MEANINGFUL IMAGE-SPACES

How can interaction with digital image collections improve image (re)findability?

&

PROJECT PHOTOINDEX

Master of Arts in Interactive Multi Media dissertation

Matthijs Rouw, BAT Interaction Design
project site: http://photoindex.thingsdesigner.com
email: matthijs@thingsdesigner.com

May – august 2005

Utrecht School of the Arts
Faculty of Arts, Media and Technology
Hilversum, the Netherlands

Thesis supervisor: Dr. B.A.M Schouten
Project supervisors: Janine Huizenga and Tom Demeyer

I. INTRODUCTION

Before the usage of digital imagery, one could find a photo by searching through structured archives at journalistic companies or browsing through stock-photo books. Stock-photo books are books that present the available photos, of which the rights belong to companies that sell these royalty-free stock photos for use in advertising. An example of a company like this is Tony Stone Images, which now only sells photos online via Gettyimages.com (Figure 1.1). Each new general volume (Figure 1.1, bottom-right) of their photo-books contained categories that were similar to other volumes. These categories were for example themes like 'Lifestyles, singles, couples, celebrations, entertainment, Christmas, weddings', 'ideas' or 'business, finance'. Next to that, there was a separate index available with references to images that resemble concepts like 'freedom', 'power' and 'other concepts'. These concepts were sub-divided in turn. 'Safety', for example, contains 'protection, comfort, safety, security'. An index with terms that describe the objects or subjects in the images, like 'boat', 'gorilla' or 'Vietnam' is available in the back of the photo book. Not every new volume of photo book contained the same sections, because sometimes there were not enough images available to fill a certain theme or concept. Every once in a while, enough images were collected for a whole volume on one single theme. They were brought together in a Specialist volume (Figure 1.1, three books at the top-right), like 'Sports & recreation'. This particular volume contains sections for kinds of sports like 'water sports' or 'motor sports' as well as a list of concepts and a subject index.

One problem with the stock photo books was the ever-growing amount of volumes. New images could not be added to older volumes or the specialist albums. The only way was printing a new general volume. The volumes contained an inconsistency regarding the categorisation structure, since the structure was dependent on the content of the new series of images. An other disadvantage to using multiple volumes for the same category is that it is virtually impossible to find 'the best' image, since a user should go through all volumes and look at all possible images. Next to that, creating a new stock photo book costs a lot of work, time and money. The advantage of stock photo books is the high browsing factor that provides much space for inspiration or a change of ideas. The printed book format is also a tactile medium, suitable for any time or location.

Digital textual documents can be found more easily than analogue ones, like books in a library. This is because of the simple fact that a computer can search and compare faster than a human. Logically speaking, this should also be the case for images or photographs. The difference between analogue structuring and digital indexing is that digital systems allow us to create more direct links to an image, where to achieve this in an analogue environment, the images would need to be reprinted several times. In other words, with the use of computers in combination with digital images and hyper-linking, more links to a single photo increases its find-ability. In the minds of people, almost anything is possible with the use of computers. The re-mediation of photography changes people's demands of searching through collections of images. A reason for this is the shift towards a digitalised visual culture. Technology provides access to cheaper image making by using affordable computers and digital photo cameras. Virtually every new hand-held device contains a photo camera as well.

Usages of new, automated processes in many different fields of work create a great flow of data. The vast amount of newly created images is part of this data flow. Satellite images, medical scans and seismologic graphs are created and stored continuously and some of these need to be able to be retrieved at will, for imaginable reasons. Not just newly created imagery needs to be accessed on a computer. The digitalisation of analogue material, one could think of textual documents as well as artworks like paintings, find their way into the digital realm for preservation and find-ability purposes. In the art world, the re-mediation of old materials introduces a creating process as well. Older works of art are accessible for creating new works of art.

There is no single solution to the problem of storing and retrieving digital images, since there are many different aspects, which determine the shape of each application. One of the factors in this problem are people. The search for an image depends on many factors. These are for example the use for the image. One person can look for a beautiful sunset image one time, when at another, the same person may be after a specific object in a specific context for informational purposes. Other factors of influence are cultural or historical background, education, field of work, etc. Every search task or goal uses its own language. Personal photos may contain visual clues that work emotionally, as ads can use a photo that communicates a different message. Different people can interpret the same image differently as well, because of the person's task or background. A forensic detective can see possible evidence of a crime in a photo, while a dry cleaner sees stains that are hard to remove. One single person can have different interpretations of the same image, depending on individual characteristics (Chapter 3.2) and her purpose for the image.

Currently there are two general approaches to the storage and retrieval problems of digital images. One of these is the use of manually entered textual image descriptions (Chapter 3). In this approach, descriptive text is added to images as metadata in order to be able to retrieve the images by querying for the added textual metadata. This added data could be technical information like time and date of creation, resolution, focal distance, etc., or a semantic description about the content of the image, like 'sky scrapers', 'New York' and 'metropolis'. Next to describing tactile objects or subjects in an image, more abstract or emotional concepts like 'colourful' and 'impressive' can be used as well. The use of text to describe the contents of images is also done because a computer is better at comparing text rather than images (More in chapter 3.3). Actually, it means that it is relatively easy to develop, because knowledge regarding text comparison has already been developed and text has a structure that can be used as a handle by the computer. In addition, manually entered descriptive texts are high-level features. When a description of an image is rich enough, the image is very well possible to turn up from a database after entering a matching query. The saying that 'Images can say more than a thousand words' expresses the need for a great amount of textual information to describe an image properly. However, manually adding textual metadata to all newly created images is a time consuming and tedious process. It is sensitive to type errors and limited by the vocabulary of, and affected by the (cultural, historical or educational) background of the person that enters the data. One person might call something a dot, as another calls the same shape a circle. One person sees orange, while another calls the same colour yellow.

The second general approach to image storage and retrieval is indexing images by computer vision. In this approach, every image submitted to the system is analysed by the computer. The computer then selects features from a wide spectrum of optional features[1] like colour, shape or structure, which best suit the automated description of the image. These directly visually derivative features are low-level features. They are measurable. The low-level feature-description, the collection of visual characteristics of an image is stored in a database. Images are retrievable from the system by looking for the stored low-level image features later on. This approach is known as Content Based Image Retrieval (CBIR).

The design of an image storage and retrieval system should mainly depend on the search tasks that a user needs or wishes to carry out. Questions regarding whether or not the user already knows what she is looking for and what the use of the image will be after it is found need to be asked, before the design and development of an image storage and retrieval system can start. It might even sometimes be necessary to retrieve images slightly different than the one sought after, for the same reason why stock photo books can be inspirational. Both manually entered descriptive metadata and CBIR approaches are solutions to technical problems rather than interaction design issues.

This dissertation provides an overview on problems regarding the interaction design of a digital image retrieval system. The theories described are useful for designing a concept that uses a different approach to storage and retrieval of digital images, compared to the two mentioned. The idea is to use advantages from both the manually entered descriptive metadata and CBIR approaches.

---

[1] Smeulders et al., *Content-based image retrieval at the end of the early years*, chapter 3. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, No 12, december 2000

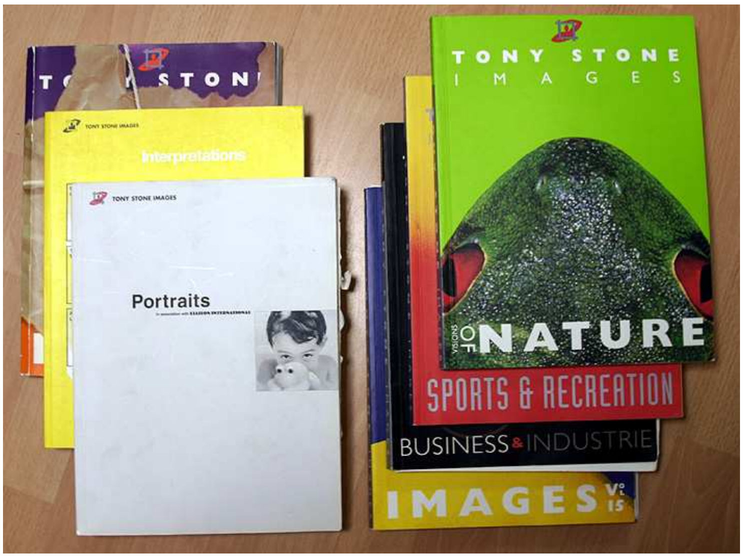

Figure 1.1 - Tony Stone stock photo books

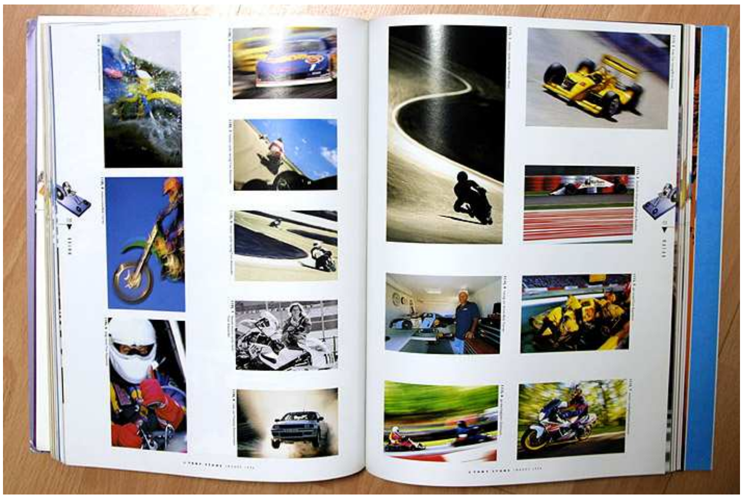

Figure 1.2 - Tony Stone stock photo book inside - Specialist edition 'Sports & Recreation

2. PHOTOGRAPY GOES DIGITAL

The initial purpose of making photos was to be able to capture and store a real life image from a certain setting at a certain time. Over time, photography went from a solely documentary medium to an art form and a way to produce and share information as well. After the creation of an analogue photo, it became duplicative, distributable and accessible.

Digital storage of photos provides ways duplicate and distribute them with more ease. The duplication of an image does not cost time, use of photo paper and chemicals, like before. Images remain in their digital form and are printed only when a hardcopy is actually needed, which saves time and money. However, the expectable advantages in digitally *accessing* images stay behind[2]. In the context of this thesis, accessibility relates to the *ease of access* or *find-ability* of a single photo in a collection, rather than the obvious availability of digitally stored imagery.

2.1 Browsing and querying

To understand ways of searching through collections of images and digital data in general, it is important to understand the difference between browsing or querying for information. Both are ways of filtering information from a pool of data.

The process of browsing is a time consuming and unreliable method for filtering information. A user has to review and compare every item from a (sub)set of data objects, to determine its usability. Usually the items are displayed in lists, tables or matrixes, which makes side-by-side comparison somewhat easier. The advantage of looking at multiple items at once saves time, because the user can glance at a list to scan for information that she is looking for. However, the presented information is reduced in detail because all items need to be scaled down in size or volume. In result, the user can easily overlook parts of important information. She also has to divide her attention between multiple items, resulting in more chance of filtering out useful information. This

[2] Santini, Simone, *Exploratory image databases*, p20. Academic Press, 2001

becomes worse when the images that need to be browsed through are distributed over several pages. Additionally, browsing takes up time.
Computers can compare information faster and more accurate (not better![3]) than people can as long as the data that needs to be compared, because they calculate faster. People can use this computational power to let the computer filter data for them. Stored data needs to be indexed before a computer can compare data items, either by hand or automatically. Next, a computer needs criteria to be able to compare data. Submitting selection criteria for filtering is called *querying*. A query can for example be 'show me all text documents about poisonous frogs'.

