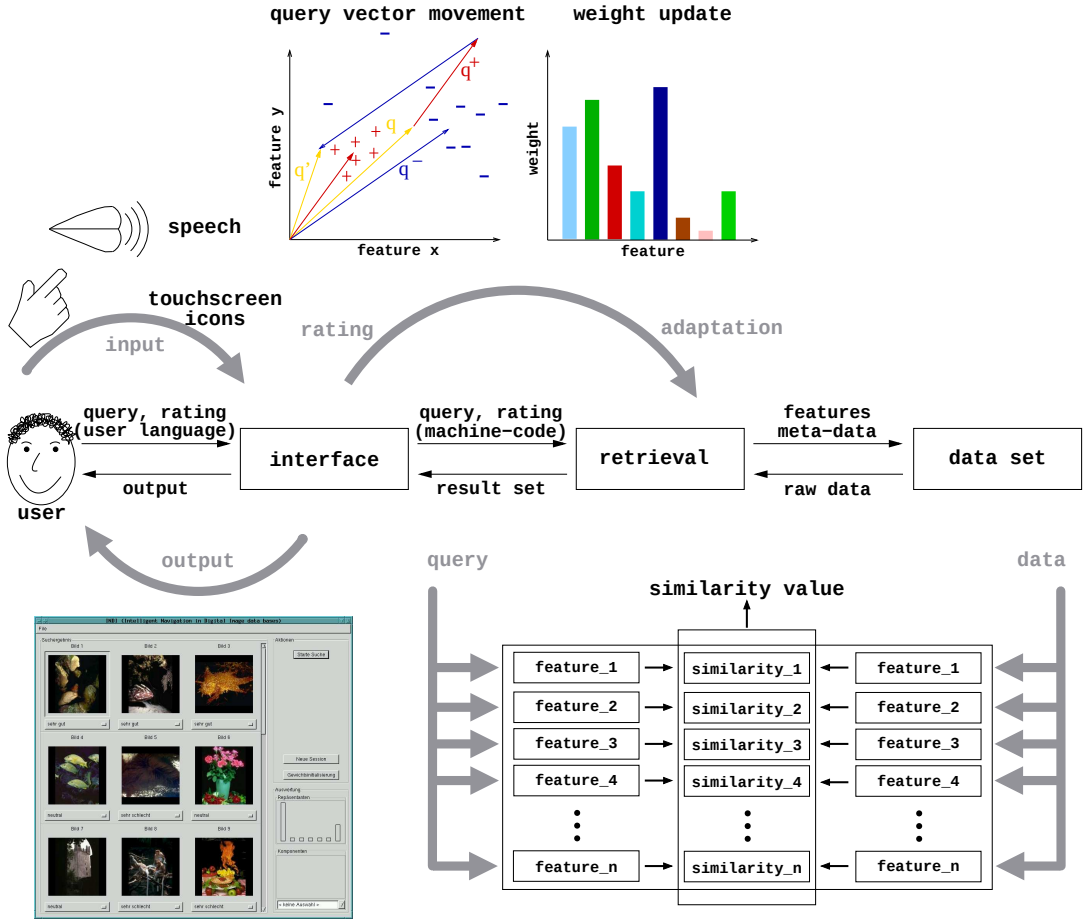


Figure 2.6: The CBIR-system INDI consists of different modules: The data input can be carried out by speech or using icons on a touchscreen. Based on a user rating query vector movement and weight update are performed to adapt the system to the user's need. A similarity search in different feature spaces determines the result image list, which is presented on the screen.

In INDI a similarity search is performed and the pictorial entities $\mathbf{x}$ are arranged according to the similarity to a given query $\mathbf{q}$. This similarity value $s(\mathbf{q}, \mathbf{x})$ is computed as a linear combination of a number of distance values:

$$s(\mathbf{q}, \mathbf{x}) = 1 - d(\mathbf{q}, \mathbf{x}) = 1 - \varepsilon \sum_{i=0}^{N_f} w_i \, d_i(f_i(\mathbf{q}), f_i(\mathbf{x})) \tag{2.1}$$

where each distance function d_i is determined by the correspondent feature f_i. Each distance value represents the distance between two pictorial entities, usually a query pictorial entity $\mathbf{q}$ and another pictorial entity $\mathbf{x}$ in one specific feature space i. The weights w_i are parameters to weight each feature space according its relevance in a specific search task. N_f is the number of used features. ε is a normalisation coefficient to scale the distances to [0...1].

The first N_r entries of the list

$$\mathbf{r}(\mathbf{q}) = [\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{N_r}] \text{ with } s(\mathbf{q}, \mathbf{x}_k) > s(\mathbf{q}, \mathbf{x}_l) \, , \, \forall \, k < l \tag{2.2}$$

build the retrieval result and are presented to the user.

A relevance feedback is performed in two ways: (1) The weights $\mathbf{w}$ in the linear combination $s(\mathbf{q}, \mathbf{x})$ and (2) the query vector $\mathbf{q}$ can be changed based on user ratings of the previous steps. In the fundamental version the implemented weight updates are motivated by [Rui et al., 1997a] and [Rui et al., 1997b]. Feature spaces wherein spatial arrangement of the pictorial entities resembles the similarity rating of the user gets higher weights than feature spaces wherein they differ.

$$\Delta\mathbf{w} \;=\; \varepsilon \sum_{j=1}^{N} \Gamma(\mathbf{x}_j) F(\rho(\mathbf{x}_j, \mathbf{r})) \tag{2.3}$$

Where $F(\rho(\mathbf{x}_j, \mathbf{r}))$ is a continuous decreasing function to filter the top of the list $\mathbf{r}$. $\Gamma(\mathbf{x}_j)$ is the user rating of image $\mathbf{x}_j$ and gets the values $\{-3, -1, 0, 1, 3\}$. $\rho(\mathbf{x}_j, \mathbf{r})$ is the position of the image $\mathbf{x}_j$ in the result list $\mathbf{r}$. ε scales $\mathbf{w}$ to the interval [0..1].

Furthermore the query $\mathbf{q}$ is adapted to the pictorial entities labelled as relevant in the current search task [J.J. Rocchio, 1971]:

$$\mathbf{q}' \;=\; \eta\,\mathbf{q} + \;\gamma \sum_{i=1}^{N^+} \mathbf{x}_i^+ - \;\beta \sum_{j=1}^{N^-} \mathbf{x}_j^- \tag{2.4}$$

Where $\mathbf{q}'$ is the query vector in the next search step, $\mathbf{x}_i^+,\ i = 1, ..., N^+$ are relevant labelled and $\mathbf{x}_j^-, j = 1, ..., N^-$ nonrelevant labelled images. η, γ and β rate the influence of the different images sets for the next query vector. For a sketch see figure 2.6 (the weights η, γ and β are not included).

In general the INDI system offers different input devices and is able to adapt to the user's need. The user interaction is the great benefit of the system. Intuitive and user friendly input modalities satisfies the user. Compared to the above presented systems this is an outstanding feature of INDI. A more flexible result set as given in PicSOM may substitute the result list and enhance the retrieval performance. Indeed blobworld and most of the other systems are also based on such result lists. The client/server architecture proposed in Viper to support comparable evaluations is transferable, whereas the retrieval approach is completely different.

Summarised INDI shares the similarity approach based on a set of image features and the ordered result list with most of the other CBIR-systems. An segmentation step is performed in few systems and a multimodal input device is very unusual.

2.2.5 AQUISAR

Since in the Trojan Room faculty room at Oxford University the first webcam had been installed and the whole world could observe their coffee maker, the number of webcams has increased enormously [EarthCam]. Researchers as well as business or private people arrange digital cameras faced to their places of interest and present the current images in the internet. An exciting application is the installation of a webcam for observing natural scenes like animals.

Usually these cameras run round the clock or at least during daytime, whereas the interesting objects act just for a short period. Therefore, most of the images taken by the webcam present just the non-interesting environment, for example an empty aquarium or an abandoned pool in the wilderness. To handle the enormous bulk of pictures resulting

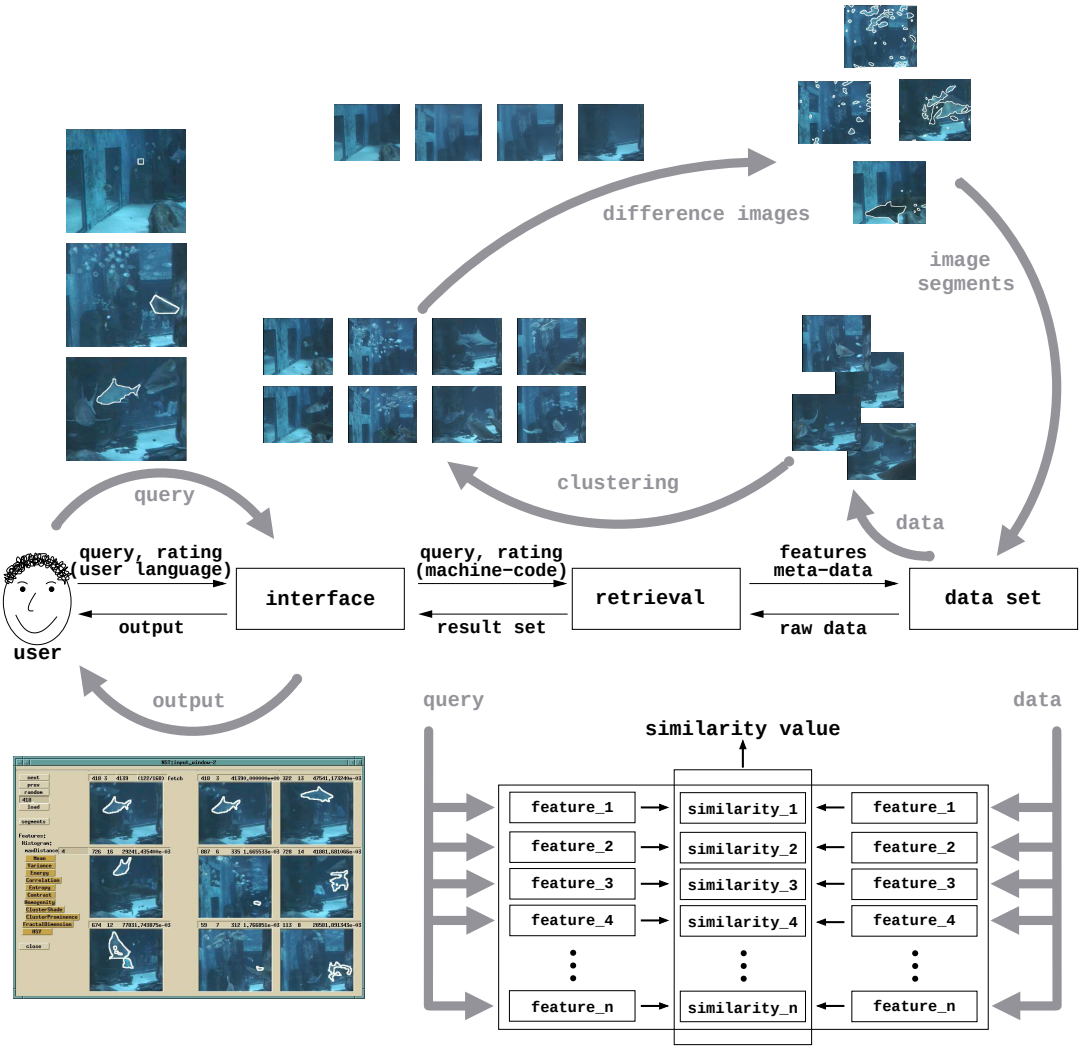


Figure 2.7: The framework AQUISAR consists of a number of modules to retrieve images in a set of underwater webcam images:
In a preprocessing step a stored set of images is grouped into four clusters with equal background. For each cluster the mean image provides a prototypical view of an *empty aquarium*. Subsequently, an image segmentation is performed based on difference images. The query images can be built from a presegmented image region, a user-defined segment enclosed by a polygon or a 15×15 square region around a selected pixel.
A similarity search in different feature spaces determines the result image list, which is presented on the screen.

from such an experiment setting, an automatic assistance to store just the relevant images is desirable.

The system AQUISAR (**Aqu**arium **I**mage **S**egmentation **a**nd **R**etrieval) [Kämpfe et al., 2004] performs the main steps necessary for retrieving interesting images in a set of images shot by the London Aquarium Webcam [London Aquarium]. Three tasks are combined in this framework: Webcam image handling, content based image retrieval and underwater computer vision.

To investigate the different approaches implemented in AQUISAR a set of images taken by the London Aquarium Webcam is stored. The set of pictorial entities encloses the orginal images shot by the webcam as well as the image regions distinguished from background covered regions. Within this set the user can look for interesting images.

To initiate the query the user can select an image as well as an image region as interesting. Furthermore he can restrain the subset of the features he considers suitable for his search.

To perform the retrieval of particular images a sequence of preprocessing steps (see figure 2.7) is implemented to calculate suitable image features:

(1) A fixed webcam takes pictures of a single scene with an unchanged background. In a set of images with the same background the image regions covered by changing entities can easily be detected via calculating difference images. For preserving the advantages of invariable backgrounds a *k-means*-cluster-algorithm group the N stored images $\mathbf{x}_i, i = 1, \ldots, N$ into clusters $C_j, j = 1, \ldots, N_{\text{pos}}$ based on the $N_{\text{pos}} = 4$ positions of the London Aquarium. The clustering is performed on the principal components belonging to the 200 greatest eigenvalues of the image autocorrelation matrix.

(2) In the next step a *region*-image $\mathbf{b}_i$ is computed, which assigns each pixel $\mathbf{x}_i^{xy}$ to a region $\mathbf{s}_{ki}$. To this end, a difference image $\tilde{\mathbf{x}}_i$ is computed first:

$$\tilde{\mathbf{x}}_i = |\mathbf{x}_i - \overline{\mathbf{x}}_j| \tag{2.5}$$

with $\mathbf{x}_i \in C_j$ and $\overline{\mathbf{x}}_j = \frac{1}{N_j} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$ is the average image of camera position j and N_j is the number of images taken from setting j. Note that each average image shows an *empty* aquarium, as can be seen in figure 2.7. From these difference images $\tilde{\mathbf{x}}_i$, label images $\mathbf{b}_i$ are computed which distinguish the background from possibly interesting coherent objects (i.e. fishes):

$$\mathbf{b}_i^{pq} = \begin{cases} k & , \text{ if } \tilde{\mathbf{x}}_i^{pq} \geq t \\ 0 & , \text{ otherwise} \end{cases} \tag{2.6}$$

where $\tilde{\mathbf{x}}_i^{pq}$ denotes the pixel value with the coordinates p, q of the difference image $\tilde{\mathbf{x}}_i$ and t is a threshold calculated iteratively on the global grey-value histogram [Ridler and Calvard, 1978]. The identifier k with $k \in [1, \ldots, K_i]$ is calculated in a preceeding step on the coherent binary objects that result from $\tilde{\mathbf{x}}_i^{pq} \geq t$ and is used to identify the various image regions $\mathbf{s}_{ki}$:

$$\mathbf{s}_{ki} = \{\mathbf{x}_i^{pq} \mid \mathbf{b}_i^{pq} = k\} \tag{2.7}$$

K_i is the number of separate regions within image $\mathbf{x}_i$ and background pixel are labelled by $k = 0$.

(3) For lack of specified features for underwater images a set of low-level features is calculated for each region. According to the physical conditions in underwater environments, texture features may be more suitable than colour. Therefore, two texture features (based on the fractal dimension and the co-occurrence matrix [Unser, 1986] respectively) and just one colour feature (empirical mean and variance of HSV[4] histograms) are implemented.

[4]Hue, Saturation, Value

The most intuitive and simple query to a webcam retrieval system is: *Show me interesting images!* This task bears two questions: *What is the meaning of interesting?* and *Which images achieve these specifications?*

For a detailed discussion regarding the term *interesting* see section 5.1. In the AQUISAR-system the presentation of an example image with a content the user considers absorbing specifies interesting images. Based on this idea, a *query by example*-framework is used. This framework is suitable to detect images in a subject observation task, where an observer wants to know when a certain animal appears. With an example image containing the requested animal he can easily search for appropriate images.

Depending on the quality of the segmentation result, the user may choose between various techniques to extract the query example $\mathbf{q}$: Choose an image region with a mouse click (clipping a small rectangle if no region met) or pick up an explicit image region by enclosing the interesting image region by a sequence of mouse clicks (see figure 2.7 top left).

To get the appropriate images the retrieval is performed as a similarity search. Therefore, the result is an ordered list $\mathbf{r}$ of the images or image regions:

$$\mathbf{r} = [\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \dots] \tag{2.8}$$

with decreasing similarity values

$$s(\mathbf{q}, \mathbf{s}_u) \geq s(\mathbf{q}, \mathbf{s}_v) \quad \forall\, u, v \text{ with } u < v$$

$\dot{s(;)}$ measures the similarity between two images. Since the used features are very general, the Euclidean distance on these features is calculated to specify the similarity. The first eight images of this list are presented in a graphical user interface.

The preprocessing steps are quite successful. In spite of similar image features the clustering in this application was able to perfectly distinguish between the four views of the London Aquarium webcam. Furthermore the mean of the images taken from the same camera setting renders a prototypical view of an empty aquarium. This can be regarded as the reference background to calculate difference images. And the unsupervised image segmentation results in segments suitable to calculate image features, although not every fish is cut out perfectly (especially very close and therefore big sharks are often detected just partially). An approximate border of an object is sufficient, since just colour and texture features are used.

The retrieval is evaluated by a precision rating. This is reasonable appropriate since a recall calculation on the desired unlimited set of webcam images is not possible. Compared to another more general CBIR-system this shows that taking the multi-angle nature of this image domain into account leads to a significantly improved retrieval accuracy. (For more details of a comparable evaluation see chapter 6).

Based on the image segmentation step the background has littleinfluence in the retrieval step. Thus AQUISAR can retrieve images with similar entities from different webcam settings, i.e. different angles of view. This striking advantage of AQUISAR motivates segmentation steps for other image retrieval systems.

2.3 Summary of Image Retrieval

Information retrieval in general is a broad field of research. Although it is a branch of information retrieval, content-based image retrieval incorporates further research tasks. While every aspect has its own attraction and would be interesting to investigate there are some striking points regarding CBIR:

A lot of retrieval approaches exist for different kinds of data. Namely text data or fixed meta data are quite suitable to retrieve based on common indexing approaches. But to apply these retrieval approaches on image data, suitable image presentations are necessary. Thus computer vision research has to be involved. This research community provides a lot of image features, low-level ones like colour and texture as well as more advanced ones which often depend on the image domain. Therefore, an analysis of possible image features is important as well as the analysis of the used image set.

Users have a lot of different intentions when retrieving images. This means that different search tasks should be supported. Furthermore users rely on their semantic interpretation of the image content. This causes the so called semantic gap since computers depend on the formal description of the images. CBIR-systems which should be accepted by users must be flexible. They should adapt to the user's need as well as to different search tasks and different image sets. Here relevance feedback is a wide spread approach.
Adaptable systems are often based on machine learning and neural network approaches. So these fields of research may be interesting for designing image retrieval systems.

A number of different image retrieval systems exist today or are under intense investigation. But which is the most suitable system? Or more detailed, which components of single systems are worth to enhance and use in future systems? After all, which systems will survive and establish the future state of the art? To answer all these important questions, the current systems and investigations have to be compared and evaluated.

Recapitulating the following aspects may be worth to investigate in more detail, whereas the user should be kept in mind.

- Which computer vision approaches can be used within CBIR frameworks?
  Which attributes of the used image sets are important to choose the right ones?
- How can the semantic gap be bridged?
  How can the user intentions regarding the single system components be modulated?
  Which approaches are suitable to adapt a system to the user's need, the search task and/or to the image domains?
- How to evaluate image retrieval systems?

In the following chapters these challenges are researched.

Chapter 3

Images and Features: Data Sets for CBIR

The image sets under consideration in CBIR-research offer miscellaneous qualities. Obviously colour photographs differ from pencil drawings with regard to image complexity. Furthermore, usually image retrieval approaches are based on a set of image features. Both sets – the given pictures as well as the used feature algorithms – are presented in this chapter.

3.1 Image Data

3.1.1 Domains

Large picture collections motivate the automation of the image retrieval processes. The feature extraction obviously depends on the image database at hand. This motivates a deeper look on the set of images under consideration, called *image domain*. Different aspects have to be kept in mind for the design of a CBIR-system [Smeulders et al., 2000]:

- Top level considerations concern the **system design** and are strongly dependent on the used data set to determine reasonable search tasks and suitable implementations of the different system modules (see section 2.2).
- The **semantic gap** influences every image retrieval approach but some image domains are affected harder than others. For example the brodatz-texture collection [Brodatz, 1966] can be suitably described by low-level (texture) features, whereas a description of a holiday photo collection depends strongly on personal memories and feelings, which cannot be expressed with simple features.
- A number of different and specialised **image features** have been developed. Most of them show good performances on particular image domains, but lack performance when applied to other domains. One example is the structure feature for detecting images of manmade objects [Iqbal and Aggarwal, 2002b]. The typical strong boundaries of manmade objects are computed based on perceptual grouping. Naturally such a feature is not suitable to describe images of completely different content. Hence the underlying image set should be kept in mind during the selection of suitable feature algorithms. To this day there is no general-purpose CBIR-system which can be applied successfully to diverse image domains.
- Furthermore, different users have different knowledge, intention and background of a particular image domain. On the one hand human experts may be involved in the

image recording process and usually they know much about the collected data. This becomes very obvious in specific image sets like medical images or deep-sea photos. On the other hand users may have different cultural backgrounds and therefore regard the same set of images differently.
In general **domain knowledge** of experts should be respected as well as literal, perceptual, physical, categorical and cultural aspects while describing an image set.
– The expected **retrieval performance** strongly depends on the image domain. Some combinations of search tasks within image sets are harder to solve than others.

In summary numerous aspects in designing a CBIR-system strongly depend on the underlying image set. A deeper analysis of the image domain as well as the a priori knowledge of common attributes of the images is helpful implementing image retrieval approaches.

How are image domains usually analysed and described? Contrary to the influence of the image set, in most documentations the underlying image set is just described by some general terms, e.g. in [Armitage and Enser, 1997]:

> "... supports a wide and general user based interested in the world of film and television, while the latter serves a much narrower range of 'expert' users interested in the specific subject domain of natural history."

In this description two well established adjectives occure: *wide* (synonymous *broad*) and *narrow*. Indeed, such an assignment is still rather intuitive, although a number of criteria and examples to rate an image domain as narrow or broad are available (see table 3.1).

Based on these attributes an explicit rating of image sets with regard to increasing broadness is not possible. Nevertheless for evaluation tasks an overall objective measure to describe image sets would be desirable, so that observations can be compared and analysed. In this context the *complexity of image databases* is proposed as a measure [Rao et al., 2002]. Initially the images are divided into sub-blocks. Then the correlation and the cross-entropy of these sub-blocks are computed over the image set. This results in a query independent rate to describe the degree of retrieval difficulty.

Developing this measure the aspects *homogeneity* and *heterogeneity* as well as the *content variety* and the *cardinality* of the data set have been taken into account. While the cardinality particularly is covered by a priori probabilities of targets the other three aspects are interesting in the context of domain properties. Homogeneity and heterogeneity can be directly related to the assignment as a narrow and a broad domain respectively. A homogeneous set resembles a narrow domain and a heterogeneous set a broad domain. In contrast to this contradictory aspects the content variety is as gradual as the broad–narrow domain classification and covers content attributes as well as semantic observations. Nevertheless, this measure is suitable to objectify the discussion of easy or difficult image sets.

Analysing example image sets using this complexity measure, Rao et al. [2002] have observed that homogeneous data sets are complex and difficult to browse, whereas heterogeneous sets are less complex and therefore easier to search. This is a contradiction to the common assumption that narrow domains (homogeneous data sets) are easier to handle than broad image domains. For example Koskela and Laaksonen [2003] state that restricted domains like trademark images are quite easy to browse. On the other hand large databases of miscellaneous images are mentioned as difficult settings. How can this

	broad domain	**narrow domain**
content variety	high	low
source of knowledge	abstract	expert knowledge
semantics	object level	detailed
ground truth	usually not given	plausible, labelled by experts
content description	subjective, superficial	objective, task dependent
scene and sensor	unknown	possible controled
application	public photo collections, news agencies	specific research databases, catalogues
tools	similarity search	classification, object detection
interactivity	high	low
evaluation	user satisfaction, qualitative	retrieval reliability, quantitative
system architecture	flexible, modular	tuned to application
cardinality	very large	medium
source of inspiration	information retrieval	object detection
homogeneity	low	high
heterogeneity	high	low

Table 3.1: A survey of broad and narrow domain attributes. See [Smeulders et al., 2000] and [Rao et al., 2002]. Unfortunately most of these attributes are quite subjective and hard to measure.

inconsistence be explained?

It should be noticed that in narrow domains the feature selection and design is detailed and adjusted to the image data whereas low level features are used to describe the pictures of broad domains. Here the success for different types of search tasks depend on the domain type.

Let's start with a look at target searches. The images of a narrow domain build a homogeneous and compact cluster somewhere in the image space (see figure 3.1 left). Carefully selected features are used to characterise the differences between two images. Consequently images are quite easy to distinguish and desired images can be found.

In contrast the pictures of a broad image domain are spread through the image space where a number of groupings can be recognised (see figure 3.1 right). Since low level features are hardly capable of distinguishing between similar images within such groups the retrieval algorithm has to be tuned well to the specific task. Hence the retrieval process takes longer and the task is rated as difficult.

In category searches, narrow image domains usually have no well distinguishable subsets or images of different categories are mixed up. Obviously it is difficult to detect such categories automatically. Actually in a broad image domain a similarity search based on an example image is simple, if the relevant image objects are grouped together.

In order to analyse the relation of retrieval complexity and data distribution the variances of the image collections are a good measure. The values of some specific data sets are listed in table 3.2 based on the principal components of a colour and a texture feature.

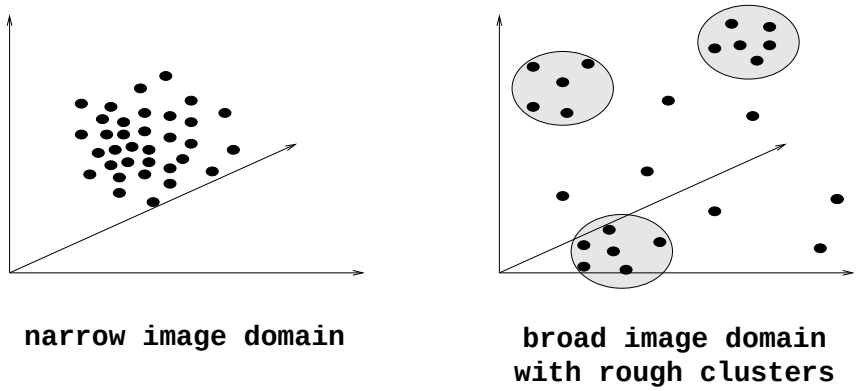


Figure 3.1: Illustration of wide and narrow image domains based on the distribution in the image space.

domain	**intuitive**	σ**(colour)**	σ**(texture)**	**sem/dist cluster**
shark cam (sec 3.2.3)	narrow	0.002 0.001	0.014 0.003	yes / yes
myMondrian (sec 3.2.2)	rather broad	0.036 0.008	0.284 0.097	yes / yes
artexplosion (sec 3.2.1)	broad	0.036 0.002	0.242 0.087	yes / no
coil [Nene et al., 1996]	rather broad	0.048 0.017	0.147 0.022	yes / yes
VisTex [VisTex, 1995]	rather narrow	0.002 0.002	0.417 0.127	yes / no
deepsea [Jaeckisch, 2004]	narrow	0.003 0.000	0.287 0.020	no / no

Table 3.2: Quantitative analysis of some image domains based on the variances (σ) in some feature spaces. The variances along the first and second principal component of a colour and a texture feature are presented. The clustering is distinguished between semantic, user recognised (sem) and feature distribution (dist) based groupings.

Thus discrepancies become recognisable based on the distribution measures. The statement that most of the narrow image domains do not offer obvious clusters is disproved by the *shark webcam*-set. Here the clustering into four subsets depending on the background is obviously a semantical grouping. At the same time this set proves the conjecture that narrow image domains have small variances in the data spaces.

Taking the variances in the different feature spaces as a hint to rate a set as narrow or broad will result in different gradings depending on the used feature. While in the colour space the rating will conincide with the intuitive description the texture would motivate a completely different labelling. The colour may be important to represent these three broad image domains. Indeed, the narrow domains show content independent from colour (textures) or at least hard to describe by colour (underwater images – see section 3.2.3).

In the texture space the *VisTex* stands out by a large variance. This set is intuitively rated as a narrow domain, which usually are assumed as compact clusters in the data space. However, it has the largest variance in the texture space. Obviously this is caused by the content of the image set, which are texture images. Texture is a special feature for this data domain.

Summarised narrow image domains do not automatically mean an easy or difficult retrieval task.

3.1.2 Categories

Category searches are desired in a lot of different situations. Consequently the objects of such a search task – the categories – are worth analysing.

Rosch et al. [1976] note that human users perceive pictures with objects on a quite rough level of abstraction. On this basic level objects of the same type have similar features and shapes. Mean images or prototypes may be suitable to describe a set of pictures showing the same object type. Similar observations are true for natural images.

In order to take advantage of these observations Torralba and Oliva [2003] analyse the statistics of such perceptual groups. They document that the visual categorisation of image sets based on second order statistics may improve computer vision tasks.

Consequently the detection of subsets within a large image collection may improve image retrieval tasks. Such groups can be defined in the following way:

> Let the given data $\mathbf{x} \in \mathbf{X}$ be grouped by an arbitrary grouping function $\Phi(\mathbf{x}) = \{\psi_1, ..., \psi_{N_x}\}$, where $\psi \in \{1, ..., N_s\}$ is a group label, N_s the number of groups and N_x the number of groups containing the document $\mathbf{x}$. Depending on the used grouping approach N_x may be limited to 1. Examples for $\Phi(\mathbf{x})$ are cluster algorithms, automatic classification approaches or semantic mappings. The result will be a number of subsets:
>
> $$\mathbf{X}_\psi = \{ \mathbf{x} \mid \mathbf{x} \in \mathbf{X} \text{ and } \psi \in \Phi(\mathbf{x}) \} \Leftrightarrow \mathbf{X}_\psi \subseteq \mathbf{X} \tag{3.1}$$
>
> Generally the whole image set $\mathbf{X}$ may be divided into a number of disjoint or not disjoint subsets $\mathbf{X}_\psi, \psi = 1, ..., N_s$.

Usually humans categorise a set of images according to different attributes or by different situations, e.g.:

- various instances of a specific object or one individual object in different orientations (e.g. the coil collection [Nene et al., 1996])
- the kind of objects, e.g. animals
- the same location or time period
- a certain event or kind of event, e.g. a birthday party
- the type, e.g. paintings, cartoons, photos, sketches
- the artist, e.g. paintings of Rembrandt
- compatible to a specific situation, like an important publication, an upcoming event or the current emotions of the user.

The grouping should ideally be invariant against cultural, sociological and other human-related influence factors [Eidenberger, 2004]. In practice an optimal grouping cannot be reached, since at least semantic categories strongly depend on user intentions and experiences. Furthermore all levels of categories may be influenced by domain knowledge and the reliability of the labelling experts.

Conceptually there are three kinds of groupings on different levels: meta-data based, groupings according to the contained objects and semantic categories. Humans use all of these levels and sometimes switch between them when grouping a set of images. The meta-data can be used for automatic grouping, whereas the grouping based on contained objects or semantics depends on feature detection algorithms. Therefore it is difficult to automate the process for different image domains.

General approaches to find categories automatically are desired. One method is discussed in section 2.1.2. Two other well known techniques to group data sets are *clustering* and *classification*, dividing the set of images in disjoint groups, called *cluster* or *classes*. But most of the common implementations can be tuned to get multiple assertions as usual in indexing. In general overlapping groups are more intuitive and realistic but automatic approaches force unique categories. In the context of CBIR-systems a clustering or classification implementation can be treated like a category search task.

Such image categories can be used to perform image retrieval tasks. A very welcome implementation is the detection of interesting pictures out of a large set. Technically, this resembles a classification

$$\Phi(\mathbf{x}) = \begin{cases} \mathbf{x} \text{ is interesting} \\ \mathbf{x} \text{ is not interesting} \end{cases} \tag{3.2}$$

where $\mathbf{x}$ is one image of the image set $\mathbf{X}$. Unfortunately the term *interesting* is not well-defined and trails a lot of research according to the semantic level information retrieval [Santini and Jain, 1996] [Hare et al., 2006]. In order to fix this challenge, disjoint subsets $\mathbf{X}_\psi \subseteq \mathbf{X}, \psi = 1, ..., N_s$ are built. The query image $\mathbf{q} \in \mathbf{X}$ determines the set of relevant images $\mathbf{X_q}$ with $\mathbf{q} \in \mathbf{X_q}$. Thus all images of the subset $\mathbf{X_q}$ are interesting with respect to the the query $\mathbf{q}$:

$$\Phi(\mathbf{x}, \mathbf{q}) = \begin{cases} \mathbf{x} \text{ is interesting regarding } \mathbf{q} & \text{, if } \mathbf{x} \in \mathbf{X_q} \\ \mathbf{x} \text{ is not interesting regarding } \mathbf{q} & \text{, otherwise} \end{cases} \tag{3.3}$$

This approach can be generalised to overlapping subsets. Each subset $\mathbf{X}_i$ may resemble a category.

3.1.3 Sequences

Data sets may have an inherent one-dimensional structure. For pictorial data those are called *image sequence*. Such a set may appear in a variety of situations, usually caused by the time span between two shots. The most obvious occurrences of image sequences are films, where the single shots can build an image set. Different picture sets of photo sessions or observation situations will be given by stretching the time spread between two shots. In recent years, this has forced a specified research field: Video retrieval [Petkovic and Jonker, 2003].

Image sequences can overbear the disadvantages of the two dimensional structure of common pictures. With a sequence of two dimensional pictures the inherent three dimensional structure of an object or a scene can be shown. For example Takaya and Choi [2001] use two dimensional TV-newscaster films to calculate three dimensional models of faces.

Usually the arrangement within a sequence is specified by a time stamp. Image sequences may be defined by:

> A data or *image sequence* $\mathbf{S}$ is a set of data pairs $\mathbf{s}_i = (\mathbf{x}_i, t_i),\ i = 1, ..., N$ where the *time stamp* t determines the order of the data point or image $\mathbf{x}$:
>
> $\mathbf{S} = \{\mathbf{s}_i\}$ with $\rho(\mathbf{s}_j) < \rho(\mathbf{s}_k) \Leftrightarrow t_j < t_k\ ,\ i = 1, ..., N\ ,\ j = 1, ..., N\ ,\ k = 1, ..., N$
>
> $\rho(\mathbf{s}_i)$ indicates the position of $\mathbf{s}_i$ in the sequence $\mathbf{S}$.

In most applications the alignment of such image sequences can be performed based on the time stamps. However it is not available in all situations. For example an unstructured photo collection has to be arranged in an sequential order and numerous CBIR-systems arrange image data in an one dimensional list to present retrieval results. So a number of questions arose regarding image sequences:

- How to describe the transformation between two succeeding pictures?

  The difference between two images depicts the essential attribute of image sequences. In [Radke et al., 2005] numerous approaches regarding distinguishing pictures are reviewed. Regarding image sequences temporal models based on pixel location are interesting. Different image comparing tasks require the detection of the background in the pictures. For that purpose a number of approaches are listed, mainly based on a mixture of Gaussians model.

  Furthermore the *optical flow* is interesting in respect to image sequences. If the pictures are recorded with a high frequence the optical flow can give important insights to align the images of a sequence. Numerous approaches to compute the optical flow are established [Beauchemin and Barron, 1995].
- Which automatic approach can find the one dimensional structure in an image set?

  The current research community offers a repertory of approaches suitable for one dimensional alignment tasks. Further algorithms are specified for analysing sequential data. Principal curves [Hastie and Stuetzle, 1989] or time series analysis [Chatfield, 2004] are just two example techniques interesting here. Furthermore a lot of neural network approaches can be modified according to one dimensional structures.
- Is it possible to specify different subsets or categories by an one dimensional structure?

  If an image set consists of a number of image sequences, can an one dimensional structure put into the whole set be used to distinguish these sequences? This will be equivalent to a classification along one direction.
- Whats about *semantic sequences*?

  Semantic alignment depends strongly on expert knowledge. Examples may be the historical ordering of buildings according to their architectural style. Other semantic alignments may be describable by content like the growing of a child or tree but usually hard to match between content and interpretation. Indeed this task has not been investigated.
- How to detect interesting things in image sequences?

  Scene observations and video surveillance tend to get interesting or important events in the observed scenes. Various approaches have been analysed regarding these tasks [Collins et al., 2000].

In general one dimensional structures within image sets are covered. Frequently the images, elements of the same sequence, are included in a larger collection. Thus the detection of these pictures is desired. To analyse this a synthetic set of image sequences is constructed (see section 3.2.2) and used for various analyses.

Figure 3.2: Example images of the artexplosion photo collection, images in one row are from the same category, namely a) underthesea b) animals c) doorswindows d) teddybears e) sunrisesunset f) venezuela g) iceland

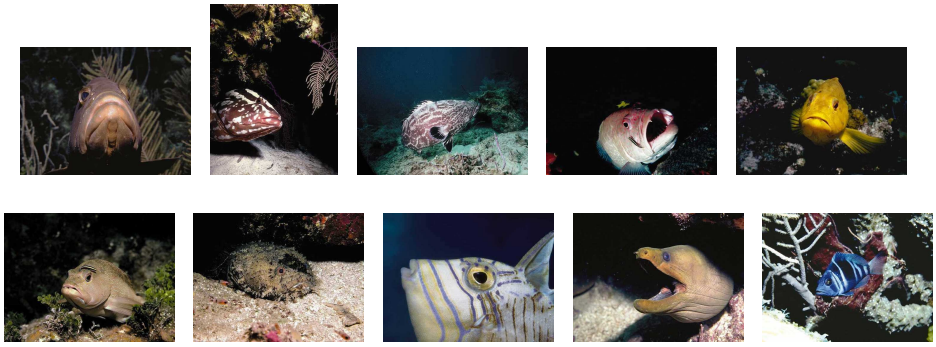


Figure 3.3: User defined subset *fish*.

3.2 Selected Image Sets

In the following section particular data sets are introduced. It has been shown that the data sets under consideration should be analysed before they are used within image retrieval research. Hence in this section the image sets used for further investigations in this work will be presented. Namely it is a subset of the artexplosion photo collection (section 3.2.1), a set of synthetically generated image sequences (called myMondrian – section 3.2.2) and a set of underwater webcam images of the London Aquarium (section 3.2.3).

The used data sets are described from the user's point of view, e.g. the visual attributes and the semantic interpretation preponderate the statistical features. The latter are defined in section 3.3.

3.2.1 Artexplosion Photo Collection

Private photo collections as well as the gigantic picture stocks of media and advertising agencies are among the most popular image sets used for CBIR applications. Consequently image retrieval systems are developed and analysed for photo collections. Unfortunately, until now no free benchmark photo collection for CBIR tasks exists. Actually the University of Washington tries to establish a ground truth database [Shapiro], but it is not as popular as the brodatz-collection [Tranden] for texture-feature analyses.

A lot of image retrieval on photo collections base on the corel image collection [corel]. Unfortunately pictures of these collection have a rather low quality. Furthermore each CBIR-system uses another subset of the corel set. Thus the corel collection cannot serve as a ground truth for image retrieval as Müller et al. [2002] convincingly demonstrate.

In this work, a set of 1499 photos of the artexplosion image collection [artexplosion] is used. This collection consists of seven thematic groups, namely underthesea (300 images, a priori probability about 0.2), animals (300, 0.2), doorswindows (300, 0.2), teddybears (100, 0.07), sunrisesunset (300, 0.2), venezuela (100, 0.07) and iceland (99,0.07). Figure 3.2 presents a selection of the used images. For some experiments, this set is reduced in respect to more specific subsets. A restricted set may be the set of *pictures showing one single fish, at the middle of the image and covering a large part of it* (see figure 3.3 fish examples).

The artexplosion image set with its thematic groups and special subsets, is appropriate for three common search tasks (target search, category search and browsing – see figure 2.1). A closer look on the examples above reveals that desired image categories may overlap

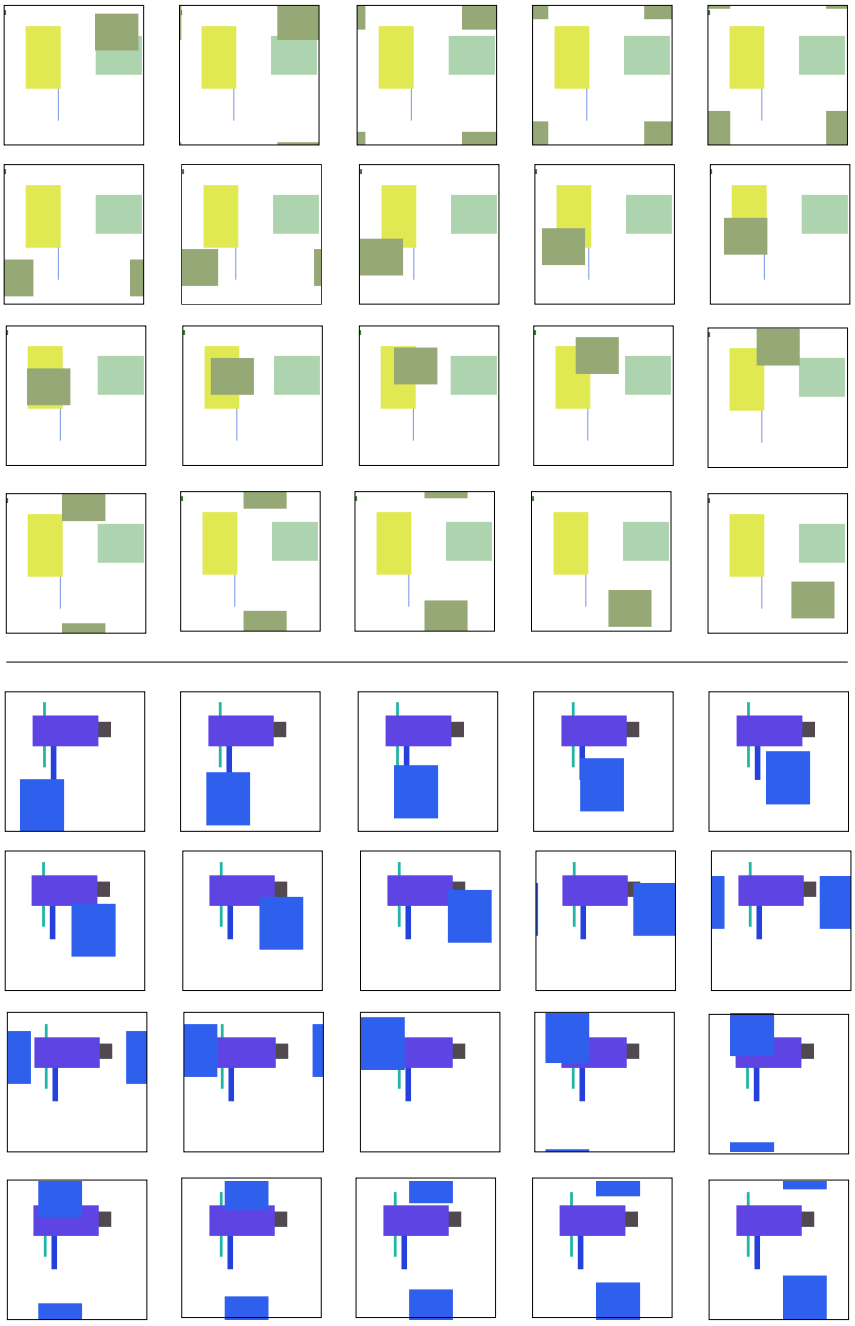


Figure 3.4: Two example sequences generated by the myMondrian-algorithm. The selection of all images belonging to one of these sequences out of the entire image set generates a category search.

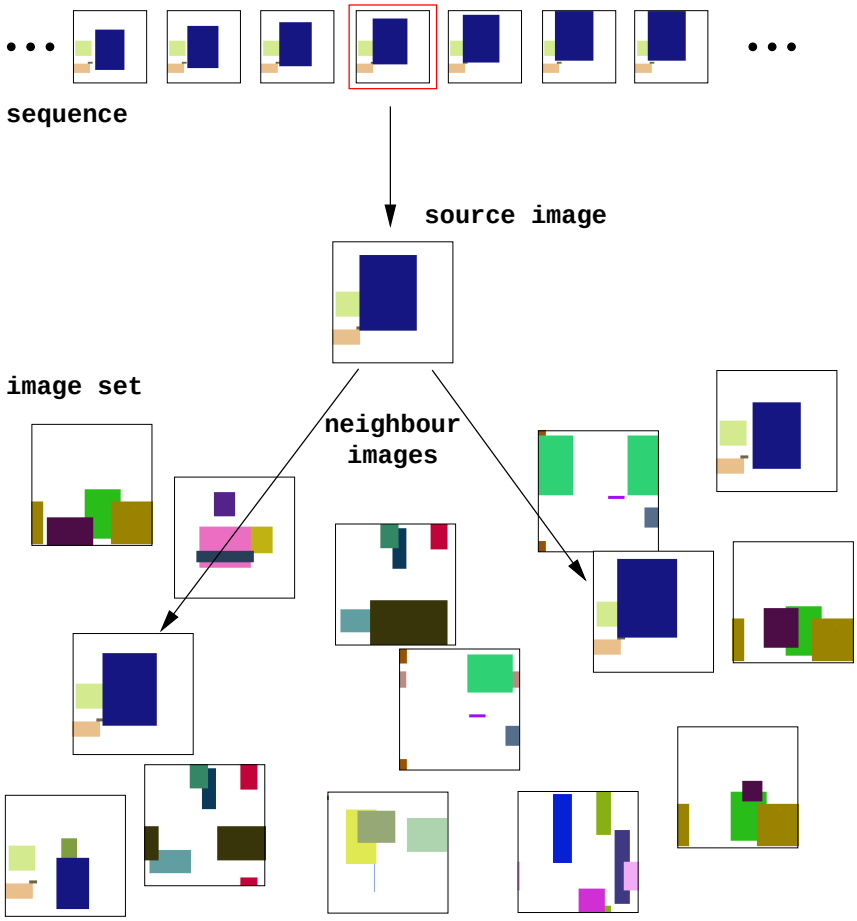


Figure 3.5: Example of an alignment task in the myMondrian image set.
Starting from the source image presented in the middle of the figure the image sequence presented above should be recovered. Therefore the neighbouring images within the original sequence have to be retrieved from the unstructured image set shown below the source image.

with different predefined categories. For example some iceland-images show horses, which are obviously also animals.

3.2.2 myMondrian Image Sequences

Artificial images with restricted and defined properties lessen the challenges of real world pictures. A ground truth set for analysing and evaluation purposes is given. Furthermore the design of the set can be tuned to a required application.

An example for such an artificial image database is the myMondrian collection, motivated by pictures of Piet Mondrian [Scheer, 1995]. The layout of these images contains a number of rectangles which are transformed within each image sequence.

There are different kinds of timeline transformations for each rectangle:

- **motion**: Objects move along defined or arbitrary directions.
- **growth**: Objects become bigger or more generally cover a greater part of the pictures.
- **colour change**: The colours of the objects changes.

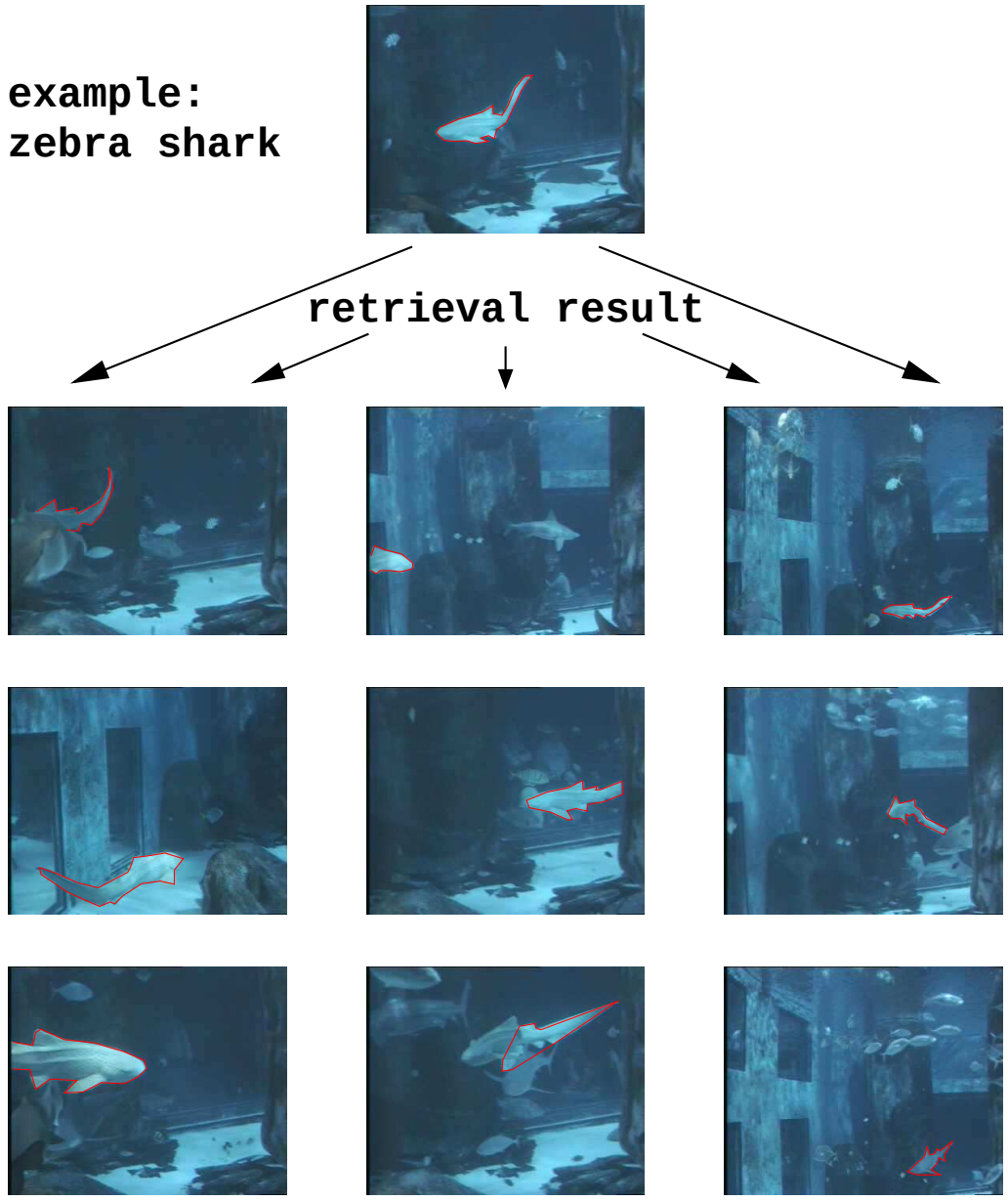


Figure 3.6: Example search in the underwater images of the London Aquarium: Search for all images showing a zebra shark.

These transformations can be realised separately or in arbitrary combinations, leading to different image sequences.

Table A.1 describes the constructed image sets.

Two image retrieval tasks are meaningful in this image set: A category search would separate the different image sequences, e.g. the two sequences presented in figure 3.4. More generally all images belonging to a required sequence are retrieved from the entire set.

Besides a target search will be performed if that image is required that follows a given image in the sequence (see figure 3.5). The repeated execution of this task will implement the alignment of an image set based on the chronological order. This task will be performed and analysed in section 4.

3.2.3 Shark Webcam of the London Aquarium

The motivation of many image processing and retrieval tasks lies in the context of interdisciplinary research. An upcoming field of research, which benefits from the progress in recording technologies is the oceanography. Since the underwater environment is rather misanthropic few pictures of underwater scenarios exists recently. Now the developments of AUV (autonomous underwater vehicle) and ROV (remotely operated vehicle) in combination with the improved camera and data storing devices offer the possibility of getting a lot of underwater pictures. Today underwater computer vision is a forthcoming field of research [Kak et al., 2000].

Underwater images depend on the special physical attributes of water: colour extinction, reflection and scattering. The main features are the absence of colour in greater depth, varying contrast, nonuniform and dim lighting and a lot of blur. Therefore, image retrieval suffers from the physical conditions as well as from the characteristics of underwater objects.

To fix some of the outdoor problems in natural science most of the research tasks are investigated in an artificial situation before exploring the real world scenario. Utilising this accepted procedure underwater images taken in an aquarium and shared via a webcam are used.

The physical conditions of underwater images complicate the application of common approaches for image segmentation like segmentation by colour. Furthermore, the interesting objects are quite different and often very hard to delineate.

Cameras arranged in a fixed position provide images with a quite similar background. This way, the detection of interesting things by a difference calculation to the background is justifiable.

In this work we consider images from the London Aquarium [London Aquarium]. The images spot a subregion of the aquarium, that contains sharks (sand tiger, brown sharks, zebra sharks), sting-rays and different sorts of fish swarms. The webcam switches between four settings (see figure 2.7 on page 21) and has an update rate of 5 seconds.

The performed search task in this image set can be described by *Give me interesting images!*. *Interesting* is hard to model as will be discussed in section 5.1. To avoid a definition of interesting the search task is approximated by a similarity search according to a selected example. Since the interesting entities are represented by image regions

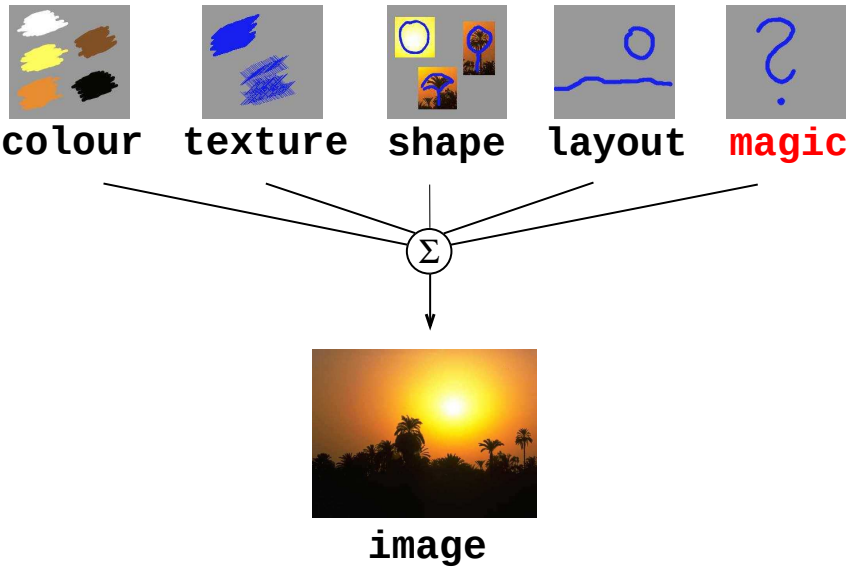


Figure 3.7: CBIR-systems usually perform a multi-level representation to describe a picture. Low-level features like colour, texture, shape or layout are suitable to represent the image primitives. But the human recognition of the content depends on semantic interpretations. For the automated system this is still *magic* [Pecenović et al., 1998].

containing the desired objects, the example will be given by a subimage. Technically such a search task is a category search, where the category is built by the image regions showing the desired object. Figure 3.6 represent an example search.

3.3 Feature Data

3.3.1 Feature Detection Approaches

The computation of suitable image features is an important step for designing a CBIR-system. Humans mostly use semantic features when describing images. Unfortunately the automatic extraction of semantic features from an arbitrary image is still unsolved.

Suitable feature algorithms are tuned to represent as much as possible of the relevant image information. They reflect human perception and intention. One example is the human world feature by Eidenberger and Breiteneder [2002]. Others are based on higher statistical levels, which certainly are very interesting for computer vision and statistical analyses. However they do not conform with human perception. Nevertheless image retrieval based on such features is satisfying. Examples are the Multi-Resolution Analysis (MRA) [Eidenberger, 2004] or ICA-based features (see section 5.2). Both kinds of higher-level features combine global with local features. Among the global higher-level features the Spatial Envelope [Oliva and Torralba, 2002] is interesting.

Base-level representation of image features are colour or texture. In general they are not sufficient to describe an image content completely, as the *magic*-module in figure 3.7 states. But they can be combined to multi-level features [Pecenović et al., 1998]. A lot of CBIR-systems perform the combination on the similarity level. An obvious advantage of such an approach is that global attributes, like a colour histogram of the entire image, can be combined with local attributes, like the shape of a striking region.

Such low-level features are most flexible and can be used in nearly every kind of image domain. Therefore these low-level features are also used in this work.

A further important reason for dealing with image features instead of raw image data in CBIR-systems is the dimension of the data. Feature algorithms are very suitable techniques to reduce the dimension of pictorial data in order to reduce the computation time for the retrieval process. Common approaches to reduce the dimension of data can be used for images, e.g. principal component analysis (PCA).

3.3.2 Used Image Features

A fixed set of image features is necessary to compare the retrieval results. The selection of a suitable distance measure depends on the feature. Hence, for each feature algorithm the used distance measure is named:

Colour

Colour is always a meaningful feature [Swain and Ballard, 1991]. Motivated by the variety of colour attributes, different colour spaces have been defined. The most popular colour space is the RGB-colour space. Three channels measure separately the colour values for red, green and blue. The features are defined as histograms for each channel. Furthermore the intensity is computed as well as a more detailed colour histogram incorporating all channels.

The second colour space used here is the HLS-colour space which is more fitting to the human perception than the RGB-colour space. Hue, lightness and saturation are measured in histograms. Additionally they are quantised in a 34 dimensional feature vector.

The histograms of the query and the database images in the RGB and the HLS-space respectively are compared by histogram intersection. For the quantised feature in the HLS-space the Euclidean distance is used.

Structure

To represent the layout of the images the disposition of the colour within the images is described. Therefore, the image is cut into subregions by a 3×3-grid as well as a 5×5-grid. The predominant colour values in the resulting image patches are represented in a feature vector. Here the IHS-colour space measuring intensity, hue and saturation is used. Four feature vectors are computed: one based on the intensity value (34 dimensional), one based on the hue value (34 dimensional), one based on the saturation value (34 dimensional) and one where all three are concatenated (102 dimensional). The Euclidean distance is used as the distance measure.

Texture

As texture feature the algorithm of [Unser, 1986] is used. It calculates a 32 dimensional texture feature based on the gray value co-occurrence matrix. Distances are measured by the Euclidean distance.

Usually the features are used separately and the combination is performed on the similarity level by linear-combinations of the respective distance values.

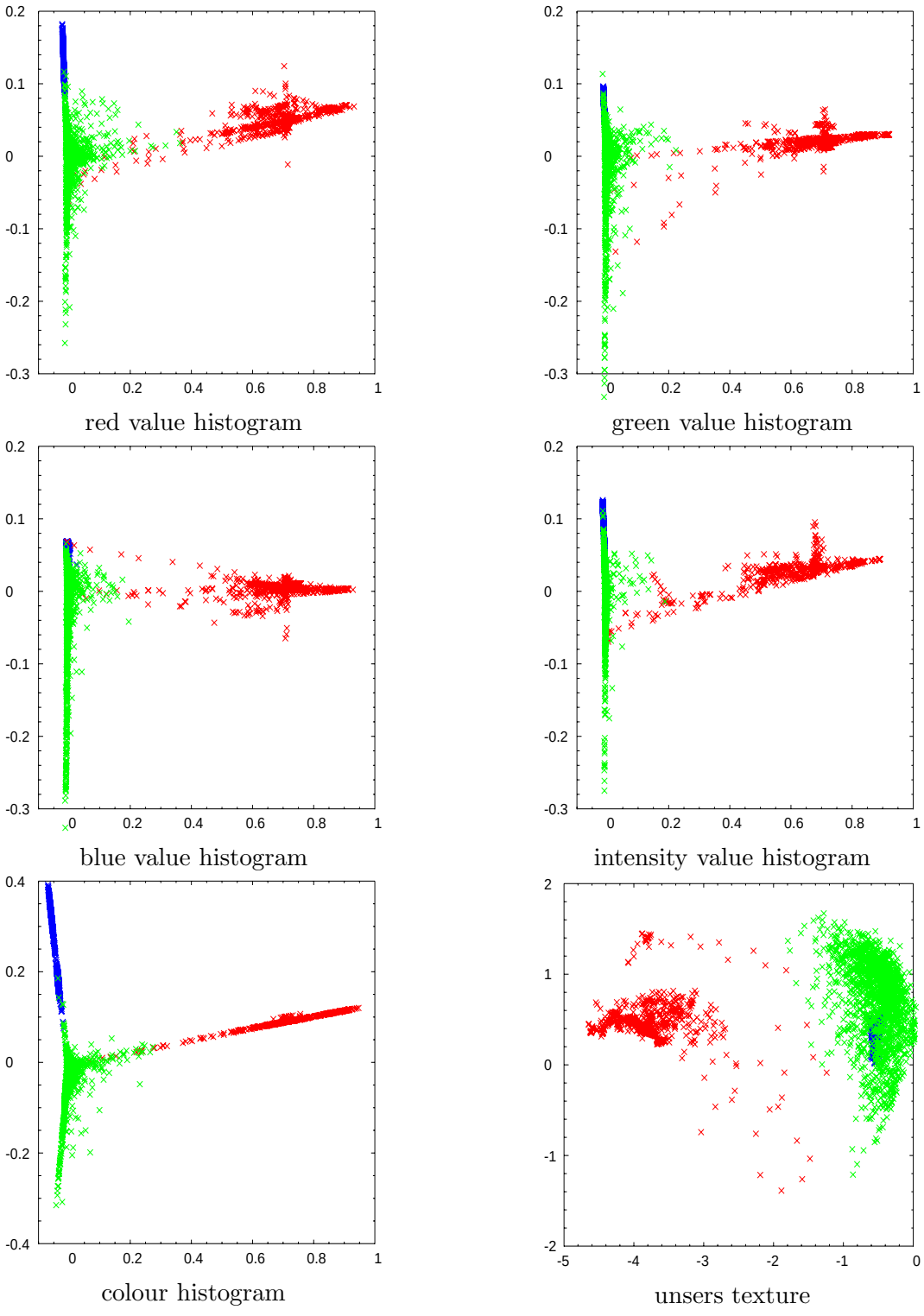
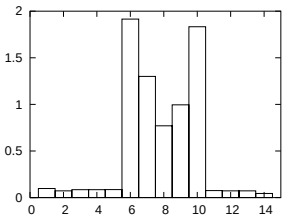
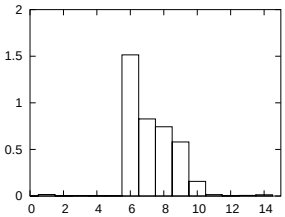


Figure 3.8: Distribution of the three domains along the eigenvectors with the two largest eigenvalues. The eigenvectors are computed in the combined data set based on different image features.
Green stands for artexplosion photos, red marks myMondrian images and blue represents pictures of the shark webcam.

	feature	the five largest eigenvalues				
1	Colour Histogram	0.098	0.016	0.007	0.004	0.003
2	Intensity Histogram	0.073	0.002	0.002	0.002	0.001
3	Red-value Histogram	0.085	0.004	0.002	0.002	0.001
4	Green-value Histogram	0.085	0.003	0.002	0.002	0.001
5	Blue-value Histogram	0.086	0.003	0.002	0.002	0.001
6	**Structure IHS**	**1.915**	**1.514**	**0.776**	**0.610**	**0.569**
7	**Structure Intensity**	**1.300**	**0.827**	**0.442**	**0.356**	**0.216**
8	**Structure Hue**	**0.771**	**0.744**	**0.499**	**0.321**	**0.218**
9	**Structure Saturation**	**0.995**	**0.580**	**0.416**	**0.376**	**0.231**
10	**Unsers Texture**	**1.833**	**0.157**	**0.027**	**0.005**	**0.003**
11	Hue-value Histogram	0.076	0.015	0.002	0.002	0.002
12	Lightness Histogram	0.073	0.003	0.002	0.002	0.001
13	Saturation Histogram	0.073	0.007	0.003	0.002	0.001
14	Quantisation of HLS	0.044	0.014	0.009	0.007	0.003



eigenvalue 1



eigenvalue 2

Figure 3.9: Eigenvalues of the different features based on the combination of the artexplosion photo collection, myMondrian and shark webcam images. In the table the largest values along each eigenvector are bold. Probably these features are more suitable for image retrieval in the whole image set than the colour features.

3.3.3 Analyses of the Used Features

In order to ensure that the used feature algorithms are suitable for image retrieval tasks the feature vectors are analysed. For base-level analyses, automatic approaches are appropriate since they are economical and reproducable. Classification tasks resembling category searches are most suitable in this case. The three domains introduced above (artexplosion, myMondrian and shark webcam) are merged to one data set to get a broad database.

Beginning with a qualitative inspection the visualisations of the feature distributions in the feature spaces are examined. For visualisation purpose a PCA is performed and the projection on the two directions of the largest variances are presented in figures 3.8, 3.10 and 3.11. Some assumptions regarding the separability of meaningful subsets in certain feature spaces emerge:

- The structure features are suitable for detecting the four camera positions of the shark webcam. Clustering this data set based on the structure features therewith is suitable.
- All feature types can distinguish between different sequences of the myMondrian set. The combined structure feature (intensity, hue and saturation) seems to be more suitable than the single structure feature. This can be recorded as an example for the advantages of feature combinations.
- In the colour feature spaces the myMondrian set varies in a quite orthogonal direction to the orientation of the other sets. This is a hint, that the most suitable feature combination depends on the used domain.
- The (semantic) categories of the artexplosion collection cannot be clustered easily in the used low-level feature spaces. Low-level features are not sufficient to describe the image content (semantic gap).

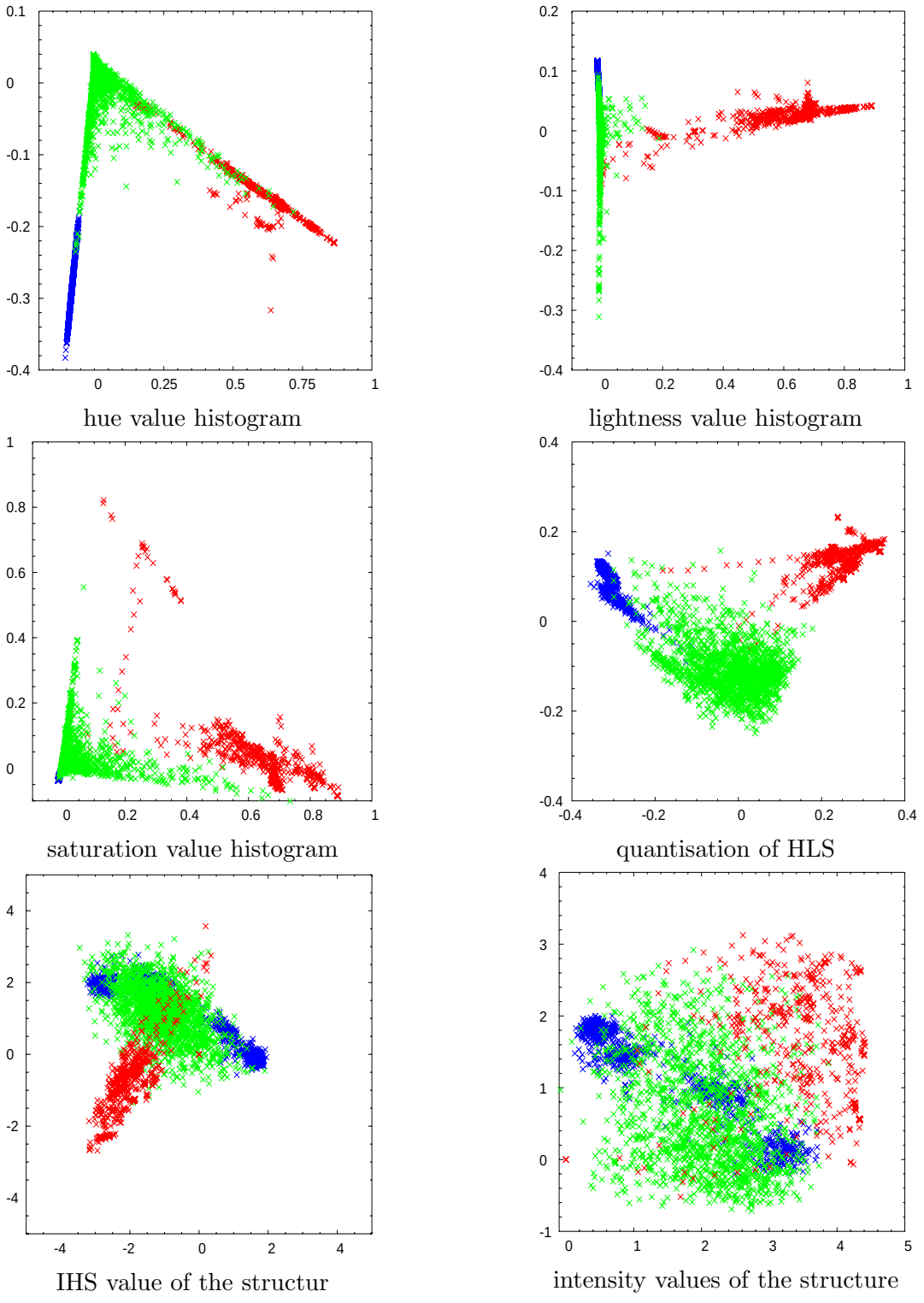


Figure 3.10: Distribution of the three domains along the eigenvectors with the two largest eigenvalues. The eigenvectors are computed in the combined data set based on different image features.
Green stands for artexplosion photos, red marks myMondrian images and blue represents pictures of the shark webcam.

	structure features								texture	
category	IHS		Intensity		Hue		Saturation			
artexplosion	−1.04	0.74	2.06	0.54	2.06	0.61	−1.18	0.73	−0.52	0.08
	1.17	0.6	0.66	**0.63**	−1.5	**0.87**	1.11	0.47	0.69	**0.23**
-underthesea	−0.9	0.85	1.82	0.54	1.95	0.5	−1.29	0.71	−0.37	0.08
	1.13	0.57	0.79	0.53	−1.33	**0.8**	1.11	0.32	0.51	**0.25**
-animals	−0.8	0.72	1.81	0.54	1.6	0.63	−1.01	0.84	−0.48	0.07
	0.87	0.56	0.95	**0.57**	−1.29	0.54	1.43	0.47	0.81	**0.11**
-doorswindows	−1.01	0.55	2.21	0.50	2.33	0.66	−1.07	0.56	−0.64	0.07
	1.02	**0.57**	1.0	**0.62**	−1.62	0.56	1.15	0.46	1.00	**0.1**
-teddybears	−0.89	0.65	2.11	0.50	2.06	0.43	−0.87	0.70	−0.55	0.05
	0.91	0.44	0.85	0.47	−1.59	**0.63**	1.3	0.35	0.88	**0.07**
-sunrisesunset	−1.52	0.74	2.21	0.5	2.19	0.54	−1.37	0.7	−0.58	0.08
	1.65	0.53	0.04	0.28	−2.14	**0.94**	0.58	0.36	0.21	**0.18**
-venezuela	−0.97	0.5	2.23	0.34	2.19	0.42	−0.91	0.53	−0.67	0.1
	1.01	0.37	0.61	**0.63**	−1.27	**0.66**	1.16	0.31	1.09	0.06
-iceland	−1.11	0.4	2.38	0.4	2.4	0.26	−1.68	0.61	−0.46	0.04
	1.54	0.19	0.15	0.27	−0.43	**0.52**	1.31	0.33	0.74	**0.05**
myMondrian	−2.13	0.43	3.53	0.49	1.19	1.03	−0.41	0.35	−3.74	0.27
	−1.03	**1.05**	1.96	**0.71**	−1.02	0.55	0.28	**0.78**	0.48	0.09
shark webcam	0.001	2.83	1.44	1.17	2.45	0.17	−0.83	1.55	−0.54	0.001
	0.9	0.82	1.2	0.41	−0.66	**0.27**	1.39	0.09	0.32	**0.01**

Table 3.3: Mean (left value) and variance (right value) along the first (first row) and second (second row) principle components of the combined data set. The bold values select categories and features where the variance along the second eigenvector exceeds that one along the first eigenvector. This indicates, that single categories have their largest variety in another direction than the entire data set.

These observations have to be supported by further quantitative analyses. An obvious measure to analyse the distribution of the data in the feature space is the variance. Therefore the eigenvalues of the combined image set are computed and the five largest ones are listed in figure 3.9 for each feature.

The results show, that the structure features and the texture feature detect better the variablity within the set. Thus they are more suitable to detect interesting subsets. This coincides with the qualitative observation in figures 3.8, 3.10 and 3.11. Regarding colour, the colour histogram seems to be the most suitable for detecting differences between images.

Since structure and texture seem to be appropriate to divide image sets into subsets, they are analysed in detail regarding single domains (see table 3.3).

One observation is that in some feature spaces and for some domains or categories the variance along the second eigenvector exceeds the variance along the first eigenvector of the combined data set. Consequently the main extension of the data within a feature space depends on the domain. Coinstantaneously the efficiency of the image features also depends on the image domain. The development and selection of domain dependent feature detection algorithms may be a consequence.

Similar to that is the task dependent feature weighting used in [Deselaers et al., 2004a]. The retrival performance based on the error rate in a classification approach is analysed. They observed that colour histograms are a good choice to describe arbitrary photographs

	colour	**structure features**				**texture**
category		IHS	Intensity	Hue	Saturation	
artexplosion	0.004	1.298	0.876	0.906	0.895	0.242
	0.002	0.969	0.480	0.737	0.481	0.087
– underthesea	0.009	1.396	0.772	0.831	0.804	0.275
	0.003	0.870	0.400	0.637	0.494	0.088
– animals	0.002	1.328	0.813	0.703	1.072	0.139
	0.001	0.875	0.506	0.658	0.509	0.052
– doorswindows	0.002	1.005	0.773	0.781	0.747	0.125
	0.001	0.897	0.657	0.730	0.513	0.051
– teddybears	0.003	1.215	0.739	0.933	0.896	0.087
	0.002	0.982	0.441	0.520	0.393	0.041
– sunrisesunset	0.008	1.452	0.765	1.016	0.898	0.195
	0.005	1.105	0.420	0.637	0.500	0.075
– venezuela	0.001	1.406	1.042	0.844	0.944	0.115
	0.001	1.035	0.439	0.665	0.447	0.059
– iceland	0.001	1.263	0.737	0.689	0.982	0.059
	0.001	0.860	0.453	0.378	0.495	0.044
myMondrian	0.071	2.502	1.072	1.193	0.982	0.905
	0.007	1.446	0.641	0.648	0.545	0.094
shark webcam	0.002	3.898	1.665	0.649	1.729	0.014
	0.001	1.250	0.481	0.268	0.587	0.003

Table 3.4: The two largest eigenvalues of the different feature spaces. PCA is computed on each subset separately.

whereas the pixel values combined with a suitable distance measure are better for medical radiographs. Therefore they confirm the demand to select image features task and domain dependently.

In section 5.2.3 the distribution of a single subset in relation to the image domain is used to evaluate the impact of a data space transformation. The developed measure compares the distances within a relevant subset with the distances to the remaining data. Independently from the transformation approach it can be observed that again the separability of different subsets depends on the used feature.

As has been shown in table 3.3 the different categories and domains show larger variances along the second eigenvector in different feature spaces. This motivates the assumption that for different data sets different features are more suitable to describe these sets. Therefore the eigenvalues in the feature spaces according to single subsets are computed and listed in table 3.4.

The category dependent eigenvalues exceed the eigenvalues computed in the entire data set (see table 3.3). Domain dependent features may be advantageous. Just the second eigenvalues in the texture feature space are smaller. This is a hint that the entire data set in the texture space constitutes a mixed and compact cluster without explicite directions for the individual domains. This is confirmed by the fact that in the PCA-space of the combined data the second eigenvalue is larger for almost all regarded subsets (see table 3.3). The visualisation of the distributions in figures 3.8, 3.10 and 3.11 indicates this. The texture feature may be adequat to describe the textures in all domains but unsuitable to perform a categorisation into the three domains.

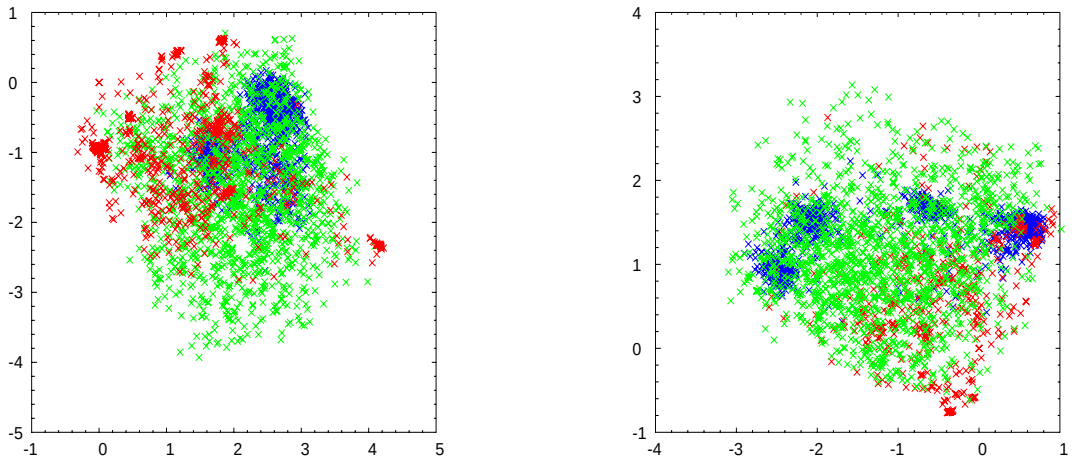


Figure 3.11: Distribution of the three domains along the eigenvectors with the two largest eigenvalues. The eigenvectors are computed in the combined data set based on different image features. Green stands for artexplosion photos, red marks myMondrian images and blue represents pictures of the shark webcam.

3.4 Summary of Image Data

CBIR tasks are applied on a variety of different data. The main types are images and feature vectors. The image set constitutes the basis of all computations. Furthermore, the user satisfaction depends on the pictures, since they search tasks and semantic interpretations are determined by the image set. Analysed image domains are photo collections, image sequences and aquarium pictures taken by a webcam. A number of low-level features are computed on these images. The distributions of the feature data differ depending on the various image domains. Thus the suitability of a single feature detection algorithm to describe the images depends on the given pictures.

Technically the retrieval performance as well as suitable feature detection algorithms depend on the image set. Usually low-level features like colour or texture are computed to represent the pictures. Most of the retrieval approaches are performed on these feature data. Thus the retrieval performance is determined by the feature detection as well as automatic clusterings or groupings of the data sets. Unfortunately, most user defined groupings are based on semantic interpretations. So low-level features have to be improved or combined to narrow the semantic gap. Semantic enhancements and adaptations to human intentions are desirable improvements.

In general, it becomes obvious that the feature detection algorithms used to describe an image domain have to be selected carefully. Attributes of the image set are important as well as search tasks and user intentions. Automatic assistances to choose the most suitable features are desirable.

Usually even the most suitable set of low-level features is not sufficient to describe real world images like the artexplosion photo collection. The semantic gap cannot be closed by them. Therefore, advanced approaches have to be developed to combine the features appropriately. The *magic*-part of image descriptions (see figure 3.7) has to be satisfied.

Chapter 4

Sequential Data Organisation and Categorisation by 1dSOMs

A ordinary approach to find a desired picture in an unstructured collection is to eye the images one by one. Humans are familar with this procedure. Nevertheless this is a time-consuming and often boring way to get a required image. Thus an appropriate sequential ordering can be helpful to support the search in image collections. An automated approach to perform this is preferable. In general, the sequential ordering of images is a desirable process step for image retrieval tasks.

The pictures of an arbitrary mixed set usually have no specific order. However, it can be assumed that similar images or consecutive pictures of a sequence are neighbours in the data space. Using artificial neural networks (ANNs) such an ordering might lie on or near a low-dimensional manifold. *Self-Organising Maps* (SOMs) [Kohonen, 1997], also known as Kohonen-Maps, are neural networks promising to perform an automated ordering of images. They preserve topological relationships to represent the similarity of consecutive images. Furthermore they adapt unsupervised. Thus the trainings step does not require a labelled data set. A similarity graph of the input data can be computed.

In this chapter a SOM based approach to perform a one dimensional alignment of images is presented. The theoretical background is introduced and several experiments are carried out in order to evaluate the approach.

4.1 Self-Organising Maps

Classical SOMs are two dimensional, spatial interacting networks, often used for adaptation and regression. The goal usually is a low dimensional data representation maintaining similarity relationships corresponding to the topological relations. In contrast to PCA as a linear approach, SOM can be considered as a non-linear extension. Based on the mapping of the statistical relationships between high-dimensional data on a low-dimensional display by preserving topological and metric relationships visualisation and abstraction of high-dimensional data are performed. Each node (often called neuron) a of a two dimensional grid A is associated with a reference vector $\mathbf{w}_a \in \mathbb{R}^D$, where D is the dimension of the input data. In the training process the reference vectors $\mathbf{w}_a$ are adapted to the input data. The SOM update rule for the weight vector of the unit a is:

$$\mathbf{w}_a(t+1) \;=\; \mathbf{w}_a(t) \;+\; h_{aa^*}(t)(\mathbf{x}(t) - \mathbf{w}_a(t)) \tag{4.1}$$

where t denotes time. The $\mathbf{x}(t) \in \mathbb{R}^D$ is the input vector randomly drawn from the data set at time t and $h_{aa^*}(t)$ the neighbourhood function around the winner neuron a^*

according to the input vector at time t. The neighbourhood function is a non-increasing function of time and of the distance between neuron a and the winner neuron a^*. It defines the region of influence that the input sample has on SOM. A decreasing function $\sigma(t)$ downsizes this region during learning. A common neighbourhood function is based on the Gaussian function:

$$h_{aa^*}(t) \;=\; \exp\left(-\frac{||\rho(a)-\rho(a^*)||^2}{2\sigma^2(t)}\right)\epsilon(t) \tag{4.2}$$

where $\rho(a)$ is the location of unit a on the map grid. Usually the neighbourhood radius is bigger at first and is decreased, e.g. linearly, to one during the training. $\epsilon(t)$ is a decreasing learning rate.

For visualisation purpose, the input $\mathbf{x}$ will be assigned to that neuron a^* which has the most similar reference vector to the input vector. The best match node a^* is given according to:

$$a^* \;=\; \arg\min_{a\in A} d(\mathbf{x},\mathbf{w}_a)\ ,\ a=1,...,N_n \tag{4.3}$$

where $d(.)$ is a distance function, usually the Euclidean distance. Using the learning algorithm outlined above the global ordering of the reference vectors will be reached in a finite number of steps. Based on the mapping of the input data to the SOM nodes the data is ordered, as well.

Since the first works on Self-Organising Maps in the 1980s [Kohonen, 1982] [Ritter and Schulten, 1986] [Ritter, 1991] SOM has become a wide spread field of research as the bibliographies [Kaski et al., 1998] and [Oja et al., 2002] show. Applications are investigated as well as improvements of the algorithm and modifications of the approach:

To overcome the discrete character of common SOMs a Parametric SOM (PSOM) is presented [Walter and Ritter, 1996]. The discrete grid positions of the SOM nodes are generalised to a continuous manifold. It is enhanced for noisy and incomplete data [Klanke and Ritter, 2005]. In [Saalbach et al., 2002] PSOM is used for classification and pose estimation of the objects in a COIL (Columbia Object Image Library [Nene et al., 1996]) image set.

A similar approach is the Continuous SOM (C-SOM) [Aupetit et al., 1999] and [Campos and Carpenter, 2000]. Basically, an interpolation step is added after the training of a common SOM grid. C-SOMs are used to perform continuous function approximations.

Usually the grids of SOMs are based on the Euclidean space. Indeed the embedding of complex highdimensional and hierarchical structures in this space is limited by the restricted size of the neighbourhood of a point. Since the hyperbolic space is better qualified to represent highdimensional neighbourhoods, Hyperbolic Self-Organising Maps (HSOM) are introduced in [Ritter, 1999]. It is used for browsing in text-databases in [Ontrup and Ritter, 2001a] and [Ontrup and Ritter, 2001b].