Quality of query and retrieval result

The quality of information (or level of satisfaction at which a user accepts the retrieved information) returned by the computer depends on (a) the domain in which a user searches, (b) the detail level of the index and (c) the detail level of the submitted query.

(a) The domain in which a user searches (or of which the information is stored)

It is impossible to find an explanation on how to install a hard disk in a computer within a library of biological science text documents. In addition, it is not logical to look for an *image* of a frog in that same database.

(b) The detail level of the index

In some systems, text documents are indexed by the text in their summary. Looking for words that appear in the entire document, and not in the summary, may not result in the retrieval of the right document, even when the document is actually stored in the system. The index is thus incomplete. Because of advances in software design, it is now possible to have full-text searches.

---

[3] Computers can never value better from worse than people can, because there is never enough information at hand for it to take into consideration. Next to that, all the required information for decision-making has a variable weight of relevance, which relies heavily on something that computers lack: emotions. An entire thesis can be written on this subject only.

Image databases that use textual descriptions, discussed further in chapter 3, are known to return irrelevant images, or not to return images that should have been returned at all. Like with the text documents, the description of the image can be wrong or incomplete. Actually, an image index can never contain enough descriptions to fully describe all visual and semantic clues in an image.
CBIR-based image databases are indexed by low-level visual features. The retrieval of images from these databases has issues because of the lack of high-level features. This is addressed in chapter 4.

(c) The detail level of the submitted query

Users need to narrow down the domain and pre-filter the information themselves before submitting the query to the system. If the query is too general, the system will return a great amount of information, which results in still having to browse through information. When a query is detailed enough, the desired information will be presented by the system. In the example of the frog-query mentioned before, it is better to enter 'poisonous frogs' than entering 'frogs' as search query. Users can refine their query when the amount of returned information items is too large, or leave out detail from the query when the selection of returned information is too narrow.

2.2 Retrieving digitally stored photos and images

A small survey amongst technology-aware students made clear that they usually look for a stored image by browsing. The creator of an image globally knows when she took the photo and browses a collection of images from around that date, to retrieve the image. Other than searching by date, the images are stored sometimes in directories or folders with a small description about the entire image set, like 'Barcelona 2005'. Indexing by means of manually adding textual descriptions to every single photo is too time consuming to these users.
In his book 'Exploratory Image Databases', Santini wrote how Ornager (1997) and Markkula and Sormounen (1998) did a study on how journalists searched for photographic material. These findings are applicable to the general photo-search

behaviour of people. When searching for an image, there are the several search tasks or goals that an individual may have. One task is the search for a specific image. In this case, the individual knows exactly what the image looks like, for example ‘The photo with me as a child with that silly orange hat in front of lake Mackenzie at Fraser Island’. Or, when speaking of a world famous image, Eddie Adams’ photo from 1968, in which a South Vietnamese police chief executes a Viet Cong suspect in the streets. An other search task is the search for any image about an event, object or subject, like photos of the carnival in Brazil or a photo of the Dutch queen. In that case, the user filters the selection of retrieved images her self. She will pick the most suitable image for her application. There is also the task of searching an image from an entire image-domain. A user can for example look for photos by satellites, images that are predominantly orange or images that are usable symbolically, like ‘happiness’ or ‘luxury’.

There are several motivations for image making and retrieval from stored collections:

Information inquiry: A person wants to know what something, someone or some place looks like, or used to look like;

Entertainment: Typically, users globally look for a category of images after which they browse the presented selection to be entertained;

Comparison: A great part of image retrieval is for comparison purposes. For example, one can think of comparing techniques in photography, determining economical valuing objects in images, comparing scientific images like medical scans or seismologic graphs or creating a face composition for law enforcement uses;

Reinforcing memories: Viewing images in order to reinforce a memory is closely related to viewing for entertainment, with a difference in emotional involvement. Looking at visuals can help to bring back memories together with the accompanying emotions;

Sharing: Photos are a good medium to share experiences with others. It is a reasonable representation of how things were at the time when and location where the photo was taken;

Illustrating ideas: When there is no room for a thousand words, an image can explain a lot, effectively;

Inspiration: Photos can work as a source of inspiration;

Navigation: drawings, graphs or photos can serve as maps for directions;
Mapping and analysis: flow-charts can help with solving complex questions;
Showing what human eyes cannot see: Microscopic images, infrared photos, satellite images of out-of-reach planets and heat maps can show information that is not visible with the human eye.

2.3 Usage scenarios

As said in the previous paragraph, image retrieval has many uses. In some cases, users know exactly what they are looking for, while in other cases, the user would like more freedom in looking for images. A description of a few different cases in which image retrieval is or can be usable for helping a user with a task follows. Every case has its characteristic uses and demands.

Car sales

When people look for a car online, the systems usually provide several variables, which are searchable for a suited car. These variables can be the brand of a car, its preferred colour or the body style of a car, like a 'convertible' or 'van'. This way, it is hard for a user to find a car of her liking if she does not know what a 'coupe' or a 'sedan' is. What if a user is looking for 'a cute little car with beautiful, shiny wheel caps'? How can a user find a car she has seen on the streets but of which she does not know the brand or the model and the year in which it has been manufactured? Visually looking for characteristics of cars might be a good way to find the right car. One could think of a system which lets the user construct a car from a separate parts. The selectable parts could be a type of front/nose (sporty, long, short), the type of backside (high, low, more or less luggage space), the car height (sports car, van), wheels, the colour, etc.

Graphic design

When an article or ad can or should feature an image, a designer looks for several specific aspects in an image. These aspects can be content or composition based. When looking for an image from the content based point of view, often a mood, colour or structure is sought after. Designers are likely to look for archetypical or metaphorical imagery since they are a good way of communicating with, because they can rely heavily on the images' semantic meaning. The best way for finding an image in this case is by browsing. The Tony Stone stock photography books used to be very well appreciated by designers. One could take a book, look for a theme and browse through the book. Next to providing an image, the books provided inspiration. Designers often end up choosing a different image than they had in mind initially.

From a compositional point of view, designers can look for other aspects within an image, which have nothing to do with the actual contents of an image. Sometimes the layout of an ad has room for an image on which the title of the article should be placed on top of. The area surrounding the title or text should not contain a heavily structured background or a certain colour. This way the image does not interfere with the readability of the text. An article or ad usually has room for text and images, and the balance between these is usually chosen during the design process. This is possibly because of the image that goes with the ad, since altering an image to fit with a text is harder to do than the other way around. Manipulating the content of an image is time consuming. Therefore, the static character of an image often determines the layout of the ad, and thus the placement of the text. Designers would profit from image searching tools with a functionality that lets them define special areas, which should (not) contain certain objects, colours or textures in images, during a search.

General consumer home image search

This user group is obvious, yet important since it contains a great amount of digital photography users. Currently, many people own a computer. They switch from analogue to digital photography because of the non-existing film and printing costs. The group consists of many different kinds of users regarding their experience level with digital media. Some are very familiar with the use of computers while others are not. It is hard to define why and how they would like to use image retrieval systems. That is why creating an application that is suitable for home consumers is extremely difficult.

Characteristically, the regular consumer with a photo camera takes many photos of friends and family. Parents document the growing-up of their children and digital photography gives them the chance to make more photos at a lower cost. Although digital photos can be printed, the majority will continue existing only in their digital form. Special events like a day out or holidays produce an additional amount of images, with different contents. Because of lower prices on digital photo cameras, more people can afford them. Teenagers, for example, bring small cameras or mobile phones fitted with cameras anywhere to take photos of friends when they go out. In addition, cameras on mobile phones continuously improve in quality. Images made with mobile phones will not only be used for mainly viewing them on small phone displays any more. In other words, the images made with these cameras will no longer be of a lesser quality than a regular digital camera and because of this, pictures taken with these phone cameras can be treated as if they were taken with a regular digital photo camera. The increasing amount of high-quality phone cameras adds to the already increasing amount of digital photos.

Programs for consumers typically use keywords to index images. ‘Tagging’ photos with keywords is time consuming and not every consumer might be familiar with the concept of tagging and keyword searching. A more natural way of looking for images is needed, since there are many regular consumers with a low level of knowledge about digital media. Thus, photos are usually stored in directories or folders by date, or named after the occasion. This is the easiest way to find back series of images. Regular consumers treat images the way they used to before. The first (analogue) photos of a newborn were put in a album, but as years passed, newer photos ended up in envelopes in drawers. It is hard to expect from people to index the extensively larger amounts of digital photos, when they did not even bother to sort their analogue photos.

The regular consumer is rather a browser than a searcher. Analogue photos can be taken out of a drawer and viewed in a social setting, stirring up memories. Digital photos are not suitable for this. One reason is that they cannot be viewed in any location. Other than that, when viewing them, every viewer needs to see the same image, whereas analogue photos can be passed on, so every viewer can take her own time to look at a picture. In

addition, every viewer can go through a different series of analogue photos. Viewing digital photos is comparable to viewing analogue slides. It takes some set-up time and the interaction moves from interpersonal to interaction with a photo-viewing object.
The question whether general consumers wish to be able to search through their digital photos needs to be asked, since the consumer is traditionally a browser. Their interaction with photo material may have to change. Tangible items, such as analogue photos, are easier to find since they take up space physically, while digital photos only take up virtual space in a computer's memory. When a drawer is full with printed photos, new analogue images need a different place, where computer memory is expandable. The photo prints are stored with more care and are subsequently easier to find. One can store an infinite amount of digital photos. In order to be able to find digital images, they need to be stored with the same care as analogue photos, rather than storing them in dated or named directories or folders. Users need to become more aware of the way they store their images.

Medical science

In medical science many images like x-rays, CAT-scans, ultrasound images, etc. are being produced. The amount of images keeps on growing strongly[4], since new medical visualisation systems are constantly improved. It is possible to quickly produce images of a patient. These images can be stored quickly as well. Access to medical images by means of CBIR can decrease the amount of time it takes to make a correct diagnosis of a patient's condition. Other than that, for clinical decision-making, as well as medical research and medical education, linked image retrieval systems can be of great value.
CBIR systems in general never work without errors. There are still problems of irrelevant images being returned by the systems and relevant images that are not retrieved. In medical systems, the error rate should of course be as low as possible. Because of this, textual indexes are still needed for image retrieval in medical applications.

---

[4] Müller, Henning et al. *A Review of Content-Based Image Retrieval Systems in Medical Applications*, chapter 2. Int J Med Inform. 2004 Feb;73(1):1-23

In medical visualisation, the difficulty with storage and retrieval of images can be even greater than in other fields, since visualisations are often multi-layered. In the case of actual 3D visualisations, the comparison or retrieval is a different story, since 3D data consists of mathematical comparable numerical vectors and coordinates.
An other problem occurs when two pictures, of lungs for example, might be similar in outline, yet one of them contains an anomalous area that possibly indicates a lung disease. It is this kind of detail which is important to medical image retrieval systems. These anomalous areas are created by nature and no two things made by nature are the same. It is possible to detect anomalous areas in images with a computer. Classifying them by computer however, is a problem. A system called the ASSERT-system, lets a physician outline the important areas in high resolution images of the lung. Specialist can only do this work, since they have the required knowledge about the subject. This makes correct indexing of these medical images rather expensive.