Apart from visualisation purposes, SOM based approaches are used for feature detection. In [Kohonen et al., 1997] as well as in [de Ridder et al., 2000] and [de Ridder et al., 2001] unsupervised procedures to compute adaptive subspaces are presented and utilised for feature detection.

As presented in section 2.2.1 Self-Organising Maps can be arranged hierarchically. The resulting approach is called Tree-structured SOM (TR-SOM) and has been successfully used in text [Kohonen et al., 2000] and image retrieval [Laaksonen et al., 2000].

Based on the mapping to the SOM grid, classification tasks can be performed. For example in [Kämpfe et al., 2001] SOM is used to classify image patches.

In order to arrange pictures sequentially a one-dimensional SOM version may be suitable. As shown in [Kohonen, 1997], for an appropriate number of learning steps t ($t \to \infty$) in a one-dimensional SOM the weights $\mathbf{w}_a$ become ordered ascendingly or descendingly. Since each weight is associated with a data vector or image, such a 1dSOM is interesting regarding picture alignment tasks.

1dSOM

1dSOM is a chain A of N_n consecutive nodes. The associated reference vectors $\mathbf{w}_a$, $a = 1, ..., N_n$ approximate the image space. The learning algorithm is the same as in the standard two-dimensional SOM above. For each image $\mathbf{x} \in X$ the the best match node a^* determines the position $\rho(\mathbf{x})$ along the sequence. The direction of the changes based on the sequential structure cannot be detected by the 1dSOM. Regarding pictures the reverse order is as good as the forward movement.

The usage of 1dSOM to align data in a sequence resembles the travelling salesman problem. Therefore different approaches are investigated. Self-Organising Maps and elastic map models are identified as suitable to solve such combinatorial optimisation tasks [Smith, 1999]. For example in [Bacao et al., 2005] this is used for path finding in marine patrol situations. Here, the optimal route to inspect critical or interesting points in the sea are desired. This resembles the approximation of a data distribution by a trajectory, since a variety of connections between interesting points are possible in a lot of situations.

Another important aspect of SOMs is the assignment of a number of data samples to one node. This can be problematic for image alignment since the pictures matching the same node still set up an unordered set. An elementary solution to this problem is an oversized 1dSOM with at least as many nodes in SOM as pictures in the sequence, $N_n \geq N$. Thus, the desired mapping of single images to each node becomes possible. Concerning the task of ordering one image sequence by one 1dSOM this is justifiable.

4.2 Experiments for Image Alignment

In order to analyse 1dSOM applied on image data, different image sets are used as well as alignment and categorisation tasks. Initially a collection of synthetically constructed picuteres is used. The myMondrian images (see section 3.2.2) offer a number of image sequences defined by moving rectangles within each sequence. Thus, these sequences offer ground truth aligmnets. The ability of 1dSOM to arrange images in the correct order is analysed based on this set. The second task to investigate is the grouping ability of 1dSOM. Therefore the pictures should be classified according to the individual sequences.

The applicability to align real world images is analysed based on an aquarium observation image set as well as on a photo collection. The first set is a collection of webcam images and should be aligned based on the moving fish. The second one is an unstructured set resembling a private photo collection. Thus, a sequential ordering to perform a slide show is interesting as well as a classification task to group images according specific attributes. Hence, both experiments performed on the myMondrian images are repeated with this real world picture set.

The experiments are performed on the low-level content based features (see section 3.3). The parameters of the used 1dSOMs are listed in the appendix in table B.1.

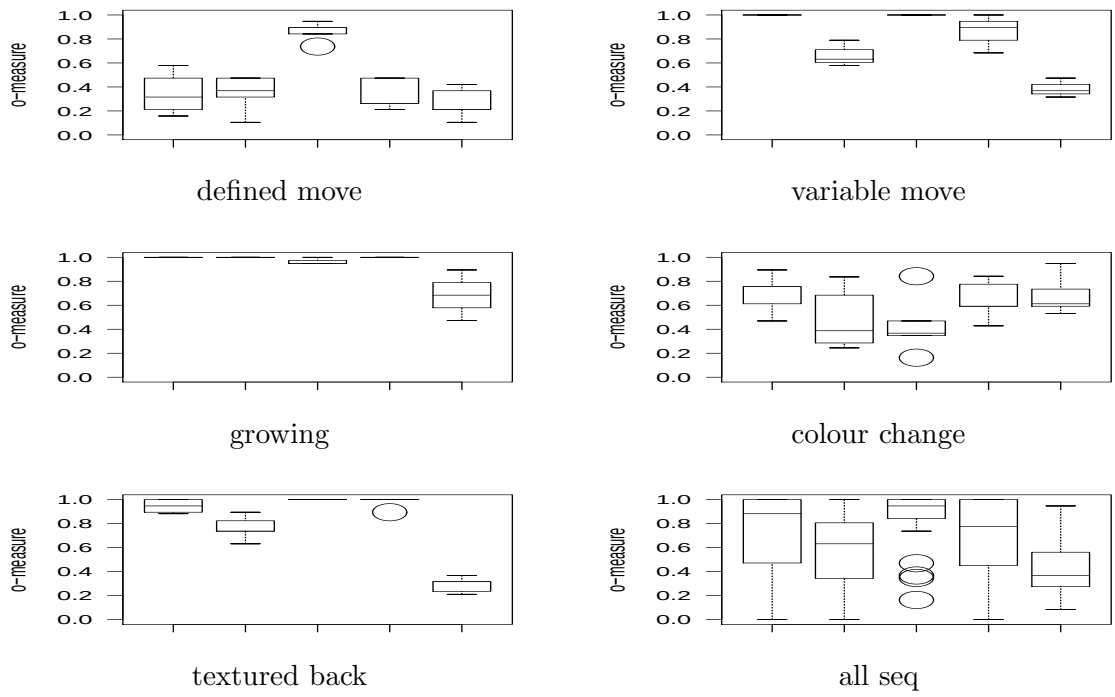


Figure 4.1: The *o*-measures of different sequence types and features. In each plot features are from left to right `HistColour`, `HistIntensity`, `StructIHS`, `HistQuant` and `UnsersTexture`. The structure feature `StructIHS` may be the most suitable image description to align any image set according an inherent sequential structure. For more detailed descriptions of the sequences see table A.1. The features are defined in section 3.3.

4.2.1 Experiment 1: Image Alignment by a 1dSOM

The aim of the first experiment is the representation of a single image sequence by a discrete 1dSOM-chain. Therefore, a single myMondrian sequence constitutes the input set to 1dSOM. After the training, each picture is attached to the best matching node. This assignment should be a bidirectional unique mapping, satisfying the definition of a function. In order to enforce this, the number of nodes must be at least the number of images in the sequence, e.g. $N_n \geq N$, in fact, $N_n \approx 2N$ is chosen. For each feature an individual 1dSOM is used. With this first approach 1dSOM can be analysed as well as the features since it can be estimated whether a sequence can be ordered correctly in a specific feature space.

Evaluation

For an automatic evaluation of the alignment along a 1dSOM trajectory the computed order is compared with the original order of a defined testset. A trained 1dSOM does not have any particular orientation. Only the topological relations are preserved. In order to evaluate the ordering the number of correct neighbour-pairs along the trajectory is measured. The *o*-measure relates this number to the number of neighbour-pairs in the

original sequence:

$$\begin{aligned} o(A, \mathbf{S}) &= \frac{\#(\text{correct pairs along the 1dSOM } A)}{\#(\text{pairs in the sequence } \mathbf{S})} \\ &= \frac{\#((\mathbf{x}_i, \mathbf{x}_j) \mid \; ||a^*(\mathbf{x}_i) - a^*(\mathbf{x}_j)|| = 1)}{N-1} \end{aligned} \tag{4.4}$$

where $\#(\xi)$ counts the objects of a set ξ and $a^*(\mathbf{x}_i)$ is the best matching node of image $\mathbf{x}_i$, N is the number of images in the sequence. $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{S}$ are neigboured images in the original sequence.

All o-measure values are collected in table B.2. In figure 4.1 these values are documented for the selected features and the various sequence considered in this study (see table A.1):

- The structure features (in figure 4.1 represented by `StructIHS`) supports the best alignment performance. This is obvious since spatial movements are detected as well as colour changes. Especially the moving of a rectangle is aligned best by the structure features. Only the colour changing sequences are not arranged appropriately.
- The texture feature (`UnsersTexture`) is not suitable to align the given sequences. Just the growing of a rectangle and the colour changes are detectable roughly since here grey value changes along vertical and horizontal directions are given.
- The colour histograms can detect the changes caused by the growing rectangle and the rectangle in front of a textured background.
- Images with a textured background can be aligned more than the sequences with a plain background. The textured background supports the detection of spatial changes.

The visual inspection of the aligned image sets (see figure 4.2) can explain the quantitative observations based on the o-measure. Additional qualitative observations are supported:

- Similar images are arranged successively, although they may appear at different sections of the original sequence. 1dSOM aligned these images differently from the original order (see figure 4.2 – structure result). This is caused by circular movements, see the green rectangle in sequence 1.
- Based on colour histogram features 1dSOMs result in groupings of pictures with overlapping rectangle and pictures without any overlap (see figure 4.2 – histogram results). Global features like colour histograms cannot detect every spatial change but describe variations of the overall colour distribution.
- The used texture feature is not suitable to align sequences with one moving rectangle. Just image pairs where the successive pictures do not show any change in the grey-level along rectangle borders may be possible to detect by matching the same node.

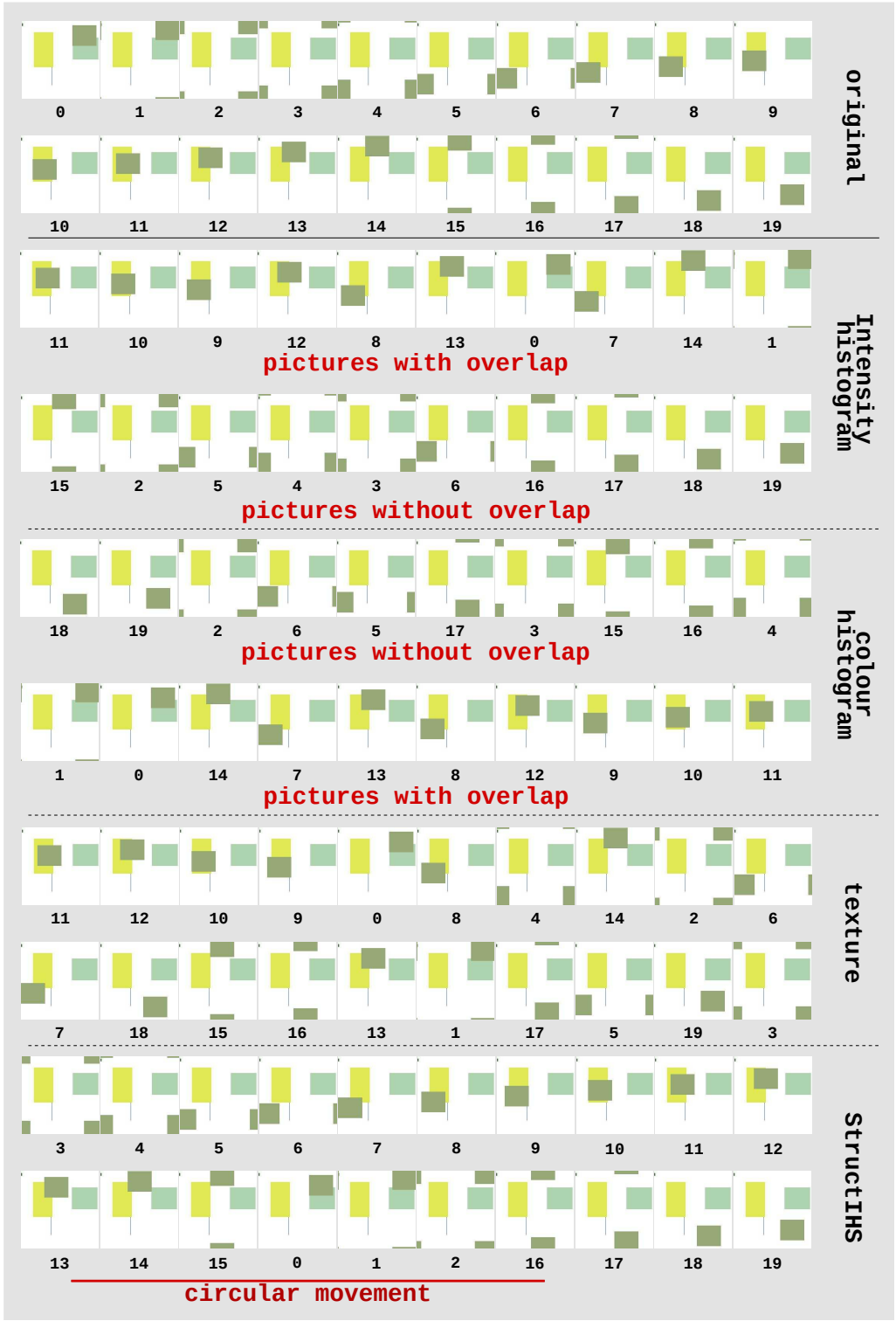


Figure 4.2: MyMondrian sequence example with one moving rectangle. Some 1dSOM-alignment results based on the named features are presented. The numbers indicates the image positions in the original sequence. For visualisation purpose each sequence is presented in two lines.

4.2.2 Experiment 2: Sequences Classification by a 1dSOM

The retrieval of image sequences faces the challenge of gaps in the sequences. In movies such breaks are caused by the cuts between two settings. Each setting constitutes a single data set interesting for further analyses. Hence the grouping of an entire image set into such subsets is an important task.

To prove the hypothesis that 1dSOM can demix different sequences, all image sequences of the myMondrian set are combined into one training set. Since each node should collect images of one sequence at least as many nodes as given sequences should be used. The sequences are described in table A.1. The feature set contains the most suitable one according to experiment presented above. Two different 1dSOM settings are used with different numbers of nodes.

Each picture is assigned to the most similar 1dSOM node of the trained 1dSOM. Then each node is labelled by the predominant sequence based on the matching images:

$$\Phi(\mathbf{w}_i) \;=\; \arg\max_{\mathbf{S}} \#^{+}(\mathbf{w}_i, \mathbf{S}) \tag{4.5}$$

where

$$\#^{+}(\mathbf{w}_i, \mathbf{S}) \;=\; \#(\mathbf{w}_i, \mathbf{S}) \;+\; \eta\, \#(\mathbf{w}_{i-1}, \mathbf{S}) \;+\; \eta\, \#(\mathbf{w}_{i+1}, \mathbf{S}) \tag{4.6}$$

$\#(.)$ counts the elements in a set, $\mathbf{S}$ is a sequence, $\Phi(\mathbf{w})$ is the label of the SOM-node and therewith a group label for all images matching this node (see introduction of groupings on page 29):

$$\Phi(\mathbf{x}) \;=\; \Phi(\mathbf{w}), \text{ with } \mathbf{w} \;=\; \arg\min_{\mathbf{w}_a \in A} d(\mathbf{x}, \mathbf{w}_a) \tag{4.7}$$

Images matching neighbouring nodes are taken into account by $\#(\mathbf{w}_{i-1}, \mathbf{S})$ and $\#(\mathbf{w}_{i+1}, \mathbf{S})$, whereas $\eta \;\leq\; 1$ adjusts the influence of the neighbourhood. Subsequently the images are labelled according to the label of their matching node.

Evaluation

The approach is evaluated quantitatively by the sequence dependent rate of correct labels resulting from the 1dSOM based separation:

$$\mathfrak{r}(\Phi, \mathbf{S}) \;=\; \frac{\#(\mathbf{x} \mid \Phi(\mathbf{x}) \;=\; \psi(\mathbf{x}) \;,\; \mathbf{x} \in \mathbf{S})}{\#(\mathbf{x} \mid \mathbf{x} \in \mathbf{S})} \tag{4.8}$$

where $\Phi(\mathbf{x})$ is the sequence label of a data point $\mathbf{x}$ according to the 1dSOM computed sequence. $\psi(\mathbf{x})$ is the original sequence label of the data point $\mathbf{x}$. Usually the sequence $\mathbf{S}$ with $\mathbf{x} \in \mathbf{S}$ dermines the label. This means $\psi(\mathbf{x}) \;=\; \mathbf{S}$ is the correct categorisation result.

The computed values are presented in figure 4.4. Viewing these boxplots results in the following observations:

- The separation rates are greater for all sequence types if more SOM-nodes are used. Obviously the separation ability of the 1dSOM increases by a longer 1dSOM chain.
- Sequences with textured backgrounds are separated better than the other ones. This corresponds to the observations in the above experiment and is caused by the dominance of the background.

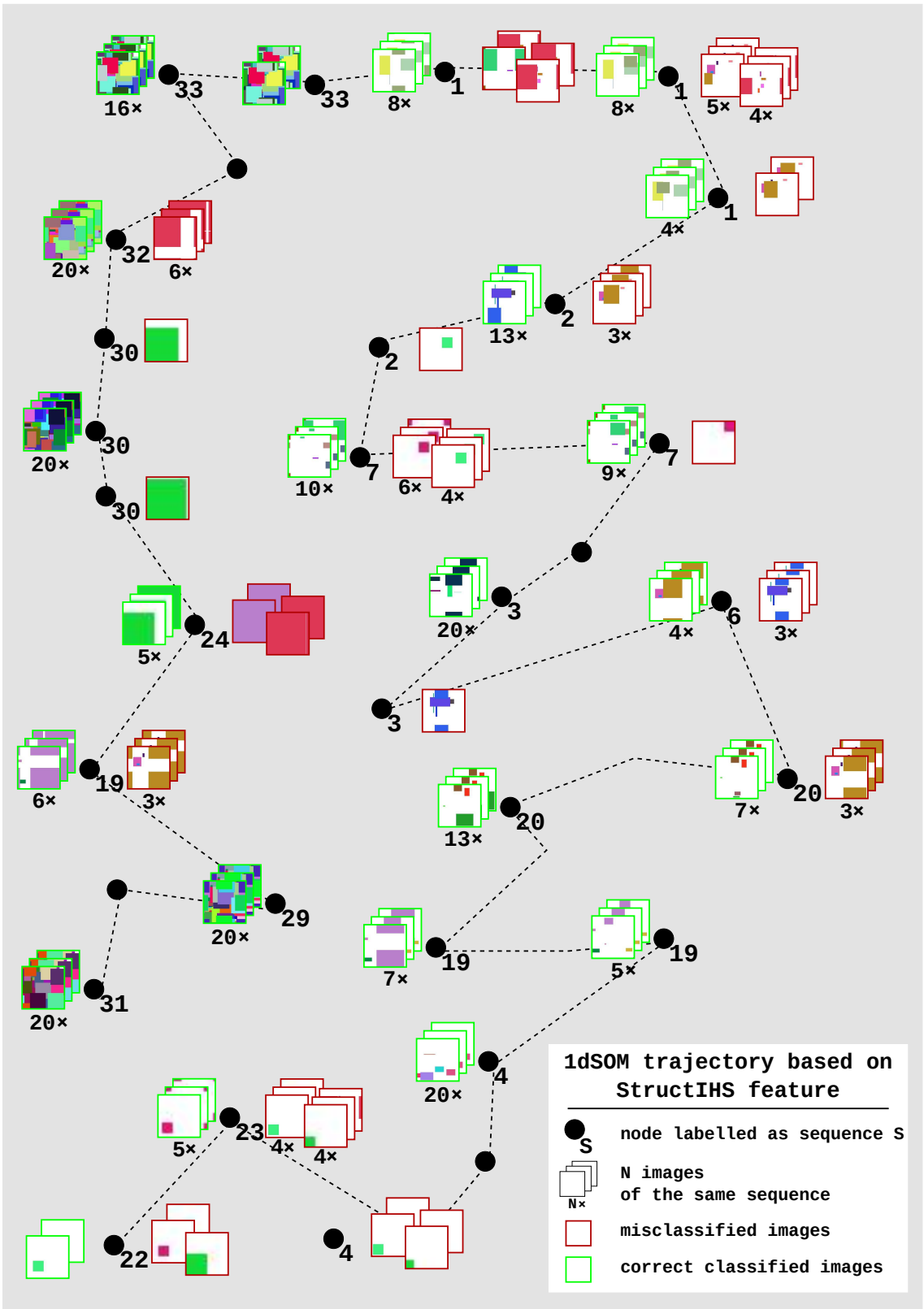


Figure 4.3: Selected nodes of the 1dSOM ($N_n = 50$) separation. Used data is the structure-feature (section 3.3) of the myMondrian images sequences (table A.1).
Pictures labelled correctly are bordered green and placed left of the matching node, pictures labelled incorrectly are bordered red and placed on the right side. The 1dSOM structure is given along the dashed line without any further topological information.

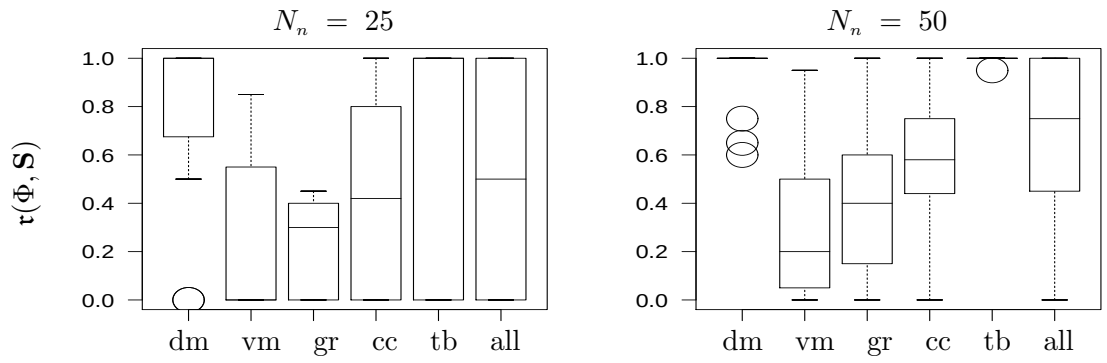


Figure 4.4: Domain dependent separation rates $\mathfrak{r}(\Phi, \mathbf{S})$ for the various sequence types in two different feature dependent 1dSOMs. Used features are colour, structur and texture to compute individual 1dSOMs. In each plot the sequence types are from left to right: defined move (dm), variable move (vm), growing rectangle (gr), colour change (cc), textured background (tb) and all sequences (all).

For a qualitative evaluation, an extract of a trained 1dSOM is presented in figure 4.3. The following observations are proven as well as some assumptions based on the separation rate $\mathfrak{r}(\Phi, \mathbf{S})$:

- Some domains get lost during the node labelling step. This occurs when the images of one sequence are spread over a number of nodes. Thus no node gets the label of this sequence. The classification of images as a member of these sequences is not possible.
- At one end of the presented 1dSOM, those images match which are coloured almost completely. Images with a lot of background match at the other end of the chain. Sequences with a growing rectangle are spread over the nodes regarding the increasing/decreasing rectangle area.
- The sequences with textured background are grouped nearly perfectly. In most of the cases no other image matches the same nodes as images with a coloured background. This supports the observation that the coloured background dominates the alignment.

4.2.3 Experiment 3: Real World Image Alignment by 1dSOMs

Given synthetic myMondrian images, 1dSOMs seem to be suitable to align pictures of a sequence in the original order as well as to separate different image sets. This motivates the application to real-world images. Therefore, a webcam image set of an aquarium observation and the mixed artexplosion photo collection are used in the following.

Piranha Aquarium

The first data set used to analyse 1dSOMs with respect to real world data are pictures of an aquarium with piranhas[1]. Since these images are collected from a webcam they contain a time stamp. Thus they present a ground truth set. Thus the 1dSOM approach is analysed according to real world pictures with known alignment.

Furthermore camera based observations are an important approach in behaviour research and aquariums are a frequent observation setup. Here 100 pictures with moving

[1]Provided by the piranha-webcam of the Natural History Museum, Fribourg, Switzerland. Since summer 2005 this webcam no longer available.

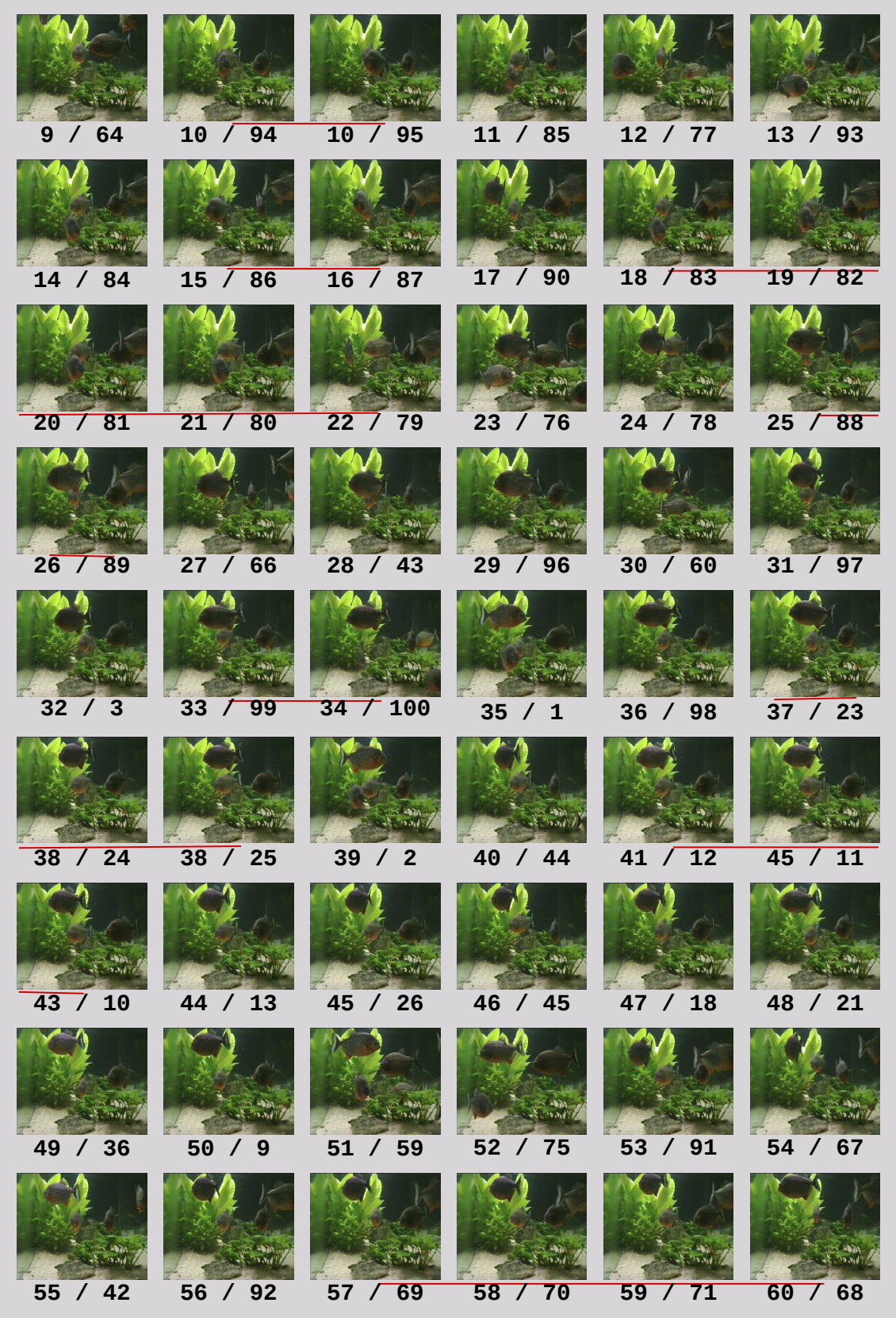


Figure 4.5: 1dSOM alignment of Fribourg piranhas based on the structure feature. The first number below each picture indicates the position based on the 1dSOM, the second one the position in the original sequence. The red bars heighlight correct orders.

	o-measure	
	1dSOM	**similarity**
colour	0.19	0.07
structure	0.20	0.03
texture	0.08	0.00

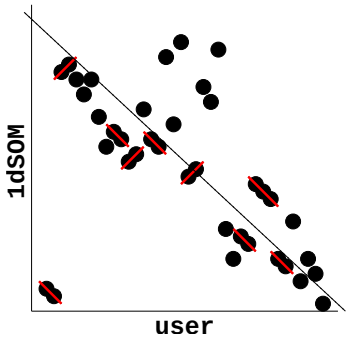


Figure 4.6: 1dSOM alignment of the Piranha Aquarium pictures. Left: The 1dSOM alignment outperforms the similarity based arrangement. Right: A comparison of the 1dSOM in the structure feature space with the user based ordering offers a rough orientation by the sloping axis. Correct neighbours along the 1dSOM are tagged.

objects, namely piranhas, are taken from a fixed camera position. This alignment task is rather challenging. Especially the sorting by hand is a difficult and tedious. Hence an automated sorting is desired. An extract of the aligned data set is presented in figure 4.5.

The evaluation is based on the *o*-measure (see equation 4.4) to get a quantitative measure for the correct alignment. To rate this value a further order is computed based on a similarity search. Therefore the first picture of the sequence is used as a query image. The result list determines the sequential order. The results are listed in figure 4.6.

As a human based ground truth a subset of the piranha pictures (40 images) is ordered by a user. The number of pictures is reduced to get a manageable set. The time to arrange the images was restricted to 15 minutes. This alignment is compared with the original order. A *o*-measure of 0.2 is obtained here. This corresponds to the highest value of the automated alignments (see figure 4.6 – structure feature). However, the 1dSOM approach has sorted 100 images. Rated by the a priori probability of the right order this means 1dSOM performs better. Furthermore the automated approach was faster than the human subject.

In order to rate the computed ordering according the human based ground truth in figure 4.6 these alignments are compared. The 1dSOM based arrangement is computed based on the structure feature. Along the axes the position in the sequence is plotted. A number of correct neighbourhood pairs is observed and a rough alignment is given by an axis from the top left to the bottom right. Using the user ordering as the ground truth increases the *o*-measure of 1dSOM to 0.27.

Although not resounding, the alignment based on 1dSOM is noticeably better than based on the similarity search. The inspection of the used data set and the result visualisation in figure 4.5 explains the moderate performance. The movement of the piranhas is rather excursive, e.g. they swim back and forth. However, 1dSOM outperforms the similarity search.

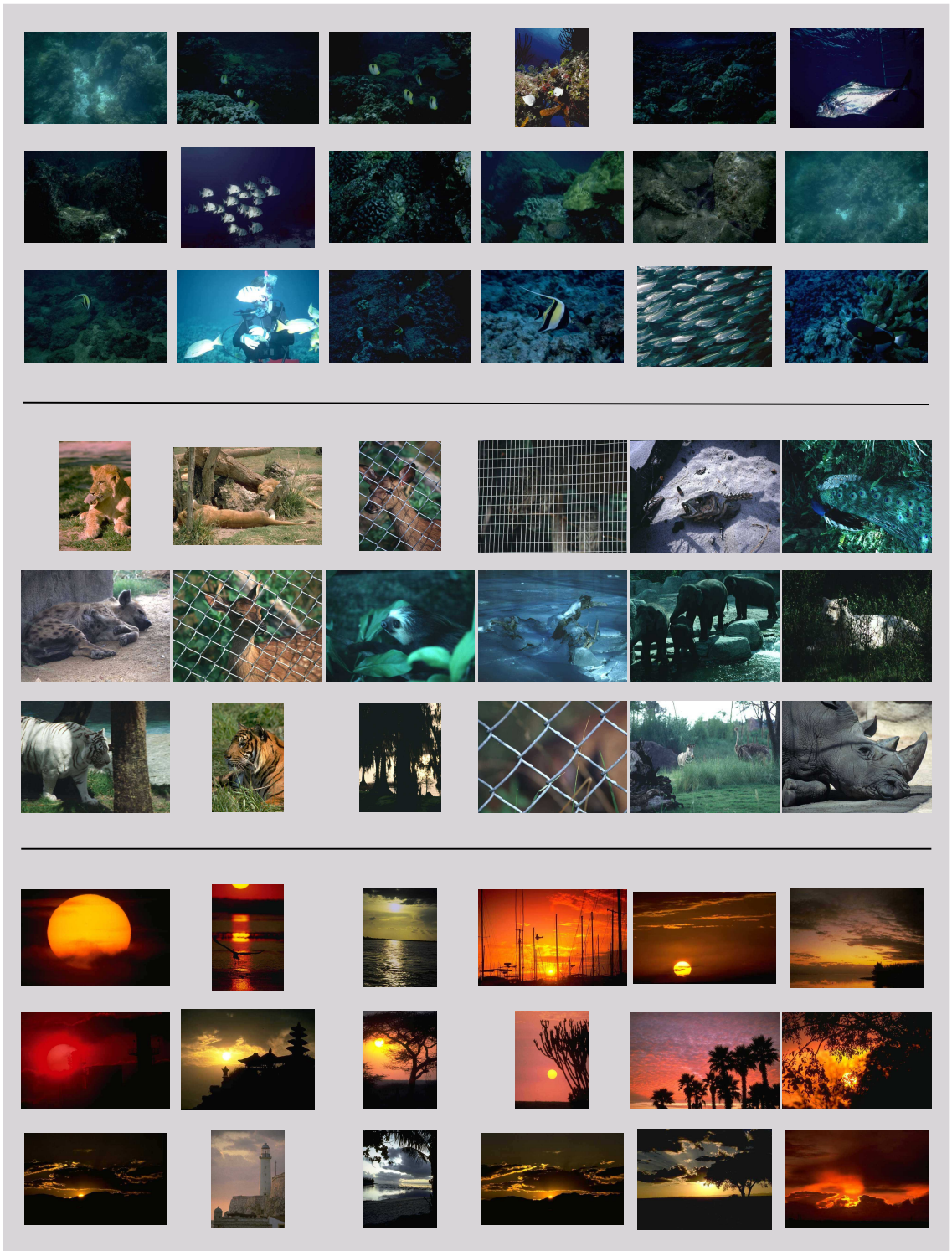


Figure 4.7: 1dSOM alignments of particular artexplosion categories based on the structure feature. For each category (from top to bottom: underthesea, animals and sunrisesunset) a single 1dSOM is used. Segments of the 1dSOMs are presented. Each 1dSOM has as many nodes as pictures given in the category ($N_n = 300$).

	separation rate $\mathfrak{r}$			
	$\mathbf{N}_n = 100$	$\mathbf{N}_n = 30$	$\mathbf{N}_n = 500$	**classifier**
underthesea	0.56	0.38	0.67	0.06
animals	0.5	0.47	0.64	0.28
doorswindows	0.37	0.37	0.59	0.31
teddybears	0.06	0	0.3	0
sunrisesunset	0.73	0.64	0.77	0.63
venezuela	0	0	0.25	0
iceland	0.6	0.72	0.52	0

Table 4.1: 1dSOM based separation of artexplosion images: evaluation by $\mathfrak{r}$-rate, see section 4.2.2. The used feature is the structure feature. As a reference a Naive Bayes Classifier with a PCA based preprocessing is used.

Artexplosion Photos

People present their unordered photo collection one by one in a slide show. Therefore the photos have to be arranged sequentially. This can be done by 1dSOM. Then the images are presented to the user in the resulting order. A concrete sequential connection between succeeding pictures is not expected. Indeed the image searches can be supported by an appropriate sequential order. Furthermore pictures of the same category should be presented successively. To analyse this the grouping task is tested. The used feature is `StructIHS` (see section 3.3). In the previous experiments it has been observed to be suitable to arrange pictures. The separation and alignment is analysed according to different N_n.

Obviously a longer 1dSOM is more suitable to separate the different image categories and 1dSOM performs better than a common classifier (see table 4.1). Furthermore, 1dSOM outperforms the classifier since it supports the sequential browsing. The presentation of a large unstructured set which is the result of a classifier is not very user friendly.

Inspecting the alignment along 1dSOM (figure 4.8 presents a section of the 500 node 1dSOM) results in the following observations:

- The subsets are spread over the whole 1dSOM. Neighbouring nodes match different categories.
- A sequential content of the images is observed. For example, the whales swim, jump and dive into water again. But a definite sequential structure is not given.

  This obeservation can be confirmed regarding category-specific 1dSOMs (see figure 4.7). In the underthesea- and the animals-set somehow similar images are neighbours. Nevertheless, no semantic sequence is given. Examining the sunrisesunset-category the course of the sun can be observed. However the sun rise and sink repeatedly along along 1dSOM.
- The computed groupings do not correspond to the a priori given categories but are explainable based on the used feature. However, the presented groupings may be interesting for the user, maybe more relevant than the semantic based categories predefined by the image collection. For example, the aquarium pictures of the animals-category are mixed with underthesea pictures.

Figure 4.8: 1dSOM alignment of the artexplosion photo set based on the structure feature: The presented section of the 1dSOM ($N_n = 500$) should be read line-by-line from left to right. The coloured boxes specify the category-label of the image.

4.3 Summary of 1dSOM Analyses

1dSOM is analysed regarding two tasks – the automated alignment and the categorisation of images. Both applications are performed on synthetic sequences as well as on real world sequences.

Regarding the alignment, the implemented 1dSOM approach is suitable to detect a sequential order in a data set. Generally it exceeds a similarity search based on an example image. However, more detailed user experiments are necessary to verify the predominance of 1dSOM.

Usually the number of used SOM-nodes is rather high. In most cases it exceeds the number of pictures to align. This over-fitting is accepted for aligning the images of one sequence, since it does not cause any problems while the image set remains unchanged. However, new pictures of the same sequence may pop up. Than the generalisation ability becomes important.

The separation of image sets to meaningful subsets has been rather successful, as well. Although not being satisfactory, it is a good starting point for the retrieval of image sequences. Hierarchical 1dSOMs may separate the whole sequence, e.g. a movie, into the different shots on the top level. A refinement within the individual voronoi cells will align the images of a shot into the right order. This may be interesting for retrieving image sequences.

In several experiments the assumption of data-dependent suitability is proved for different feature detection algorithms by analysing the 1dSOM alignments. Indeed combinations of different features may enhance the performance. Probably the weighted combination in the distance computation, as done in many image retrieval approaches, may be promising. A modification of the used distance function and the SOM-learning algorithm would be interesting. The strategy to use different features and combine them on a higher level is proven.

For further enhancements one dimensional versions of the advanced self-organising approaches presented in the introduction of this chapter, may be interesting.

Chapter 5

Adaptation to the User: Relevance Feedback

Common algorithms to represent or compare images do not match the human perception and interpretation. To narrow this semantic gap is an important task in information and image retrieval. Therefore, the systems have to adapt to the user. The other way round users have to teach the retrieval systems. Such procedures are known by the term *relevance feedback*.

Numerous approaches to implement relevance feedback in image retrieval feature one common attribute: They offer a set of labelled images. Usually the user rates pictures retrieved by a system as relevant and non-relevant. Such labelled sets can be used to compute a representation adapted to the user's intention. They present training sets for machine learning setups.

One approach to compute prevailing attributes of a data set is the Independent Component Analysis (ICA). With respect to image retrieval tasks this is applied in two frameworks: Data spaces are transformed based on ICA. This should offer a representation more suitable appropriate to a specific retrieval session. Moreover, ICA is included in a classification approach to enhance category searches.

This chapter gives an overview over different relevance feedback aspects and approaches. Then ICA is introduced and applied for data space transformations as well as in a classification framework. Finally ICA is analysed based on the given image data.

5.1 Relevance Feedback

The most important factor within an image retrieval task is the human user. He should be satisfied by the performance of the system. Unfortunately, men perceive and interpret images diversely, depending on their intention, experience, circumstance and temper or even unmotivated on a transient view. On the other hand CBIR-systems depend on defined mathematical descriptions of data spaces and retrieval algorithms. Thus different ways of dealing with image data arise: While men compare images based on their semantic interpretation, the system's responses result from numerical values. A wide gap between the high-level semantic concepts and the low-level features is observed and known as the *semantic gap*.

Since computer scientists usually should not modulate people, they have to tune the

systems to resemble the user and narrow the semantic gap. To get a mathematical description the human search behaviour has been researched for decades. Especially psychologists tried to model the human way of comparing images. Common models are based on discriminant learning and multidimensional scaling [Wyckoff, 1952] [Cowan, 1968]. Although still very popular these approaches suffer from the unpredictable user. Therefore, a general and a priori implementation of human interpretations and similarity ratings is not possible. CBIR-systems should approximate the user's behaviour. They should adapt to different users and different search tasks based on ratings of preceding retrieval sessions. The systems are trained to simulate a human like perception.

Since the subjective user level and the technical system level have to be combined, relevance feedback is based on two stages: At first some ratings of the retrieved images with respect to the query $\mathbf{q}$ are collected from the user. This rating can be defined as a tuple

$$\Gamma \;=\; (\mathbf{x}_i, \gamma)_{\mathbf{q}} \tag{5.1}$$

where $\mathbf{x}_i$ is an individual image and γ the user rate of this image.

Then the performance of the CBIR-system is modified based on the ratings. Therefore, a lot of different implementations are established. In common information and text retrieval two basic techniques of relevance feedback exist – *query expansion* and *term re-weighting* [J.J. Rocchio, 1971].

In the MARS framework [Rui et al., 1998] these approaches are transformed for image retrieval straightly. A relevance feedback approach is proposed that adapts a set of weights. These weights are part of a multimedia object model and determine the impact of the different features to the similarity value. Based on a heuristic model positive and negative ratings influence the outcome of the retrieval in single feature spaces.

The second relevance feedback technique implemented in MARS is the query-vector-movement. Analogously to the well-known approach of the early SMART Information Retrieval system [J.J. Rocchio, 1971] the intention is to move the query $\mathbf{q}$ towards relevant objects and away from non-relevant ones [Baeza-Yates and Ribeiro-Neto, 1999]. The computation of the new query vector $\mathbf{q}'$ bases on vector additions:

$$\mathbf{q}' \;=\; \eta\,\mathbf{q} \;+\; \gamma\,\mu^{+} \;-\; \beta\,\mu^{-} \tag{5.2}$$

where μ^{+} is the mean of the relevant rated images and μ^{-} the mean of the rejected ones, respectively. The INDI-system (see section 2.2.4) incorporates analogue relevance feedback approaches.

Similar to the query-vector-movement approach [Rui et al., 1997b][Kämpfe et al., 2002] [Baeza-Yates and Ribeiro-Neto, 1999] is the feature space warping proposed in [Bang and Chen, 2002]. Instead of the query the prototype vectors of rated images are moved in the feature space. Relevant objects are moved to the query and non-relevant ones are shifted away. The result will be a clustering based on the performed search sessions. The datapoints are shifted while the space remains unchanged. This will cause a changed database and a kind of long-term learning is performed in this way.

This basic framework to adapt the system's performance to the user's intentions offers various critical tasks:

- In which way does the user give his rating?
- How to use the rating?

- How to implement the similarity computation?
- Should the feature vectors be fix?
- Should the system remember preceding search sessions?

These questions are discussed in the following.

5.1.1 The User Rating

On the technical level an extensive and detailed rating of the retrieval results would be desired to adapt the system based on a broad data set. On the other hand users are lazy. Naturally the rejection of all non-relevant images will be done. Most of the users will avoid rating large sets of pictures by a number of detailed levels. Therefore, some CBIR-systems ask to select the images the user prefer, e.g. PicSOM (section 2.2.1). Regarding the MARS-system a tradeoff is suggested [Rui et al., 1997b]: Five levels (absolutely non-relevant, non-relevant, undecided, relevant, very relevant) for a restricted number of images are offered for the relevance feedback rating.

Depending on the used approaches positive and negative ratings or just the positive ratings are used. Oositive ratings improve the retrieval result predominantly in the first iteration [Franco et al., 2004]. In the next steps positive ratings have less influence since usually the retrieved images satisfy the adapted parameters quite good. Significant changes in the parameter set will not occur. Hence, in these steps negative ratings may be important to filter the non-relevant images and tune the parameters to reject this ones. Unfortunately, in many situations the number of non-relevant images outperforms the number of relevant images and destroys the success of the preceding iterations.

Motivated by the observation that positive feedback causes improvements only in the first iteration, in [Kherfi et al., 2002] a splitted implementation is proposed: Based on positive ratings after the initial retrieval step the data space is clustered to detect all images containing the desired features. In the following iterations the negative feedback is exploited to refine the result set. In this step images containing undesired features are rejected. Therewith the problem of common features, which possess the relevant as well as the non-relevant images, is considered.

5.1.2 Interestingness and Relevance of Images

The most user-friendly search task is the retrieval of *interesting* and *relevant* images. Some marginal differences between interesting and relevant may exist, nevertheless in the following just the term *interesting* is used to embrace both. *Relevant* may be interpreted as *interesting with respect to the desired query*. Nevertheless, the definition of these predicates is fuzzy on the user level, while the system level requires a determined specification [Mizzaro, 1997]. Summarised the user rating of an image as interesting depends on a number of different factors:

- Search task: Obviously the desired image determines the relevance of all retrieved images, since this is the object of the image retrieval.
- Context: Depending on the usage different images are interesting. For example a picture of a horse as the query image could represent a farm as well as a race-course. In the first situation images with animals may be relevant while in the second a picture with a lady wearing a large hat could be interesting. In general, the acceptability of any image can only be judged in context of applications or concrete tasks.

- The user: People are different. Therefore, they rate images differently. Especially the experience of the user determines his rating of the retrieved images.
- Preceding search tasks: Each performed retrieval session with a CBIR-system influences the user. Based on the systems performance in preceding sessions he learns how the system acts and which pictures are in the collection. Hence, he adapts his ratings to these observations.
- Data quality: Usually people are looking for pictures with a specified content. Furthermore, the quality of the image affects the interestingness and images with the same content may be less interesting if the quality is bad.
- Unspecific things: Human behaviour is unpredictable. An arbitrary effect can cause an unexpected rating. Thus all user models depend on probability considerations.

Nevertheless, automatic approaches to retrieve interesting objects from data collections are required in image retrieval. Intuitively the desired pictures should get a label *interesting* and the CBIR-system should select them by this predicate.

Obviously such an approach resembles a classification task. In [Santini and Jain, 1996] an implementation of such a procedure based on a *Fuzzy Feature Contrast* model is proposed. Here the truth value of a predicate applied on an image is defined as the is-element-relation of the image to the set of all images showing this predicate. This concept motivates an unsophisticated approach to rate interesting images based on appropriate sets. Detecting such sets of interesting images will be the task to solve. This resembles the categorisation challenge mentioned in section 3.1. Again this cannot be performed in advance but has to be integrated in the process interactively.

A further approach based on relevance scores attached to each picture is proposed in [Giacinto and Roli, 2004]. Each image gets a relevance score based on the scores of his nearest neighbours. This resembles a nearest neighbour search based on a query-by-example since a similarity computation is required to detect the pictures determining the relevance score of the single pictures. Relevance is a local property. Relevant images can be retrieved from different unconnected regions of the data space.

One important challenge regarding nearest neighbour searches is the computation time. Wu and Manjunath [2001] propose an efficient computation of nearest neighbours by changing metrics. The retrieval acts in two steps: A broad filtering of possible candidates based on similarity and an improved feature filtering in the results of the first step. Since the first step is more time consuming it's implementation is modified. An adaptive search strategy based on upper bounds of the similarity improves and accelerates the filtering of similar images.

5.1.3 Similarity Models

Most of the CBIR-systems are based on a similarity search initiated by a query-by-example. These approaches support relevance feedback by a query refinement as well as by adaptable similarity computation. Hence, the similarity computation is the most important step within query-by-example systems. Usually it is based on common metrics like the Euclidean distance. Proposals to use multidimensional scaling and similarity definitions by metric-based distance measures occur [Beals et al., 1968] [Micko and Fischer, 1970]. A lot of psychological studies have observed that the human way of comparing images does

not fulfil the common metric axioms. The validity of self-similarity, minimality, symmetry and triangle inequality is at least questionable [Santini and Jain, 1996].

In order to analyse this, a formalism is stated that regards mathematical distance functions (perceived similarity) $d(\mathbf{x}_i, \mathbf{x}_j)$ as well as the human cognition (judged similarity) $\delta(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are images or stimuli. A generalisation function $g(.)$ describes the correlation between mathematical description and human recognition. Since various psychological studies have confirmed that the judged difference is a linear function of the judged similarity by a slope of -1 [Tversky, 1977] this discussion of similarity models does not distinguish between similarity and distance.

In early works $g(.)$ is assumed to be a suitable monotonically non-decreasing function [Santini and Jain, 1996]. Indeed the discrepancy between the characters of perceived and judged similarity becomes obvious. Therefore, Thurstone and Shepard (see [Santini and Jain, 1996], page 5 and following) modified the similarity model and analysed the generalisation function $g(.)$. Based on some probability theoretic considerations they analysed exponential as well as Gaussian like generalisation functions. However, this model suffers from the assumption, that similar stimuli always cause similar or equal responses from the user. This does not solve the discrepancy between human cognition and mathematical models but shift the problem from the similarity discussion to the response definition.

Nevertheless, the generalisation function $g(.)$ may be a good starting point for relevance feedback considerations. Since relevance feedback should approximate the perceived similarity to the judged similarity the generalisation function has to be learned based on the user ratings. The goal of relevance feedback can be described as:

> Given a function $d(\mathbf{x}_i, \mathbf{x}_j)$ to describe the perceived similarity by a geometrical or mathematical model and a function $\delta(\mathbf{x}_i, \mathbf{x}_j)$ that satisfies the human similarity judgment, the goal of relevance feedback may be to find a generalisation function $g(.)$ that achieves
>
> $$\delta(\mathbf{x}_i, \mathbf{x}_j) \;=\; g(d(\mathbf{x}_i, \mathbf{x}_j)) \tag{5.3}$$

Another approach to define similarity models is based on the description of images by the presence or absence of concrete predicates. The result is a parametrised set-theoretic distance function [Santini and Jain, 1996] based on user ratings like “X is more P than Y”, where X and Y are objects of the same type (e.g. images) and P is a predicate describing such objects. The parameters of this distance function should be learned based on the user rating. Such set-theoretic similarity models conflict with the common geometric distance models. The implementation of distances in the Riemann space offers a geometrical interpretation of the set-theoretic similarity models [Santini and Jain, 1996].

5.1.4 Adaptable Features

The selection of suitable feature algorithms is an important task. Although dependent on the given retrieval situation and the particular user this is usually a preprocessing step and not influenced by any interaction. This may be appropriate for global low-level image features like colour. Nevertheless, even these features do not coincide with the human perception. Furthermore, local features are desired to detect relevant objects within the images. For example in the retrieval of a specific car the background may be absolutely non-relevant. Unfortunately the problem of image segmentation or object detection is

not solved in general. To handle this rough and unsatisfying segmentation results can be implemented adaptable. Therewith a reasonable basis for local image descriptors is given.

This attends the perceived image content. Men rate images predominantly by their semantic [Eidenberger, 2004]. The goal is to find semantic features, but as mentioned above the automated detection of semantic features is still unsolved. Hence, the high-level visual phenomena are described by high-level statistics [Qiu et al., 2003]. Optimised features [Wood et al., 1998] are defined and adapted to extract *global* properties pertaining to relevant images [Giacinto and Roli, 2004].

Another task of relevance feedback is the selection of suitable and unsuitable features concerning a specific search task [Cowan, 1968]. The example in [Kherfi et al., 2002] shows that the retrieval of images containing a red car may depend on a shape feature (the predicate *shape of a car*) and a colour feature (the predicate *red*). Therefore these two features would be suitable to execute this query. Nevertheless, a red car has to be distinguished from a blue car. If this image is rated as non-relevant by the user the predicate *blue* is rated as unsuitable with respect to this query. That's correct. Nevertheless, the predicate *shape of a car* would be rated as unsuitable as well. This conflicts with the desired query. Obviously the rating of features as suitable and unsuitable based on user-ratings is not trivial.

5.1.5 Short-term and Long-term Relevance Feedback

A further character of human behaviour influencing relevance feedback is the lack of consistency: Regarding the same retrieval task a user may act today in one way and tomorrow in another way. Such an unpredictable behaviour complicates the question how long a CBIR-system should remember the learned performance or former ratings. Two relevance feedback strategies emerge from this considerations:

(1) An intra-query or short-term learning is performed during each search session separately. Assumed a parameter set to adapt, this means that each search session starts with the same initial parameters. Ratings and adaptations of earlier sessions are irrelevant. Thus the system has to be trained for a repeated search session again and again. The adaptation to the user's need and the simulation of human behaviour is limited to the achievements feasible in a few learning iterations.

(2) To lessen these drawbacks the adaptation process can incorporate experiences of preceding search session. These can be user ratings of various retrieval results or adapted parameters of the retrieval algorithms. This inter-query or long-term learning has to be tuned carefully. Performing a lot of learning steps based on a small set of ratings or search examples will result in an over-fitted system. The retrieval of new pictures will become improbable.

The approach presented in [Wood et al., 1998] combines on-line and off-line training for relevance feedback. It performs three steps: In a preprocessing step the images are partitioned and the features are computed. The second step is the query. Relevance feedback is performed as a LVQ (Linear Vector Quantisation) clustering based on positive and negative ratings. In the query phase new examples can be inserted to avoid local minima and facilitate a double clustering. This is motivated by the observation that

relevant images are spread in the data space and will generate different clusters, they are not concentrated in one cluster. The third step is executed between arbitrary queries. In this off-line step an RBF-network is trained to get a library of object classes. Based on these objects advanced queries are supported.

Another CBIR-system that combines on-line and off-line learning is the PicSOM framework (see section 2.2.1). During the retrieval sessions the relevance labelling of the pictures is adapted based on the convolution of the Self-Organising Maps [Koskela, 2003]. In order to take advantage of preceding retrieval sessions a *relevance feature* is computed. Motivated from the vector space model to describe textual documents features common for relevant rated pictures are collected in a term-by-document matrix. The dimensionality of this matrix is reduced based on latent semantic indexing.

5.1.6 Relevance Feedback as an Optimisation Problem

The definition of relevance feedback can be diminished to an optimisation problem. This is proposed straight forward in [Vasconcelos, 2004] by implementing image retrieval as the minimisation of the retrieval error. Therefore they use a Bayes Classifier and postulate to minimise the density estimation error. A similar approach will be presented in section 5.3.

[Rui and Huang, 2000] verifies feature weighting and query-vector-movement by deriving an optimised learning for relevance feedback. As a conclusion neural networks and support vector machines (SVMs) are proposed. The task would be to minimise the difference between the human rating and the computational result. Relevance feedback is implemented as a supervised learning task. Unfortunately, the common optimisation approaches need labelled data sets greater than that ones available in image retrieval situations.

For example PicSOM (section 2.2.1) and the proposal in [Williamson, 2001] are based on self-organising topographical networks. While that are unsupervised methods the RBF learning of the query vector proposed in [Wood et al., 1998] is a supervised approach. In [Ko and Byun, 2002] a probabilistic neural network for multi-class learning is proposed to learn the link between high-level concepts and low-level features. The weights of each feature are used as the weights for the network.

The approach presented in [Carkacioglu and Vural, 2002] uses a neural network to perform a nonlinear data space transformation. The features are transformed in a similarity space. The cost function is based on the minimisation of the intra-class distances and the simultaneous maximisation of the inter-class distances. Therewith the approach to learn a data space transformation based on given user ratings is motivated. The user's need is approximated by representing the data in a space where the data distribution resembles the human perception. While Carkacioglu and Vural [2002] use a Multi Layer Perceptron (MLP) any optimisation technique is worth to analyse. For example in [Giacinto and Roli, 2004] and [Franco et al., 2004] relevance feedback is performed by a PCA based approach to compute different projection spaces. PCA is applied on different subsets namely sets of relevant objects and different clusters of non-relevant ones.

A further approach using PCA in relevance feedback steps to transform the data is presented in [Peng and Bhanu, 2001]. PCA is used to decorrelate the image features

before the feature weights are updated. For this purpose PCA is computed locally on the set of images already presented to the user. The neighbours of the query determine the principal components to transform the data before performing the next retrieval step. The transformation is applied on a subset of the entire image database. More data points are transformed than used to compute the principal components and their number is determined at the beginning of a search session. This causes a strict preselection of possibly interesting images by the initial nearest neighbour search step. Since just this selected data set is transformed during the different retrieval steps the remaining data stays in the original data space. A similarity computation based on the transformed query and the adapted feature weights is not valid for this data. New image data cannot be inserted. Most of the search session end in local minima.

Similar approaches are based on the *Independent Component Analysis* (ICA) to find a suitable transformation matrix. This will be introduced and analysed in the following sections.

5.2 Relevance Feedback Based on Independent Component Analysis

Relevance feedback by data space transformations assumes that a data space exists, that is more suitable to represent the given data. Such a data space may be defined by statistically independent directions. These can be computed by ICA.

5.2.1 Data Space Transformations

The user level of relevance feedback usually results in at least one set of data labelled by attributes representing the user's intention. Computationally the data should be grouped with respect to these labels. Therefore, on the system level relevance feedback can be implemented as a transformation of the data into a more user-friendly data space. Any axis within a data space is called *suitable data direction*, if the projection of the data onto this axis enhances the desired grouping.

Adapted from the definition that each image is relevant, if it is element of a somehow described set, a definition of relevance feedback based on data grouping is feasible. Let the given data $\mathbf{x} \in \mathbf{X}$ be grouped by an arbitrary grouping function $\Phi(\mathbf{x}) \in \Psi$, where $\Psi = \{\psi_i\}$ is a set of group labels with $\psi_i \in \{1, ..., N_c\}$. Usually the grouping is limited to a unique assertion $\Phi(\mathbf{x}) = \psi_i$, where $\Phi(\mathbf{x})$ may be a cluster algorithm, an automatic classification approach or a semantic mapping. For details regarding image groups see section 3.1. A rating algorithm $\Re(\Phi(\mathbf{x}))$ is defined to evaluate the computed groupings (e.g. classification rates) or the usability (e.g. computation time or user satisfaction).

Based on this, relevance feedback can be defined as any transformation

$$\mathfrak{F}(\mathbf{x}) \;=\; \mathbf{z}$$

which enhances the grouping. Therefore,

$$\Re(\Phi(\mathbf{z})) \;\geq\; \Re(\Phi(\mathbf{x}))$$

should be true. Figure 5.1 exemplifies this approach. The task is to find such a transformation and a definition of suitable data directions.

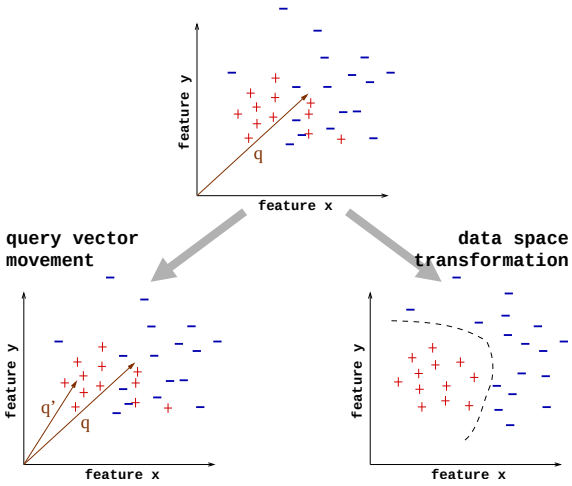


Figure 5.1: Two different relevance feedback approaches replacing a weight adaptation: The query-vector-movement (left) shifts the query closer to the center of relevant data points. The data space transformation (right) looks for a data representation that better separates the different groups.

The Principal Component Analysis (PCA) is one approach suitable to compute such transformations based on statistical attributes. It is a common method to find a reasonable linear transformation of a given data set $\mathbf{X}$. Applications are data analysis and compression. Based on a data set $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ – zero mean assumed – the symmetric covariance matrix is computed by

$$\mathbf{C} = \langle \mathbf{X}\mathbf{X}^T \rangle \tag{5.4}$$

Then its eigenvectors $\mathbf{e}_i$ and the corresponding eigenvalues λ_i are computed. In doing so

$$\mathbf{C}\mathbf{e}_i = \lambda_i \mathbf{e}_i \quad , \quad i = 1, ..., N \tag{5.5}$$

is true. Therewith the directions of largest variance are detected.

Let

$$\mathbf{E} = (\mathbf{e}_1, ..., \mathbf{e}_N)^T \tag{5.6}$$

be the matrix with the eigenvectors $\mathbf{e}_i$ as rows. Transforming a data vector $\mathbf{x}$ will be performed by

$$\mathbf{y} = \mathbf{E}\mathbf{x} \tag{5.7}$$

Thus $\mathbf{y}$ is a point in the data space defined by the eigenvectors. It is assumed that the eigenvectors are ordered regarding descending eigenvalues ($\lambda_1 > \lambda_2 > ... > \lambda_N$). Usually a subset of the eigenvectors is used, namely those eigenvectors related to the largest eigenvalues. $\mathbf{E}^K$ represents the matrix with those eigenvectors: $\mathbf{E}^K = (\mathbf{e}_1, ..., \mathbf{e}_K)^T$, $K \leq N$.

Using PCA for relevance feedback is determined as

$$\mathfrak{F}(\mathbf{x}) = \mathbf{E}^K \mathbf{x} \tag{5.8}$$

where $\mathbf{E}$ is based on the set of relevant rated images and $K \leq N$ determines the number of used eigenvectors.

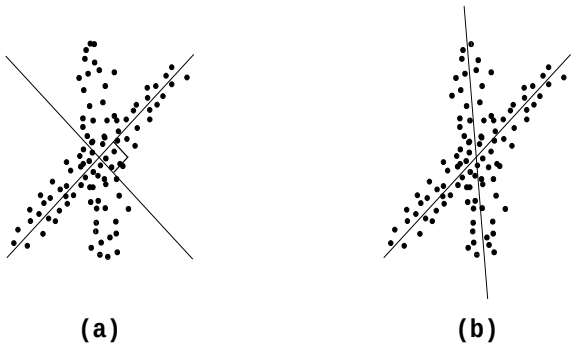


Figure 5.2: PCA detects orthogonal projection axes (a). Unfortunately the main axes of a data distribution may not be orthogonal. ICA detects statistically independent axes, which probably better represents the data (b).

PCA decorrelates the data and detects faithful representations. Second-order statistics describe the components and orthogonal directions are computed. Unfortunately the most important directions of a data distribution are not necessarily orthogonal. Approaches that detect directions representing higher-order characteristics of the data set may be preferable [Hyvärinen, 1999]. So do *Independent Component Analysis* (ICA). It offers an approach to estimate a linear transform based on almost statistically independent factors. These are not necessarily orthogonal to each other and can be more meaningful than the principal components (see figure 5.2).

Applied on an image set ICA will cover the distribution of all pictures in the collection. Contrary to that each image retrieval task desires just a subset. Therefore, ICA is analysed with respect to the characteristics of relevant labelled subsets. A category or subset dependent transformation will be performed. Based on that transformation matrix the entire data set is analysed regarding the differentiation of the individual subsets. Perhaps the distribution in the ICA transformed space is better for image retrieval in conjunction with the specific user.

5.2.2 ICA Theory and Algorithm

The first works concerning ICA were motivated by the well-known signal processing task *Blind Source Separation* (BSS) [Comon, 1994]. Based on heuristic observations it is assumed, that each observed signal $\mathbf{x}$ is a (weighted) combination of source signals $\mathbf{s}$. Since usually the original sources $\mathbf{s}$ are unknown but wanted an automated reconstruction from the observations is required. Such a demixing process is known as BSS. Figure 5.3 visualises this task.

In order to describe this model, the observed signals are assumed to be a linear combination of the sources. Thus a mixing matrix $\mathbf{A}$ represents the combination process and the task is to find the (pseudo-)inverse to demix the observations. $\mathbf{W}$ represents this demixing matrix. Most ICA approaches restrict $\mathbf{W}$ to a square matrix.

The basic assumption established in ICA is the independence of the source signals. Given that all components of the source vector $\mathbf{s}$ are statistically independent its probability density can be stated as a product of the marginal densities.

In general it is assumed that

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \eta \qquad \text{with } p(\mathbf{s}) = \prod_{i=1}^{N} p_i(\mathbf{s}_i) \text{ and } \eta \text{ is noise} \tag{5.9}$$

and ICA estimates a demixing matrix $\mathbf{W}$ to invert this linear transform:

$$\mathbf{u} = \mathbf{W}\mathbf{x} \tag{5.10}$$

with approximately $p(\mathbf{u}) = \prod p_i(\mathbf{u}_i)$ and $\mathbf{W}\mathbf{A} = 1$, $\mathbf{u}_i$ is row vector i of $\mathbf{u}$.

Related to this is the aim of Bell and Sejnowski [1995]. They intend to perform a redundancy reduction in a given data set and revise the effects of unknown filters $\mathbf{A}$. Such blind deconvolution also tries to remove noise from observed data.

This description of statistically independent sources is the starting point of ICA. It is motivated by the observation in the visual system of mammals that cortical feature detectors perform a redundancy reduction process [Barlow, 1961] [van Hateren and van der Schaaf, 1998] [Park et al., 2002]. The description of observed signals by a factorial code is supported by this. The goal of ICA is to find a set of statistically independent components that span the space of the input images.

ICA is introduced to find suitable representations of multivariate data [Hyvärinen, 1999] [Hyvärinen et al., 2000]. It is used to minimise the statistical dependencies between the components of a representation. ICA is analysed from a statistical point of view. Only the essential structure of the data should be captured. The redundancy reduction models aspects of the early processing of sensory data in the brain. Furthermore, ICA is used in feature detection and exploratory data analysis (projection pursuit). Similar to that data exploration task Everson and Roberts [1999] motivate ICA for modeling and understanding empirical data.

ICA is a statistical approach to solve signal processing tasks. Fiori [2001] motivates this with the statement that the stimuli important in signal processing are no deterministic but stochastic excitations. Signal processing as well as statistics influence ICA and take advantages of it. ICA aids the modeling and understanding of empirical data [Everson and Roberts, 2000].

Various approaches for computing the decorrelation matrix $\mathbf{W}$ exist [Everson and Roberts, 1999] [Amari et al., 1996] [Comon, 1994] [Lee et al., 2000]. In general they consist of two important modules [Hyvärinen, 1999]: At first a *contrast function* is defined to describe the desired characteristics of the separated outputs $\mathbf{u}$. Statistical independence is the most obvious criterion here. For the separated outputs $\mathbf{u}$ the optimum of this contrast function should be reached. Therefore an *optimisation algorithm* is necessary.

Usually the maximisation of statistical independence constitutes the challenge. Nevertheless, many approaches [Basu, 2000] [Hyvärinen, 1999] describe the independence between two variables based on the differential entropy $H(\mathbf{x})$ to define the contrast function. This derivation rests upon the *mutual information* $I(\mathbf{x}, \mathbf{y})$ between two random variables $\mathbf{x}$ and $\mathbf{y}$.

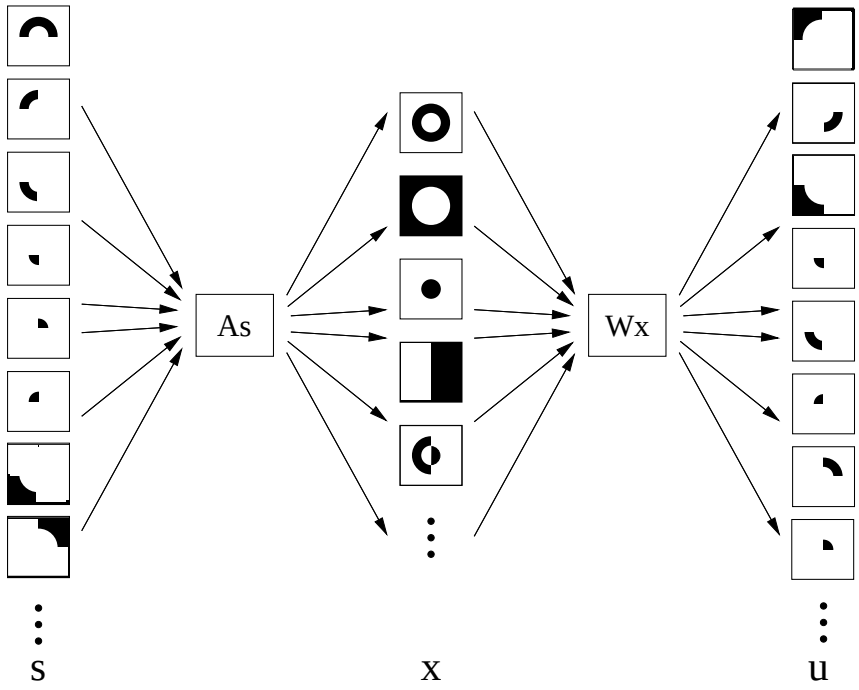


Figure 5.3: Basic assumption of ICA: Observations $\mathbf{x}$ are mixtures of independent sources $\mathbf{s}$. The task is to find a demixing matrix $\mathbf{W}$ (decorrelation matrix).

Since statistical independence means that one random variable contains no information about the second variable, $I(\mathbf{x}, \mathbf{y})$ would be a suitable contrast function. The definition of $I(\mathbf{x}, \mathbf{y})$ in terms of the entropy $H(\mathbf{x})$

$$I(\mathbf{x}, \mathbf{y}) \;=\; H(\mathbf{x}) \;-\; H(\mathbf{x}|\mathbf{y}) \tag{5.11}$$

with

$$H(\mathbf{x}) \;=\; -\int p(\mathbf{x}) \;\log\; p(\mathbf{x})\; d\mathbf{x} \tag{5.12}$$

motivates the entropy maximisation. $p(\mathbf{x})$ is the probability density.

The algorithm used in this work performs this way. It is based on the INFOMAX-algorithm [Bell and Sejnowski, 1995] [Bartlett et al., 1998]. The idea is to maximise the entropy of an auxiliary variable $y = g(\mathbf{W}\mathbf{x})$ where $g(.)$ is a sigmoid squashing function. When $\mathbf{W}$ is such that the probability $P(y)$ resembles a uniform density (which is characterised by a maximal entropy), it factorises and so does the corresponding density for the linearly transformed variables $\mathbf{W}\mathbf{x}$. Gradient ascent on the entropy yields the learning rule which can be set into a computationally more advantageous form by using the *natural gradient* [Amari et al., 1996]:

$$\triangle \mathbf{W} \propto (\mathbf{1} + (1 - 2\mathbf{y}) \cdot \mathbf{x}^T) \cdot \mathbf{W} \tag{5.13}$$

While the application of learning rule (5.13) for the training images $\mathbf{x}_i$ would be feasible, it has been observed in [Bartlett et al., 1998] and [Kämpfe et al., 2001] that the replacement of the $\mathbf{x}_i$ by a subset of the eigenvectors $\mathbf{e}_j,\ j = 1, \ldots, K$ of the covariance matrix $\langle \mathbf{X}\mathbf{X}^T \rangle$ leads to a better learning performance. Geometrically, this means that ICA components

are sought in a subspace that captures the data variation except for a small part that corresponds to the neglected eigenvalues. At the same time, this offers a simple approach to compute only a limited number K of independent components from a much larger number N of image patches.

Such preprocessing is necessary in most of the ICA approaches if less components are desired than input data are given. Usually the demixing matrix $\mathbf{W}$ is assumed to be quadratic and therewith $N = K$ holds.

Furthermore the replacement of the eigenvectors implicates the *prewhitening* or *sphering*, which is proposed in [Bell and Sejnowski, 1997] and [Karhunen, 1996] as a beneficial preprocessing step. Thus the covariance matrix becomes the unit matrix. Effects of second-order statistics are removed and the input is constituted by mutual uncorrelated vectors with unit variance. Since uncorrelatedness is a weaker condition than independence this preprocessing checks if ICA could be favourable [Basu, 2000]. After performing a prewhitening step the separation matrix $\mathbf{W}$ can be assumed as orthogonal.

Approaches to sphere the data may be learning a decorrelation matrix $\mathbf{V}$ according to [Laheld and Cardoso, 1994]

$$\mathbf{V}_{t+1} \;=\; \mathbf{V}_t - \mu_t(\mathbf{v}_t\mathbf{v}_t^T - \mathbf{I})\mathbf{V}_t$$

based on the covariance matrix as in [Bartlett et al., 1998]

$$\mathbf{V} \;=\; 2 \cdot \langle \mathbf{X}\mathbf{X}^T \rangle^{-\frac{1}{2}}$$

or is any other transformation that decorrelates the data. However, the most favourable prewhitening is PCA as in [Borgne and Guerin-Dugue, 2001], [Takaya and Choi, 2001] and [Kämpfe et al., 2001]. It includes a data compression with an optimal mean-square-error, removes noise and can be used to estimate the number of independent components.

5.2.3 ICA Based Data Space Transformations

ICA computes prevailing attributes of the relevant set. Relevance feedback aims at improving automated data grouping. The ICA based transformation should offer a space suitable to separate relevant from non-relevant data. This approach is analysed regarding different image data. Therefore, different sets are divided into relevant and non-relevant subsets. Based on the relevant images the data spaces are transformed and the data distributions are examined. User sessions are simulated based on small subsets representing selected queries and search tasks respectively.

Experiment Setup

Based on the set $\mathbf{X}$ of N $D-$dimensional data vectors, e.g. images or image features, a number of subsets $\mathbf{X}_c \subseteq \mathbf{X},\ c = 1, ..., N_s$ is composed. Each set comprises a group of data to be separated by an automated grouping algorithm. In a CBIR-situation this means that each set is relevant in a category search. $\mathbf{X}_d = \mathbf{X} \setminus \mathbf{X}_c$ describes the set of non-relevant data according to the query for $\mathbf{X}_c$.

The grouping corresponds to the definitions and predefined classifications given in section 3.2 extended by additional user defined subsets. Three image sets are used to analyse the relevance feedback qualities of ICA:

1. The myMondrian set (see section 3.2.2) is developed to investigate the separation of different sequences. Therefore, the given sequences (see table A.1) are grouped into a number of set pairs, where each pair consists of a relevant and a non-relevant image set.
2. The artexplosion photo collection (see section 3.2.1) provides semantic image categories which are used as ground truth groups. The separability of one category from the remaining ones is analysed as well as the pairwise separability of two categories.
3. The detection of user defined subsets is related to a real retrieval session. Therefore a number of groups is composed by a semantic interpretation as proper subsets of the predefined categories. Table C.1 describes these subsets.

In order to perform relevance feedback, ICA is applied on a subset $\mathbf{X}_c$ and a category- or subset-dependent transformation matrix $\mathbf{W}_c^K$ is computed. K stands for the number of independent components and the dimension of the resulting data space respectively. $K = 10$ is set since Wendt [2002] has observed that in the used feature spaces a small number of principal components is sufficient to describe these image sets. To avoid the missing of an independent component, K is set a little bit above the necessary number of principal components. For readability K is omitted in the following equations.

Using the independent components $\mathbf{U}_c = \mathbf{W}_c\mathbf{X}_c$ the relevance feedback is defined as the transformation:

$$\mathfrak{F}(\mathbf{X}, \mathbf{X}_c) \;=\; \mathbf{U}_c\, \mathbf{x}_i \qquad \forall\, \mathbf{x}_i \in \mathbf{X} \tag{5.14}$$

where $\mathfrak{F}(\mathbf{X}, \mathbf{Y})$ is defined as relevance feedback for searching in set $\mathbf{X}$ depending the rated set $\mathbf{Y}$. More generally $\mathbf{Y} = \Gamma \;=\; (\mathbf{x}_i, \gamma)_{\mathbf{q}}$ holds (see definition (5.1)). The used features and the suitable distance measures have been defined in section 3.3.

Evaluation Framework

A transformation of the data set based on ICA computation may stretch the distribution of these data. Then the average of the distance values become larger and a differentiation between the images within this set would be more clearly. Such an effect can be measured by the fraction of the intra-category distance mean after and before the ICA-transformation:

$$\mathfrak{R}(\mathfrak{F}(\mathbf{X}_c)) \;=\; \frac{\langle d(\mathbf{X}_c, ICA)\rangle}{\langle d(\mathbf{X}_c, f)\rangle} \tag{5.15}$$

where $d(.)$ is a distance measure suitable to compare the data points and normalised to [0..1]. The computation is performed in different feature spaces separately. f indicates the used feature. If the ICA based transformation improves the differentiation $\mathfrak{R}(.)$ will be greater than one. Figure 5.4 represents the distance rates of the performed ICA based transformations.

Futhermore the separability to the remaining data has to be evaluated. After a relevance feedback transformation the relevant group should be separated better from the others. In the best case the relevant data is close together – compared to the remaining data – and distant to the non-relevant data (see figure 5.1). The mean of the distances within the relevant group should be less than the distance of relevant objects to non-relevant ones. This can be measured by the fraction of the intra-category distances d_{intra} to the inter-category distances d_{inter}:

$$\mathfrak{D} = \frac{d_{intra}}{d_{inter}} \tag{5.16}$$

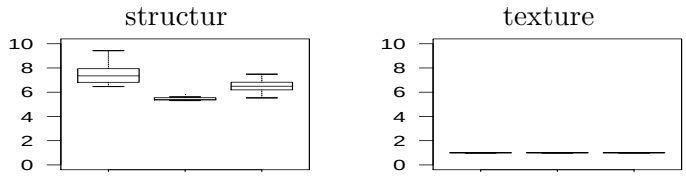


Figure 5.4: Influence of the transformation on the intra-category distance in the different feature spaces. Vertical axes depict $\mathfrak{R}(\mathfrak{F}(\mathbf{X}_c))$ according to equation 5.15. A value greater than one indicates a broader distribution after ICA transformation. The data sets are from left to right myMondrian, artexplosion collection and artexplosion subsets.

This ratio should be less than one. Relevance feedback should reduce this ratio furthermore, whereas the separability is the better the smaller $\mathfrak{D}$ is. Therefore, distances between the data points can measure the relevance feedback impact. Hence the rating function can be defined in general as

$$\mathfrak{R}(\mathfrak{F}(\mathbf{X})) \;=\; \frac{\mathfrak{D}(\mathfrak{F}(\mathbf{X}))}{\mathfrak{D}(\mathbf{X})} \tag{5.17}$$

The used relevance feedback approach is ICA. Indicating this as well as the dependence on the desired subset $\mathbf{X}_c$ and the given feature the rating function would be specified by

$$\mathfrak{R}(\mathbf{X}_c, \mathfrak{F}(\mathbf{X}), f) \;=\; \frac{\mathfrak{D}(\mathbf{X}_c, \mathbf{X}_d, \mathrm{ICA})}{\mathfrak{D}(\mathbf{X}_c, \mathbf{X}_d, f)} \tag{5.18}$$

where $\mathbf{X}_c, \mathbf{X}_d$ are two categories or subsets and f denotes the considered feature space. ICA denotes the ICA-transformed feature space. One group, usually $\mathbf{X}_d$, may be the entire data set. A value less than one represents an enhancement of the data distribution in the feature space.

This relevance feedback evaluation is based on distances between data points. As quoted regarding image features (see section 3.3.2) a lot of distance measures d^f appropriate for CBIR-tasks exist. In the following the features f determines which one is used (see section 3.3.2). In ICA space the Euclidean distance is used. $f(\mathbf{x})$ represents the feature vector of image $\mathbf{x}$. Various definitions of the intra- and inter-category distances d_{intra} and d_{inter} are possible. To allow for this, two different ratios are defined:

(1) The *mean-ratio* compares the means of all point-to-point distances:

$$\mathfrak{D}^{(1)}(\mathbf{X}_c, \mathbf{X}_d, f) \;=\; \frac{d^{(1)}_{intra}(\mathbf{X}_c, f)}{d^{(1)}_{inter}(\mathbf{X}_c, \mathbf{X}_d, f)} \tag{5.19}$$

where

$$d^{(1)}_{intra}(\mathbf{X}_c, f) \;=\; \frac{1}{N_{c-1}^2} \sum_{i=1}^{N_c-1} \sum_{j=i+1}^{N_c} d^f(f(\mathbf{x}_i), f(\mathbf{x}_j)) \tag{5.20}$$

$$d^{(1)}_{inter}(\mathbf{X}_c, \mathbf{X}_d, f) \;=\; \frac{1}{N_c \cdot N_d} \sum_{\mathbf{x}_i \in \mathbf{X}_c, \mathbf{x}_j \in \mathbf{X}_d} d^f(f(\mathbf{x}_i), f(\mathbf{x}_j)) \tag{5.21}$$

These values are presented in figure 5.5.

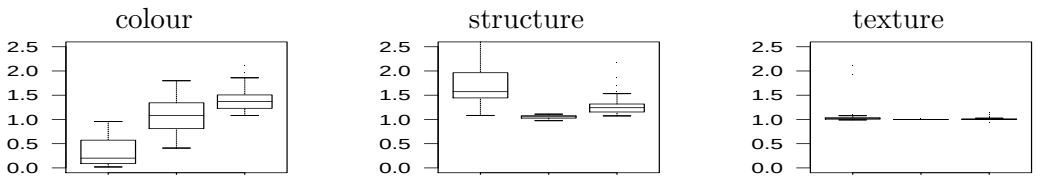


Figure 5.5: Influence of ICA transformation on the category separation in the different feature spaces. The comparison is computed by a mean based ratio $\mathfrak{R}(\mathfrak{F}(\mathbf{X}_c))$ (see equation 5.19). A value less than one indicates an enhancement. The data sets are from left to right myMondrian, artexplosion collection and artexplosion subsets.

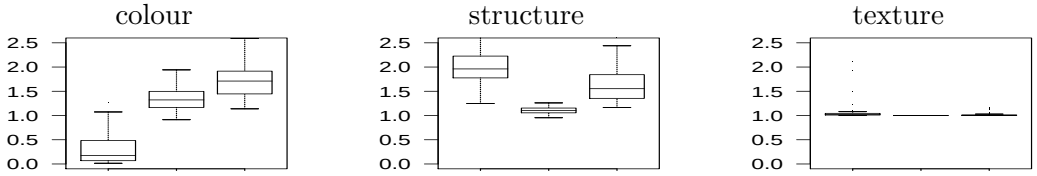


Figure 5.6: Influence of ICA transformation on the category separation in the different feature spaces. The comparison is computed by a mean based ratio $\mathfrak{D}^{(2)}(\mathbf{X}_c, \mathbf{X}_d, f)$ (see equation 5.22). A value less than one indicates an enhancement. The data sets are from left to right myMondrian, artexplosion collection and artexplosion subsets.

(2) The *variance-ratio* is motivated by the aim of compact clusters. The distance to the mean of the relevant subset prefers a compact cluster around the mean and widely spread nonrelevant data. A variance-motivated measure represents this:

$$\mathfrak{D}^{(2)}(\mathbf{X}_c, \mathbf{X}_d, f) = \frac{d^{(2)}_{intra}(\mathbf{X}_c, f)}{d^{(2)}_{inter}(\mathbf{X}_c, \mathbf{X}_d, f)} \tag{5.22}$$

where

$$d^{(2)}_{intra}(\mathbf{X}_c, f) = \sqrt{\frac{1}{N_c} \sum_{i=1}^{N_c} (d^f(f(\mathbf{x}_i), \tilde{\mu}^f_c))^2} \tag{5.23}$$

$$d^{(2)}_{inter}(\mathbf{X}_c, \mathbf{X}_d, f) = \sqrt{\frac{1}{N_d} \sum_{j=1}^{N_d} (d^f(f(\mathbf{x}_j), \tilde{\mu}^f_c))^2} \tag{5.24}$$

$$\tilde{\mu}^f_c = \frac{1}{N_c} \sum_{i=1}^{N_c} f(\mathbf{x}_i) \tag{5.25}$$

$\tilde{\mu}^f_c$ is the mean of the relevant data. These values are presented in figure 5.6.

5.2.4 Observations

In general, ICA transformation does not have any impact in the texture feature space. Although a separability of the different data domains based on the texture has been observed in chapter 3, a separability of subsets within the used domains (artexplosion photos and myMondrian sequences) is not supported.

This may be caused by using the texture as a global feature. Applied on image regions representing objects, texture may possibly be more suitable to separate the pictorial entities. Indeed this requires a successful image segmentation.

Regarding the other features, the ICA transformation supports various observations. The distance ratios show that the data set, which generates the input to ICA, is spread more broadly after the transformation. The differentiation within this set is improved therewith.

To perform relevance feedback ICA is computed on a subset and the transformation is applied on the entire set. This does not gain the desired effect of better separability. In various situations the ICA transformation worsens the separability of the relevant set from the non relevant data.

What are the reasons for the failure of the ICA transformation to serve the purpose of relevance feedback? For the used features the distance ratios do not confirm a better separability in the considered data sets. The non-relevant data may be pressed into the relevant based ICA space and the differences may get lost. Furthermore the relevant sets are much smaller than the non-relevant sets. However, just the attributes of a small part are represented in the ICA transformation.

On the other hand the separability is increased for the intensity histogram feature in the myMondrian image set. This may be an exception. Or it may be a hint, that an ICA based relevance feedback is successful under special circumstances.