Architecture and art

Digital catalogues are created from old and new artwork. Paintings are digitalised so they can be preserved and spread. Photography is the medium that is used to store three-dimensional works of art. These works can be small installations in a museum or even entire buildings, in other words, architecture. These works all contain characteristic features, which can be used for retrieving the image of the artwork. Especially when a viewer wants to know more about an artwork, because the only thing they know about the artist is that particular artwork itself. Looking for explicit features in the artwork would then be the easiest way to find more information about the artwork and its creator.
This is because the search method is brought back to the corresponding domain. Art is, except for poetry and other art in textual form, a medium that communicates mainly in the visual domain. Looking for works of art would therefore logically be done visually, instead of searching for a textual description. However, art makes use of high semantic meanings, metaphors and symbolism. These concepts are hard to communicate without the use of text, unless understood by the viewer. For example, the dog in Jan van Eyck's

painting *The Marriage of Giovanni Arnolfini and Giovana Cenami*, 1434, symbolises faithfulness and loyalty between husband and wife.

The age of digital imagery also provides ways to manipulate, combine and recreate existing art, as well as to create new art. Combinations of existing art-pieces or derivative works contain semantic connections with their originals. It could be preferable to retrieve originals together with derivative works and vice-versa in the field of art history for example.

2.4 Issues regarding user's image retrieval desires

Developing image storage and retrieval systems for images within pre-defined context can provide usable solutions. An example is PARISS by the University of Leuven (Belgium) and CWI Amsterdam. In a short presentation, the sought after image actually turned up but the program is developed to process textile patterns. It is not developed for images with irregular patterns, like photos of cars. Thus, by pre-defining a context for a new image retrieval system with its stored images, functionality is limited.

Interpretation issues do not only exist because of people having different interpretations. The ambiguity of images is the main reason for possible different interpretations. A single image can be useful because of its esthetical value, because its layout suits the space it needs to fill in an ad, because it has journalistic value, because it brings up memories, because it explains a difference, etcetera, all at once. This is only about the entire image. There are also separate objects, colours, structures or shapes that can be interesting for some reason (Figure 2.1).

It is also possible to describe an image without knowing what it actually looks like. For these search tasks, a system based on image browsing is more suitable. A characteristic for a browsing system is that it returns images that differ from what the user expects. In this case, this is a feature rather than an unwanted result. When a user does not know exactly what she is looking for, she likes to get inspiration from the returned images. Search task or goal needs can even change during an exploratory search for images, when the images that are returned by the system inspires the user to think differently. This makes the development for image retrieval images even harder.

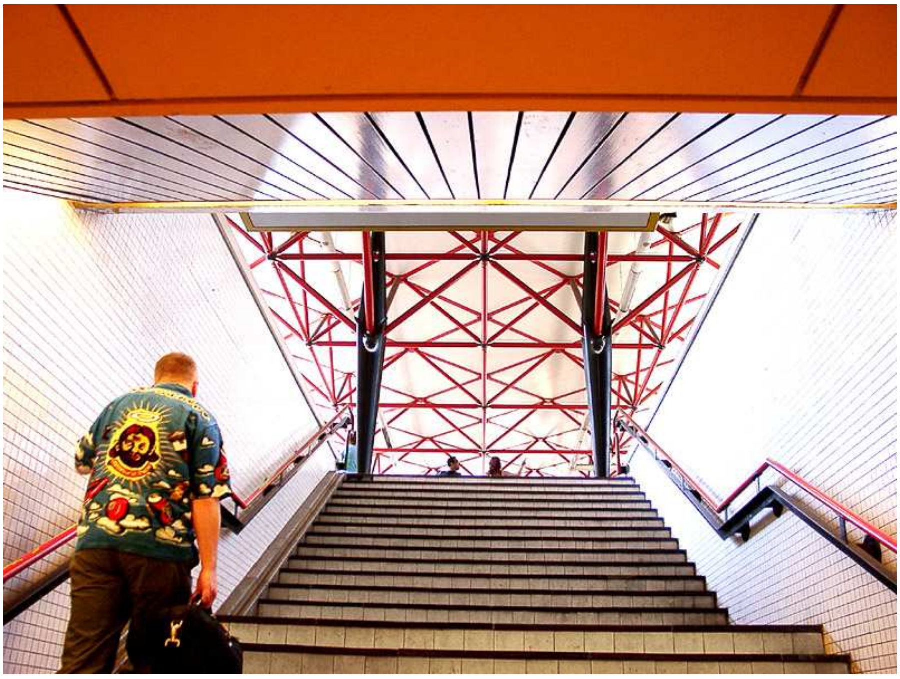

Figure 2.1 - The ambiguity of images. This image can be interesting for several reasons. It can be an example of a certain roof-construction. It can be printed with an article about religious prints on clothing. It can be used to release the fantasy of viewers by explaining that the vertical black bars as eyes, the stairs, and stair handles as arms resemble the shape of a creature.

3. APPROACH: MANUALLY ENTERED DESCRIPTIVE METADATA

Manually entered descriptive metadata (MEDM) is a chosen term for the process of describing the contents of an image with text. It is a method for categorising images. The descriptive text regularly has the form of separate words, which are known as ‘keywords’ or ‘tags’. This way of indexing makes it possible to place images in a context. As a result, images can be retrieved by looking for the textual references to, and thus contexts of the images.

Search-by-keyword systems may suffice in serving the user when a searched image is a specific picture about a certain subject, event, object or person. However, as said before, manually entering descriptive data is not a solution people like to use. It is merely useful for a small selection of images, which the creator of the image wants to get to other’s attention. The Flickr website (www.flickr.com, recently acquired by Yahoo!) for example, is a very popular site for exposing and sharing images that utilises keywords for indexing images. The site gives users the opportunity to upload imagery and add keywords that they think are relevant in order to find the images with a keyword search.

Flickr mainly serves the following purposes:

For sharing general photos from events of any kind, such as holidays or parties. Sharing photos online provides accessibility to photos for friends and family at any location with a computer and access to the internet (
Figure 3.1);

As storage space for images, often humoristic images created by thirds, for use in web-logs;

As storage space for non-photographic images like graphs or screen-shots for sharing information (Figure 3.2);

As portfolio or exhibition space for artistic photography and illustrations (Figure 3.3);

As an online community discussion forum, mainly related to photography (Figure 3.4);

Theoretically, the more extensive the descriptions of an image, the higher its find-ability. On this site however, one usually uploads an ‘exposure-worthy’ selection of the entire image collection by the author. This way, the amount of uploaded images that need a manually entered description is kept small at every batch of uploads. The images hardly contain useful descriptions when a user uploads a larger amount of images. In that case, all images from the same batch contain the same keywords like ‘birthday party’ and ‘2005’. This does link the photos to each other but it does not provide an option to find a specific image from the series. In other words, when using keywords for large amounts of images, details within images are not important enough to describe. This however, *is* needed to be able to make an image uniquely find-able.

Keyword searches are characteristically searches for a ‘sunset’ on a ‘beach’, the ‘Queen (of the) Netherlands’ or a ‘Callithrix Callithrix’ or ‘Pigmy Marmoset’ monkey. Textual search queries will return one or more images about the queried subject and a satisfyingly suiting image is picked from a larger set of images presented by the system. The user is interested in the subjects or objects within the image and works as a filter herself. To make an image findable by keyword search, one needs to think about how the target audience, or the searching user, would look for an image. The general context of the image library and the jargon used for the keywords is determined by the system’s user and vice versa. Tagging an image collection is useless when the tags are written in a different language than the user expects or wishes to use. There are as many ways of describing images as there are images, search tasks or goals.

It is also difficult to tag images for other people to find. If someone would want an image of a cat, that person could choose the plural word ‘cats’ in a search query, since she is looking for images of cats in general, not ‘an image with a cat’. However, an image with a cat is very likely to be tagged with the word ‘cat’. Adding the tag ‘cats’ to the image, means that the author of the image classifies the image, next to tagging it according to the visual clues in the image. After that, people who look for a ‘cat’, as well as people who look for ‘cats’ can find the image. Someone who looks with the keyword ‘feline’ however cannot. The only way to make an image in a MEDM system as findable as possible is to add as many relevant tags as possible. The description of an image should depend on the way the indexer thinks the finder will search.

Manually entering descriptive metadata is a well usable indexing method for organisations like stock photography agencies such as Getty Images, formerly known as Tony Stone. Their clients still use the same searching method as they did when they used stock photo books (Chapter 1). To the agencies, investing time in extensively describing images is necessary. If they would not do this, visitors of the archives would not be able to find the images that suit their expectations. As a result, the images would not be sold, which is the core business of stock photography agencies in the first place.

3.1 From semiotics to semantics

A photo itself does not have a meaning. It is merely a rectangular shape with coloured amorphous blotches of various sizes. When looking at an image, we interpret and compare the blotches to objects or situations that we have seen and identified before. The cumulative of all visual clues in a photo gives us the ability to constitute context and meaning. These interpretations can be different to each individual because of language, cultural background, or psychological state. This will be addressed later on in this chapter.

Adding semantics to a photo by adding keywords is a powerful way to make images retrievable from a stored collection. Semantics is the study of meaning of symbols. In linguistics, the symbols like words and word combinations like sentences, derive meaning from conventions, interpretation and context.

It is possible to form a picture in mind, before we would start a database search, without describing it in text. Visually, like a short movie rather than in textual form, it is possible to remember that a flame is hot and that it burns achingly when touched. Memories like these are not part of semantics, but of semiotics. Semiotics is the study of signs, like the yellow and black lines on a wasp's back to indicate danger, or the odour of spoiled food, which tells us not to eat it. After the first touch of a flame, a different flame will still be recognised as a flame since it transfuses the same signs (colour, temperature, movement, emission of light). The following part of this paragraph will lead along semiotics to semantics.

When memories like the burning of a flame need to be communicated, especially when the actual subject about which is communicated is not at hand, some kind of language is needed. Communicating about the flame and its heat can be done with the use of a photo or drawing combined with sign language. To prevent another person from burning herself on a real flame, one could do a performance of what would happen if she would touch the flame, assuming she knows the concept of photos being a two-dimensional representation of the real world. Conventions in a language like performance are quite natural. (The term convention may not even be appropriate for this kind of language). Animals can communicate by using sound and body language and a dog can even communicate with a cat to a certain extent.

Fortunately, spoken language helps us to communicate more effectively. Sounds we produce to communicate with can have meaning derived from a natural sound. Pointing at (a picture of) the flame and saying 'ouch' says enough and the sound or word in English is not very different from other languages. Other, more complicated words have meaning because of conventions. Certain words sound a certain way and have one or several certain meanings because at some point people agreed to this. Still, for communicating, one does not need to know how to write. In classical Greece and Rome, literacy was often only a privilege for the upper class. It was not until the major upcoming of printing press in the late 18th century, followed by the introduction of compulsory schooling in the 19th century, that literacy became available to the masses.

With the use of *written* language, we use *symbols* to describe something, without having to resort to drawing or otherwise producing literal images, gestures or sound. This is semantics: the use of symbols for a meaning. Other examples of symbols that describe meaning, and consequently have (their own) semantics as well, are Chinese calligraphy and Egyptian hieroglyphics.

Written language can be used to communicate about things that are not tangible as well, which is a great advantage of its use. Words like 'empty', 'love', 'tasty' or 'size' are hard to depict in images. Images can give clues about, show differences between, or represent things related to words like these, but they cannot communicate the exact meaning of these words.

The main advantage of using text to index images in retrieval systems:
Written language is useful when a message needs to be sent across a distance, or transmitted over time. In the case of image databases, an image is stored at one single place, to be able to access it any time. Since text is useful for communicating across time and distance, it is a logical choice to use it for describing stored images.
Other than that, developing software for searching through text is easier and requires less computational power in comparison to applications for comparison of shapes, colours and textures in an image.
Text can also describe the type of an image, like ‘black and white photo’, ‘aquarelle painting’ or ‘flowchart’.

3.2 Interpretation issues

MEDM is quite usable for image databases where the domain of the image collection is narrow and people, who have the time and experience for this, do the filtering of the images. For describing medical images, one requires different knowledge and experience than comparing company logos for copyright-related issuses, which requires special training as well. With a narrow domain comes a defined language space, or jargon. Users of systems within a domain know of conventions and textual terms, which relate to the system’s context.
When the image retrieval system needs to store, index and retrieve images from a broad range of subjects, issues arise. These issues come forth from interpretations that are influenced by personal, cultural, educational and psychological experience.

Personal interpretation

A mood or state of mind can influence the way an individual interprets text or images. According to Ervin Goffman[5], a person plays different roles in different locations and in different social circles. With this way of behaving comes a mental state. At home, one uses a different jargon than in a professional environment. A single individual can choose

[5] E. Goffman, *Presentation of Self in Everyday Life*. Knopf Publishing Group, 1972

different words to store or search for the same subject, depending on where she is. The search task or goal can even change during an exploratory search for images, when the images that are returned by the system inspires the user to think differently.
Every other individual has her own interpretations. One person may use the word ‘circle’, as the other might choose for ‘ball’ and yet another for ‘sphere’. When photos are tagged for personal use, this is not so much of an issue. There is a high chance that a user chooses the same word as index tag for a photo as she would for retrieving that photo from the system. In this case, she might even relate the tag directly to that photo only - not to other images with a ‘circle’ in it.

Cultural infulences

What someone thinks is important, relies highly on the cultural background of that person. The environments in which people live affect their thoughts, norms, values, aesthetic valuation, and consequently interpretations and name conventions. According to the Sapir-Whorf hypothesis, the Inuit have over thirty perceptions, and thus words, for types of snow. Peter Gordon researched the language of the Pirahã tribe in Brazil, which allegedly only has three counting words; one, two and many. Both theories are criticized, but not for the fact that surroundings determine the need or lack of need for words to be able to express oneself. This can also be interpreted as a level of education. The Inuit are highly educated or experienced in the ‘science of snow’ and its uses and the Pirahã tribe has not learned to count because they have no need for it.

Language

Language is also an issue with MEDM-based image databases. Before the development or first use of a tagged-based image database, the system’s main user language needs to be determined. This way, every user is expected to use the same language, which is less a problem when the system used in a single country.
Tag-based image databases are very useful for the internet. This is mainly because searching in and selecting from textual databases is fast and compatible with web-based

databases. Online image databases are available to many countries since internet is intrinsically accessible from all over the world.
English is the main used language online. Still, there are many countries of which the residents' level of English (spelling) skills is not high. This makes English based systems hard to use to non-English native speakers. At Flickr, non-English native speakers regularly tag images in two or more languages. Misspelled keywords for indexing photos (because of a type-error or lack of language skills) result in erroneous image retrieval.

Recapitulating, differences between people results in differences in interpretation and needs. One way of closing down the gap between people's interpretations is getting as much information from as many different people as possible. This way, a general 'truth' or interpretation can be formed. In addition, users will have to agree to this system and set their minds to think more general, instead of using their personal interpretation. It becomes impossible to let users interact within the same system when the difference between people's interpretations becomes too great.

> *Image retrieval in general, could possibly be done more efficiently when possible errors, because of lexical differences, are avoided. Therefore, one single, universal language like Morse code should be introduced.*

3.3 Talking with words when we mean images

When a user does not precisely know what she is looking for, when layout of an image is important, or when the sought after result actually changes during the search, keyword searchable systems are less suitable. In that case, the image as an entire graphical entity becomes more important. The emphasis shifts to the entire image, rather than the depictions of objects and subjects within the image. The user is looking for a certain layout, colour or texture.
MEDM-based image retrieval systems rely on the cohering structure of text, because images lack the intrinsic structural coherence of linguistics. This lack of coherence comes

from the ambiguity and subjectivity of images (Figure 2.1)[6]. Language and text lean on general conventions, which makes sure that words mean the same to (virtually) everyone. The context of a text in which words are used reinforces the correct interpretation of words. The structure in text helps the computer to get a grip on indexing images and to form a link between people's and the computer's interpretation. It is exactly because of the lack of structure in images why MEDM-based image retrieval systems still do not suffice in finding images, since adding text to images is more of a bug fix to the problem rather than an actual solution. It is like wrapping a photo in a piece of paper and writing a description of the photo on the paper. Text and images do not mix well, since they come from a different domain.

Modality of words is less an issue within (lingual) communication in comparison to the ambiguity of images. The translation or re-interpretation is done automatically by people. A message still comes across perfectly when one person prefers using the word 'circle' but another uses the word 'sphere'. Both terms are interpreted the same. To a computer however, these two words are completely different of course.

> *Using Manually Entered Descriptive Metadata for indexing images is communicating within the textual domain, about the entirely different, visual domain. It is like describing a scene to a blind person (textual domain) with sign language (visual domain), with a deaf person as the person who describes the image (ambiguity, personal interpretation of someone who's experience lies within the visual domain rather than the lingual domain).*

In order to find textual documents one can use characters and words. In order to find images one needs to use visual language - colours, shapes, placement, orientation, special organisation, etc.

Content-Based Image Retrieval is an approach to image retrieval in the visual domain. This, considering what has been said, looks like a better, more promising approach.

---

[6] Dr. B.A.M. Schouten. "*Giving Eyes to ICT! Or: How Does a Computer Recognise a Cow?*". University of Amsterdam, 2001, p 14

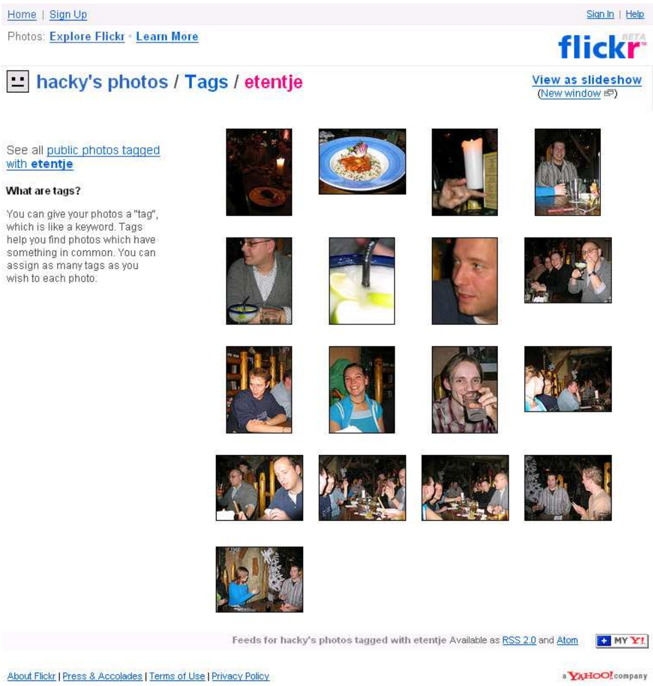

Figure 3.1 - Flickr is used for sharing general photographs

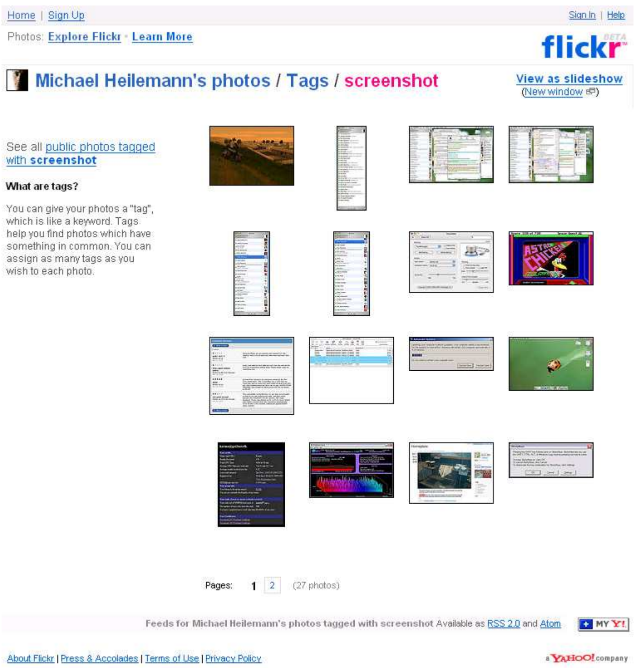

Figure 3.2 - Flickr is used for storing third party graphics

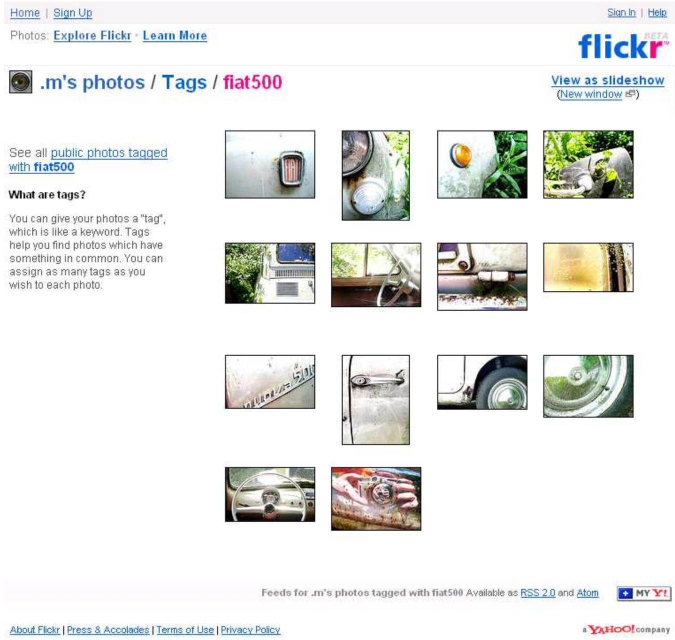

Figure 3.3 - Flickr used for artistic photography

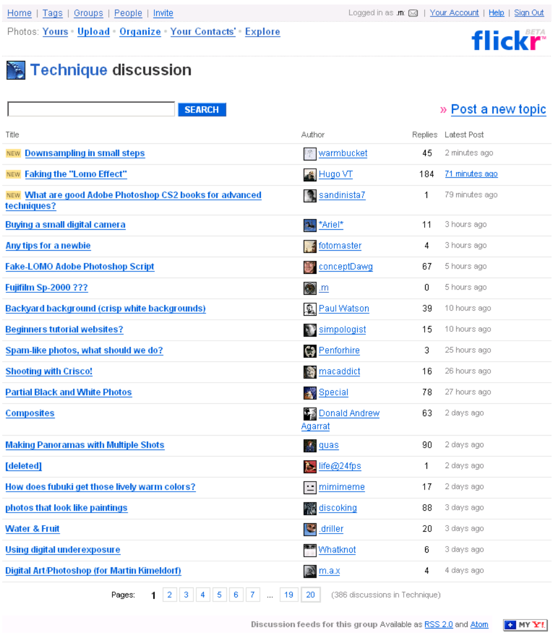

Figure 3.4 - Flickr used as online community

4. APPROACH: CONTENT-BASED IMAGE RETRIEVAL (CBIR)

Content-based image retrieval, a sub-domain of computer vision, is a system in which a computer analyses an image to extract visual features. These features are known as low-level features. This means that the features are visually detectable, without assigning any semantic meaning to them. With CBIR, images can be indexed relatively quickly and easily. The user uploads an image and the system indexes it automatically. However, the many image feature extractions in CBIR are hard to develop in comparison to MEDM-based systems. The features are also hard to understand for end users. In addition, lack of semantics gives bad results after querying for an image from a CBIR system. This lack of semantics in CBIR-based systems is known as a semantic gap (chapter 4.4).