The data of the relevant set which constitutes the basis for the ICA computation becomes more spread by an ICA based transformation. Hence, such a transformation may be profitable for the entire set. However, this will not be a relevance feedback.

5.3 Combining ICA with Naive Bayes Classification

In the previous section ICA is applied to perform data space transformations as relevance feedback steps. For this purpose CBIR-frameworks provide a set of relevant images $\mathbf{X}_c$. ICA computes an individual representation of this underlying set. The goal of the retrieval would be the detection of this set out of the entire database. This resembles a classification task. Hence ICA can be regarded as a preprocessing step in a classification approach. The relevant rated image set $\mathbf{X}_c$ builds the training set.

It has been observed in [Bell and Sejnowski, 1995] that ICA provides a sparse, peaky distrubuted representation for data of the same class like the training set. Computing ICA on a single class results in a representation that shows high probabilities for the target class and low probabilities for data not belonging to the target class. Thus, the usages of density based classifiers is motivated. Since a category search can be implemented as a common classification task this is promising for image retrieval applications.

The Optimum Bayes Classifier is an often used classification approach, e.g. [Peng and Bhanu, 2001] use it for image retrieval. Basically it relies on the probability densities of the classes and classifies each data point according to the most probable class. However, the underlying probability densities are usually unknown and have to be estimated. In most cases they are assumed to be Gaussian. Alternatively the different feature dimensions have to be independent. Both cannot be guaranteed [Rish, 2001]. Nevertheless, this old-fashioned Naive Bayes Classifier [Domingos and Pazzani, 1997] has recently celebrated a great comeback in information retrieval tasks [Bressan and Vitrià, 2002b].

To enhance the density estimation an ICA preprocessing step may be suitable. This should detect data directions which are more suitable to represent the independent proba-

bility distributions. At first they are on definition as statistically independent as possible. This requirement of the Naive Bayes classifier usually is assumed to be given but seldom proven. Furthermore, they result in probability distributions most suitable to distinguish the interesting class from the others as in [Bell and Sejnowski, 1995] is outlined. We shall call the combination of ICA with a Naive Bayes Classifier *icaNbayes*.

A similar approach has been developed in [Lee et al., 2000]. The class-conditional probabilities and the ICA parameters are determined simultaneously. Therefore, the learning rule of ICA is modified to learn the classification parameters. Therewith this approach performs an unsupervised clustering of the data. This is interesting for an initial retrieval step or an a priori organisation of an unknown image data set. Against it relevance feedback requires the possibility to insert labelled training data and therefor prefer a kind of supervised classification.

In [Zhou et al., 2001] ICA is used as a preprocessing step to factorise histogram-like image features. Thereby they reduce the features to a suitable dimension. The factorised data build the input of a classifier in an object detection task. Again ICA is computed on the entire data set. A relevance feedback adaptation as supported by the class dependent ICA is not possible here.

The theoretical derivation of icaNbayes is presented in the following. After that it is applied on synthetically constructed data sets. In this artificial setting advantages and problems should become obvious and the performance of the approach can be analysed. Finally it is applied on the image data used above.

5.3.1 The icaNbayes Approach

A category search can be implemented as a grouping function (see section 3.1.2). Usually any implementation of such classification approaches is based on the probability distributions of the different groups. The class with the largest probability is assumed to be that one a data point belongs to. This means a classification $\Phi(\mathbf{x})$ is the mapping of a data point $\mathbf{x}$ to the class with the greatest probability:

$$\Phi(\mathbf{x}) = c, \quad \text{if} \quad p_c(\mathbf{x})\ P(\mathbf{X}_c)\ \geq\ p_e(\mathbf{x})\ P(\mathbf{X}_e) \quad \forall\, e \neq c \tag{5.26}$$

where c and e are instances of the group label $\psi \in \{1, ..., N_s\}$ (see definition 3.1). $p_c(\mathbf{x})$ and $p_d(\mathbf{x})$ are the probability density functions of class $\mathbf{X}_c$ and $\mathbf{X}_e$ respectively. $P(\mathbf{X}_c)$ and $P(\mathbf{X}_e)$ are the a priori probabilities of the classes and can be neglected for equiprobable classes.

This classification approach satisfies the definition of the Bayes Classifier which minimises the classification error:

$$\Phi_{Bayes}(\mathbf{x}) = \arg\max_{\psi} P_\psi(\mathbf{x}) \tag{5.27}$$

where $P_\psi(\mathbf{x})$ is the probability of class ψ.

However, image data is of a very high dimensionality. This complicates the estimation of the underlying probability densities. To accelerate this computation usually

class-conditional independence of the dimensions is assumed. Therefore, the Naive Bayes Classifier is defined as:

$$\Phi_{Naive}(\mathbf{x}) \;=\; \arg\max_{\psi} \prod_{d=1}^{D} P(x_d|\mathbf{X}_\psi)P(\mathbf{X}_\psi) \tag{5.28}$$

where D stands for the dimension of the data space. For equiprobable classes it reduces to

$$\Phi_{Naive}(\mathbf{x}) \;=\; \arg\max_{\psi} \prod_{d=1}^{D} P(x_d|\mathbf{X}_\psi) \tag{5.29}$$

In order to deal with the problem of unknown probability density functions the probability density may be approximated by factorisation of statistical independent sources. While in most cases the independence is not proved or even ignored, this attribute hints to a well known approach to find the particular underlying functions: ICA aims to detect statistically independent components of a data set.

To justify the Naive Bayes Classifier a data transformation should be inserted before the classification step. This transformation should result in class conditional independent data dimensions. ICA can achieve this (see equations 5.9 and 5.10). Thereby it is important that ICA is computed in the different classes individually. This satisfies a class-conditional independence. A general independence of the entire data set would not suffice [Bressan and Vitrià, 2002a,b].

The projection of the data $\mathbf{x} \in \mathbb{R}^D$ on the class dependent ICA sources $\mathbf{U}_c$ will be given by:

$$\upsilon = \Upsilon(\mathbf{x}, c) = \mathbf{U}_c\ \mathbf{x} \tag{5.30}$$

where $\mathbf{U}_c \in \mathbb{R}^{K\times D}$ is a matrix of the independent sources computed on set $\mathbf{X}_c$ and $\upsilon \in \mathbb{R}^K$ with $K \leq D$. For readability the class labels are neglected where unambiguous.

This ICA based transformed data set $\upsilon_i,\ i = 1, ..., N$ yields independent features. Furthermore, the class dependent computation of ICA ensures the required class conditional independent features. Thus, the overall probability density $p(\upsilon)$ can be factorised.

The Naive Bayes Classifier is applied on the transformed data set. Therefore, ICA is performed in a preprocessing step on a subset of the data which resembles a single class.

Capitalising on ICA to estimate the class conditional probabilities $P(\upsilon_d|\mathbf{X}_c)$ in the Naive Bayes Classifier induces a modified classification rule:

$$\Phi_{ICA}(\mathbf{x}) = \arg\max_{c} \prod_{d}^{D} \nu_d p_d(\upsilon)\ P(\mathbf{X}_c) \tag{5.31}$$

Therewith the classification in the original D-dimensional data space is replaced by a classification in the ICA based data space. This induces an ICA based Naive Bayes Classifier *icaNbayes*:

$$\Phi_{icaNbayes}(\mathbf{x}) = \arg\max_{c} \sum_{d}^{D} \sum_{k}^{K} \log p_c(\upsilon_{dk}) + \mathcal{N} \tag{5.32}$$

for equiprobable classes, where K is the dimension of the ICA based transformed data space and $\mathcal{N}$ is a normalisation factor $\mathcal{N} = D \cdot (\int p(\upsilon)d\upsilon)^{-1}$.

This approach consists of two fairly independent steps – the training of the class dependent independent components and the Bayesian classification according to [Bressan et al., 2001]. Various approaches to compute the source matrix $\mathbf{U}$ exists (for an overview see section 5.2.1 or [Basu, 2000]). Here the INFOMAX-algorithm motivated by [Bell and Sejnowski, 1995] as described in section 5.2.2 is used.

The Bayes classification step requires the probability density $p(\upsilon)$. Therefore, the single class-conditional probability densities $p_c(\upsilon)$ are computed by a common Kernel-density-estimation [Bishop, 1995]:

$$\tilde{p}_c(\upsilon) \;=\; \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{D/2}} \;\exp\left(-\frac{(\upsilon - \mathbf{y}_n^c)^2}{2h^2}\right), \quad c = 1, ..., N_s \tag{5.33}$$

where h is the kernel width and D the dimension of the data. $\tilde{p}_c(\upsilon)$ implies the projection into the space $\upsilon = \Upsilon(\mathbf{x}, c)$ and $\mathbf{y}_n^c = \Upsilon(\mathbf{x}_n, c)$.

Based on the classes and categories defined for the different data sets ICA is performed. Subsequently the projection parameters are used to transform new objects, called test objects. Regarding the relevance feedback task the objects may be image feature. After the transformation the appearance probabilities in the different classes are determined. The class which leads to the highest probability is considered to be the true class of the object.

5.3.2 Experiments on Synthetic Data

To present a specific classification approach a data set would be nice, that can be classified by this approach but not by other commonly prevalent classifiers. Most of the known classification approaches rely on orthogonal axes and use the distributions of the data along these axes for classification. ICA computes axes which can be non-orthogonal. Hence, a data set is interesting which cannot be classified based on the orthogonal Cartesian data axes but based on more data specific, perhaps non-orthogonal axes.

The class dependent ICA should detect these axes. Thus the data sets are constructed based on non-orthogonal directions. The desired attributes of the data can be defined as:

$$p_c(\mathbf{x}) \quad \text{small if} \quad \mathbf{x} \notin \mathbf{X}_c \tag{5.34}$$

where the class dependent probability densities $p_c(\mathbf{x})$ are computed along the data specific axes. $\mathbf{X}_c$ denotes the treated class.

The sketch (figure 5.7) of such a data set shows that the data space regions described by the source axes and the related probability density functions should not contain any data points of the other class. Based on a barbell like probability density function (figure 5.7 (a)) along each data axis a two-dimensional data set will form a set of blobs in the data space. The main direction should be chosen non-orthogonal. The barbell like probability density is implemented by two Gaussian bells along each axis. The intersection of two humps with respect to two axes defines the data blobs (figure 5.7 (b)).

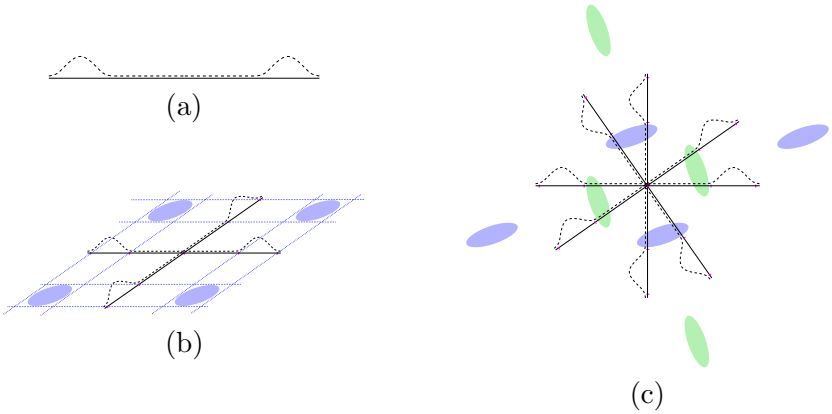


Figure 5.7: Construction of the barbell data set: (a) Along each axis a barbell-like probability density is assumed. (b) Based on two axes four blobs arise. (c) Constructing two data classes with the same mean but rotated by an angle β results in an XOR-like classification task.

barbell

The particular construction of class $\mathbf{X}_c$ is described by:

$$\mathbf{s}_1 = \mu + \lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \mathbf{s}_2 = \mu + \lambda \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix} \tag{5.35}$$

where $\mathbf{s}_1$ and $\mathbf{s}_2$ are the main directions of the data distribution. $\mu = \binom{0}{0}$ is the center. $\lambda \in \mathbb{R}$ is a random variable with

$$P(\lambda) = \frac{1}{2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\lambda - r)^2}{2\sigma^2}\right) + \frac{1}{2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\lambda + r)^2}{2\sigma^2}\right) \tag{5.36}$$

r is the distance from the mean to the center of the humps. This data distribution is two dimensional ($D = 2$ and $N_s = 2$) and the angle between the axes $\alpha(\mathbf{s}_1, \mathbf{s}_2)$ measures 45° or 0.79 radian.

For analysing the classification approach a number of samples of this data distribution is constructed. In each case one class is created according to equation 5.35, whereas the variance $\sigma(\lambda)$ of the humps varies. For classification a second class $\mathbf{X}_d$ is desired. This class is constructed based on the density described in equation 5.35 but rotated by an angle β. The given classification task resembles the XOR-problem. Figure 5.7 (c) presents a set of two classes.

Construction parameters are presented in table 5.1. To analyse icaNbayes classification this is applied on these sets and the result is compared with the results of the Naive Bayes classifier. The classification results of these sets with two classes in the 2d space are presented in table 5.1.

barbell 3d

Based on the definition above further data sets are constructed by increasing the number of classes as well as the dimension. Again non-orthogonal construction axes should show that icaNbayes outperforms the common Naive Bayes classifier.

Table 5.2 presents a number of data sets, their construction parameters and the classification results for icaNbayes and common Naive Bayes. These data sets are labelled as "barbell_X".

data set	β	$\sigma(\lambda)$	$\mathfrak{r}(\Phi)$	
			Φ_{Naive}	$\Phi_{icaNbayes}$
barbell_1	1	0.1	0.54	0.65
barbell_2	1	0.1	0.45	0.60
barbell_3	1	0.1	0.40	0.6
barbell_4	0.8	0.1	0.43	0.61
barbell_5	0.5	0.1	0.50	0.69
barbell_6	1	0.2	0.49	0.58
barbell_7	1	0.05	0.50	0.75

Table 5.1: Classification rate $\mathfrak{r}(\Phi)$ of icaNbayes on various barbell data sets. β is the rotation of $\mathbf{X}_d$ related to $\mathbf{X}_c$ measured in the radian and $\sigma(\lambda)$ is the variance of the single gaussian humps.

helix 3d

ICA detects linear directions. However, a lot of data sets have interesting directions which are non-linear. Thus the behaviour of icaNbayes under these conditions may be interesting. The given data sets are based on a fragmented and squeezed helix around a skew axis in the common Cartesian data space.

The construction of each class is based on:

$$\mathbf{x} \;=\; \mu \;+\; \mathbf{r}_1 \cdot \begin{pmatrix} \sin(\lambda_1) \\ \cos(\lambda_1) \\ \lambda_1 \end{pmatrix} \;+\; \mathbf{r}_2 \cdot \begin{pmatrix} \lambda_2 \\ \sin(\lambda_2) \\ \cos(\lambda_2) \end{pmatrix} \;+\; \mathbf{r}_3 \cdot \lambda_3 \;+\; \eta$$

where μ is the mean, $\mathbf{r}_i$ a scaling factor along the different axes and η a random noise. λ_i is a random variable with a unit distribution of range σ_i. The used parameter values are given in table C.2. The data sets are labelled with "helix_X" in table 5.2.

correlated blobs

Most data sets have an intrinsic data dimensionality smaller than the observed dimensionality. However, the direction in the same dimensionality may be important, as well. Thus three directions are chosen arbitrarily in the original Cartesian data space. Along two axes trigonometric-like distributions are defined for a limited interval to get the blobs. The third direction is chosen Gaussian to get three-dimensional data. Two classes are constructed based on the following definition in the "blobs" data set.

$$\begin{aligned} \mathbf{x} \;= & \mu \;+\; \mathbf{r}_1 \cdot \cos(\lambda_1) \\ & +\; 0.5 \cdot \mathbf{r}_1 \cdot \lambda_1 \;-\; \mathbf{r}_1 \;+\; \mathbf{r}_2 \cdot \cos(\lambda_2) \\ & +\; 0.5 \cdot \mathbf{r}_2 \cdot \lambda_2 \;-\; \mathbf{r}_2 \;+\; \mathbf{r}_3 \cdot \lambda_3 \;+\; \eta \end{aligned}$$

where μ is the mean, $\mathbf{r}_i$ a scaling factor along the different axes and η a random noise. λ_i is a random variable. The used parameter values are given in table C.2. The data set is labelled with "blobs" in table 5.2.

data set	dim	N_s	$\mathfrak{r}(\Phi)$ Φ_{Naive}	$\mathfrak{r}(\Phi)$ $\Phi_{icaNbayes}$
barbell_8	2d	3	0.34	0.47
barbell_9	3d	3	0.49	0.75
barbell_10	3d	3	0.33	0.88
barbell_11	3d	3	0.38	0.67
helix_1	3d	3	0.82	0.89
helix_2	3d	3	0.77	0.86
blobs	3d	2	0.83	0.90

Table 5.2: Construction parameters and classification rate $\mathfrak{r}(\Phi)$ of icaNbayes for various data sets.

Observations

The classification results of icaNbayes are compared with the results of a common Naive Bayes classifier. Therefore, the classification rates $\mathfrak{r}(\Phi_{icaNbayes})$ and $\mathfrak{r}(\Phi_{Naive})$ (see equation 4.8) are computed:

$$\mathfrak{r}(\Phi, \mathbf{X}_c) \;=\; \frac{\#(\mathbf{x} \mid \Phi(\mathbf{x}) \;=\; \psi(\mathbf{x}) \;,\; \mathbf{x} \in \mathbf{X}_c)}{\#(\mathbf{x} \mid \mathbf{x} \in \mathbf{X}_c)}$$

where Φ denotes the used classifier – icaNbayes $\Phi_{icaNbayes}$ and common Naive Bayes Φ_{Naive} respectively. $\psi(\mathbf{x})$ is the correct class of data point $\mathbf{x}$. Tables 5.1 and 5.2 presents the classification rates $\mathfrak{r}(\Phi)$ of the synthetic data sets.

For the two dimensional barbell-data sets as well as for the various three-dimensional sets it can be observed that icaNbayes outperforms the common Naive Bayes classifier. Hence, an application of icaNbayes on real world image data appears interesting and worth to analyse.

5.3.3 Experiments on Image Data

As done for the preceding tasks – 1dSOM and ICA based data space transformations – each image domain $\mathbf{X}$ is divided into a number of subsets $\mathbf{X}_c, c = 1, ..., N_s$. To analyse the classification task disjoint subsets are used.

The first application of the icaNbayes-classification approach on image data is the myMondrian set (see section 3.2.2). Classification objects are the individual image sequences as well as the sets of different sequence types (see table A.1). Furthermore the artexplosion photo set (see section 3.2.1) is used. The predefined categories (underthesea, animals, doorswindows, sunrisesunset) should be detected. For both image domains a structure, a texture and a colour feature are used.

The evaluation is based on the classification rate $\mathfrak{r}(\Phi)$. Table 5.3 lists the classification rates of the entire artexplosion set into the equiprobable categories. Figure 5.8 presents the rates of the myMondrian type classification.

The impact of ICA to the Naive Bayes classification is measured by the difference between the classification rates of the icaNbayes and the common Naive Bayes classifier $\mathfrak{r}(\Phi_{icaNbayes}, \mathbf{X}_c) - \mathfrak{r}(\Phi_{Naive}, \mathbf{X}_c)$. The ranges of these differences over different classification objects are presented in figure 5.9.

category	classification rate $\mathfrak{r}(\Phi)$					
	icaNbayes			Naive Bayes		
	Structure	Texture	Colour	Structure	Texture	Colour
underthesea	0.45	0.11	0.03	0.19	0.11	1
animals	0.05	0.40	0.14	0.25	0.40	0
doorswindows	0.28	0.76	0.90	0.19	0.76	0
sunrisesunset	0.57	0.72	0.24	0.75	0.72	0
overall	0.27	0.40	0.26	0.28	0.40	0.2

Table 5.3: Classification rates of icaNbayes and common Naive Bayes artexplosion classification. The entire data set should be classified into the predefined categories.

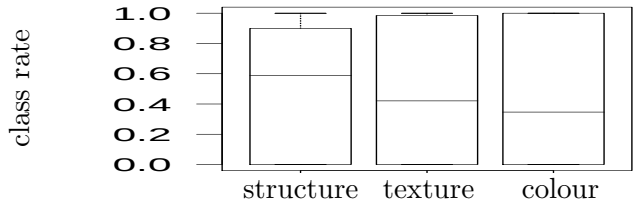


Figure 5.8: Classification rates $\mathfrak{r}(\Phi_{icaNbayes})$ of the myMondrian-type classification. All permutations and numbers (from two sets upto all five sets) of different types are addressed.

Furthermore, the precision $Pr(\Phi, c)$ of the classification is measured by the ratio of correctly classified data points with respect to each class:

$$Pr(\Phi, c) \;=\; \frac{\#(\mathbf{x} \mid \Phi(\mathbf{x}) \;=\; \psi(\mathbf{x}) \;,\; \mathbf{x} \in \mathbf{X}_c)}{\#(\mathbf{x} \mid \Phi(\mathbf{x}) \;=\; c)} \tag{5.37}$$

c indicates the considered class. The precision values of the artexplosion and the myMondrian type classification are listed in tables 5.4.

Inspecting the icaNbayes classification results in the following insights:

- The classification of the individual myMondrian sequences is improved by icaNbayes compared to the common Naive Bayes classifier.
- The myMondrian sequence type classification is enhanced only for the structure feature.
- The texture feature is most suitable to detect the myMondrian class with coloured background. All images of this class have been identified while just five false positives occur. The class with a growing rectangle is again passed over completely.
- For the other two features the myMondrian type classification is dominated by the class with colour changes. All images are classified as members of this set.
  Maybe this is the explanation why the ICA transformation does not give any impact on the sequence type classification, which means that the colour changes of this sequence type are too dominant and all other distinctive features are eliminated by the ICA transformation
- A more detailed analysis of the classification result offer the precision values. Looking at table 5.4 displays that some classes are misseed completely. Regarding the myMondrian types just the texture feature is suitable to detect different classes. Structure and colour classifies every input as an element of a colour changing sequence. This

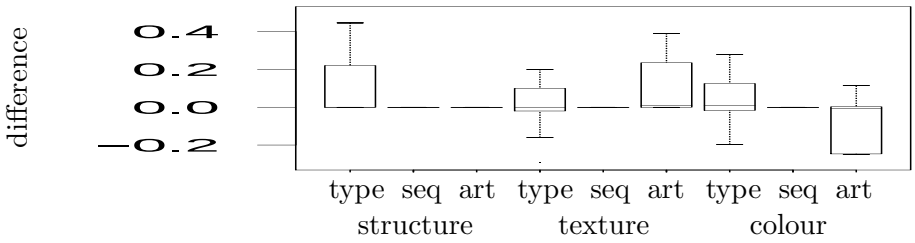


Figure 5.9: Difference of the classification rate of the icaNbayes and the common (PCA-based) Naive Bayes classifier $\mathfrak{r}(\Phi_{icaNbayes}, \mathbf{X}_c) - \mathfrak{r}(\Phi_{Naive}, \mathbf{X}_c)$. Values > 0 indicates that the ICA-preprocessing enhances the classification. Considered classification objects are the separations of thow subsets.

category	$Pr(\Phi_{icaNbayes})$		
	Struct	Text	Colour
underthesea	0.22	0.60	0.60
animals	0.24	0.30	0.23
doorswindows	0.24	0.33	0.25
sunrisesunset	0.38	0.60	0.36

category	$Pr(\Phi_{icaNbayes})$		
	Struct	Text	Colour
def move	–	0.20	–
var move	–	0.47	–
growing	–	–	–
colour change	0.40	0.50	0.40
text back	–	0.87	–

Table 5.4: Precision $Pr(.)$ of icaNbayes classifications for artexplosion categories (left) and myMondrian sequence types (right). The "–" indicates that a class is missed completely. No data point is detected as a member of this class.

may be caused by dominance of the colour in these features. The colour changing class is mainly described by the colour. Thus, ICA may detect this as the important quality and all other images are ranged in this description.

- The artexplosion classification is enhanced for the structure feature. In contrast to that the results for the colour feature degrade.
- Viewing the results of the Naive Bayes the artexplosion photo set seems to be distributed equally in the original colour intensity space. This may be differentiated by ICA and the classification becomes a little bit more diverse.
  This may cause the worsening of the classification rate $\mathfrak{r}(\Phi)$.
- The texture feature does not show any impact from the ICA based transformation. This coincides with the observations in the preceding sections.

5.3.4 Summary icaNbayes

The insertion of ICA into a classification framework to detect data directions most suitable to represent the required classes seams to be promising. The analyses regarding different synthetic data sets confirms this assumption since the classification rates outperforms those of a common Naive Bayes Classifier. However, these data sets where designed to emphasise the advantages of ICA. They are based on non-orthogonal construction axes. Real world data often do not offer orthogonal description axes as well. Thus the proposed icaNbayes classifier is applied on various image data. The task was to detect the predefined subsets. This approach shows initial benefits although it lacks further improvements.

Since the icaNbayes-classifier was satisfying applied on the optimised data sets, this approach may not be rejected completely. The data set to classify should be analysed in

advance regarding the suitability of the ICA preprocessing. In the following section some analyses are presented to explore the performance of the used ICA approach with respect to the given data sets.

5.4 Analyses of the ICA Based Relevance Feedback

Various questions arose with respect to the apparent shortcomings of ICA applied on relevance feedback tasks:

- Do the independent components achieve the expected attributes of non-orthogonality and statistical independence?
- Do the given data distributions satisfy the required qualities determined by the used ICA approach?
- What happens to the non-relevant data by the transformation computed on the relevant data? Is it allowed to transform a large data set by a transformation computed on a very small subset?

These questions are analysed in the following subsections.

5.4.1 Analysis of the Independent Components

In order to perform an ICA based relevance feedback, independent components are computed on the a priori defined image subset. However, the analysis of the transformed data shows that this approach does not improve the separability of arbitrary categories and subsets. To explain this observation the independent components are analysed with respect to the attributes expected, non-orthogonality of the different components and statistical independence. The first quality is measured by the angle between two independent components and the second one by the mutual information between them.

ICA is assumed to specify better the underlying data set than PCA. Basically, this should be achieved by the possibly non-orthogonal directions of the computed components. To verify that an ICA is reasonable in the given image data the angles between the different independent components are analysed.

The distribution of the angles measured in the radian of the unit circle between two components are presented as boxplots in figure 5.10. The values hardly differ from the value of $\pi/2$ for a right angle. In this case, ICA does not have any benefit compared to PCA.

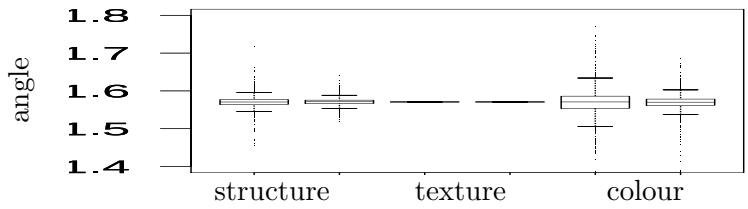


Figure 5.10: Distribution of the angles between two independent components in the different feature spaces. In each case the myMondrian (left) and the artexplosion (right) image sets are used. The angle is measured in the radian.

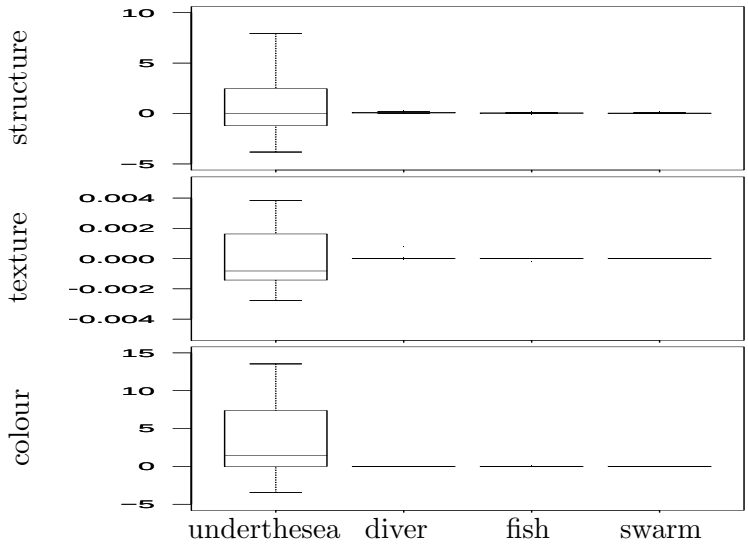


Figure 5.11: Mutual information between the independent components in the category underthesea and corresponding subsets. Note that the range of the presentation is optimised according to the particular feature.

The mutual information (see equation 5.11) between two data sets describes the statistical dependence between them. Thus, it is computed for the components of the required subsets. Selected mutual information values are presented in appendix C (tables C.3 to C.6). Figure 5.11 presents the feature dependent distributions of all mutual informations for the category underthesea.

Depending on the analysed data sets and used features and based on the mutual information different observations are confirmed:

- A small number of components are statistically independent among each other (see tables C.4 and C.5).
- Nearly no component pair is statistically independent (see table C.6).

Concerning to the myMondrian set the first observation is given based on the texture feature. Contrary, with the structure feature or the colour feature statistical independence cannot be shown for all data sets. Especially the sequences where just the colour changes do not support ICA. The other sequences perform statistically independent components. However, the number of appropriate independent components may be smaller than the used number.

Basically these observations hold for the artexplosion set, as well. Furthermore, the user defined subsets of the artexplosion collection get projection axes which are less dependent than the entire categories. The analysis of the mutual information confirms this (see figure 5.11).

An enhancement of the statistical independence between the components is not observed by comparing the mutual information between the output of ICA and the result of PCA. This approach may not be suitable to detect interesting directions within these image sets.

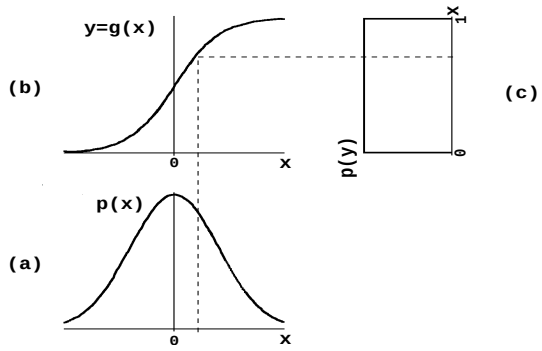


Figure 5.12: Squashing function of the INFOMAX-algorithm: Based on the probability density function of the input variable $\mathbf{x}$ (a) a non-linear squashing function $g(\mathbf{x})$ is defined as the cumulative density function of the input (b). In that way $\mathbf{x}$ is mapped on a new random variable $\mathbf{y}$ so that $\mathbf{y} = g(\mathbf{x})$ has uniform density over $[0, 1]$ (c).

5.4.2 Used Feature Data

The ICA approach requires some prerequisites. For example, at most one input channel is allowed to be Gaussian distributed. Otherwise the computed components will conform to the output of PCA. Contrary to that, Qian et al. [2002] state that most of the approaches to learn or optimise a function for relevance feedback assume (one-dimensional) Gaussian distributed relevant images. For example, the MindReader approach [Ishikawa et al., 1998] is based on the single Gaussian distribution model. This would contradict an ICA approach for relevance feedback, since ICA allows at most one Gaussian input. Indeed, they doubt the validity of this assumption.

Furthermore the squashing function $g(\mathbf{x})$ (see page 74) chosen in that way that the learning rule (see equation 5.13) becomes computationally simple. The stated sigmoid function is the Fermi function $g(x) = 1/(1 + e^{-\beta x})$, where β determines the slope of the function (see figure 5.12). Therefore, the underlying data set should be super-gaussian. Sub-gaussian data cause problems with this ICA approach. To explain this Fiori [2001] has analysed the behaviour of the pdf[1]-matching neurons proposed in [Bell and Sejnowski, 1995]. The core of that approach is the sigmoid squashing function. Thus, it takes an important portion of the analysis. Based on the dependencies between the probability densities of the input and output variables an activation function similar to the standardly used sigmoid is rated as suitable and mathematically tractable.

In [Bell and Sejnowski, 1995] is conjectured that the inference of the individual entropies mentioned above occurs merely in the case of sub-gaussian probability densities. Newer papers (e.g. Lee et al. [2000]) take the infeasibility of INFOMAX for sub-gaussian data as a matter of fact. Nevertheless, this is a good starting point to analyse the INFOMAX-algorithm applied on the given image feature data.

Super-gaussian distributed means that the fourth-order kumulant, the kurtosis, is greater than zero:

$$kurt(\mathbf{x}) \;=\; E\{\mathbf{x}^4\} \;-\; 3 \cdot \{E\{\mathbf{x}\}\}^2 \;>\; 0 \tag{5.38}$$

Super-gaussian distributions are peaky in relation to Gaussian distributions. Contrary to

[1] probability density function

feature	artexplosion categories	artexplosion user sets	myMondrian
structure	0.93	0.93	0.59
texture	0.27	0.72	0.61
colour	0.28	0.51	0.31

Table 5.5: Rate of feature dimensions where the kurtosis is < 0. These dimensions of the data sets offer sub-gaussian distributions and may cause problems while computing ICA.

that, sub-gaussian distributions have an kurtosis less than zero and are flat in relation to gaussian distributions, e.g. an uniform distribution is sub-gaussian. The INFOMAX-algorithm requires super-gaussian data. This means that the kurtosis should be greater than zero.

Depending on the image features and the data sets different rates of the dimensions show sub-gaussian distributions (see table 5.5). Especially the structure feature offers nearly completely sub-gaussian data. Examining the artexplosion categories and user defined subsets stands out, that the user defined subsets have more sub-gaussian directions. Perhaps this is caused by the smaller data sets. In general, the colour feature has less sub-gaussian directions than the other used image features. ICA should achieve better results in this feature space.

5.4.3 Influence of the Class Dependent ICA on the Remaining Data

User based relevance feedback naturally offers very small sets of rated data. Regarding the entire image collection, these sets present a portion under one percent (e.g. in INDI usually 0.6 percent or for PicSOM about 0.03 percent). This means PCA or ICA based relevance feedback transformation is computed on a small data set whereas the transformation is executed on a larger set. This may cause difficulties regarding the computation of the transformation matrix as well as the validity for the unlabelled data.

One problem regarding the transformation matrix is named in [Rui and Huang, 2000]: A PCA-transformation requires a full weight matrix $\mathbf{W}$ to perform feature (-component) weighting. The computation of such a matrix asks for at least as many (rated) image examples as the given feature space has dimensions ($\#(\gamma_i) \geq N_f$, where γ_i are the rated images). Since most of the features are of large dimensions and users rate just a small number of images, this will not be satisfied in most of the image retrieval situations. This affects the ICA approach if the data is pre-whitened with PCA. Therefore, such a transformation is not generally suitable.

Image retrieval is the detection of a somehow defined and usually quite small subset within a mostly very large set. Thus, the problem of sets with different cardinalities is inherent in the image retrieval task. The most obvious situation where this emerges is the initial query-by-example. Based on one data point similar data should be detected. This means that the image retrieval is the identification of a subregion of the data space, usually neighbouring to the query. If the images located in this subregion are promising but not satisfactory, a larger region with equivalent properties should be found. This region may be located distant to the initial query. Thus, local attributes of the neighbourhood around the query are interesting for image retrieval and relevance feedback. Using these attributes to enhance the further retrieval steps seems tempting. Unfortunately they do not hold for

the entire data set. Squeezing arbitrary data into a structure describing a small subset may destroy important topological relationships in other data space regions distant from the query.

Nevertheless, a better description of the local neighbourhood around the query and the relevant data is desired. A benefiting representation will offer a stretched distribution of the relevant set. However, how does such a transformation influence the distribution of the remaining data?

In general, the projection in a data space with more suitable directions should show a small approximation error. For the relevant data this is guaranteed by the chosen algorithms. However, what's about the approximation error for the non-relevant data? A good approximation of the original data in another data space keeps as much information coded in the data distribution as possible.

Both of the described qualities regarding data transformations based on relevance feedback rely on the variance of the data distributions. Thus, the variance is analysed with respect to different sets before and after an ICA based transformation computed on the relevant set. The comparison of the different values results in the following observations:

- Nearly all sets show a larger variance after the ICA transformation. Especially for the structure feature the increase is enormous (see figure 5.13 second row).
- Although most of the feature spaces show larger variances for the (larger) non-relevant sets, a lot of ICA transformations cause a switch in this relation. For example the elephant-set in the structure space has a smaller variance as the associated non-relevant set, but after the ICA transformation computed on the elephant-set the variance of the non-relevant data is smaller as the variance of the relevant (see figure 5.13 down right).
- Using the colour feature the categories underthesea and sunrise show larger variances for the original data than for the remaining data set. This is fortified by the ICA transformation (see figure 5.13 first row).
- Based on the structure feature stands out, that the smaller variances of the unrated set in relation to the rated set after performing ICA occurs especially for the small user-defined subsets (see table C.7).
- Regarding the texture feature, the variances hardly change by the ICA transformation. The expected assumption holds that the variance of the non-rated set is larger than that of the used subset (see table C.7).

5.4.4 Summary

ICA based relevance feedback has been analysed regarding three items – the computed components, the underlying data distributions and the application of the transformation to the non-rated data.

The computed components are observed to be orthogonal in most cases. Therefore, the result of ICA resembles principal components. An outstanding enhancement compared to a PCA-based approach will not be given. However, such a relevance feedback should be possible in general.

Another attribute which is interesting regarding the components is the mutual information between two components. Thereby the statistical (in)dependence is measured and shown in most of the cases. Sometimes a few components are slightly depend. This may

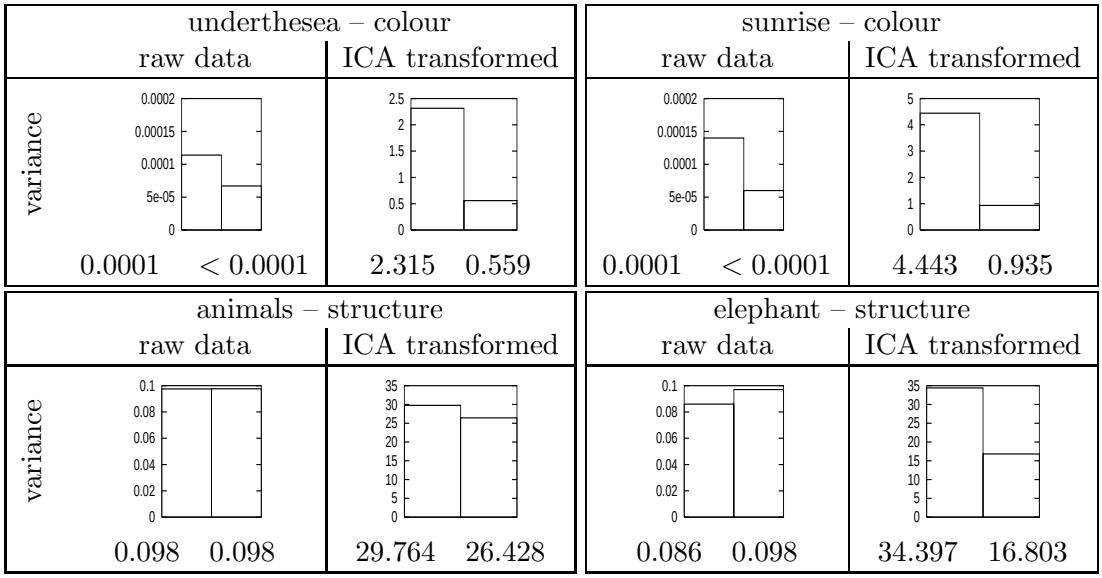


Figure 5.13: Variances of selected artexplosion categories and subsets before and after an ICA-transformation. The left bar of each graph represents the variance of the relevant rated subset and the right bar the variance of the remaining, non-relevant data. The values are quoted below each graph. Note that the ranges are optimised in each graph separately.

be a hint that the number of computed independent components can be smaller. Nevertheless, this do not affect the further usage of these components in a crucial manner. However, the structure feature and regarding colour changes of the myMondrian sequences the colour feature show a critical amount of dependence between the components. Thus, in these cases ICA is not suitable.

In order to analyse the suitability of the used ICA algorithm for the given data, the distributions are explored regarding their gaussianity. Since no Gaussian distributions are observed ICA in general may be feasible. Nevertheless, the required super-gaussianity is not guaranteed for all data sets. Especially the small user-defined sets offer sub-gaussian distributions. Comparing the prospects based on the sub-gaussian data with the observations on the ICA transformed data sets (see section 5.2.3), explains the results. The structure and the texture are not generally suitable for this approach. For the colour intensity only the artexplosion categories and the myMondrian sequences are tolerably practical. Thus an ICA based relevance feedback approach may be problematic since user ratings usually produce small trainingssets for ICA.

A further problem may be the reliability of the computed independent components. As Comon [1994] stated, the computation of independent components is inherently non-unique. At least a scale factor and a permutation of the components cannot be detected by the common approaches. To lessen this drawback Comon [1994] forces ICA to be unique by the demand of some requirements. The columns of the demixing matrix $\mathbf{W}$ should be of unit norm, the covariance of the observations $\mathbf{x}$ should be ordered decreasingly and the largest values of the $\mathbf{W}$ columns should be positive. Figure 5.14 shows that based on one synthetic data set different transformations may be computed.