4.1 Seeing, comparing and distinguishing

People are able to do something else that computers cannot, which is a great issue in the context of CBIR. They are able to see one depiction and recognise transformed versions of that same object (Figure 4.1). This is the Invariance-property of the Gestalt theory. In addition, they can imagine what an object looks like on the side that is not within view. Next to that, they can translate certain characteristics from one object to another, enabling them to categorise the object the very first time they see it. When a person sees an Old English Sheepdog (Figure 4.2), it is recognised as a dog. When the same person sees a photo Chihuahua (Figure 4.3), she identifies it as dog as well, regardless of the fact that it has a different colour, shape and absolute size. The orientation and pose of both animals do not even need to be the similar to be able to identify both as dogs. A cat however, features similarities in comparison to a dog (Figure 4.4). It also has four legs, a tail and two ears, yet people recognise it as a cat, not a dog.

4.2 The computer is wrong

The best-known examples of extractable image features are colour, shape and texture. In case of colour, we can think of the overall colour of an image, or the colour of separate shapes within the image. Features are extracted from every new image that is introduced to the system, by means of mathematical algorithms. The extracted image features are then stored in a database for each image. Storing as many as possible relevant features increases the chance of an image to be found. A user can then retrieve images from the system by looking for image features. The user defines visual criteria for the retrieval of images she is looking for: she queries (chapter 2.1). Querying in CBIR can be done in several ways of interacting with an image retrieval system, as described chapter 4.6. The combination of all indexed features of a single image should create a unique description of that image. This single image should be retrieved by the system when a user defines an exactly matching feature-combination as a query.

Literally, an endless amount of optional low-level features could be extracted from images. When indexing a new image by means of CBIR, the computer has to decide which features are most useful and which are less relevant for indexing and later retrieving an image. Some features are more important than others are. Features that seem important because of size, amount of extractable information and redundancy are marked as such. Features extracted from smaller parts in an image, which contain less information are treated accordingly.

As said earlier, all features from one image together form a link to that single image. Photos contain coherence between features. For example, after a deduction of dimensions the following clues can all tell us we are looking at ‘sky’:

it generally sits in the upper half of an image;

it can have a variable size or shape;

it often has a blue colour, sometimes purple, pink or orange;

if blue, a yellow or orange circle could be seen within that blue shape;

the circle itself (if it actually is the sun, not a ball) would never be in the lower half if the blue shape were in the upper half;

the circle’s position within the blue shape is variable;

The computer cannot detect this coherence between all variable features, since recognition is based on semantics. Even if we would teach the computer these variables, would it be able to understand the concept of 'sky' when it is reflected in water (Figure 4.5)? What happens when the sky is the larger part in an image, but a small silhouette in the lower-left bottom is relevant to the user? As Dr. B.A.M. Schouten states, it is impossible to select the most important feature automatically, since the most important feature is never the same[7]. This is not just the case at the indexing stage of images, but also at the time of retrieval. The importance of a feature depends on the user's task or goal and her personal interpretation (chapter 3.2).

> *The user will have to tell the system what is important, what is similar and what is different to her, to make it possible for the computer to retrieve what a user is looking for. Interaction is the key to satisfying image retrieval results.*

4.3 People are wrong

People can miss-interpret colours from an image. When the computer-measured colour of a wall is grey, a person interprets the wall as white. Edward H. Adelson elaborates this observable fact with his Checker Shadow Illusion[8]. In the left-hand-side image in Figure 4.6, the shades of square A and B seem to have a different shade of grey. The image on the right-hand side shows that the actual colour of both squares is equal.
Comparison by people is done differently than comparison by a computer. Like with a personal vocabulary (chapter 3.2), the amount of detail a user can detect depends on the personal experience with a sub-domain of images. An antiquarian for example, can see whether an old looking piece of furniture is actually antique, just old furniture or a new piece of furniture that is forged to look like antique (Figure 4.7). Another good example is the value of a Lomo photo camera. To someone who is not a photography enthusiast, a Lomo LC-A camera (Figure 4.8) is an old camera which still holds 'old fashioned' 35mm

---

[7] Dr. B.A.M. Schouten. "*Giving Eyes to ICT! Or: How Does a Computer Recognise a Cow?*". University of Amsterdam, 2001, p 22.

[8] E.H. Adelson. "*Checker Shadow Illusion*", 1995. Figure 4.6

film and produces bad photos with wrong colours and an image-masking vignette around the edges (Figure 4.9). In the Russian Republic, these cameras are lying around in people's cupboards. To Lomo photography enthusiasts however, this is a highly appreciated piece of creative mechanics, loved because of the saturated colours and the beautiful vignette. These cameras are sold for as much as US$200.

One image in either JPEG format or GIF file-format looks the same to people, but very different to a computer, since they are two different file-types. To show this, an image is printed in both file compression formats in Figure 4.10 and Figure 4.11. They are exaggeratedly compressed, in order to be able to see difference between both. The content of both the images is the same, however the images are actually quite different (mainly in colour). People need to program applications in such a way that the computer sees the images in Figure 4.10 and Figure 4.11 as equal. In the field of CBIR, these problems could be solved. What if the same image is in black-and-white? At one time the colour is important, another time the image is.

4.4 The semantic gap

Chapters 4.2 and 4.3 show the difference in interpretation between people and computers. These differences cause the existence of the semantic gap. The semantic gap exists at occasions where a denotation of a subject or object does not match the semantic concept of the same subject or object in the mind of a person. In other words, in CBIR, the problem is that low-level features that are extracted from an image by the computer are stored as such. No meaning other than visual features are stored with the index. To the computer, the shape of an airplane can be similar to the shape of a whale. Both can be grey blotches, surrounded by a blue coloured area. To people, the blue backgrounds mean a lot. Together with the shape and other features of the object and the blue area, users determine whether they are looking at an airplane in the sky or a whale in the ocean.

4.5 Adding semantics

Natural language processing (NLP) is part of artificial intelligence science, combined with linguistics. One approach to address the semantic gap problem in CBIR, discussed in chapter 4.4 is through this science. The idea is to get computers to interpret language the way people do. In practise, people could textually describe an image. The computer would then try to interpret this text though NLP. If the text says something about the image, it probably says the same or something similar about an image that looks similar. The difficulty with NLP is that a single word can have different meanings. Other than that, the same thing can be said with different kinds of sentences. Small differences in pieces of text can mean large differences in meaning. These are some problems that make it hard for a computer to correctly interpret human language, especially when language is used in a metaphorical way. Other than that, NLP is only useful to CBIR when the images within the system are described extensively. Text-to-speech technologies can be of great help here.

An other way of adding semantics to CBIR is by combining it with MEDM. Next to low-level features, high-level descriptions are added to the images.

That darn ambiguity

The ambiguity of language interpretation and personal vocabulary is not an issue within CBIR. However, image ambiguity is still a problem. Unless we create computers with character, that can think like humans, and are able to understand what people actually mean when they communicate, we need to leave the interpretation of visual matter to people.

> *Computers are good at calculating, which that is what they should be used for. People can in turn do what they are good at themselves: interpreting, decision-making and determining relevance,.*

Since interpretation by people is ambiguous as well, we need to get them to speak the same language in the visual domain. One way to achieve this is to go back to basics of visual communication and lose all information that may cause misinterpretation. By using geometrical shapes, without naming them, things can never be interpreted. A child from China understands geometrical shapes in the same sense as a child from the Netherlands. It is one of the basic elements in the growing-up phase in a human's life. Children can recognise shapes before they can even name them (Figure 4.12).

> *The advantage of CBIR is that it does communicate within the visual domain. Its disadvantage is that it works with ambiguously shaped blotches, instead of identified objects.*

An application like Blobworld works with coloured shapes that are extracted from images. This, however still makes a shape ‘a blotch with a certain colour’, unless the blotch is described textually, which brings back all the issues of MEDM. One way to decrease the ambiguity of coloured blotches in images by flattening dimensions. To get rid of different interpretations in images, the notion of ‘sky’ should become something like ‘a blue area in the upper part of an image, often accompanied by a yellow circle and white blotches’ instead of something that can have many colours and shapes.

4.6 Interaction and CBIR

As said, when someone is looking for a concept like ‘comfort’, a semantic-lacking CBIR-based approach fails without help of language or user interaction. There are several different approaches to retrieving images from CBIR systems. Much information on optional interaction models and their usability is available from published CBIR surveys[9, 10]. Several well-known and remarkable models will be addressed here.

CBIR systems are always queried by features. This paragraph emphasises on the query by selecting features by the user. The most basic way to interact with a CBIR system is by directly selecting features which the user thinks to be relevant in comparison to the image

[9] B. Johansson, *A Survey on: Contents Based Search in Image Databases*. Linköping University, 2000

[10] R.C. Veltkamp, M. Tanase, “*Content-Based Image Retrieval Systems: A Survey*”. Utrecht University, 2002.

she is looking for. A photo with a red car in the foreground could theoretically be retrieved by selecting red as a dominant-colour-feature for querying. Naturally, images of red fish, red brick walls and photos taken in a discotheque with red lighting, are possible to be retrieved as well (assuming they are actually stored in the database).

A problem with querying by feature is whether a user realises or knows which features she can query for. If the system is based on colour dominance only, she can impossibly look for a black and white photo (colourless) or the bark of a tree (structure). On the other hand, when an endless amount of features are at her disposal, she might not understand which features could return the image she is looking for. To prevent the user from having to work with features directly, several interfaces have been developed for querying. Two of the most used will be discussed below.

Query by example

CBIR systems are often fitted with Query By Example (QBE) interfaces. These interfaces let the user provide an example image that can be used for comparison by the system. This can be done by means of selecting an image that is already present in the database or by uploading a new image. The system can then look for images with similar features that will be presented to the user. Preferably, the user can give relevance feedback to the system by interacting with the retrieved selection. She can point out which images are relevant to her and which are not. Next, she re-submits the query, refining her search and narrowing down the diversity of the retrieved image collection.