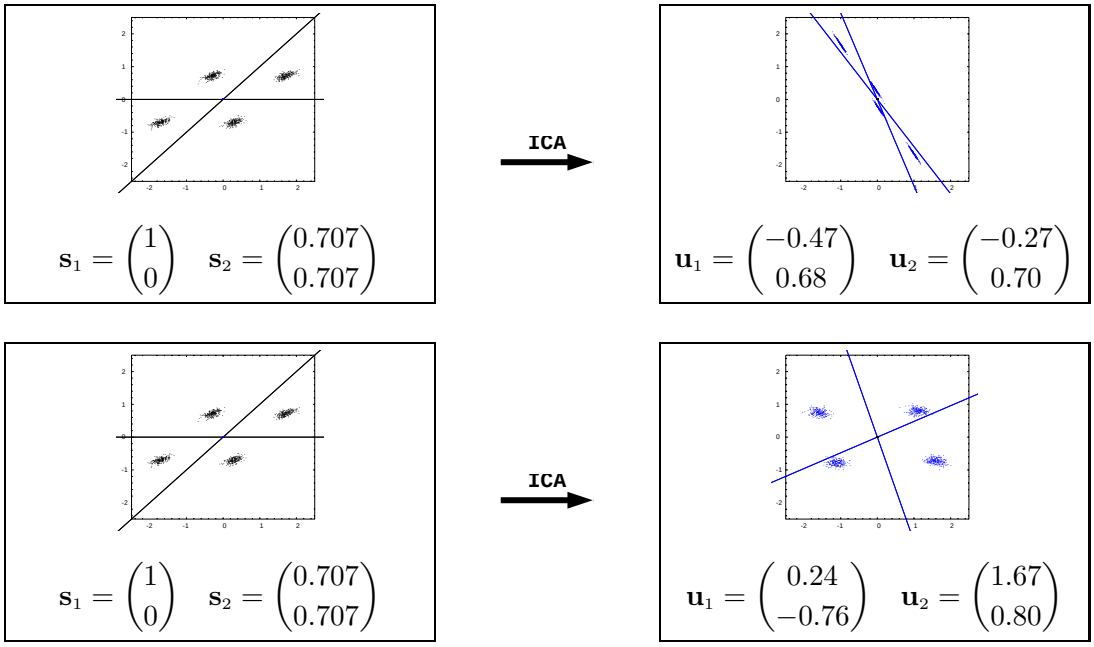


Figure 5.14: Analysis of the INFOMAX-ICA-approach based on the synthetic barbell data set. Since ICA is not reproducable the transformed data distinguish from each other although they are computed on the same data set.

The last point is the data transformation in general. ICA based transformations stretch the distributions of the data which should enhance the separability of the data. The compact distributions are proved by little changes in the variances of the texture data together with the orthogonal independent components of this data. However, the smaller variances of the not rated set after the transformation are conspicuous. Most clearly for the user defined subsets in the structure space holds the relation that the variance of the non-relevant data is clearly smaller than the variance of the relevant set. While this feature is observed to be suitable to detect interesting subsets (see section 3.3), this ability gets lost by squeezing all data points into the local distribution.

Summarising the observations regarding the structure feature shows, that this is suitable for image retrieval, but ICA based relevance feedback is not reasonable here. On the other hand the colour feature is good but less good than the structure feature for image retrieval. Here ICA based relevance feedback causes enhancements. Thus the suitability of different relevance feedback approaches depends on the used feature algorithms.

The suitability of ICA to enhance data distributions with respect to image retrieval tasks is neither substantiated nor disproved generally. The result of the transformation depends strongly on the given data distributions. Since each image set can be represented based on a number of different feature algorithms (see section 3.3) a representation can be chosen that supports an ICA computation. Furthermore, different ICA implementations exist. Depending on the given data the suitability of the algorithms may vary.

Chapter 6

CBIR Evaluation

In general, Moore's Law should be kept in mind: *An information system will not be used when it's more trouble than it is worth* [Moore, 1960]. Obviously suitable evaluations can help to reduce the trouble. Indeed CBIR is a very miscellaneous task. Thus the evaluation of such systems is miscellaneous, as well. Therefore a lot of different approaches and frameworks appeared. For example the performance of retrieval systems is analysed and the user acceptance is considered. In order to improve the various CBIR processing steps suitable evaluation setups have to be used. Numerous approaches are reviewed to motivate and support valid evaluations.

6.1 Motivation and Challenges

Why evaluate?

CBIR evaluation is motivated from different starting points. On the one hand, the presentations of new CBIR-systems require objective descriptions of their performances. Meaningful evaluations are essential to ensure improvements related to prior implementations. Therefore, a lot of publications include various analyses of their CBIR-systems, e.g. [Aslam and Savell, 2003] [Koskela et al., 2001a] [Müller et al., 2004] [Liu et al., 2001] [Black Jr et al., 2002].

On the other hand, CBIR is used as an evaluation tool itself. In the computer vision community different tasks are researched, e.g. feature detection and image segmentation. These approaches have to be evaluated with respect to an application. Thus image retrieval frameworks are used to show the benefit of these implementations, e.g. [Koskela et al., 2001a] [Heczko et al., 2000] [Sumengen and Manjunath, 2005] [Min et al., 2004] [Carson et al., 2002]. Using CBIR as an evaluation tool does not necessarily desire a further evaluation of the CBIR step but requires a standardised CBIR evaluation framework.

What to evaluate?

Three levels should be distinguished to define the evaluation object [J.J. Rocchio, 1971]: (1) the internal evaluation of a single system, (2) the external evaluation comparing different systems and (3) the evaluation of the real-life applicability and the user acceptance of a system.

(1) Concerning an **internal evaluation** (or *physical* performance evaluation [Santini, 2000]) every processing step can be analysed separately. The individual modules depend on the implementation. In terms of CBIR typical ones are feature detection, image segmentation, similarity measures or clustering (see the different modules presented in section 2.2).

Relevance feedback and retrieval performance affect the whole CBIR-system. Thus they have to be evaluated holistically. In doing so the situation should be constant, whereas the system changes. Performance and usability of a system depend on the image domain and the given data set. Therefore, a framework to analyse this is desired (see chapter 3).

(2) At least two different CBIR-systems are compared in an **external evaluation**. Based on a fixed data set and using a determined search task every attribute considered in an internal evaluation can be used for this purpose. Most of these components are exchangeable and independent from the surrounding system.

Flexibility and performance depend heavily on the complete system. To measure the performance basically the data set and the search task should be fixed. In [Santini, 2000] this is called *contextual evaluation*. Indeed external evaluations are performed seldom.

A system should be evaluated in different situations to analyse the flexibility and the generality. Data sets and queries should vary. This can be integrated in a holistic performance measure to compare different systems.

(3) On the most advanced level the **real-life applicability** and the **user acceptance** of a CBIR-system are evaluated. Therefore, system attributes concerning the user are important. First of all the consumer satisfaction should be measured, since this determines whether a system will be used. To reach this the user's need is relevant as well as the usability of the system. The last one depends heavily on the interface design. Therefore, rating attributes determined in the field of human-computer interaction are needed .

How to evaluate?

In general the merit of any approach or algorithm can only be judged in context of applications or concrete tasks. Furthermore the search task determines the evaluation strategy in a lot of situations. The search task determines the answer of an important question regarding the evaluation: Can a CBIR-system be evaluated automatically or are user experiments necessary? An automatic evaluation will support a quantitative measurement of the retrieval, whereas user experiments are more qualitative [Large et al., 2001].

A category search resembles a classification. Thus the same evaluation strategies can be used. Therewith an automated evaluation is possible and a lot of quantitative measures are available.

Quantitative measures support comparing different systems. Obviously this is more difficult based on user experiments. Experiments respect the user as the most important factor and the success of iterative image retrieval systems depends on the individual user. Therefore, user satisfaction is the evaluation target. Nevertheless, this is a qualitative measure and automated evaluations are not possible. User experiments are the state of the art for evaluating browsing.

On the other hand, in [Cox et al., 1996] is stated that both comparability and methodology lack in user experiments. Thus a qualitative evaluation is counterproductive. The gap between technical algorithms and unpredictable user behaviour occurs again. A sur-

vey of different user studies [Large et al., 2001] supports this thesis. User satisfaction does not correlate with performance, if anything it depends on the user experience. This is rated as a *decontextualised evaluation* [Santini, 2000].

The evaluation based on a target search is proposed to cover all retrieval situations [Cox et al., 1996]. Target search is rated as the most global search task. This means, if a system performs well in a target search, it probably will do so in other search tasks.

6.2 Performance Measures

The most holistic evaluation object regarding automated systems is the *performance*. Superficially considered, this describes how good a system operates or how far it acts as expected. Regarding image retrieval systems, this means *Does a system find the pictures it should find?*

Based on this question well known performance measures established in information retrieval are interesting. In the early 1960s the information retrieval community was faced with the same challenge as the image retrieval community in the 1990s: *How to evaluate and compare information retrieval approaches?* Thus, in the Cranfield Project [Cleverdon et al., 1966] *precision* and *recall* were proposed. The intention therein is to count the relevant documents in the set of retrieved documents. To get a meaningful measure, they are rated to the number of retrieved documents and the number of existing relevant documents, respectively[1].

$$\text{precision}(i) \;=\; Pr(i) \;=\; \frac{\#(\text{retrieved relevant images})}{\#(\text{retrieved images})} \;=\; \frac{N_i^+}{i} \tag{6.1}$$

$$\text{recall}(i) \;=\; Re(i) \;=\; \frac{\#(\text{retrieved relevant images})}{\#(\text{relevant images in database})} \;=\; \frac{N_i^+}{N^+} \tag{6.2}$$

with $\#(\Phi)$ determines the number of objects in a set Φ, N_i^+ is the number of relevant images within the first i retrieved images and N^+ the number of relevant images in the entire collection.

In [Baeza-Yates and Ribeiro-Neto, 1999] is stated that precision and recall depend on linear ordered result lists. This drawback is overcome by a definition of precision and recall based on sets presented in [Narasimhalu et al., 1997]:

$$Pr \;=\; \frac{||R_{\mathbf{q}}^E \cap R_{\mathbf{q}}^I||}{||R_{\mathbf{q}}^E||} \qquad Re \;=\; \frac{||R_{\mathbf{q}}^E \cap R_{\mathbf{q}}^I||}{||R_{\mathbf{q}}^I||} \tag{6.3}$$

where R_q^E is the set of objects relevant with respect to system E and $R_{\mathbf{q}}^I$ is the set of objects labelled as relevant in the ground truth set. $\mathbf{q}$ represents the query.

Precision and recall describe different qualities of a retrieval result. Furthermore, they act contrary. High precision often comes along with low relevance and vice versa. An extreme example is to retrieve all images in the collection. Then the recall value is one. However, the precision is very low and a meaningful retrieval is not performed. To set these measures into relationship usually both are presented in a precision-recall-diagram. Examples are given in figure 6.1.

[1] The definitions in this chapter do not distinguish between documents and images. Usually it is assumed that the data are images.

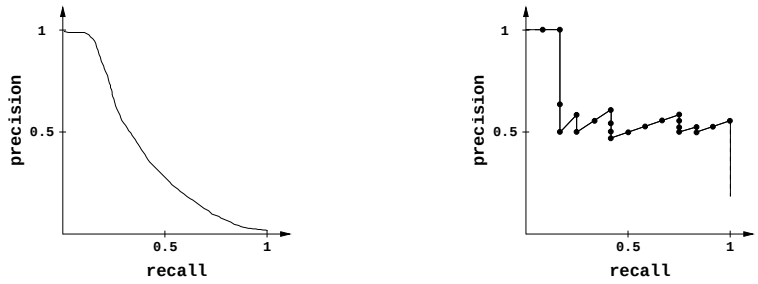


Figure 6.1: Example of a typical precision-recall-graph. With increasing recall usually the precision decreases. The right figure exemplifies a precision-recall-graph of a typical CBIR-system. Each dot represents a retrieved image of the ordered result list. Relevant images cause an amount of both recall and precision. This effects the upturn steps in the graph.

Until today precision-recall based measures are the most popular approaches to evaluate retrieval results. For example, they are used in the TREC conference[2]. To take into account different aspects of retrieval tasks and systems various precision-recall based measures are used:

- $Pr(10)$, $Pr(30)$, $Pr(N^+)$ (precision after the first $n = \{10, 30, N^+\}$ retrieved documents, where N^+ is the number of relevant documents in the collection)
- averaged precision (see equation 6.5)
- $Re(i)$ for $Pr(i) = 0.5$
- $Re(1000)$ (recall after 1000 retrieved documents)
- rank of first relevant document
- retrieval efficiency (see page 102)
- $Pr(i)$ and $Re(i)$-graphs where i is the number of retrieved documents

The last item gets special interest with respect to image retrieval. It considers that the number of relevant objects in the collection may be above the number of retrieved objects. High relevance values cannot be achieved in this situations. This often occurs in image retrieval. *Precision over scope* is used as a name for this measure [Rui and Huang, 2000].

Nevertheless, precision and recall measurements to evaluate CBIR-systems have some drawbacks. A lot of modified measures based on precision and recall are developed to handle them:

Precision and recall are batch-mode-measures

In [Large et al., 2001] as well as in [Baeza-Yates and Ribeiro-Neto, 1999] is noticed that the common precision-recall-graphs neglect any kind of interactivity. Thus relevance feedback impact is not recognised by this evaluation measure. To overcome this drawback, a scalar value for each retrieval step would be helpful. For example, the TREC competition requires some scalar values (see above). A well established measure is the *equivalence point* where

[2]At the TREC conference an information retrieval competition has been established for a number of years. See section 6.4.3 on page 113

$Pr(i) = Re(i)$. The intersection point of the bisecting line rates both values equivalently. However, it often fails when the number of retrieved objects is less than the number of relevant images in the collection.

A further scalar measure is developed with respect to the INDI system. Since predominantly the relevance feedback should be evaluated the measure should be compared over a number of successive iterations. It should be maximal if all retrieved images are relevant (in a category search this means the images are in the desired class) and minimal if no relevant image is retrieved. More relevant images should lead to a higher value.

Precision and recall are combined in one value by computing their product. Thereby the maximum value is used in each iteration step k to evaluate the order of the relevant images in the result list. The so called *maximum precision-recall* $pr(k)$ is calculated by:

$$pr(k) \;=\; \max\{Pr(k,i) \cdot Re(k,i)\}, \quad i = 1, \ldots N^{+} \tag{6.4}$$

where

$$Pr(k,i) \;=\; \frac{N^{+}_{k,i}}{i} \qquad \text{and} \qquad Re(k,i) \;=\; \frac{N^{+}_{k,i}}{N^{+}}$$

$N^{+}_{k,i}$ represents the number of relevant images retrieved in session k within the first i retrieved images and N^{+} specifies the number of relevant images in the database. This measure is used in the INDI based evaluation examples.

Precision and recall are computed for each query individually

CBIR-systems should be evaluated based on a number of different queries. Therefore, averaging measures are suitable:

In [Baeza-Yates and Ribeiro-Neto, 1999] the average of the precision at each recall level is proposed to evaluate an algorithm over all test queries:

$$\overline{Pr}(r) \;=\; \sum_{\mathbf{q}=1}^{N_q} \frac{Pr_{\mathbf{q}}(r)}{N_q} \tag{6.5}$$

where r is a recall level, N_q is the number of used queries and $Pr_{\mathbf{q}}(r)$ is the precision of query $\mathbf{q}$ at recall level r.

The R-precision RP averages the precision for R retrieved documents where R is the number or relevant documents in the collection. Usually this is normalised by the a priori precision [Rummukainen et al., 2003]:

$$p_c \cdot RP \quad \text{with } p_c = \text{a priori of class } c$$

$$\text{and} \quad RP \;=\; \frac{1}{R} \sum_{i=1}^{R} Pr(i)$$

i represents the number of retrieved images.

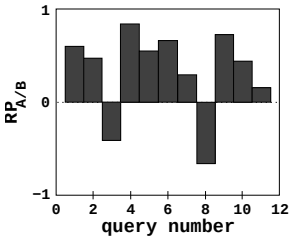


Figure 6.2: Example of a precision histogram with synthetic data. $RP_{A/B}$ is the difference of the R-precsions between system A and B regarding one query.

Inspecting precision histograms two algorithms can be compared based on a number of queries.

$$RP_{A/B}(\mathbf{q}) = RP_A(\mathbf{q}) - RP_B(\mathbf{q})$$

where $RP(\mathbf{q})$ is the R-precision of query $\mathbf{q}$. The values are presented as bar histograms, one bar for each query (see figure 6.2). Thus a fast visual inspection of two algorithms is possible.

Technical features like cost, time or interface handling are neglected

Precision and recall are called *incomplete* [Large et al., 2001] since technical and psychophysical aspects are ignored. For example, the response time is interesting. In [Bouteldja et al., 2006] the CPU time is used to analyse different retrieval strategies (a tree-structured nearest neighbour search is compared to a sphering retrieval).

The user satisfaction is neglected in precision-recall based performance evaluation, although the user is the most important factor. In [Large et al., 2001] is stated that the user acceptance of results with bad precision depends on the database magnitude. Moreover, in [Cox et al., 1996] is observed that a subjective success may cause an objective failure. Pictures of aircrafts are rated as interesting in the search for a picture with an eagle since in both cases a large part of the picture is covered with sky. Thus, the user should be involved while evaluating a system.

In [Baeza-Yates and Ribeiro-Neto, 1999] a user oriented measure is proposed:

$$\text{coverage} = \frac{|R_k|}{|U|} \qquad \text{novelty} = \frac{|R_u|}{|R_u| + |R_k|}$$

where R_u are the documents relevant, retrieved and **u**nknown to the user, R_k relevant, retrieved and **k**nown to the user, and U the a priori relevant objects.

General doubts on automatic evaluations are expressed in [Santini, 2000]. The author mistrusts the applicability of common experiment setups to technological systems. Experiments for testing a theory are important and well stated in natural science. Applied in technology and engineering they cause problems. Technology has to interact with its social environment. Experiments locked in a laboratory are impossible.

The importance of precision and recall may vary

The user often requires a set of specific images. Usually he does not know how many relevant images are in the collection. The demand of maximum recall supposes a detailed knowledge of the data [Baeza-Yates and Ribeiro-Neto, 1999]. Thus the user may be satisfied if an appropriate number of relevant images is found. The recall value of such a retrieval session is unimportant. However, false positives in the result set may bother and a high precision is desired.

In other situations the user may be interested in all images relevant for a specific query. Perhaps he accepts nonrelevant images in the result set if the number of false negatives is low. Here the recall is the important value and the precision may be ignored.

The importance of false negatives and false positives is addressed in [Cox et al., 1996]. The authors state that false negatives are worse than false positives since only target testing can check if the desired image is found. The rate of relevant pictures in the result set may be good although the pictures which should be found are missed. However, the detection of a complete set of images with respect to a specific query is not considered.

In general precison and recall should always be presented both since they measure different qualities. Furthermore, different retrieval situations require precision and recall differently. Thus in [Narasimhalu et al., 1997] a weighted combination of precision and recall is proposed as a quality measure:

$$Q \;=\; w_{Pr} \cdot Pr \;+\; w_{Re} \cdot Re \tag{6.6}$$

where w_{Pr} and w_{Re} are the weights.

More measures to weight precision and recall are proposed in [Baeza-Yates and Ribeiro-Neto, 1999]: The **harmonic mean**

$$F(j) \;=\; \frac{2}{\frac{1}{Re(j)} + \frac{1}{Pr(j)}} \tag{6.7}$$

assumes to get a high value only if both precision and recall are high. Detecting the max of the harmonic mean may result in the best possible retrieval approach compromising between precision and recall.

The **E-measure** is based on a user defined weight b to rate the importance of precision and recall:

$$E(j) \;=\; 1 \;-\; \frac{1 + b^2}{\frac{b^2}{Re(j)} + \frac{1}{Pr(j)}} \tag{6.8}$$

$b = 1$ is the inverse of the harmonic mean, $b < 1$ is used if recall is more important, and $b > 1$ if precision is more important.

Furthermore the computation of recall is often difficult [Large et al., 2001]. The number of relevant objects in the data collection is considered. Therefore the data set has to be known. For an internet search this is impractical. In addition the relevance of each object has to be defined a priori. This contradicts the flexibility of a system. The users intention may vary and an a priori labelling is not possible.

Although called *old-fashioned* [Draper et al., 1999] and a number of drawbacks are known, precision-recall-diagrams are still an important and often used measure. The popularity may be the most important reason for this. Researchers are familiar with it and can read it easily without further training. The selection of some quite up-to-date publications proves this observations:

- In [Hare et al., 2006] predominantly the semantic gap is discussed. However, feature detection and retrieval approach are evaluated based on precision-recall-graphs.
- In [Rao et al., 2002] an averaged (over all images in the database) precision-recall measure is used to show that the performance depends on the given image set.
- The precision over the number of retrieved images is used to compare dissimilarity measures in an image retrieval application in [Puzicha et al., 1999].
- In [Müller et al., 2003] an example of comparing two systems based on precision-recall is presented.
- Further different precision-recall derivatives (partly integrated above) are proposed in [Müller et al., 2001b].
- In recent years an image retrieval evaluation event was initiated that compares image retrieval systems based on mean average precision [Clough et al., 2005a].

Further performance measures are presented in [Müller et al., 2001b]. The rank of the best match is proposed as well as the average rank of relevant images. The error rate

$$\text{error} = \frac{\#(\text{retrieved non-relevant images})}{\#(\text{retrieved image})}$$

is interesting if false positives are very bothering.

The retrieval efficiency is a more complex measure. If less images are retrieved than relevant ones are in the database it is the precision. If more images are retrieved it is recall. This measure mixes two different well known measures. Hence, it is confusing.

In [Koskela et al., 2001a] a measure to evaluate single content descriptors is presented. Originally the τ-measure is used as an overall performance measure to analyse the entire retrieval process. In [Laaksonen et al., 2000] the suitability of different image features and their combinations is evaluated based on this measure.

The number of pictures presented to the user until all pictures of the desired category are retrieved is counted. They are weighted by the a priori probability of the category. The performed search task is a target search.

$$\tau \;=\; \text{number of images presented until the target is found} \tag{6.9}$$

In general the performance measures depend on testbed and query set [Rao et al., 2002]. And they are influenced by human subjectivity – at least in the labelling process of the relevant objects.

6.3 Internal Evaluation of Single Modules

While the entire CBIR-system may be evaluated based on a common performance measure, single system modules should be evaluated individually based on specific features. This may be completely independent from any image retrieval task. Nevertheless, the impact on the retrieval has to be analysed.

6.3.1 Evaluation of Feature Detection

Feature detection is an important step within CBIR (see section 3.3). Thus the algorithms have to be chosen carfully. However, the suitability of individual features strongly depends on the user behaviour. For example in user experiments it has been observed that a test person fails in a target search since he rates images based on a feature not implemented in the system [Cox et al., 1996]. The observed situation was a positive rating based on the shape of a flamingoe's neck whereas no shape feature was included. Indeed, few objective evaluation procedures exist to rate the suitability of feature detection algorithms concerning image retrieval.

One analysis of image features is presented in [Heczko et al., 2000]. Based on the relationship between effectiveness and efficiency different algorithms are compared. Therefore, the effectiveness is measured based on retrieval results, namely the rank of relevant images in a similarity list. The efficiency is quantified by the dimensionality of the feature vectors.

Texture features are reviewed in [Wagner, 1999]. To compare the different algorithms their performance is documented in *computation time* and *recognition rates* based on various texture image sets.

Both publications show that the evaluation of feature detection algorithms usually depends on a specific task. Regarding image search, this is obviously the performance of the retrieval step. Thus in [Laaksonen et al., 2000] the suitablility of different image features is evaluated based on the τ-measure (see equation 6.9). Originally, this measure is defined to analyse the whole retrieval process. Furthermore, the performance of individual features is described by the *observed probability*. This is defined as the probability of having objects of the same class as nearest neighbours. Although independent from a specific retrieval task this measure is tuned to the PicSOM system.

Usually feature detection is evaluated by the performance of the whole system. Thus a lot of evaluation measures take into account the special characteristics of the retrieval approaches. At the same time the suitability of features depends on the given image set. Thus they have to be selected based on the domain (see chapter 3).

6.3.2 Evaluation of Image Segmentation

The automatic segmentation of images often is evaluated visually. The developer or user looks at the segmentation results and verbalises a qualitative rating.

For example Blobworld [Carson et al., 2002] is a CBIR approach that depends on segmentation results. The main focus of their research is the a priori segmentation of images. Nevertheless, the evaluation of this step is performed visually. Segmentation errors are explained based on this qualitative evaluation. Furthermore, the segmentation is justified by comparing retrieval results based on the segmentation with those based on global colour histograms. A quantitative evaluation of the segmentation is not presented.

A further example for visual inspection is the presentation of a markov tree based segmentation approach designed for image retrieval tasks [Shaffrey et al., 2002b]. The segmentation is performed unsupervised. Thus new pictures can be inserted into a database without further segmentation tuning.

Segmentation algorithms usually are based on a number of parameters. In order to tune them, a performance measure is necessary. This requires a ground truth data set

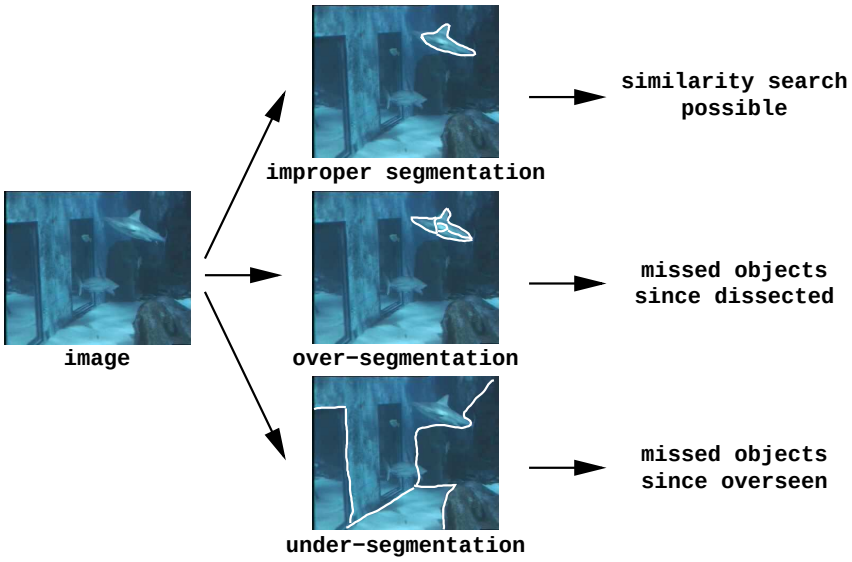


Figure 6.3: Different challenges of segmentation. While evaluating the segmentation, these should be considered. Only the segmentation borders representing the named challenge are shown.

to rate the individual segmentation approach. Furthermore, the segmentation should be compared with other algorithms. Therefore, further evaluation measures may be necessary.

A commonly used evaluation approach is based on the image retrieval performance to rate the segmentation steps. This is usually combined with a visual inspection. For example, the evaluation of the segmentation results in AQUISAR starts with a visual evaluation. The retrieval is based on a similarity computation of feature vectors. Actually just texture and colour are computed. No shape or contour detection is implemented in the framework. Thus approximated object borders are sufficient (see figure 6.3).

In other frameworks a classification task is used to analyse the segmentation step [Min et al., 2004]. The comparison of computed regions with the ground truth is based on different oberlap levels. The parameters of the suitable levels are detected by optimum classification results. This automatically tuned segmentation algorithm is compared with manually tuned parameters. Again, the comparison is based on correct classification rates.

In [Sumengen and Manjunath, 2005] a curvature based segmentation algorithm is proposed. The evaluation measure compares different curvatures. The overall evaluation is based on the harmonic mean F (see equation 6.7), which is usually used to analyse image retrieval performance.

While the evaluation measure used in [Sumengen and Manjunath, 2005] is based on image retrieval applications further objective measures have been proposed independent from retrieval tasks.

In [Mezaris et al., 2003] a segmentation evaluation approach based on a ground truth set is presented. The error measure computes the overlap of the detected regions with the ground truth segments. Depending on the distance to the original boundary of the segments each pixel assigned to the wrong region is weighted and counted. Over- and under-segmentation (see figure 6.3) are considered both.

The intention in [Unnikrishnan et al., 2005] is to compare a number of different segmentation algorithms. The resulting set of image segmentations is compared with a hand-labelled reference set. The comparison is based on the probability that a pixel has the same label in different segmentation results.

Another approach is proposed in [Shaffrey et al., 2002a]. A user experiment is performed to compare the results of different segmentation approaches. The subject has to select the segmentation result he prefers out of each pair of segmentations. Such psychophysical experiments are very fruitful. However, they are very expensive and time-consuming.

6.3.3 Evaluation of Relevance Feedback

The well known performance measures precision and recall are modified to scalar values (see section 6.2). Thus an improvement by succeeding relevance feedback iterations can be documented. For example in [Heesch and Rüger, 2003] the evaluation of relevance feedback is based on the average precision at different iteration steps. Nevertheless, further evaluation measures are desirable. A straight forward measure is based on a target search. A lower number of iterations required to get the target image exposes a better approach.

In [Müller et al., 2000b] and [Müller et al., 2000a] is proposed that an image browser benchmark should evaluate how far a CBIR-system narrows the semantic gap between low-level visual image features and high-level semantic. They propose an evaluation framework based on the number of relevant objects. Therefore, a ground truth data set is required. The influence of the human user is assumed to be important.

In general the relevance feedback performance indicates the flexibility of an image retrieval system. Such an adaptability to the user may be analysed based on the search path in the data space. However, this is a qualitative analysis and time-consuming. Further qualities interesting concerning the adaptiveness are parameter changes influenced by the relevance feedback and the adaptability to different image domains.

Summarised the evaluation of different modules important in a CBIR-system is often based on the retrieval performance. Furthermore they are tuned to the special qualities of the used CBIR-system. Hence, different steps are evaluated together. Predominantly this is done for interacting steps. Two evaluation tasks according the INDI-system exemplify this.

6.3.4 Evaluation of Region Based Ranking in INDI

Image retrieval research often starts with a search based on global image features. On the other hand, the retrieval for pictures showing one specific object is a popular search task. Since object detection is not solved in general today, a region based search may be a good workaround. In a relevance feedback framework this requires a region based ranking. Such a naive and rough approach is interesting for initial experiments regarding image retrieval using local image features.

Based on the INDI-system region based retrieval and rating is evaluated. The performed search is a target search. The task is named by an object covering one image region. The regions are built by a uniform 3×3 grid applied on the each picture.

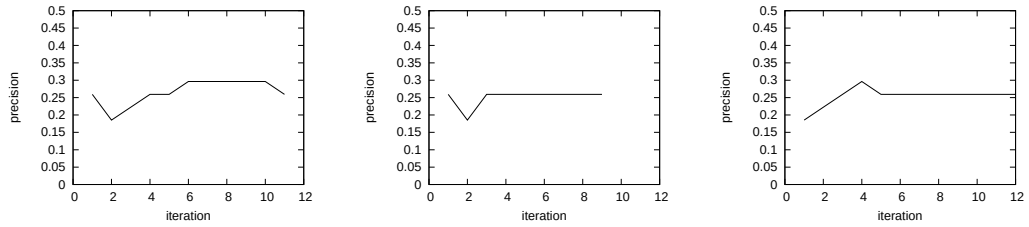


Figure 6.4: Precision over iteration for a flower search. Left: The adaptation is based on the entire picture. A region based user rating is performed. Centre: Region based adaptation. Right: Only global features are used.

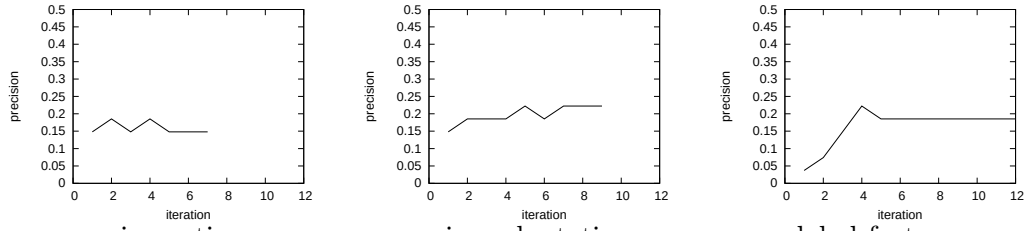


Figure 6.5: Precision over iteration for an underwater search. Left: The adaptation is based on the entire picture. A region based user rating is performed. Centre: Region based adaptation. Right: Only global features are used.

This setting fortifies different questions: Which image part does the user rate – the most interesting region or the entire image? Does the region based search improve the retrieval results? Which features should be utilised by the relevance feedback step – local features or global features?

Different versions of the INDI-system are implemented to analyse these tasks: (1) The image regions are used to compute the result list, whereas the relevance feedback is based on the global features. This is done to reduce the requirements to the user since the region selection requires more user interaction in the rating step. (2) Retrieval and relevance feedback are based on the image regions. This is the desired region based search. (3) To prove the usage of regions, this is compared with a retrieval approach without computing any local features.

The evaluation is based on precision and recall values. The precision (see figures 6.4 and 6.5) is documented separately since in the INDI-system false positives are worth than false negatives. Furthermore the maximum precision-recall $pr(i)$ for each iteration i is presented (see figures 6.6 and 6.7).

It could be observed that some search tasks are improved by the region based search. The underwater search (see figures 6.5 and 6.7) is one example for this. Others – like the flower search (see figures 6.4 and 6.6) – actually are better based on the entire images. Furthermore, the more complex rating of the image regions overstrains the patience of the users. In this situations they terminate the search after a few retrieval steps.

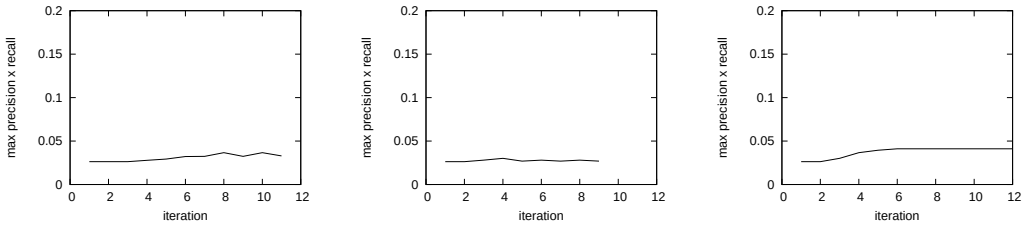


Figure 6.6: Maximum of (precision×recall) over iteration for a flower search. Left: The adaptation is based on the entire picture. A region based user rating is performed. Centre: Region based adaptation. Right: Only global features are used.

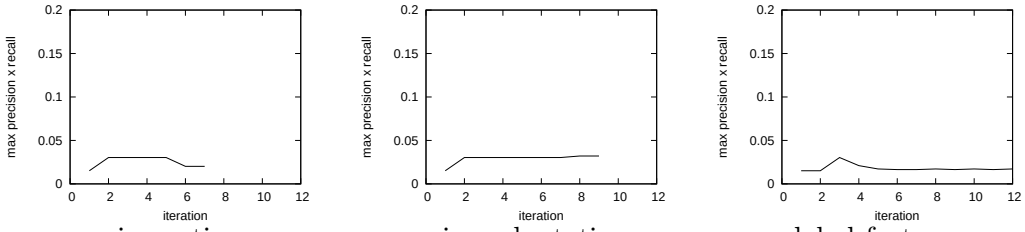


Figure 6.7: Maximum of (precision×recall) over iteration for an underwater search. Left: The adaptation is based on the entire picture. A region based user rating is performed. Centre: Region based adaptation. Right: Only global features are used.

In general this shows that different steps often depend on each other. A region based approach with relevance feedback is just suitable if the user interface fortifies a region based rating. Furthermore, it shows that the impact of an image segmentation (or more generally cutting into smaller patches) depends on the given image set.

6.3.5 Evaluation of the Weight Adaptation in INDI

Relevance feedback usually is evaluated based on a performance improvement or a comparison to other approaches. Nevertheless a stand-alone analysis of a relevance feedback implementation will result in interesting observations.

The weightadapation (see equation 2.3 on page 20) is an optional step for relevance feedback in the INDI image retrieval framework. This step can be evaluated with respect to its influence of the retrieval results or more precisely the retrieval performance. This is measured by the maximum precision-recall $pr(i)$ (see equation 6.4). Thus the value is plotted regarding the iteration step i to document the weight adaptation (see figures 6.8, 6.10 and 6.12).

A further question is, does the system reach a fixed state, e.g. the optimum for a specific query. This can be evaluated by the absolute weight changes in the single steps (see figures 6.9, 6.11 and 6.13).

Furthermore the specific attributes of the weight adaptation should be analysed. This may be in relation to the motivation of the weight adaptation step. The suitability of an image feature to detect relevant images is assumed to be different. At least it may depend on the query and the user intention. Thus different weight distributions are expected for different retrieval sessions. This can be observed by a simple plot (see figure 6.14).

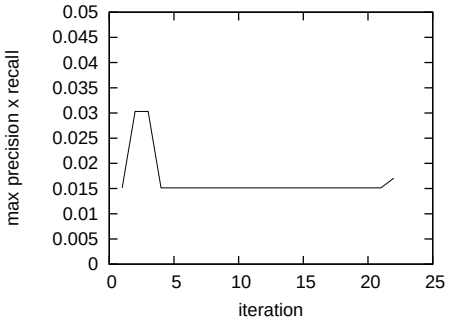


Figure 6.8: Maximum of precision×recall for an underwater search. Only positive ratings are used.

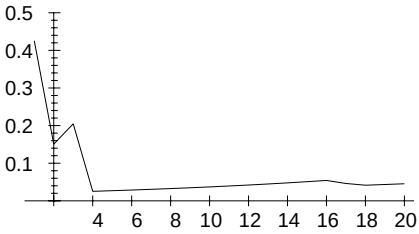


Figure 6.9: Absolute weight changes for a search based on positive ratings.

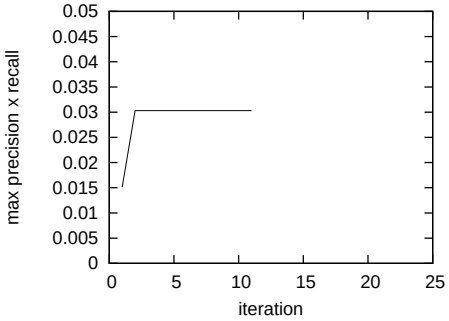


Figure 6.10: Maximum of precision×recall for an underwater search. Positive and negative ratings are used.

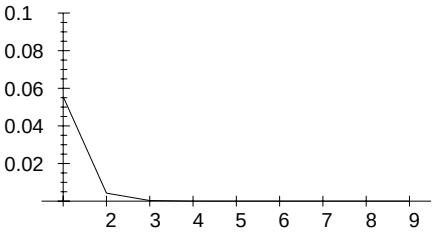


Figure 6.11: Absolute weight changes for a search based on positive and negative ratings.

Further aspects are the number and ranges of the utilised rating levels. Therefore, the measures presented above are computed for a relevance feedback using positive ratings only (figures 6.8 and 6.9), positive and negative ratings as + and – (figures 6.10 and 6.11) and a splitted strategy where the first five iterations use the positive ratings only and starting with the sixth iteration negative ratings are used as well (figure 6.12 and 6.13).

In general it has been observed that the retrieval result is improved by this relevance feedback approach. The weight changing depends on the rating strategy. A striking change is self-evident at that point where the rating is extended to negative ratings. Otherwise the weights reach fix values after a small number of iterations if both – negative and positive – ratings are used. Utilising positive ratings only causes recognisable weight changes for a lot of iterations. This causes a performance benefit after a long while. However, the user has to show a lot of patience. The different weight settings after a number of retrieval steps show that the suitability of different features depends on the given image set. An extended analysis based on a lot of feature sets may show whether some features are improper in general. The different rating strategies show that negative ratings are important. They gain the most effect if negative ratings are given after a number of positive ratings.

Thus different aspects conserning relevance feedback, feature detection and user interaction are supported by such experiments.

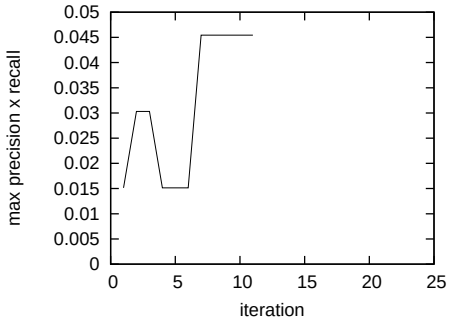


Figure 6.12: Maximum of precision×recall for an underwater search. 5 steps are rated only positive then positive and negative ratings are used.

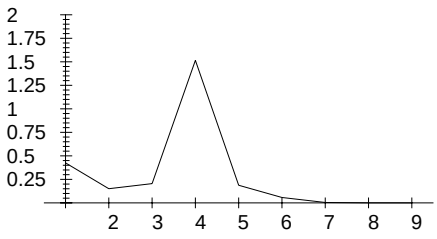


Figure 6.13: Absolute weight changes for a search beginning with positive ratings. Later on negative ratings are used, too.

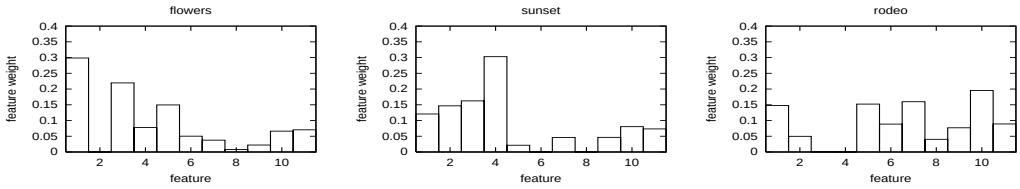


Figure 6.14: Feature weights after ten search steps of the INDI system. The order of the features from left to right is four structure, one texture and six colour features.

6.4 External Evaluation – Comparison of Systems

A very good external evaluation would be the comparison of a system with a human user, in [Santini, 2000] named as *decontextualised evaluation*. This requires extensive user experiments. Based on more automatic evaluation strategies two or more CBIR-systems may be comparable. To obtain a meaningful comparison it is essential that a suitable data set is used and the queries are defined. Thus this section begins with a deeper discussion on the ground truth data sets. Thereafter approaches and examples of comparing CBIR-systems are presented.

6.4.1 Defining Ground Truth Data Sets

A public set of images is an important prerequisite to support comparative evaluations of different CBIR-systems. This set should offer a labelling according to different semantic image sets. Thus a lot of research is based on the Corel Photo Collection [corel]. Unfortunately this set is not public and has been very expensive. Actually it is not obtainable. Added by the drawbacks presented in [Müller et al., 2002] this image set should be rated as improper as a common ground truth. Consequently, some groups start to collect images and offer them to establish a ground truth collection, e.g. the University of Washington [Shapiro]. However, non is established in the community today and image retrieval research still lacks of common database or benchmark.

Discussions on such ground truth data sets arose years ago with respect to information retrieval. The most important observations and proposals are based on a study of the British Library Research and Development Department [Jones and van Rijsbergen,

1976]. They proposed the today well establish *pooling* approach [Jones and van Rijsbergen, 1975] to get a ground truth labelling of a very large data collection. The requirements of information retrieval evaluation experiments are investigated, whereas the main focus is the underlying test data. In addition to the requirements some approaches to gain these attributes are developed. The user is the most important factor to get a labelled data set of an appropriate size. Unfortunately a hand-made labelling is too time-consuming and expensive. Imagine a database of 10,000 images. No human user can be asked to rate all images concerning any arbitrary query. To shorten this process it is proposed to built a *pool* of documents. Therefore, a query is put to each of the given retrieval systems and the set of all documents retrieved by at least one of the systems builds the pool. The remaining documents are assumed to be non relevant. The user has to rate only the documents in this pool. This approach is still used to get the relevance labelling of a large data collection, e.g. the retrieval competitions TREC and CLEF use it.

A similar approach is presented in [Aslam and Savell, 2003] as an evaluation framework without relevance judgement.

$$\mathrm{SysSim}(\mathrm{Sys}_1, \mathrm{Sys}_2) \;=\; \frac{\mathrm{Ret}_1 \cap \mathrm{Ret}_2}{\mathrm{Ret}_1 \cap \mathrm{Ret}_2} \tag{6.10}$$

with Ret_i is number of retrieved documents in system i. Similarities of retrieval results are assumed to be correct retrievals. The proposal corresponds to the pooling definition. Based on this measure [Aslam and Savell, 2003] observed that the common evaluation frameworks (if no good explicite relevance judgement is given) does not rate the systems performances. They rather rate the popularity.

In [Liu et al., 2001] different approaches for getting ground truth datas are summarised: (1) Synthetic images will reduce any noise or technical disruption. This approach is used especially for texture images. The human user is neglected. (2) The direct labelling is a straight-forward simulation of the human recognition. The drawbacks and pooling as an established solution are presented above. (3) As a further process to get a ground truth set is the keyword annotation mentioned. However, keywords are incomplete to represent images (see section 2.1.3).

6.4.2 Comparison of Systems

The comparison of two CBIR-systems is a straight forward approach to present the benefits of the individual ones. One situation where two CBIR-systems may be compared are presentations of individual (new) image retrieval approaches. The comparison with another, commonly known or at least publicated earlier system, will support the presentation of the individual benefits. The comparison with an earlier version of the same CBIR-system may be the most simple approach to expose the benefits of the new version. Few comparisons of different and independently developed systems are published:

GIFT/Viper vs. Histogram Intersection

At the University of Genf in addition to the CBIR-system GIFT/Viper [GIFT] an evaluation framework is developed. The main aspect is the communication protocol MRML to implement a client/server architecture. This approach should be established in an evaluation event called *benchathlon*. In [Müller et al., 2003] the suitablity of this evaluation framework is presented based on the comparison of Viper with a histogram intersection approach to retrieve images.

The used data set is the Washington image collection. For the comparison of the two retrieval approaches a number of performance measures (see section 6.2) are presented. In order to evaluate the relevance feedback, the precision is plotted for a number of iterations. Although it was just a secondary intention to show benefits of Viper this can be observed on the presented comparison.

PicSOM vs. GIFT/Viper

PicSOM should be compared with another CBIR-system [Rummukainen et al., 2003]. Therefore, it was modified to be able to communicate using MRML. Since GIFT/Viper is the only system using MRML this is used as the comparative systems. Therefore GIFT/Viper is used in a black-box fashion.

Computation time and storage requirements are compared marginally. Indeed PicSOM is clearly faster and requires less storage. For evaluating the performance a PicSOM specific measure is used. That is based on assumptions GIFT does not achieve. Namely GIFT presents single pictures repeatedly. Furthermore precision-recall-diagrams are used. Thus the images in each iteration step determines the performance. Overlap of the non-relevant images with the preceding iterations is acceptable although obviously not very user-satisfying.

No clear winner in the performance based on the six analysed classes could be elected. Nevertheless, some insights regarding the PicSOM system are gained. Thus classes where PicSOM lose (horses, planes, cars) are detected. The reason for that is still unknown. However, it becomes obvious that a variety in the used classes is important to get a fair benchmark. The same is true for different performance measures.

INDI vs. PicSOM

Starting from the INDI point of view a comparative evaluation is based on the *pr*-measure. To get comparable values at least the underlying image set and the search task have to be equal. Further variable things like the used image features or the rating levels for the relevance feedback are taken as system specific things. For deeper analyses they should be equal. Indeed a first rough evaluation compares the overall performance.

A set of example queries is used in the already used artexplosion image collection. The relevance feedback ratings are based on the predefined categories. Positive and negative ratings (as + and –) are used for *is in the same category as the query* and *is not in the same category*. Figure 6.15 presents an example plot of the averaged maximum precision-recall.

For the PicSOM system a better performance can be observed compared to the INDI system. INDI depends strongly on the initial configuration and the retrieved images build one cluster in the feature space. PicSOM retrieves images based on relevance values. Thus results of different map regions are possible and the retrieved images may come from different clusters. However more detailed insights are not possible here.

Further analyses are necessary if two systems are compared on a performance measure only since the systems have different striking attributes. PicSOM computes the relevance based on TR-SOMs (see section 2.2.1). INDI focus on the user interaction. To cover the special qualities of these systems usually they are evaluated in completely different ways: PicSOM presents convolution maps (see figure 6.16) to visualise the relevance landscapes. INDI arranges user experiments to analyse the interface (see figure 6.17).

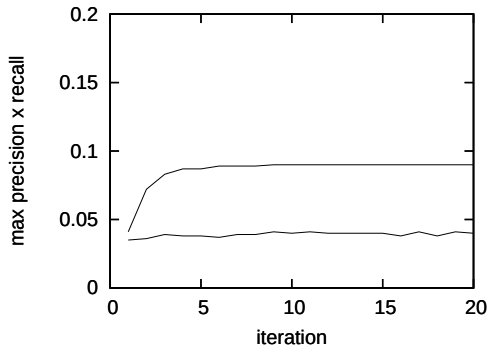


Figure 6.15: The averaged maximum precision-recall *pr* for an underwater category search with INDI (lower line) and PicSOM (upper line). PicSOM adapts better and outperforms INDI.

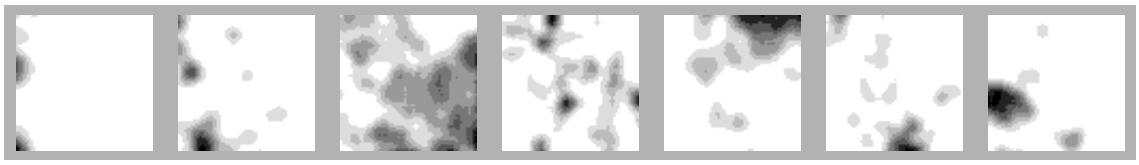


Figure 6.16: Convolved SOMs of the seven artexplosion categories (from left to right: underthesea, zoo, doorsandwindows, teddybears, sunrise, venezuela and iceland). Used feature is a colour layout feature. The dark map regions contain the relevant images with respect to the different categories.

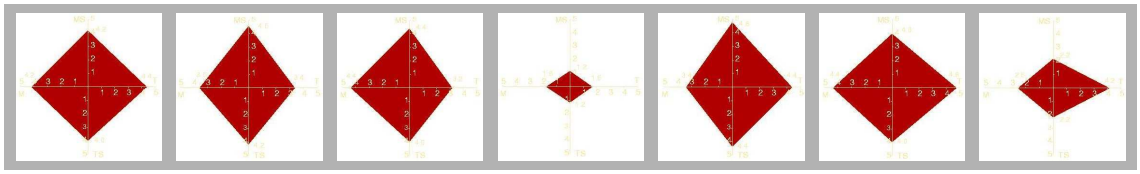


Figure 6.17: Evaluation of the user handling in INDI. Different interaction modalities are investigated regarding (from left to right) accomodation, efficiency, handling, nasty, fun, learn and patience. The interaction modalities are mouse+speech, touchscreen, touchscreen+speech and mouse. The extend along each axis represents how far the user has agreed the corresponding thesis in a questionair. (For details see [Bauckhage et al., 2003])

AQUISAR vs. INDI

The benefits of the image retrieval system AQUISAR (section 2.2.5) are exposed in comparison to the INDI system (section 2.2.4). Table 6.1 presents the precision of both systems for retrieving aquarium images based on the same query set. Figure 6.18 shows extracs of some retrieval results.

The striking advantage of AQUISAR is that it can retrieve images with similar entities from *different* webcam settings, i.e. angles of view. The results of INDI contain just images taken from *the same* angle of view as the reference image. This disadvantage of a conventional CBIR-system like INDI is rooted in the fact that the main part of each aquarium image is covered by the background. Therefore, the surrounding is dominant for calculating the result lists. The comparison shows that the used INDI release depends

	mean	variance	min	max
AQUISAR	0.65	0.04	0.25	1
INDI	0.48	0.04	0.13	0.88

Table 6.1: The achieved precision $Pr(i) = \frac{N_i^+}{i}$ is presented. A human user has rated the results and determined the number of interesting images N_i^+ for $i = 8$ retrieved images. For a comparison the same has been done for the CBIR-system INDI.

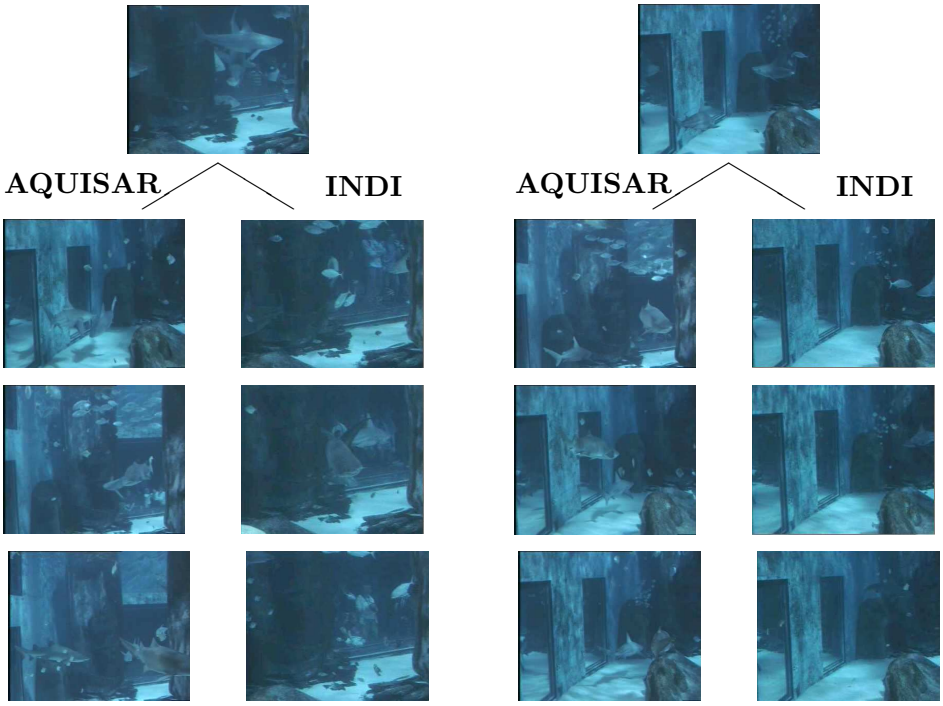


Figure 6.18: Retrieval results of AQUISAR and INDI.

strongly on the background of the images. This is overcome by the segmentation step of AQUISAR. Thus the intergration of a segmentation module into INDI is motivated by such a comparison. Indeed the relevance feedback, which INDI offers, is not used.

6.4.3 Image Retrieval Evaluation Events

The next step after comparing two CBIR-systems is to inforce evaluation events. In various research fields competitions to elect the most promising approaches are performed. Especially in text retrieval such a workshop has been established. From this well known TREC conference (Text REtrieval Conference [TREC]) emerged the TRECVID-workshop [TRECVID]. TREC evaluates text retrieval systems and video data offer a lot of textual data – speech and captions. Thus common text retrieval and text retrieval evaluation frameworks are used to evaluate video retrieval. Indeed TRECVID is no image retrieval contest.

In [Müller et al., 2004] an overview and discussion of different evaluation events is presented. At the time of that publication the image retrieval community still lack an established evaluation event. As a possible reason is stated that image retrieval systems do not perform well enough to be evaluated. However, comparative evaluations are necessary to improve them. A further problem for establishing such an event is that no ground

truth image data set has been established yet. Nevertheless, some promising approaches emerged in the last years:

MIRA

In the years 1996–1999 an EU project with the title *Evaluation Framework for Interactive Multimedia Information Retrieval Applications*[3] was realised. The results were proposed on a workshop on evaluation of multimedia information retrieval [Draper et al., 1999]. An important finding was that common evaluation frameworks developed in the second half of the 20th century are tuned to analyse static textretrieval systems. Current multimedia retrieval systems require new evaluation frameworks since they involve the user and are interactive. The evaluation task changed from matching concrete documents to user satisfaction. New system modules like user interfaces became important. The well know precision-recall measures are old-fashioned. The quantitative evaluations should be substituted or at least added by qualitative evaluations based on user experiments. An evaluation framework for interactive and multimedia information retrieval applications was proposed.

benchathlon

The benchathlon was proposed in the Internet Imaging Conference at SPIE West. Under this name different works emerged to implement an evaluation framework [Müller et al., 2003] [benchathlon]. Under the acronym BIRDS-I (Benchmark for Image Retrieval using Distributed Systems over the Internet) an initial suggestion was given at the Internet Imaging Conference 2000 to propagate client/server architectures for CBIR-systems [Gunther and Beretta, 2001]. A first contest was performed at Internet Imaging 2001. Müller et al. fortified the client/server idea. The XML-based communication protocol MRML (Multimedia Retrieval Markup Language) is introduced at SPIE Photonics East [Müller et al., 2000b].

Based on these initial works some sessions are performed in the following years [benchathlon]:
2000: Based on ample discussions the intention to establish a CBIR evaluation framework called benchathlon emerged.
2001: The first benchathlon event was performed. The contribution of BIRDS-I enables comparative evaluations.
2002: A number of contributions concerning to CBIR were submitted. Requirements for a standard CBIR benchmark were confirmed. As a concept the collection of data, software and image retrieval related publications were stated.
2003: Again a number of contributions related to image retrieval were submitted. The insight, that an annotated ground truth data set would be the most important thing to establish a standard CBIR benchmark was achieved.

Since 2003 no benchathlon competition has been performed. The most recent documentation published on the benchathlon webside[4] are from the Internet Imaging Conference 2003. The benchathlon seems to be droped off. But why?

[3] EU project number EP 20.039
[4] `www.benchathlon.net/events/index.html`

The benchathlon emerged from the desire of a standard CBIR benchmark. To perform this technical solutions have been proposed. Although conceptually intuitive and computationally simple the client/server architecture as well as the MRLM communication protocal have not prevailed. Research CBIR-systems are often implemented en bloc or emerge dynamically. Therefore, the adaptation to the requirements for participating the benchathlon competition is an obstacle. Thus no appreciable competition took place and the benchathlon session at the Internet Imaging Conference remains as a discussion panel.

ImageCLEF

The well known competition for text retrieval evaluation TREC have caused a number of spin-offs. One example is the Cross Language Evaluation Framework (CLEF). This workshop focusses on text retrieval using queries and documents in different languages.

Images are inherently independent from languages. Nevertheless, a lot of digitally stored images provide textual annotations. Retrieving images based on such captions requires a text retrieval. Since the images are independent from the language this text retrieval should be cross-language. Thus a workshop *Cross-Language Retrieval in Image Collections (ImageCLEF)* has been established in 2003 [Clough et al., 2004] and extended in the following years.

The focus of this workshop lie on an ad-hoc image retrieval of common photographs as well as on medical imaging. Combinations of textual and visual queries are supported. With respect to the medical images one task is the retrieval based on textual and visual features. A further task is the automated annotation of the images. This resembles a classification task [Clough et al., 2005a]. Since 2004 an interactive retrieval task is adjoined and automatic as well as manual relevance feedback is supported.

The main conclusion of the actually three performed and the one announced ImageCLEF is that the image retrieval community wellcomes an image retrieval evaluation event to compare and discuss actual developments and outcomes. Although the combination of textual and visual retrieval shows the best retrieval performance (measured in the mean averaged precision, see section 6.2) still a partitioning between researchers of both fields are observed.

In 2004/2005 ImageCLEF was the only image retrieval evaluation event [Clough et al., 2005b]. So, why seems the ImageCLEF to be successful whereas the benchathlon fails?

The ImageCLEF emerged from an established evaluation workshop. Therewith a lot of experiences of performing evaluation events were available. Thus the presented data were restricted to the images with the related captions and some relevance labelling by experts. The participants were invited to send their retrieval results and the annotation results, respectively. Therewith an evaluation can be performed quite easy and without any modifications of the given systems.

Furthermore the ImageCLEF is presented and announced at various events [Müller et al., 2005] [Müller et al., 2004] [Clough et al., 2004]. The results of each ImageCLEF competition are summarised and documented [Clough et al., 2005b] [Clough et al., 2005a] including some proposals and prospects for the coming event. Participants published their results and observations achieved at ImageCLEF. For example the FIRE system is evaluated based on ImageCLEF in 2004 [Deselaers et al., 2004b] and 2005 [Deselaers

et al., 2005]. The publications of the developments and conclusions based on these events propagate the visibility of the system as well as of the evaluation competition.

6.5 User Experiments

The user is the most important factor concerning image retrieval systems. He should be satisfied with the retrieval results. Thus, the evaluation process should incorporate him. Since the user cannot be simulated appropriately, user experiments are the most capable approach to measure his satisfaction. Only real life situations involving real user can rate the real life usage of a system.

Different aspects of image retrieval can be evaluated based on user experiments. As discussed in section 6.4.1 the definition of ground truth data sets is important but difficult. An automated labelling is not possible. Therefore pooling (see equation 6.4.1) has been established to reduce the rating amount for the user.

In [Santini, 2000] another approach is presented to get a user based ground truth. Instead of using the user to label a data set and evaluate the retrieval steps with respect to this labelling his behaviour is taken as the ground truth. Different implementations of a system are compared with the corresponding actions of a user. The discrimination of categories may be an obvious example for such a comparison with the user. In an experiment the subjects can group a set of images and these groups can be compared with different clustering or classification implementations.

A further task to investigate in user experiments is the user behaviour. The user has to interact with the system. For example the relevance feedback steps require a rating of the already retrieved images. Thus the ordinary user behaviour should be known. Patience and accuracy are characters important for a successful relevance feedback. The number of rated images in each step, the number of rating levels or the number of succeeding relevance feedback steps are important measures concerning the user rating in a relevance feedback approach.

Obviously the rating as well as the queries have to be entered. Therefore, the interface is important. Current technical developments offer different modalities to interact with automated systems. Thus these modalities should be analysed with respect to their user acceptance. Especially multimodal interfaces as the INDI system offers are interesting (see section 2.2.4 and figure 6.17).

A lot of assumptions are necessary to simplify the experiment setup. Otherwise the variety of possible observations and determinations cause a burst of the number of required test persons [Cox et al., 1996] [Santini, 2000]. Therewith the type of the users should be attended. Experience, age, education or gender will cause differences in the behaviour. Thus the realisation of an experiment would not be practicable. Possible assumptions may be that all persons act equally, only one feature determines the selection of an image, only features of presented images or the target are important or that the probablity to select a picture is a linear function of the image score.

The search task in user experiments usually is a target search. Therewith other search situations may be covered [Cox et al., 1996]. To simulate different user situations the target could be presented in different ways: (1) A continued presentation on the monitor

beside the retrieval interface will be an artificial situtation. If the target is available a retrieval would be unnecessary. However, this presentation will support the retrieval of the right target. (2) A short presentation at the beginning of the search session simulates the target picture in the memory of the user. This situation is more realistic. (3) The most realistic task covers the presentation of a distored copy of the target image. Thus a fading or vague memory of the target would be simulated.

The evaluation of user experiments is carried out subjectively as well as objectively. A lot of user depending objects are necessarily subjective. These could be quantified based on questionaires [Large et al., 2001]. The answers on such usability questions may be presented in glyphs as figure 6.17.

An objective evaluation is based on transaction logs [Müller et al., 2001b] to measure the performance of real user sessions. Quantitative values like the amount of time, the number of interaction or the number of relevance feedback steps could be used to get an objective evaluation measure for user experiments. Especially the taken time may be a good measure to compare different systems based on user experiments. Other measures depend on the system design, e.g. the number of interactions can be measured in mouse klicks, retrieval steps or the number of displayed pictures.

User experiments are a powerful evaluation tool. Indeed they are time consuming. Thus the number of publications presenting user experiments in image retrieval tasks undercut that of automatic evaluations. Usually extended requirements or proposals to benefit from user observations are expressed. For example in [Santini, 2000] a framework to exploit the user to get the ground truth is proposed. User experiments are used as a *measurement device* for visual information systems. The presented evaluation examples are the comparisons of different similarity measures and models with the similarity ratings of the user. Obviously such a setup has to be developed and converted to the individual evaluation task. Furthermore, they are very time consuming. For example, the presented experiments requires repetitions with the same persons after two weeks.

In [Black Jr et al., 2002] user experiments are presented to get a ground truth for similarity ranking. The computed similarity value is compared with the user ranking.

The PicHunter CBIR-system [Cox et al., 1996] is evaluated performing extensive psychophysical experiments [Papathomas et al., 1998]. The performance is evaluated by image comparisons. The user has to mark which image is more similar to a target. Based on such a rating the retrieval results can be evaluated. Furthermore, different versions of the system are compared based on user experiments.

In [Large et al., 2001] different user study findings are surveyed. An important observation is that user satisfaction and system performance does not correlate. If anything the performance depends on the user experience.

In an early study the user's need has been analysed [Armitage and Enser, 1997]. Therefore, query formulations were collected from different picture archives. These intuitive formulations were analysed respective different search tasks.

Furthermore, single steps of an image retrieval could be evaluated based on user experiments. In [Shaffrey et al., 2002a] image segmentation for image retrieval applications is evaluted in this way. The users should compare different segmentation results to find the most suitable algorithm with respect to the given images.

6.6 Summary of CBIR Evaluation

Various aspects of evaluating CBIR-systems are reviewed. This results in some outstanding observations:

- Precision-recall based performance measures are still very popular for retrieval evaluations.
- Different retrieval approaches often require specialised evaluation frameworks.
- In order to enforce comparative evaluations of CBIR-systems, an evaluation competition is desired.
- User experiments are desirable but expensive.

In general, the evaluation of CBIR approaches is a challenging task. A universal evaluation framework is not possible and evaluation guidelines are helpful to design an evaluation setup but have to be adapted to the respective CBIR-system.

Chapter 7

Summary and Outlook

Content-based image retrieval is a broad field of research, incorporating various uptodate challenges. Thereby different research disciplines are important and influence the new developments. For example, computer vision provides algorithms to describe images, information retrieval offers methods for indexing and searching data, machine learning presents adaptable approaches and psychology analyses the user behaviour. Based on these varying approaches one issue emerges again and again: The *semantic gap* between the human based semantic interpretation and the technical description of image contents. Approaches to narrow this are desired. Therefore, systems have to adapt to the user and machine learning methods can be helpful.

In order to develop such systems, two starting points are possible: (1) Based on psychological research human behaviour can be analysed. With those insights the user's intentions and behaviours can be described. The goal would be a mathematical description of human behaviour to implement in automated systems. (2) Such mathematical definitions and algorithmic descriptions present another possible starting point. Based on established automated and stand alone approaches modifications are desired to satisfy the user's need.

Summary

Most of the popular CBIR-systems follow the second way. So does the thread of this work. Starting with an extended overview of information and image retrieval research various challenges regarding CBIR are outlined. In general, various questions embrace the CBIR challenges and are discussed in this work, namely computer vision insights to describe images based on low-level features, learning approaches to adapt a CBIR-system to the user and the evaluation of the different retrieval steps. Different systems are reviewed in this work. Furthermore, two approaches are developed: The INDI-system representing a system focussed on the user interaction and the AQUISAR-framework developed to analyse approaches regarding image retrieval in webcam setups.

Developing from automatic approaches to human like behaviour the description of images is the reasonable starting point. Low-level features are analysed depending on various image domains (a photo collection, synthetic image sequences and a set of aquarium web-cam images). The suitability of the feature detection algorithms differs depending on the used images. Furthermore, the semantic is not covered by those image features. Thus improvements are required and advanced adaptable processing steps are motivated.

One possible step to realise human like behaviour regarding image collections is to align them in a sequence. Common situations to use this are slide shows. Numerous CBIR-systems perform the presentation of retrieved pictures sequentially. Therefore, one dimensional Self-Organising Maps (1dSOMs) are analysed with respect to image sequences.

1dSOM is suitable to align image sequences, synthetic ones as well as real world image sequences. Especially regarding the difficult task of arranging webcam images of an aquarium the gives helpful impact to assist the user. In general, the 1dSOM based image arrangements resemble the human interpretation. Furthermore, pictures can be grouped based on 1dSOM if they are element of the same sequence. This can be used for *shot detection* and hence be helpful in video retrieval.

Usually the given approaches to retrieve images are successful on a certain level. Indeed they often do not satisfy the user's need. Thus the systems have to be tuned to approximate the user's intention. Common approaches to implement such user adaptation are based on *relevance feedback*. Therefore, the users have to rate the images a system has retrieved. Utilising these rates the system is trained to resembles the human way of comparing images.

Various approaches to support this are introduced. The relevance of images is related to their interestingness, similarity models are presented and different methods to perform the relevance feedback are discussed. One approach to approximate the human recognition of images is to transform the data space. Therefore the most suitable directions of the data space should be detected. This motivates to use an independent component analysis (ICA), which is introduced and implemented to transform the given data sets. The used ICA algorithm is based on the INFOMAX approach.

Therewith specific attributes of relevant data sets are computed, namely the independent components to represent important directions within the data spaces. Based on these components the image collections are transformed and the new data distributions are analysed with respect to relevant and non-relevant image groups.

Furthermore, ICA is used to enhance a Bayes Classifier. Since a category search can be implemented as a classification this is absorbing regarding image retrieval applications. The density estimation is improved by ICA to get the statistically independent directions a Naive Bayes Classifier rely on. This icaNbayes classification is introduced and used on a synthetic data set. Additionally it is applied on the image collections.

ICA data space transformation as well as the icaNbayes classification of the given data sets are feasible but not satisfactory. Therefore, ICA computation on these data sets is analysed more detailed.

At first the computed independent components were analysed. They do not achieve the expected attributes. Namely the main directions are orthogonal to each other. Thus they do not give any improvement compared to a common PCA. Indeed the application of ICA depends on the used features. For some features it is suitable, for others not. Therefore, the distributions of the different feature data are analysed. The used INFOMAX approach requires data sets satisfying defined distributions. In general holds that this algorithm fails for data which is not super-gaussian.

Based on the relevance feedback applications ICA is computed on a small subset. The resulting transformation is applied on the entire data set. Thus ICA input differs from the transformed data. This cause undesired effects on the differentiation of the relevant sets from the non-relevant sets, namely the differentiation worsens. However, computed on the

same set as is transformed ICA causes improvements regarding the data distributions.

Adapting to a user means that the automated approaches have to be evaluated with respect to human users. Furthermore, even the evaluation of retrieval systems based on objective and quantitative measures is important but difficult. Different frameworks are reviewed. The challenging observation is that although required by many researchers today no common evaluation framework is established. Various CBIR competition events were proposed and initiated but few competitions were proceeded.

Furthermore, comparative evaluations based on CBIR performance are important since CBIR performance is often used to evaluate single computer vision approaches.

Outlook

CBIR-systems incorporate numerous different approaches. In many cases various independent modules perform individual processing steps. Thus a lot of starting points for future works exist.

Based on the observations in this work such a research task can be the development of advance image features. Tuned to restricted image sets sophisticated feature detection algorithms would be interesting to represent domain specific attributes. The goal may be to implement semantic based image descriptions. Motivated from text retrieval and presentations at the ImageCLEF workshop [Clough et al., 2004] hybrid systems may be prosperous to enhance image retrieval. The combination of textual descriptions with content-based features may approximate semantic based image retrieval. Additionally psychological insights regarding user behaviour may be prosperous to improve user friendly image representations.

The 1dSOM alignment of images can be applied to various image retrieval tasks. Namely the detection of video sequences or the development of a content-based movie retrieval system can be based on the 1dSOM image alignment. Therefore, the grouping along the sequence may be the starting point to retrieve meaningful image sequences. Then deeper levels of a tree-structured SOM may be used to represent the content of different sequence episodes.

Furthermore, the relevance feedback offers numerous challenges for subsequent research. For example the data space transformation to approximate the human perception of images may be analysed further on. Especially the implementation in more complex CBIR-frameworks and the usability in real-world situations have to be evaluated. Starting points may be other transformation algorithms, e.g. different approaches to implement the independent component analysis.

In general, the evaluation of CBIR-systems and image retrieval implementations is still important.

Appendix A

myMondrian Sequences

name	**sequence**	N	**dim**	**class**	**alteration**
defined move	1	20	100×100	1	fixed step size 15 right, 10 up (or the other way round)
	2	20	100×100	1	
	3	20	100×100	1	
	4	20	100×100	1	
	5	20	100×100	1	
var move	6	20	100×100	2	fixed direction, variable step size and extension
	7	20	100×100	2	
	8	20	100×100	2	
growing	9	20	100×100	3	just the extension changes
	10	20	100×100	3	
	11	20	100×100	3	
var move	12	25	100×100	2	1 rectangle, fixed moving
	13	20	20×20	2	
colour change	14	20	20×20	4	1 square, just extension
	15	50	100×100	4	1 square, just blue changes
	16	50	100×100	4	1 square, just green changes
	17	50	100×100	4	1 square, just red changes
	18	50	100×100	4	1 square, just rgb changes
textured back	19	20	100×100	5	1 rectangle, fixed direction and step, no extension, coloured background
	20	20	100×100	5	
	21	20	100×100	5	
	22	20	100×100	5	
	23	18	100×100	5	

Table A.1: Details of the generated myMondrian image sequences.

Appendix B

1dSOM Parameters and Results

experiment	N_s	ϵ_{init}	ϵ_{final}	η_{init}	η_{final}	steps
exp 1	50	0.9	0.01	8	1	10000
exp 2 (a)	50	0.9	0.01	12	1	10000
exp 2 (b)	25	0.9	0.01	8	1	10000
exp 3 - aquarium	200	0.9	0.01	30	1	100000
exp 3 - artexplosion_30	30	0.9	0.01	30	1	100000
exp 3 - artexplosion_100	100	0.9	0.01	30	1	100000
exp 3 - artexplosion_500	500	0.9	0.01	50	1	100000

Table B.1: Parameters of the 1dSOM experiments.

domain	*o*-measure													
	HistColour	HistBlue	HistGreen	HistRed	HistIntensity	StructHS	StructIntens	StructHue	StructSat	UnsersTexture	HistHue	HistLig	HistSat	HistQuant
defined move	0.37	1.00	0.37	0.42	0.42	0.89	1.00	0.95	0.84	0.21	0.37	0.47	0.42	0.26
	0.21	0.16	0.21	0.37	0.16	0.95	0.95	0.68	1.00	0.42	0.16	0.16	0.16	0.26
	0.21	0.32	0.21	0.32	0.42	0.84	0.89	0.58	0.74	0.21	0.26	0.26	0.26	0.26
	0.68	0.63	0.63	0.68	0.47	0.74	0.84	0.53	0.74	0.42	0.63	0.58	0.58	0.53
	0.53	0.37	0.58	0.58	0.53	0.84	0.74	0.84	0.68	0.16	0.68	0.42	0.42	0.53
var move	1.00	0.68	0.79	0.84	0.58	1.00	1.00	1.00	1.00	0.37	0.63	0.58	0.58	0.68
	1.00	0.95	0.79	0.79	0.68	1.00	0.95	0.53	1.00	0.53	0.58	0.58	0.42	0.89
	1.00	1.00	1.00	1.00	0.84	1.00	1.00	0.84	1.00	0.42	0.84	0.84	0.84	1.00
growing	1.00	1.00	1.00	1.00	1.00	0.95	1.00	0.84	0.95	0.95	1.00	1.00	1.00	1.00
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.79	1.00	0.58	0.89	0.84	0.79	1.00
	1.00	1.00	1.00	1.00	1.00	0.95	1.00	0.95	0.89	0.68	1.00	1.00	1.00	1.00
var move	0.04	0.04	0.04	0.04	0.04	0.92	1.00	0.54	1.00	0.17	0.04	0.04	0.08	0.04
	0.26	0.21	0.21	0.32	0.32	0.95	0.95	0.63	0.95	0.37	0.26	0.37	0.21	0.21
colour change	0.89	0.53	0.95	0.84	0.74	0.84	0.95	0.53	0.89	0.95	0.53	0.84	0.74	0.84
	0.76	0.27	0.16	0.16	0.27	0.18	0.12	0.20	0.31	0.63	0.45	0.39	0.22	0.43
	0.47	0.18	0.10	0.22	0.39	0.37	0.12	0.37	0.10	0.53	0.41	0.37	0.20	0.59
	0.76	0.24	0.29	0.37	0.84	0.37	0.33	0.33	0.24	0.76	0.76	0.65	0.39	0.78
	0.61	0.45	0.37	0.39	0.31	0.47	0.29	0.55	0.29	0.61	0.55	0.24	0.39	0.59
textured back	0.89	0.84	0.84	0.84	0.89	1.00	0.95	0.95	1.00	0.47	0.79	1.00	0.84	1.00
	1.00	0.84	0.74	0.79	0.79	1.00	1.00	1.00	1.00	0.26	0.79	0.74	0.74	1.00
	0.95	0.68	0.68	0.74	0.68	1.00	0.89	0.89	1.00	0.37	0.79	0.79	0.84	1.00
	1.00	0.58	0.74	0.58	0.74	1.00	1.00	1.00	1.00	0.37	0.79	0.84	0.53	0.89
	0.88	0.82	0.76	0.82	0.82	1.00	1.00	1.00	1.00	0.29	0.88	0.82	0.76	1.00

Table B.2: Evaluation of 1dSOM experiment 1: The alignment of the myMondrian sequences is analysed based on the *o*-measure (see equation 4.4).

Appendix C

ICA – Data and Results

category	N_c	subset 1	N_1	subset 2	N_2	subset 3	N_3
underthesea	300	fish	10	swarm	7	diver	17
animals	300	elephant	21	monkey	33	lion	25
doorswindows	300	storefront	45	church	18	ruin	9
teddybears	100	one bear	17	two bears	57	more bears	26
sunrisesunset	300	round sun	51	yellow Sky	29	skyline	17
venezuela	100	one person	11	one building	10	coastline	4
iceland	99	ship	3	horses	3	seaside	8

Table C.1: User defined subsets of the artexplosion image collection.

data	μ	$\mathbf{r}_1$	σ_1	$\mathbf{r}_2$	σ_2	$\mathbf{r}_3$	σ_3
helix_1	$(14,15,17)^T$	$(1,1,1)^T$	2π	$(1,1,1)^T$	2π	$(2,0.5,0)^T$	10
	$(14,15,17)^T$	$(1,1,1)^T$	2π	$(1,1,1)^T$	$\frac{1}{2}\pi$	$(2,0.5,0)^T$	10
	$(17,20,19)^T$	$(1,-1,1)^T$	2π	$(1,-1,1)^T$	$\frac{1}{2}\pi$	$(2,0.5,0)^T$	10
helix_2	$(14,15,17)^T$	$(1,1,1)^T$	$\frac{1}{2}\pi$	$(1,1,1)^T$	2π	$(2,0.5,0)^T$	10
	$(14,15,17)^T$	$(1,1,1)^T$	2π	$(1,1,1)^T$	$\frac{1}{2}\pi$	$(2,0.5,0)^T$	10
	$(17,20,19)^T$	$(1,-1,1)^T$	2π	$(1,-1,1)^T$	$\frac{1}{2}\pi$	$(2,0.5,0)^T$	10
helix_3	$(16,17,17)^T$	$(1,1,1)^T$	$\frac{1}{2}\pi$	$(1,1,1)^T$	2π	$(2,0.5,0)^T$	10
	$(14,15,17)^T$	$(1,1,1)^T$	2π	$(1,1,1)^T$	$\frac{1}{2}\pi$	$(2,0.5,0)^T$	10
	$(15,16,17)^T$	$(1,-1,1)^T$	2π	$(1,-1,1)^T$	2π	$(2,0.5,0)^T$	10
blobs	$(3,4,5)^T$	$(-0.5,-1.3,1)^T$	17	$(5,20,-1.5)^T$	8	$(1,1,1)^T$	1
	$(5,6,7)^T$	$(-0.5,-1.3,1)^T$	17	$(5,20,-1.5)^T$	8	$(1,1,1)^T$	1

Table C.2: Construction parameter of the helix and the blobs data sets. Each row represents one class of the respective data set.

–	2	3	4	5	6	7	8	9	10
1	-0.0002	0.0014	0.0004	-0.0000	-0.0000	0.0000	-0.0000	-0.0000	0.0000
2	–	-0.0000	-0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
3	–	–	0.0000	-0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
4	–	–	–	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	–	–	–	–	0.0000	0.0000	-0.0000	0.0000	-0.0000
6	–	–	–	–	–	0.0000	0.0000	0.0000	0.0000
7	–	–	–	–	–	–	0.0000	0.0000	0.0000
8	–	–	–	–	–	–	–	0.0000	-0.0000
9	–	–	–	–	–	–	–	–	0.0000

Table C.3: Mutual information of myMondrian dataset 4 – HistIntensity feature space

–	2	3	4	5	6	7	8	9	10
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	–	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	–	–	0.0000	0.0000	-0.0000	-0.0000	0.0000	0.0000	-0.0000
4	–	–	–	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	–	–	–	–	0.0000	0.0000	0.0000	0.0000	0.0000
6	–	–	–	–	–	-0.0000	0.0000	0.0000	-0.0000
7	–	–	–	–	–	–	0.0000	0.0000	0.0000
8	–	–	–	–	–	–	–	0.0000	0.0000
9	–	–	–	–	–	–	–	–	0.0000

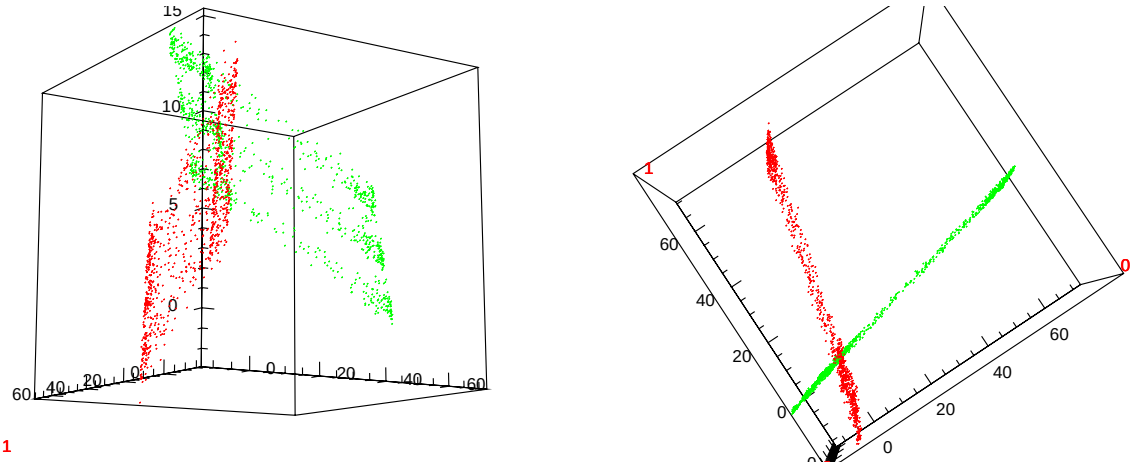
Table C.4: Mutual information of myMondrian dataset 3 – UnsersTexture feature space

–	2	3	4	5	6	7	8	9	10
1	0.0109	0.6291	0.0024	0.0018	0.0001	-0.0002	0.0011	0.0001	0.0003
2	–	0.0108	0.0001	-0.0003	0.0000	-0.0000	-0.0001	0.0000	0.0001
3	–	–	0.0031	0.0029	0.0002	-0.0002	0.0017	0.0001	0.0003
4	–	–	–	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
5	–	–	–	–	-0.0000	0.0000	0.0001	0.0000	-0.0000
6	–	–	–	–	–	0.0000	-0.0000	0.0000	0.0000
7	–	–	–	–	–	–	0.0000	0.0000	-0.0000
8	–	–	–	–	–	–	–	0.0000	-0.0000
9	–	–	–	–	–	–	–	–	0.0000

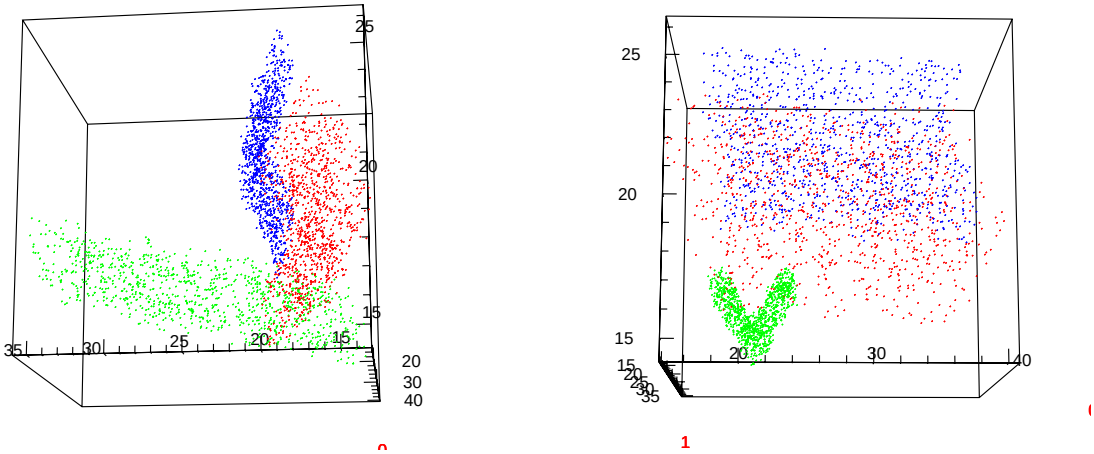
Table C.5: Mutual information of myMondrian dataset 7 – HistIntensity feature space

–	2	3	4	5	6	7	8	9	10
1	8.4871	4.2961	4.2393	4.9096	11.1845	7.0654	11.8903	2.5192	0.1936
2	–	3.0091	2.9219	4.4596	10.3575	6.2260	8.7273	1.7474	0.1696
3	–	–	3.7195	2.4352	3.7102	2.8413	4.1292	1.1480	0.0680
4	–	–	–	2.3062	3.7643	2.6254	4.1241	1.0651	0.0643
5	–	–	–	–	5.6044	4.2487	5.0623	1.7366	0.2205
6	–	–	–	–	–	7.8816	11.4588	2.3354	0.2610
7	–	–	–	–	–	–	7.1908	1.8584	0.2335
8	–	–	–	–	–	–	–	2.5068	0.2280
9	–	–	–	–	–	–	–	–	0.0608

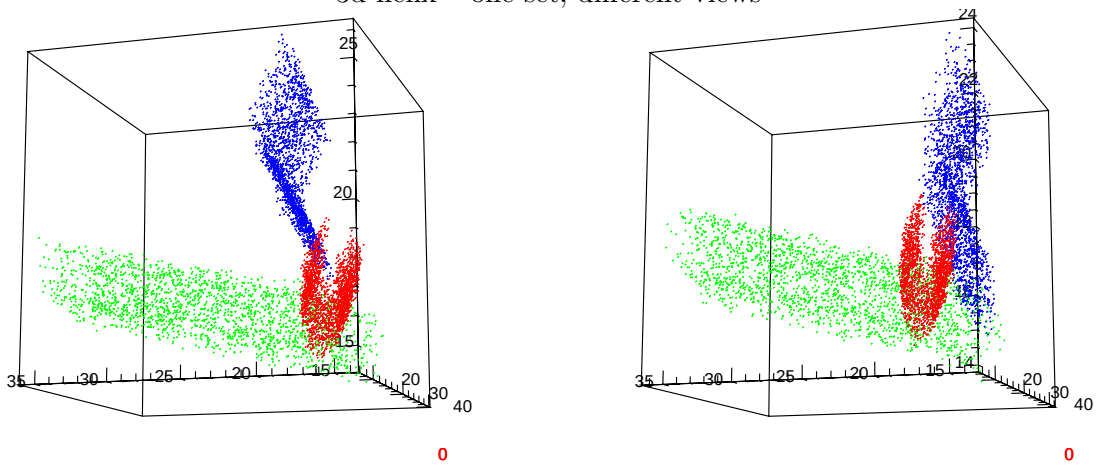
Table C.6: Mutual information of myMondrian dataset 25 – StructIHS feature space



3d correlated blobs – one set different views



3d helix – one set, different views



3d helix – two sets

Figure C.1: Synthetic Data Sets.