Other than just providing information about whether the entire image is relevant or not, users are helped by being able to point out areas within images that are usable for querying. More detail in the query can be provided if the user is able to supply information about the amount of relevance (weight) of these features.

Query by sketching

Some systems let the user query for an image by sketching a drawing. Although this sounds like a good interface, it is hard to use for people who cannot even get a real life looking object down with a pen on paper, let alone by using a mouse. It is also

understandable that query-by-sketching is hard to use when the images in the database contain details that are hard to reproduce by hand, for example X-Ray images, which contain many gradients.
In the CBIR survey conducted by Utrecht University[10], students stated that the sketching interface can be useful but is often used erroneously. To give an example, drawing filled polygonal shapes would be less suitable for a database with images of tropical fish. Users would be more helped with a system that lets the user 'paint' coloured patterns, regardless of the shape of the fish. In addition, no fish will ever turn up from the system, when the painting would be applied as a query in a system that contains architectural images, no matter how well and detailed the coloured pattern is drawn. Ideally, the sketch interface should enable the user to draw what she needs, but disable her what she does not need.

> *The quality of the created query image should not rely on the drawing skills of the user as.*

Existing applications

QBIC IBM

QBIC features two modules for image retrieval: colour search and layout search. The strengths of the colour search models is that it allows the user to choose colours to query for. Next, the colours' weight can be adjusted by changing the box-size of the selected colour. To a certain extent, this module works well. It does however, turn up irrelevant images as well.
The layout search module is about creating squares and circles in a single colour. Paintings in the database are hardly ever contain geometrically shaped areas with a single bright colour. It is hard to use this engine for retrieval by layout. This engine would be better useful for looking for paintings in a Mondriaan collection. A demo of QBIC is available at rhe site of the Hermitage museum: http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English

Blobworld

Blobworld uses feature extractions that are detected as a singularly coloured blotch within an image. A user can select the shape, or blob from an indexed image to look for images that contain similar blobs. The strength of Blobworld is that it features weight ajustment of the selected blob and the weight of the background. The weakness is that it is only possible to select one single blob at the time. To increase find-ability, the images are stored in pre-defined sub-categories (Animals, People, Flowers, Ocean Scenes, Outdoor Scenes, Manmade Objects). This seems weird for a system that wants to prove image find-ability by visual features.

PARISS

PARISS learns from the user by her interaction with the system. In an iterative process, a user can create clusters of images. The user separates the relevant images from the irrelevant images by clustering images. The system reads the clustering in comparison to low-level features and tries to cluster the other images within the system according to the clusters defined by the user. In the screenshot (Figure 4.13), every dot represents an image, which can be displayed and moved to a cluster of images by dragging and dropping. Every image can be ‘closed’ again, turning it into a red dot.

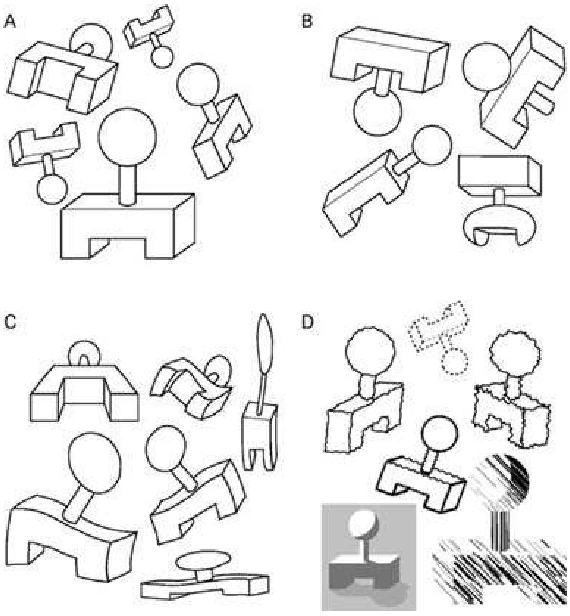

Figure 4.1 - Gestalt laws - Invariance.

Objects in C and D can be recognised as deformed versions of the object in A.

Objects in B contain similar features, yet are different objects.

Figure 4.2 – Old English Sheepdog.

Figure 4.3 – Chihuahua - Looking quite different from a Sheepdog, yet it is also a dog.

Figure 4.4 – Looking quite similar to the Sheepdog yet it is not a dog

Figure 4.5 - Not an image of sky

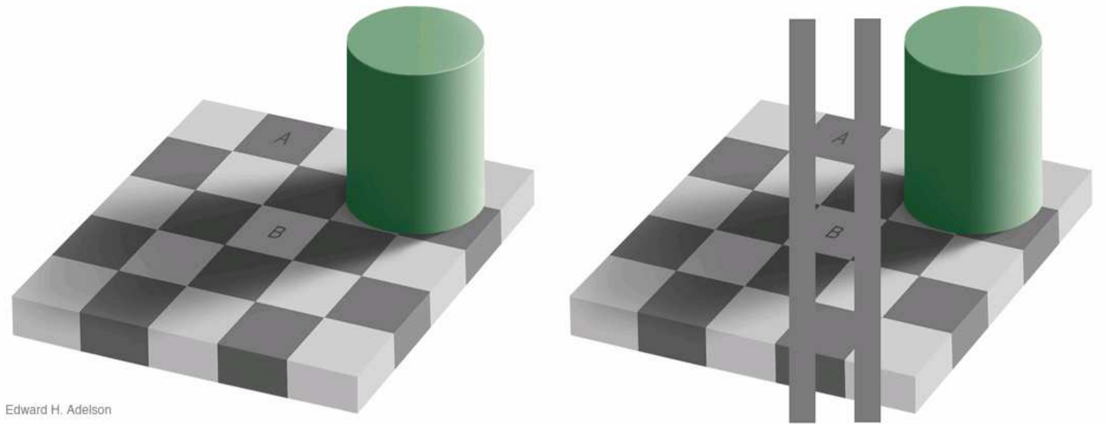

Figure 4.6 – Checker Shadow Illusion, courtesy of E.H. Adelson

Figure 4.7 – Antique? Replica?

Figure 4.8 - An original Russian Lomo LC-A camera. Junk to some, a valuable item to others

Figure 4.9 - A photo made with a Lomo LC-A camera. Highly saturated colours and a vignette.

Figure 4.10 – An image in (heavily compressed) JPG format

Figure 4.11 - The same image as in Figure 4.10, compressed (heavily) in GIF format

Figure 4.12 – Children learn to recognise shapes before naming them.

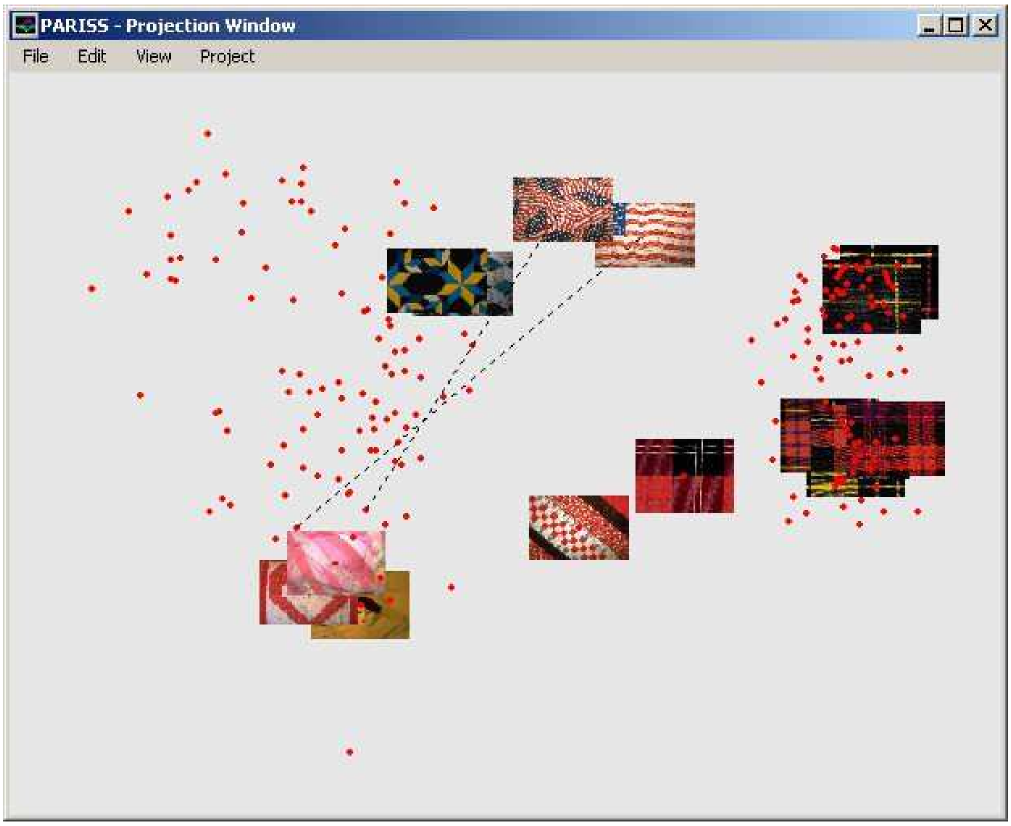

Figure 4.13 - Screenshot of the PARISS user interface

5. PROJECT: PHOTOINDEX

The Photoindex project is the practical part of this dissertation. The theory in this document forms the foundation for the Photoindex project. The project is an attempt to combine advantages of, and avoid disadvantages from both MEDM and CBIR, into a new approach. Photoindex is not the answer to all image-retrieval problems; nevertheless, it can be a stepping-stone for new approaches in image retrieval systems, with higher levels of usability and user-friendliness. It presents my approach on closing the semantic gap.

5.1 Concept

Like every image retrieval system, the goal of Photoindex is to retrieve images in collections of digital images. This is done by increasing the find-ability of archived photos, in a way that lies closer to the human perception, than computer vision (CBIR). The computer is merely used for calculating differences between a limited set of data, which consists of predefined, well comparable variables.

The main target-audience for Photoindex consists of users with a high level of awareness regarding image layout, colour and composition. This level of awareness is the common language they use, so that indexes by other people are interpreted correctly. Typically, these are for example graphic designers and photographers. Photoindex is useful for personal photo indexing by general users as well, since it relies on indexes that are defined by the user herself. The idea is that when a user is confronted with the way an image is indexed, like being aware of putting an analogue photo in a certain drawer. This process of letting the user decide which features she can remember the image by, is done by using a mixture of low-level features that can contain high-level, semantic information. This information is designed in such a way that it is hard to misinterpret, since it resides in the visual domain, rather than the domain of linguistics.

The approach to indexing the photos in the system, is letting the user trace the photo with a predefined set of vector-based shapes and symbols. These shapes which will be addressed by the term Indexing-symbols from here on. The interpretation of the Indexing-symbols remains the task of the user. A yellow, circular shape in the upper half of the photo is nothing but a shape with an id and a colour variable to the computer. In the mind of the user this can be interpreted as the sun, however this will never be named as such within the system. There is a very high chance that the same Indexing-symbol fur 'sun' will be used by other users as well, because of the limitated amount of available Indexing-symbols. Further workings of Photoindex will be explained in the following chapters.

5.2 The idea

The idea behind Photoindex is that every person interacting with the system uses the same language; the basic visual language with shapes as language elements. These elements are even understandable to a person that cannot read. Like Morse code, this general, unified language should ideally not suffer from differences in local (lingual) language, cultural background, educational level and personal interpretation (Chapter 3.2). The makeup and elements of the language are to described in the following paragraphs.

Simple, abstract shapes

The program is based on the concept of query by sketch. Photoindex gives the user the opportunity to look for two-dimensional shapes in an image, only not before that shape has been defined by the user before. These two-dimensional shapes can only be a square, circle or triangle. After being dragged on a photo, these shapes (Indexing-symbols) can be transformed to a certain extent. Users can apply changes in position, scale (in x or y axis or both) and rotation to modify the Indexing-symbols (Figure 5.1). This way, the user does not require any sketching-skills for indexing and more importantly, for querying.

The advantage of using vector shapes over sketching is that the shape, which is defined as Indexing-symbol, is always the same as the shape drawn by the user. There is no need to sketch a perfect circle to retrieve a circular shape, whether the system had indexed a circle (Chapter 4.2), or the user recognises it as such (Chapter 4.3). Neither the computer, nor the user is 'right'. They have in a way, unconsciously agreed upon the 'right' shape. Actually, the *user* agrees with the decision she made for the used shape at the time of indexing the image; the system has no idea of the shape's semantic relation to the photo.

Archetypes

In order to infuse the system with a certain level of high-level features (semantics), users have a set of archetypical shapes at their disposal. The narrower the available set of symbols, the less chance of misinterpreting indexed features. In other words, instead of providing the user with a symbol for 'sea', a symbol for 'faucet water' and a symbol for 'rain', the user can only use a symbol for 'water'. An Indexing-symbol behaves exactly the same way as the geometrical objects discussed before. The appearance of the symbol is chosen in such a way that it is as abstract as possible (Chapter 3.1) and recognisable to virtually any user. The user can mark water in a photo by dragging the 'water' Indexing-symbol on top of the photo, and scaling it until it covers the water-area in the photo. Following, an overview of available Indexing-symbols in Photoindex:

For the project's demo version, these are all of the optional symbols. They should suffice for basic image indexing however; usability testing might prove a need for sub-categories. For this version, there has deliberately been chosen not to provide more Indexing-symbols, since the fewer choices a user has, the less ambiguity in photo-indexes can occur. Less choice also keeps the interface clean, which increases the usability of the program.

The last two Indexing-symbols ('water' and 'sky') would be of better use as shape-fill tools. This would give the user more freedom of creating circular-shaped water as a pond for example, when used in combination with the geometric symbols. This has not been implemented in this version of the Photoindex because of technical limitations. However, this should not be a problem for the proof of concept of indexing photos with Photoindex, since the user *can* define water, there is only less detail.

Of each Indexing-symbol that has been placed on the photo, a user can also choose a 'density' value. This is to be used when a large amount of shapes, like a group of people, need to be dragged on the photo. By clicking the '+' or '–' button (Figure 5.1), the user can increase or decrease the Indexing-symbol's density. This way, the user does not need to drag a large amount of 'people'-Indexing-symbols on the photo. When clicking the '+' button next to a 'person'-Indexing-symbol, the amount of shapes shown within that single object will increase. The object will remain a single object however, with a higher 'density-value'.

Limited amount of colours

The Indexing-symbols can be coloured after being dragged on the photo. Users can only apply a small amount of different colours to the Indexing-symbols, for the same reason why there is a limited amount of Indexing-symbols: fewer colours give less chance of interpretation differences. As a result, there is a higher probability of matching a queried object with an indexed feature. As visible in the figure, the user can choose two shades of blue – light and dark. In combination with the 'sky'-Indexing-symbol, the following semantic combinations can be constructed:

rectangle + sky + light blue = day

rectangle + sky + dark blue = night

rectangle + sky + orange = sunset

circle + water + green = dirty pond

(The technical limitation of not being able to fill geometrical Indexing-symbols in the developed Photoindex version has been neglected in order to be able to give a detailed example)

5.3 Indexing

As said, indexing photos is done by tracing a photo by means of dragging two-dimensional, vector-based Indexing-symbols on top of the photo. Photos in the system can only be retrieved after they have been indexed. To start indexing, the user chooses 'INDEX' in the image-browsing module of Photoindex. After that, all images that have been uploaded, but have not been indexed yet, are presented to the user. The user can now select the image she wishes to index (Figure 5.4).

Clicking on one of the un-indexed photo will activate the indexing interface of Photoindex, with the photo that has just been selected by the user. Next, from the bar in the top of the indexing module, a user can drag and drop Indexing-symbols on top of the photo (Figure 5.5). A maximum of ten objects can be dragged on the photo, for usability's sake. The idea is to quickly index an image, not to draw a new one. Other than that, limiting the amount of possible Indexing-symbol on the image stimulates the user to choose an Indexing-symbol's *density*, rather than cluttering up the photo's index with many of the same Indexing-symbols.

The interface does not contain a 'save' button. Instead of letting the user store indexed photos, every Indexing-symbol on the photo, together with its properties (scale, rotation, position and colour) is stored in the database whenever the user de-selects that particular Indexing-symbol.

This interface, when programmed on a different platform, could be expanded with some interesting features for expanding functionality of the indexing module. One is a real-time check for similar indexes. This way, duplicate images can be pointed out to the user. An other, more interesting feature would be the addition of automatic feature recognition, like in Blobworld. The system could then learn which choices the user makes in indexing photos and relate the user's choices to shapes that can be extracted from the image automatically. That way the system can place shapes on the photo, right before a user starts indexing a new photo. The longer a user operates the system, the more accurate the pre-placed objects are. A disadvantage of such an expansion would be the lower awareness of features in a photo, since the user does not intentionally define shapes any more.

5.4 Querying

For the sake of usability and consistency, the querying interface looks the same as the indexing interface. The only difference is that there is no image on the background and that the interface features an extra button for submitting the query. After submitting the query, thumbnails of the retrieved images will be presented in the photo-browsing window (Figure 5.4). To prevent the user from having to browse through a large amount of retrieved images, the system will only present photos that have the highest similarity in comparison to the query-drawing.

Rating system

The presentation order (rank) of retrieved images is determined by a rating system. The more similar a photo in the system is in comparison to the query-drawing, the higher it will be placed in the array of retrieved images.

The rank of a photo's match is determined in seven steps. The comparison happens in a sequence of steps, from comparing Indexing-symbols down to matching details like the rotation of the Indexing-symbols. After each comparison-step, only the high-scoring images will 'proceed' to the next step. This will result in a higher score for images that are the most similar to the query-drawing. The query will be compared to the photo-indexes in the database in the following order of steps:

Similarity of selected Indexing-symbols. If a 'person' and an 'animal' have been drawn in the query, all images containing 'persons' will be retrieved in the background and awarded with points for their match-rating. Photos with a 'person' *and* an 'animal' will receive a higher rating. Photos within a certain scoring-range will be tested for the next query-match.

Occurrence of matched Indexing-symbols. After comparing the kind of Indexing-symbols in the photo, the system checks whether the amount of Indexing-symbols in a photo-index matches the amount in the query-drawing. The more exact the match, the higher the extra points for the photo's rating.

Colour of matched Indexing-symbols. The colours of the Indexing-symbols in the query drawing are compared to those of the indexed the photos that are within a certain scoring range, after the previous comparison-step.

The density of Indexing-symbols. All Indexing-symbols in the query-drawing are compared to the density of the Indexing-symbols that are placed on the photos.

The position comparison of Indexing-symbols;

The scale comparison of Indexing-symbols;

The rotation comparison of Indexing-symbols.

The more query-steps an indexed photo passes, the higher its accumulative match rating will be, the higher in (match) rank it will be presented to the user.

5.5 Editing

The editing interface looks the same as the indexing and querying interface. When the system would be used by several people, all image-indexes can be modified or refined by every user. Possible missing features or misinterpretations in photo-indexes could be altered by others. This way, a more general opinion is formed about the way a photo should be indexed, creating a 'community-based truth' (chapter 3).

5.6 Technical approach

At first, the idea was to develop the system for use on a handheld computer with touch-screen operation, because these are likely to be fitted with high-resolution cameras in the future, or digital cameras could be fitted with larger touch-screen displays. The thought was to simulate the indexing-process at the moment of a photo's creation, by means of touch-screen. This plan was discouraged by the HKU mentors because it would be too difficult to develop for this platform. The additional value of usability testing of this approach was questionable as well.

Therefore, the choice for the PC with Macromedia Director as development platform has been made, for creating the user interface. The storage of photo indexes and the comparison of query-drawings with the indexed photos is developed in PhP with MySQL as a database system. Both director and PhP/MySQL were familiar platforms, which made development less difficult than it would have been if other platforms were chosen.

5.7 SWOT analysis

A Strength, Weaknesses, Opportunities and Threats analysis of Photoindex is listed below.

Strengths

Photoindex communicates within the visual domain. Therefore, it does not suffer from possible misinterpretations of indexes like communication within the textual domain and mistranslations between textual descriptions and visual clues.

The program uses basic shapes and symbols that are easily understandable.

There is no difference between the querying-language and the indexing-language of the system. This makes things easier for both the system's user and developer.

Photoindex features a different approach to closing the semantic gap, providing possible interesting insights on developing image retrieval systems in the future.

Weaknesses

It is not yet possible to combine a texture and a colour with an Indexing-symbol. Therefore, it would not be possible to draw a tiger for example ('animal' + orange + stripes).

Because of a narrow selection of Indexing-symbols, it becomes hard to create better-distinguishable objects. There is one Indexing-symbol available for 'transportation'. The means of transportation (by air, by sea, etc) cannot be defined with this symbol.

Opportunities

Adding sub-sets of Indexing-symbols could give users more possibilities in creating higher-detailed indexes. An example could be 'transportation', with optional sub-Indexing-symbols for 'plane/by air', 'ship/by sea', 'car/by road', etc.

By replacing the Indexing-symbols with Indexing-symbols from a sub-category, the system can become usable for specialised applications, like indexing photo-collections with only cars for car-sales purposes.

Threats

Strong development in text-to-speech combined with MEDM based systems makes it easier to enter textual descriptions, taking away one of the major disadvantages of MEDM based systems.

Automatically indexing by MEDM would become available when in the future all text-based systems would be linked. An example could be a n automatic registration of the GPS location where a photo has been created, in combination with a date, could retrieve a lot of information on the location and possible reason for taking that photo, by matching it with activity-agendas online.