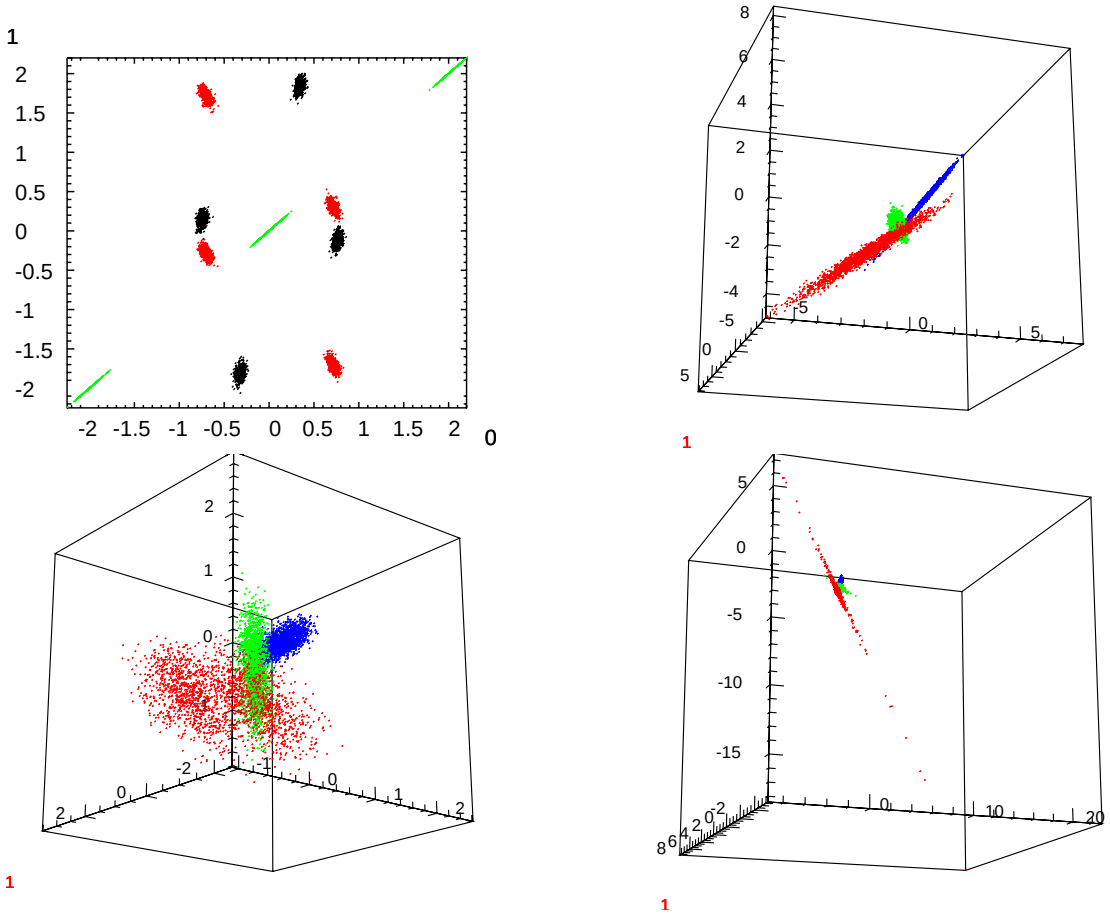


Figure C.2: Synthetic Data Sets 2.

category	**texture**		**structure**		**colour**	
	$\mathbf{X}_c$	$\mathbf{X}_d$	$\mathbf{X}_c$	$\mathbf{X}_d$	$\mathbf{X}_c$	$\mathbf{X}_d$
underthesea	0.039	0.033	28.220	26.349	2.3152	0.5592
	0.012	0.010	0.095	0.100	0.0001	0.0001
diver	0.046	0.038	39.135	19.057	0.5403	0.2075
	0.014	0.012	0.090	0.094	0.0001	0.0001
fish	0.030	0.039	36.685	14.211	7.0608	1.1483
	0.009	0.012	0.077	0.094	0.0003	0.0001
swarm	0.042	0.039	41.486	13.744	7.0233	0.2064
	0.013	0.012	0.088	0.095	0.0002	0.0001
animals	0.021	0.039	29.764	26.428	0.5626	0.2744
	0.007	0.012	0.098	0.098	0.0000	0.0001
elephant	0.011	0.021	34.397	16.803	0.2843	0.2109
	0.003	0.007	0.0863	0.0972	0.0000	0.0000
monkey	0.012	0.022	32.910	20.025	0.2558	0.2306
	0.004	0.007	0.089	0.098	0.0000	0.0000
lion	0.011	0.022	32.027	18.683	0.7942	0.0987
	0.003	0.007	0.0873	0.0974	0.0000	0.0000
doorswindows	0.019	0.037	29.068	27.225	0.9673	1.1157
	0.006	0.011	0.098	0.099	0.0000	0.0001
storefront	0.013	0.020	30.049	19.185	0.0873	0.0596
	0.004	0.006	0.093	0.098	0.0000	0.0000
church	0.024	0.016	32.706	17.475	12.1649	0.1751
	0.008	0.005	0.076	0.098	0.0003	0.0000
ruin	0.009	0.019	42.746	15.072	0.0383	0.0391
	0.003	0.006	0.091	0.098	0.0000	0.0000
sunrise	0.0316	0.029	31.050	24.584	4.4431	0.9347
	0.010	0.009	0.087	0.099	0.0001	0.0001
roundSun	0.024	0.033	33.970	25.056	2.1535	2.2442
	0.008	0.0102	0.083	0.087	0.0001	0.0002
yellowSky	0.033	0.031	39.434	25.557	3.5542	2.1716
	0.010	0.010	0.093	0.087	0.0001	0.0001
skyline	0.021	0.032	42.328	20.770	4.9453	0.6071
	0.007	0.0101	0.093	0.087	0.0002	0.0001

Table C.7: Mean of the variances of the different artexplosion categories and subsets. The first row presents the values after an ICA-transformation based on the relevant subsets $\mathbf{X}_c$ and the second row the variances in the original data space. $\mathbf{X}_d$ is the set of non-relevant data.

Bibliography

S. Amari, A. Cichocki, and H. Yang. A New Learning Algorithm for Blind Signal Separation. Advances in Neural Information Precessing Systems 8, 1996. MIT Press.

amazon. http://www.amazon.com. last visited 04/19/2006.

D. Anastassiou. Visualseek. http://www.columbia.edu/cu/record/23/20a/search.html, 2005. last visited 04/18/2006.

L. Armitage and P. Enser. Analysis of User Need in Image Archives. *Journal of Information Science*, 23(4):287–299, 1997.

J. A. Aslam and R. Savell. On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments. In J. Callan, G. Cormack, C. Clarke, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 361–362. ACM Press, July 2003.

M. Aupetit, P. Massotte, and P. Couturier. C-SOM: A Continuous Self-Organizing Map for Function Approximation. In *Proceedings of Intelligent System and Control*, 1999.

F. Bacao, V. Lobo, and M. Painho. The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computers & Geosciences*, 31:155–163, 2005.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

H. Y. Bang and T. Chen. Feature Space Warping: An Approach to Relevance Feedback. In *Image Processing. 2002. Proceedings. 2002 International Conference on Publication*, 2002.

H. Barlow. Possible principles underlying the transformation of sensory messages. In W. Rosenblith, editor, *Sensory Communication*. MIT Press, 1961.

M. S. Bartlett, H. M. Lades, and T. J. Sejnowski. Independent Component Representation for Face Recognition. In *Proceedings of the SPIE*, volume 3299: Conference on Human Vision and Electronic Imaging III, pages 528–539, 1998.

S. Basu. ICA: A Critical Review of Three Prominent Approaches. Technical report, 2000.

C. Bauckhage, T. Käster, M. Pfeiffer, and G. Sagerer. Content-Based Image Retrieval by Multimodal Interaction. In *Proceedings of the 29th Annual Conference of the IEEE Industrial Electronics Society*, pages 1865–1870, Roanoke, VA, 2003.

R. Beals, D. H. Krantz, and A. Tversky. Foundations of Multidimensional Scaling. *Psychological Review*, 75(2):127–142, 1968.

S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Computing Surveys*, 27(3):433–466, 1995.

A. J. Bell and T. J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

benchathlon. http://www.benchathlon.net/. last visited 04/18/2006.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

J. A. Black Jr, G. Fahmy, and S. Panchanathan. A method for evaluating the performance of content-based image retrieval systems. In *Proceedings of the Fifth IEEE Southwest Symposium on Image Analysis and Interpretation*, page 96. IEEE Computer Society, 2002.

bodleian. Bodleian Library, University of Oxford. http://www.bodley.ox.ac.uk/. last visited 04/18/2006.

H. L. Borgne and A. Guerin-Dugue. Sparse-Dispersed Coding and Images Discrimination with Independent Component Analysis. In *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation, ICA 2001*, 2001.

N. Bouteldja, V. Gouet-Brunet, and M. Scholl. Evaluation of strategies for multiple sphere queries with local image descriptors. In *Proceedings of Multimedia Content Analysis, Management and Retrieval 2006 SPIE*, volume 6073, 2006.

M. E. Bowden, T. B. Hahn, and R. V. Williams, editors. *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, Pittsburgh, Pennsylvania, 1998.

M. Bressan, D. Guillamet, and J. Vitrià. Using an ICA Representation of High Dimensional Data for Object Recognition and Classification. In *Proceedings of IEEE Conference on Computer Vision and Patern Recognition, CVPR 2001*, pages 1004–1009, 2001.

M. Bressan and J. Vitrià. Improving Naive Bayes using Class-Conditional ICA. In *Proceedings of Advances in Artificial Intelligence - IBERAMIA 2002, 8th Ibero-American Conference on AI*, volume 2527 of *Lecture Notes in Computer Science*, pages 1–10, Seville, Spain, November 2002a. Springer.

M. Bressan and J. Vitrià. Independent Component Analysis and Naive Bayes Classification. In *Visualization, Imaging, and Image Processing (VIIP 2002)*, Malaga, September 2002b.

P. Brodatz. *Textures: A Photographic Album for Artists & Designers*. Dover, New York, 1966.

M. M. Campos and G. A. Carpenter. Building adaptive basis functions with a continuous self-organizing map. *Neural Processing Letters*, 11:59–78, 2000.

A. Carkacioglu and F.-Y. Vural. Learning Similarity Space. In *Proceedings of the IEEE International Conference on Image Processing, ICIP 2002*, 2002.

C. Carson. blobworld, 2004. last visit 03/31/2006.

C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image Segmentation Using ExpectationMaximization and Its Application to Image Querying. *IEEE Trans. PAMI*, pages 1026–1038, 2002.

S.-F. Chang. Webseek. http://persia.ee.columbia.edu:8008/. last visited 04/18/2006.

C. Chatfield. *The analysis of time series: an introduction*. Boca Raton: Chapman & Hall CRC, 2004.

C. Cleverdon, J. Mills, and E. Keen. Factors determining the performance of indexing systems, vol. 2: Test results. Technical report, Aslib Cranfield Research Project, Cranfield, England, 1966. Technical report, http://www.itl.nist.gov/iaui/894.02/projects/irlib/pubs/cranv2/cranv2_index/cranv2_toc.html.

P. Clough, H. Müller, T. Deselaers, M. Grubinger, T. Lehmann, J. Jensen, and W. Hersh. The CLEF 2005 Cross-Language Image Retrieval Track. In *CLEF Workshop (2005)*, 2005a.

P. Clough, H. Müller, and M. Sanderson. The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, and B. Magnini, editors, *Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Lecture Notes in Computer Science (LNCS)*. Springer, Heidelberg, Germany, 2005b.

P. Clough, M. Sanderson, and H. Müller. A proposal for the CLEF Cross-Language Image Retrieval Track 2004. In *Conference for Video and Image Retrieval (CVIR 2004)*, 2004.

R. T. Collins, A. J. Lipton, and T. Kanade. Introduction to the Special Section on Video Surveillane. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), August 2000.

C. Colombo, A. D. Bimbo, and P. Pala. Semantics in visual information retrieval. *IEEE Multimedia*, 6(3):38–53, 1999.

P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36: 287–314, 1994.

convera. Visual Retrieval Ware and Excalibur. http://www.excalib.com. last visited 04/18/2006.

corel. *Corel GALLERY^{TM} Magic 65000*. Corel Corp., 1600 Carling Ave., Ottawa, Ontario, Canada K1Z 8R7.

T. M. Cowan. An Observing Response Analysis of Visual Search. *Psychological Review*, 75(3):265–270, 1968.

I. J. Cox, M. L. Miller, T. P. Minka, T. Papathomas, and P. N. Yianilos. The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.

I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Target Testing and the PicHunter Bayesian Multimedia Retrieval System. In *Advanced Digital Libraries ADL'96 Forum*, Washington D.C., 1996.

D. de Ridder, J. Kittler, and R. Duin. Probabilistic pca and ica subspace mixture models for image segmentation. In *British Machine Vision Conference*, pages 112–121, 2000.

D. de Ridder, O. Lemmers, R. Duin, and J. Kittler. The Adaptive Subspace Map for Image Description and Image Database Retrieval. *Lecture Notes in Computer Science*, 1876:94–103, 2001.

T. Deselaers, D. Keysers, and H. Ney. Features for Image Retrieval: A Quantitative Comparison. In *DAGM 2004, Pattern Recognition, 26th DAGM Symposium, Tübingen, Germany*, volume 3175 of *Lecture Notes in Computer Science*, pages 228–236, 2004a.

T. Deselaers, D. Keysers, and H. Ney. FIRE — flexible image retrieval engine: ImageCLEF 2004 evaluation. In *CLEF Workshop (2004)*, 2004b.

T. Deselaers, T. Weyand, D. Keysers, W. Macherey, and H. Ney. FIRE in ImageCLEF 2005: Combining Content-based Image Retrieval with Textual Information Retrieval. In *CLEF Workshop (2005)*, 2005.

P. Domingos and M. J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

S. W. Draper, M. D. Dunlop, I. Ruthven, and C. J. van Rijsbergen, editors. *Mira 99: Evaluating interactive information retrieval (MIRA-99)*, 1999. Electronic Workshops in Computing.

J. P. Eakins. Towards intelligent image retrieval. *Pattern Recognition*, 35:3–14, 2002.

J. P. Eakins, P. Briggs, and B. Burford. Image Retrieval Interfaces: A User Perspective. In *proceedings of CIVR 2004*, pages 628–637, 2004.

EarthCam. http://search.earthcam.com. last visited 04/18/2006.

E. N. Efthimiadis. Online Public Access Catalogues: Characteristics of the Literature. *Journal of Information Science*, 16(2):107–112, 1990.

H. Eidenberger. A new perspective on visual information retrieval. In *Proceedings of the SPIE*, volume 5304: SPIE Electronic Imaging Symposium, San Jose, 2004.

H. Eidenberger and C. Breiteneder. Semantic Feature Layers in Content-based Image Retrieval: Implementation of Human World Features. In *Proceedings IEEE International Conference on Control, Automation, Robotic and Vision*, Singapore, 2002.

P. Enser. Pictorial information retrieval. *Journal of Documentation*, 51(2):126–170, 1995.

R. Everson and S. Roberts. Particle Filters for Non-Stationary ICA. In M. Girolami, editor, *Advances in Independent Component Analysis*, Perspectives in Neural Computing, pages 23–41. Springer, 2000.

R. M. Everson and S. J. Roberts. ICA: A flexible non-linearity and decorrelating manifold approach. *Neural Computation*, 11(8):1957–1983, 1999.

S. Fiori. Some Properties of Bell-Sejnowski PDF-matching Neuron. In *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation, ICA 2001*, pages 194–199, 2001.

A. Franco, A. Lumini, and D. Maio. A new approach for relevance feedback through positive and negative samples. In *Proceedings of the IEEE International Conference on Image Processing, ICIP 2004*, 2004.

S. Geisler, O. Kao, and T. Bretschneider. Analysis of cluster topologies for workload balancing strategies in image databases. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications*, pages 874–880, 2001.

G. Giacinto and F. Roli. Nearest-Prototype Relevance Feedback for Content Based Image Retrieval. In *Proceedings of the IEEE International Conference on Image Processing, ICIP 2004*, 2004.

GIFT. http://www.gnu.org/software/gift/. last visited 04/18/2006.

google. http://www.google.com. last visited 04/18/2006.

M. E. Graham and J. P. Eakins. ARTISAN : a prototype retrieval system for trade mark images. *Vine*, (107):73–80, 1998.

N. Gunther and G. Beretta. A Benchmark for Image Retrieval using Distributed Systems over the Internet BIRDS-I. Technical report, HP Labs, Palo Alto, San Jose, 2001.

J. Hare, P. Lewis, P. Enser, and C. Sandom. Mind the Gap: Another look at the problem of the semantic gap in image retrieval. In *Proceedings of Multimedia Content Analysis, Management and Retrieval 2006 SPIE*, volume 6073, 2006.

C. Haslego. History of the Camera, 2005. http://ezinearticles.com/?History-of-the-Camera&id=18736, last visited 10/19/2005.

T. Hastie and W. Stuetzle. Principal Curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

M. Heczko, D. A. Keim, and R. Weber. Analysis of the Effectiveness-Efficiency Dependence for Image Retrieval. In *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.

D. Heesch and S. Rüger. Performance boosting with three mouse clicks - Relevance feedback for CBIR. In *25th European Conference on Information Retrieval Research (ECIR, Pisa, 14-16 Apr 2003)*, Springer LNCS 2633, pages 363–376, 2003.

A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *Proceedings of the Semantic Web Workshop at SIGIR-2003*, 26th Annual International ACM SIGIR Conference, Toronto, Canada, 2003.

T. S. Huang. Mars. http://www-db.ics.uci.edu/pages/research/mars/index.shtml. last visited 04/18/2006.

M. Hussain and J. P. Eakins. Visual Clustering of Trademarks Using a Component-Based Matching Framework. In *Proceedings of CIVR 2004*, pages 141–149, 2004.

A. Hyvärinen. Survey on Independent Component Analysis. *Neural Computation Surveys*, 2:94–128, 1999.

A. Hyvärinen, P. O. Hoyer, and M. Inki. The Independence Assumption: Analyzing the Independence of the Components by Topography. In M. Girolami, editor, *Advances in Independent Component Analysis*, Perspectives in Neural Computing, pages 45–62. Springer, 2000.

IBM. QBIC. http://wwwqbic.almaden.ibm.com/. last visited 04/18/2006.

Q. Iqbal and J. K. Aggarwal. CIRES: A System for Content-bases Retrieval in Digital Image Libraries. In *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 205–210, Singapore, December 2002a. Invited session on Content Based Image Retrieval: Techniques and Applications.

Q. Iqbal and J. K. Aggarwal. Retrieval by classification of images containing large man-made objects using perceptual grouping. *Pattern Recognition*, 35(7):1463–1479, July 2002b.

Y. Ishikawa, R. Subramanya, and C. Faloutsos. MindReader: Querying databases through multiple examples. In *Proceedings 24th Int. Conf. Very Large Data Bases, VLDB*, pages 218–227, 24–27 1998.

N. Jaeckisch. Characterisation of the mega-epibenthic community along a depth gradient in the deep Fram Strait (Arctic Ocean). Master thesis, University of Hamburg, 2004.

R. Jain, editor. *Proceedings of the US NSF Workshop Visual Information Management Systems*, 1992.

J. J.J. Rocchio. *The SMART retrieval system.* Prentice-Hall, 1971.

K. S. Jones and C. van Rijsbergen. Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, 1975.

K. S. Jones and C. van Rijsbergen. Progress in Documentation. *Journal of Documentation*, 32:59–75, 1976.

A. C. Kak, V. Murino, and A. Trucco. *Special issue on underwater computer vision and pattern recognition*, volume 79 of *Computer Vision and Image Understanding.* Elsevier Science Inc., July 2000.

T. Kämpfe, T. Käster, M. Pfeiffer, H. Ritter, and G. Sagerer. INDI – Intelligent Database Navigation by Interactive and Intuitive Content-Based Image Retrieval. In *IEEE 2002 International Conference on Image Processing, Rochester, USA*, pages 921–924, 2002.

T. Kämpfe, T. W. Nattkemper, and H. Ritter. Combining independent component analysis and self-organizing maps for cell image classification. In B. Radig and S. Florczyk, editors, *Pattern Recognition*, Lecture Notes in Computer Science 2191, pages 262–268. Springer-Verlag, 2001.

T. Kämpfe, T. W. Nattkemper, and H. Ritter. AQUISAR: Image Retrieval in Underwater Webcam Images. In *Proceedings of 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Lisboa, Portugal, Apr 2004.

J. Karhunen. Neural Approaches to Independent Component Analysis and Source Separarion. In *Proceedings of the 4th European Symposium on Artificial Neural Networks (ESANN'96), Bruges, Belgium*, pages 249–266, April 24-26 1996. (Invited paper).

S. Kaski, J. Kangas, and T. Kohonen. Bibliography of Self-Organizing Map (SOM) Papers: 1981– 1997. *Neural Computing Surveys*, 1:102–350, 1998.

T. Käster, M. Pfeiffer, C. Bauckhage, and G. Sagerer. Combining Speech and Haptics for Intuitive and Efficient Navigation through Image Databases. In *Proceedings of International Conference on Multimodal Interfaces (ICMI'03)*, pages 180–187. ACM, 2003.

M. L. Kherfi, D.Zinou, and A. Bernardi. Learning from Negative Example in Relevance Feedback for Content-Based Image Retrieval. In *Proceedings of IEEE/IAPR International Conference on Pattern Recognition (ICPR)*, Quebec, Cananda, August 2002.

S. Klanke and H. Ritter. Psom$^+$: Parametrized self-organizing maps for noisy and incomplete data. In *Proceedings of the 5th Workshop on Self-Organizing Maps (WSOM 05)*, Paris, France, September 2005.

B. Ko and H. Byun. Probabilistic Neural Networks Supporting Multi-class Relevance Feedback in Region-based Image Retrieval. In *Proceedings of IEEE/IAPR International Conference on Pattern Recognition (ICPR)*, Quebec, Cananda, August 2002.

T. Kohonen. Clustering, Taxonomy and Topological Maps of Patterns. In *Proceedings of the Sixth International Conference on Pattern Reconginiton*, pages 114–128, Munich, Germany, 1982.

T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, 2 edition, 1997.

T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585, 2000.

T. Kohonen, S. Kaski, and H. Lappalainen. Self-organized Formation of Various Invariant-feature Filters in the Adaptive-subspace SOM. *Neural Computation*, 9(6):1321–1344, 1997.

M. Koskela. *Interactive Image Retrieval Using Self-Organizing Maps*. Helsinki University of Technology, Espoo, 2003. Dissertations in Computer and Information Science, Report D1.

M. Koskela and J. Laaksonen. Using Long-Term Learning to Improve Efficiency of Content-Based Image Retrieval. In *Proceedings of Third International Workshop on Pattern Recognition in Information Systems (PRIS 2003)*, 2003.

M. Koskela, J. Laaksonen, and E. Oja. Comparison of Techniques for Content-Based Image Retrieval. In *Proceedings of SCIA2001*, 2001a.

M. Koskela, J. Laaksonen, and E. Oja. Self-organizing image retrieval with MPEG-7 descriptors. In *Proceedings of IR2001*, 2001b.

M. Koskela, J. Laaksonen, and E. Oja. Implementing Relevance Feedback as Convolutions of Local Neighborhoods on Self-Organizing Maps. In *Proceedings of International Conference on Artificial Neural Networks (ICANN 2002)*, Madrid, Spain, August 2002.

M. Koskela, J. Laaksonen, M. Sjöberg, and H. Muurinen. PicSOM experiments in TRECVID 2005. In *Proceedings of the TRECVID 2005 Workshop*, pages 267–270, November 2005.

J. Laaksonen, M. Koskela, S. Laakso, and E. Oja. PicSOM - content-based image retrieval with self-organizing maps. *Pattern Recognition Letters*, 21:1199–1207, 2000.

J. Laaksonen, M. Koskela, S. Laakso, and E. Oja. Self-organizing maps as a relevance feedback technique in content-based image retrieval. *Pattern Analysis & Applications*, 4:140–152, 2001.

B. Laheld and J.-F. Cardoso. Adaptive source separation with uniform performance. In *Signal Processing VII: Theories and Applications (Proc. EUSIPCO-94), Lausanne: EURASIP*, volume 2, pages 183–186, 1994.

A. Large, L. Tedd, and R. Hartley. *Information Seeking in the Online Age: Principles and Practice*. Sauer, München, 2001.

T. Lee, M. Lewicki, and T. Sejnowski. ICA Mixture Models for Unsupervised Classification of Non-Gaussian Classes and Automatic Context Switching in Blind Signal Separation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1078–1089, 2000.

W. Liu, Z. Su, S. Li, and H. Zhang. A Performance Evaluation Protocol for Content-Based Image Retrieval Algorithms/Systems. In *Proceedings IEEE CVPR Workshop on Empirical Evaluation in Computer Vision*, Kauai, USA, December 2001.

London Aquarium. http://www.londonaquarium.co.uk/. last visited 04/18/2006.

H. P. Luhn. Automated intelligence systems: Some basic problems and prerequisites for their solution. In E. Tomeski, R. Westcott, and M. Covington, editors, *The clarification, unification & integration of information storage & retrieval proceedings of February 23rd 1961 symposium: Management Dynamics*, pages 3–20, New York, 1961.

W. Y. Ma and B. S. Manjunath. NeTra: a toolbox for navigating large image databases. *Multimedia Systems*, 7(3):184–198, May 1999.

B. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2000.

S. Marchand-Maillet. Viper. http://viper.unige.ch/. last visited 04/18/2006.

metacrawler. http://www.metacrawler.com. last visited 04/18/2006.

V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Still Image Objective Segmentation Evaluation using Ground Truth. In *5th COST 276 Workshop*, pages 9–14, 2003.

H. C. Micko and W. Fischer. The metric of multidimensional psychological spaces as a function of the differential attention to subjective attributes. *Journal of Mathematical Psychology*, 7:118–143, 1970.

J. Min, M. Powell, and K. W. Bowyer. Automated Performance Evaluation of Range Image Segmentation Algorithms. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, 34(1), February 2004.

T. P. Minka. photobook. http://vismod.media.mit.edu/vismod/demos/photobook/. last visited 04/18/2006.

S. Mizzaro. Relevance: The Whole History. *Journal of the American Society of Information Science*, 48(9):810–832, 1997.

C. N. Moore. Moore's law, or why some retrieval systems are used and others are not. *American Documentation*, 11(3), July 1960.

MRML. http://www.mrml.net. last visited 04/18/2006.

H. Müller. Jäger des verlorenen Fotos - Das GNU Image Fining Tool für Linux in der Praxis. *c't*, 6:252–257, 2002.

H. Müller, P. Clough, W. Hersh, and A. Geissbuhler. IMAGECLEF 2004–2005: Results, Experiences and New Ideas for Image Retrieval Evaluation. In *CBMI 2005 - Fourth International Workshop on Content-Based Multimedia Indexing*, 2005.

H. Müller, A. Geissbuhler, S. Marchand-Maillet, and P. Clough. Benchmarking image retrieval applications. In *Proceedings of the Seventh International Conference on Visual Information Systems*, San Francisco, USA, September 8-10, 2004.

H. Müller, S. Marchand-Maillet, and T. Pun. The Truth about Corel - Evaluation in Image Retrieval. In *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*, pages 38–49, London, UK, 2002. Springer-Verlag.

H. Müller, W. Müller, S. Marchand-Maillet, T. Pun, and D. Squire. A Framework for Benchmarking in CBIR. *Multimedia Tools and Applications*, 21(1):55–73, 2003.

H. Müller, W. Müller, S. Marchand-Maillet, D. Squire, and T. Pun. Strategies for positive and negative relevance feedback in image retrieval. In *Proceedings of the International Conference on Pattern Recognition (ICPR'2000)*, 2000a.

H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun. Automated Benchmarking in Content-Based Image Retrieveal. In *International Conference on Multimedia and Exposition, ICME 2001, Tokyo, Japan*, 2001a.

H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals. *Pattern Recognition Letters (Special Issue on Image and Video Indexing)*, pages 593–601, 2001b.

H. Müller, D. M. Squire, W. Müller, and T. Pun. Efficient access methods for content-based image retrieval with inverted files. In *Multimedia Storage and Archiving Systems IV (VV02),(SPIE Symposium on Voice, Video and Data Communications), SPIE Proceedings*, 20–22 September 1999.

W. Müller, S. Marchand-Maillet, H. Müller, and T. Pun. Towards a fair benchmark for image browsers. In *SPIE Photonics East, Voice, Video, and Data Communications*, Boston, MA, USA, nov 5–8, 2000b.

E. V. Munson and Y. Tsymbalenko. To Search for Images on the Web, Look at the Text, Then Look at the Images. In *Proceedings of the First International Workshop on Web Document Analysis*, pages 39–42, September 2001.

A. D. Narasimhalu, M. S. Kankanhalli, and J. Wu. Benchmarking Multimedia Databases. *Multimedia Tools and Applications*, 4:333–356, 1997.

S. Nene, S. Nayar, and H. Murase. Columbia Object Image Library (COIL-100). Technical report, Department of Computer Science, Columbia University, February 1996.

S. Newsam, B. Sumengen, and B. Manjunath. Category-Based Image Retrieval. In *Proceedings of the IEEE International Conference on Image Processing, Special Session on Multimedia Indexing, Browsing and Retrieval*, pages 596–599, 2001.

artexplosion. *Art Explosion Photo Gallery*. Nova Development Corporation, 23801 Calabasas Road, Suite 2005 Calabasas, California 91302-1547, USA.

M. Oja, S. Kaski, and T. Kohonen. Bibliography of Self-Organizing Map (SOM) Papers: 1998–2001 Addendum. *In Neural Computing Surveys*, 3:1–156, 2002.

A. Oliva and A. B. Torralba. Scene-Centered Description from Spatial Envelope Properties. In *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pages 263–272, London, UK, 2002. Springer-Verlag.

J. Ontrup and H. Ritter. Hyperbolic self-organizing maps for semantic navigation. In *Advances in Neural Information Processing Systems 14*, 2001a.

J. Ontrup and H. Ritter. Text categorization and semantic browsing with self-organizing maps on non-euclidean spaces. In L. D. Raedt and A. Siebes, editors, *Proceedings of PKDD-01, 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 338–349. Springer, LNAI 2168, 2001b.

T. V. Papathomas, T. E. Conway, I. J. Cox, J. Ghosn, M. L. Miller, T. P. Minka, and P. N. Yianilos. Psychophysical Studies of the Performance of an Image Database Retrieval System. In *Proceedings of SPIE*, 1998.

S.-J. Park, K.-H. An, and M. Lee. Saliency map model with adaptive masking based on independent component analysis. *Neurocomputing*, 49((1-4)):417–422, 2002.

Z. Pecenović, M. Do, S. Ayer, and M. Vetterli. New Methods for Image Retrieval. In *Proceedings of ICPS'98 Congress on Exploring New Tracks in Imaging*, September 1998.

J. Peng and B. Bhanu. Independent Feature Analysis for Image Retrieval. *Pattern Recognition Letters*, 22(1):63–74, 2001.

M. Petkovic and W. Jonker. *Content-Based Video Retrieval: A Database Perspective*. Kluwer, August 2003.

J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of IEEE International Conference on Computer Vision(ICCV-1999)*, pages 1165–1173, 1999.

F. Qian, M. Li, L. Zhang, H. Zhang, and B. Zhang. Gaussian mixture model for relevance feedback in image retrieval. In *Proceedings IEEE International Conference on Multimedia & Expo (ICME)*, 2002.

G. Qiu, L. Ye, and X. Feng. Fast image indexing and visual guided browsing. In *CBMI 2003, Third International Workshop on Content-Based Multimedia Indexing*, 2003.

R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image Change Detection Algorithms: A Systematic Survey. *IEEE Transactions on Image Processing*, 14(3), March 2005.

A. Rao, R. K. Srihari, L. Zhu, and A. Zhang. A method for measuring the complexity of image databases. *IEEE Transactions on Multimedia*, 4(2):160–173, June 2002.

T. W. Ridler and S. Calvard. Picture thresholding using an iterative selection method. *IEEE Trans.*, SMC-8:630–632, Aug. 1978.

I. Rish. An empirical study of the naive Bayes classier. In *IJCAI-01 workshop on "Empirical Methods in AI"*, 2001.

H. Ritter. Learning with the Self-Organizing Map. In T. Kohonen, editor, *Artificial Neural Networks*, pages 379–384, Amsterdam, 1991. Elsevier Science Publishers.

H. Ritter. Self-organizing maps in non-euclidian spaces. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 97–110. Amer Elsevier, 1999.

H. Ritter and K. Schulten. On the Stationary State of Kohonen's Self-Organizing Sensory Mapping. *Biological Cybernetics*, (54):99–106, 1986.

E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

Y. Rui and T. Huang. Optimizing Learning in Image Retrieval. In *CVPR 2000*, pages 236–245, 2000.

Y. Rui, T. Huang, and S. Mehrotra. Content-Based Image Retrieval With Relevance Feedback in MARS. In *IEEE International Conference on Image Processing, ICIP'97*, Santa Barbara, CA, 1997a.

Y. Rui, T. S. Huang, and S. Mehrotra. Relevance Feedback Techniques in Interactive Content-Based Image Retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, 1998.

Y. Rui, S. Mehrotra, and M. Ortega. Automatic matching tool selection using relevance feedback in MARS. In *International Conference on Visual Information Retrieval*, 1997b.

M. Rummukainen, J. Laaksonen, and M. Koskela. A efficiency comparison of two content-based image retrieval systems, GIFT and PicSOM. *Image and Video Retrieval, Second International Conference, CIVR 2003, Urbana-Champaign, IL, USA, July 24-25, 2003, Proceedings*, 2728, 2003.

A. Saalbach, G. Heidemann, and H. Ritter. Parametrized SOMs for Object Recognition and Pose Estimation. In *Artificial Neural Networks - ICANN 2002*, pages 902–907. Springer, 2002.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

S. Santini. Evaluation Vademecum for Visual Information Systems. In *Proceedings of SPIE: Storage and Retrieval for Image and Video Databases VIII*, volume 3972, 2000.

S. Santini and R. Jain. The use of psychological similarity measure for queries in image databases. http://www-cse.ucsd.edu/users/ssantini., 1996. Technical report, Visual Computing Laboratory, University of California San Diego, 1996.

T. Scheer. *Piet Mondrian, Komposition mit Rot, Gelb und Blau : eine Kunst-Monographie.* Insel-Verlag, 1995.

searchengineswatch. http://www.searchengineswatch.com. last visited 04/18/2006.

J. Sedding and D. Kazakov. WordNet-based Text Document Clustering. In *Proceedings of the Third Workshop on Robust Methods in Analysis of Natural Language Data (ROMAND)*, 2004.

C. W. Shaffrey, I. H. Jermyn, and N. G. Kingsbury. Psychovisual Evaluation of Image Segmentation Algorithms. In *Proceedings of ACIVS 2002 (Advanced Concepts for INtelligent Vision Systems)*, September 9–11 2002a.

C. W. Shaffrey, N. G. Kingsbury, and I. H. Jermyn. Unsupervised Image Segmentation via Markov Trees and Complex Wavelets. In *Proceedings IEEE International Conference On Image Processing (ICIP)*, September 2002b.

L. G. Shapiro. http://www.cs.washington.edu/research/imagedatabase/groundtruth/. last visited 11/24/2005.

H. Simon and U. Verstegen. p r o m e t h e u s Das verteilte digitale Bildarchiv für Forschung & Lehre – Neuartige Werkzeuge zur Bereitstellung von verteiltem Content für Wissenschaft und Forschung. *Historische Sozialforschung, Sonderheft: Elektronisches Publizieren & Open Access*, 29(107):247–257, 2004.

M. Sjoberg, J. Laaksonen, and V. Viitaniemi. Using image segments in picsom cbir system. In *SCIA03*, pages 1106–1113, 2003.

A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1379, 2000.

K. A. Smith. Neural Networks for Combinatorial Optimization: A Review of More Than a Decade of Research. *INFORMS Journal on Computing*, 11(1):15–34, 1999.

D. M. Squire, W. Müller, H. Müller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In *The 11th Scandinavian Conference on Image Analysis*, pages 143–149, Kangerlussuaq, Greenland, jun 7–11 1999.

B. Sumengen and B. S. Manjunath. Edgeflow-driven variational image segmentation: Theory and performance evaluation. Technical report, Vision Research Lab, UCSB, May 2005.

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

K. Takaya and K.-Y. Choi. Detection of facial components in a video sequence by independent component analysis. In *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation, ICA 2001*, pages 266–271, 2001.

TASI. A review of image search engines. http://www.tasi.ac.uk/resources/searchengines.html, October 2004. last visited 04/18/2006.

A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, (14):391–412, 2003.

T. Tranden. Brodatz Textures. http://www.ux.his.no/~tranden/brodatz.html. last visited 04/28/2006.

TREC. http://trec.nist.gov/. last visited 03/29/2006.

TRECVID. http://www-nlpir.nist.gov/projects/trecvid/. last visited 03/29/2006.

A. Tversky. Features of Similarity. *Psychological Review*, 84(4):327–352, July 1977.

R. Unnikrishnan, C. Pantofaru, and M. Hebert. A measure for objective evaluation of image segmentation algorithms. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '05), Workshop on Empirical Evaluation Methods in Computer Vision*, June 2005.

M. Unser. Sum and difference histograms for texture classification. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, PAMI-8(1):118–125, 1986.

J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc.R.Soc.Lond.*, 265:359–366, 1998.

N. Vasconcelos. Minimum Probability of Error Image Retrieval. *IEEE Transactions on Signal Processing*, 52(8):2322–2336, August 2004.

U. Verstegen. prometheus – Das verteilte digitale Bildarchiv für Forschung & Lehre. *zeitenblicke 2*, 1, 05 2003.

VisTex. http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html, 1995. last visited June 2005.

T. Wagner. Texture analysis. In B. Jähne, H. Haussecker, and P. Geissler, editors, *Handbook of Computer Vision and Applications*, chapter 12, pages 275–308. Academic Press, 1999. volume 2.

J. Walter and H. Ritter. Rapid Learning with Parametrized Self-Organizing Maps. *Neurocomputing*, 12:131–153, 1996.

W. Wang, Y. Wu, and A. Zhang. SemView: A Semantic-sensitive Distributed Image Retrieval System. In *4th National Conference on Digital Government Research (DGO 2003)*, Boston, May 18–21 2003.

B. H. Weinberg. Why indexing fails the researcher. In *Proceedings of the 50th Annual Meeting of the American Society for Information Science*, October 1987.

H. H. Wellisch. *Indexing from A to Z.* Bronx, NY, Wilson, 1991.

V. Wendt. Untersuchung von Clusterverfahren zur Integration in ein bestehendes Bildsuchsystem. Master's thesis, Bielefeld University, 2002.

J. R. Williamson. Self-Organization of Topographic Mixture Networks Using Attentional Feedback. *Neural Computation*, 13(3):563–593, 2001.

M. E. J. Wood, N. W. Campbell, and B. T. Thomas. Iterative Refinement by Relevance Feedback in Content-Based Digital Image Retrieval. In *ACM Multimedia 98*, pages 13–20, Bristol, UK, September 1998. ACM.

P. Wu and B. S. Manjunath. Adaptive nearest neighbor search for relevance feedback in large image databases. In *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 89–97, New York, NY, USA, 2001. ACM Press.

L. Wyckoff, Jr. The role of observing responses in discrimination learning. *Psychological Review*, 59:431–442, 1952.

yahoo. http://www.yahoo.com. last visited 04/18/2006.

K. Yanai and K. Barnard. Probabilistic Web Image Gathering. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 57–64, New York, NY, USA, 2005. ACM Press.

K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–408, New York, NY, USA, 2003. ACM Press.

X. Zhou, B. Moghaddam, and T. S. Huang. ICA-based Probabilistic Local Appearance Models. In *Proceedings of International Conference on Image Processing (ICIP'01)*, 2001.

L. Zhu, A. Zhang, A. Rao, and R. Srihari. Keyblock: An approach for content-based image retrieval. In *Proceedings of ACM Multimedia 2000*, pages 157–166, Los Angeles, California, USA, Oct 30 – Nov 3 2000.