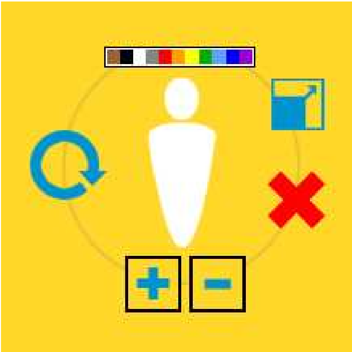

Figure 5.1 - The 'people' Indexing-object with
rotate, scale, colour-selection, density and delete tools

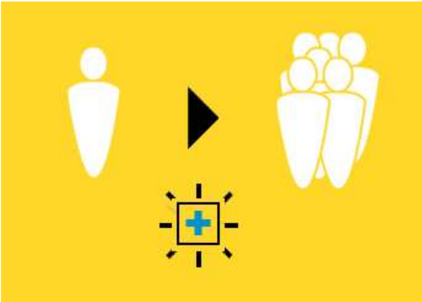

Figure 5.2 - Increasing an Idexing-symbol's density

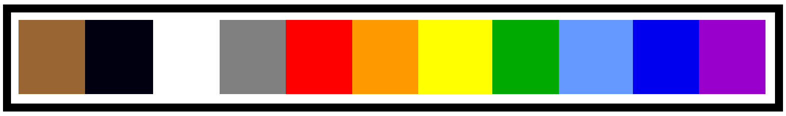

Figure 5.3 - Available colours

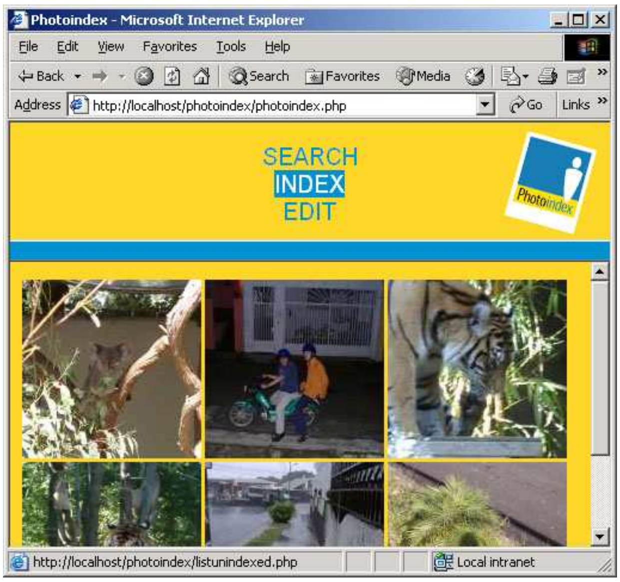

Figure 5.4 – The user can select a photo she wishes to index

Figure 5.5 - Indexing a photo

6. CONCLUSION

With current and new technology, applied to applications for both professional as well as private use, the amount of produced images is growing fast. However, technologically speaking, the manner in which images are stored for retrieval stays behind. The more images produced, the harder it becomes to retrieve a specific image from the ever-growing collection.

Retrieval of images can be done either by browsing or by querying. For browsing, in general, images are scaled down and presented in multiple numbers at once to the user. This way it is possible to compare several images at the once to save time. In addition, it makes side-by-side comparison possible. The user misses details as result of scaling down the images in size. She also needs to divide attention between all the presented materials, with the risk of overlooking images. Not only is the process of browsing inaccurate, it is time consuming as well.

The computer could do the comparison of images for people much faster, on the condition that the image index (a description of an image's characteristics that is readable by a computer) is detailed enough. A user can instruct the computer to retrieve an image from a collection by entering search-conditions that possibly match a certain image's index. Entering these search conditions and instructing the computer to look for images by filtering according to these conditions, is called querying. A highly detailed image index should provide the possibility for a detailed image search.

There are many problems regarding indexing images. For one, it is difficult to decide which part of an image should be described, since images are ambiguous. For example, a photo of children at a sunset beach can be interesting because of the depicted children or because of the colours in the sunset sky. Points of interest like these vary per photo, per image domain, per user and per search task. A work-around for this problem is to narrow the down the domain of stored images to a collection within a pre-determined context. This way, all images in the collection are potentially interesting and the details in images can be described rather than figuring out whether the image is interesting in the first place.

Next, assuming the points of interest are determined well, descriptions of these points of interest create difficulties as well, because of possible differences in interpretation. One person calls the sun yellow; another might feel it is white. Other than that, the descriptions need to be readable and comparable by a computer.

To achieve computer readability, two main methods of indexing images are used. One is by manually describing images in a textual format. The other is utilisation of computer vision, where a computer indexes images by detecting features like colours, shapes and textures in an image.

Describing an image with text needs to be done for every single image by a person by hand. If a photo contains enough manually entered descriptive metadata keywords, it has a high chance of being able to be retrieved. However, it takes much time to enter a sufficient amount of suitable keywords. It is virtually impossible to find this sufficient amount of suitable keywords, since "An image says more than a thousand words". The suitability of the keywords depends totally on the user that looks for an image. While less ambiguous than the semantic content of images, textual image descriptions suffer from ambiguity as well. Image searchers all speak different languages and use different words for the same visual clues. One person might choose the word 'circle' as another might choose 'sphere' to describe the same visual element in an image. The language a searcher 'speaks' depends on her local language, her cultural background, her vocabulary and her search task. When an image's index does not contain sufficient keywords, one can only retrieve an image if one knows how it is stored. A sought after image would not be retrieved by the system when the 'wrong' keywords are used for a search. This could be an advantage, since finding just a few suitable keywords would not take up too much time to enter. But then again, which words would be suitable and not miss-interpretable at the same time?

Using keyword descriptions for describing images is not right when images need to be retrieved efficiently, without having to browse after querying. This is because text lies within an entirely different domain in comparison to images. The act of *seeing* (not interpreting) a photo cannot be performed erroneously (assuming the person is not sight impaired or (colour) blind). Without naming a shape in a photo, everybody sees the exact same shapes, structures and colours in an image. Inaccuracy can only happen with the *interpretation* of (visual features within) images. The language within the visual domain it self does not suffer from ambiguity. Therefore, indexing images could be done better by indexing visual features instead of using textual descriptions. *Image retrieval should take place by communicating within the visual domain.*

Content-Based Image Retrieval, a sub-domain of the science of computer vision, utilises the non-ambiguity of sight. By extracting and storing visual features from images, known as low-level features, users of CBIR-based systems are able to find images by looking for these features. However, computers and people 'speak' the same language, but in a different dialect, regardless of the fact that the language of the (correct) visual domain is used. To people, both a perfectly round shape and a less-perfectly round shape can be interpreted as (or translated to) a circle. Computers however, see either one shape or another. It is up to the user to tell the computer whether she thinks the shape is a circle or not. The computer's task should be nothing more than looking for similarity in indexes that have been previously constructed by users. In addition, a problem is still, like with textually describing images, to determine *which* elements in an image are important to the user at which occasion. The only way to find this out is to 'ask' the user. *The key to higher accuracy in image retrieval lies in the interaction of the user with the system.* Users point out important elements and determine what these elements look like. This way, the computer only needs to compare indexes, while the user is actually aware of how the photo is indexed. This is similar to remembering in which drawer an analogue photo is stored.

After pointing out interesting elements in images, they need to be described. In the Photoindex project, this is done by assigning a shape or an archetypical symbol (like a symbol for 'human' or 'animal') to an element in the photo. By providing the user with a small set of optional elements and symbols for describing elements, miss-interpretation is less possible. In other words, the few optional shapes and symbols force the user to make a choice in describing the photo. Less choice and decision making means a flattening of dimensions, which results in less interpretation issues during the retrieval of images.

To index a photo in Photoindex, the shapes and symbols can be placed on top of a photo. They can also be rotated, scaled and coloured. All shapes and symbols on a photo together can form a context. The position relative to each other provides clues about the indexed photo. A circle placed on the photo could describe a ball or a hole for example. When it is coloured yellow, it can no longer be a hole, but it could still be a ball or the sun. When the yellow circle is placed in the upper half of the image, on a blue rectangle with the symbol for 'sky', the yellow circle in scene has a high chance of becoming interpreted as being the sun.

Photoindex is a way to show that and decision making about interesting parts in images and leaving interpretation of images should be done by people. One advantage of this is that people actually remember how images are indexed. More importantly, users index photos by creating pictures in their minds with objects that have no meaning themselves, preventing misinterpretation of the objects themselves. These objects are at the same time readable and comparable by the computer.

7. LITERATURE LIST

L. Armitage and P. Enser, *Analysis of User Need in Image Archives*. Journal of Information Science, vl. 23, no.4, 1997.

C. Bauckhage, T. Käster, M. Pfeiffer and G. Sagerer, *Content-Based Image Retrieval by Multimodal Interaction*. Industrial Electronics Society, 2003. IECON '03. The 29th Annual Conference of the IEEE
Volume 2, 2-6 Nov. 2003 Page(s):1882 - 1887 Vol.2

E. Goffman, *Presentation of Self in Everyday Life*. Knopf Publishing Group, 1972

J. Hunter, *Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology*. International Semantic Web Working Symposium (SWWS). Stanford. July 2001

B. Johansson, *A Survey on: Contents Based Search in Image Databases*. Linköping University, Dept. of Electrical Engineering, Computer Vision Laboratory. December 8, 2000

T. Kato, T. Kurita, N. Otsu, and K. Hirata, *A Sketch Retrieval Method for Full Color Image Database – Query by Visual Example.* Proc. IEEE-IAPR-11, pp. 530-533, Aug.-Sept. 1992.

A. Mojsilović, J, Kovačević, J. Hu, R.J. Safranek, S.K. Ganapathy, *Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns.* IEEE Transactions on Image Processing, vol. 1, pp. 38 54, Jan. 2000

S. McDonald, and J. Tait, *Search strategies in Content-Based Image Retrieval.* ACM Press New York, NY, USA, 2003

H. Muller, N. Michoux, D. Bandon, A. Geissbuhler, A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int J Med Inform. 2004 Feb;73(1):1-23. Review

S. Ornager, Image Retrieval: *Theoretical and Empirical User Studies on Accessing Information in Images*. 60th Am. Soc. Information Science Ann. Meeting, vol. 34, pp. 202-211, 1997

S. Santini, Exploratory Image Databases – Content-Based Retrieval. Academic Press, 2001.

S. Santini and Ramesh Jain, *User interfaces for emergent semantics in image databases*. In Proceedings of the 8th IFIP Working Conference on Database Semantics (DS-8), Rotorua (New Zealand), Jan. 1999

Dr. B.A.M. Schouten. Giving Eyes to ICT! Or: How Does a Computer Recognise a Cow. University of Amsterdam, 2001.

A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, *Content-Based Image Retrieval at the End of the Early Years,* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, No. 12., Dec. 2000

C.-H. Wei, C.-T. Li, and R. Wilson, *A Content-based Approach to Medical Image Database Retrieval*, in Database Modeling for Industrial Data Management: Emerging Technologies and Applications, ed. by Z. Ma, Idea Group Publishing, 2005. (Accepted)

8. IMAGE COURTESY

Following, a list of images and there respective owners, if known. Images that have not been listed here are owned by me.

Figure 1.1 Figure 1.2
Tony Stone Images

Figure 3.1 Figure 3.2 Figure 3.3 Figure 3.4
Yahoo! - Flickr.com

Figure 4.1
Gestalt invariance Lehar S. (2003) The World In Your Head, Lawrence Erlbaum, Mahwah, NJ. p. 53, Fig. 3.5

Figure 4.2
Mark Derrick
http://www.flickr.com/photos/thederricks/31464733

Figure 4.3
Chris Besset
http://www.flickr.com/photos/12569534@N00/19809496/

Figure 4.4
Larissa
http://www.flickr.com/photos/lara68/23137992/

Figure 4.5
Ryan Paul Maxwell
http://www.flickr.com/photos/rmaxwell/

Figure 4.6

E.H. Adelson. "*Checker Shadow Illusion*", 1995. Accessed august 16, 2005.

Source and explanation available online at:

http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html.

Figure 4.7

Unknown

Figure 4.8

Unknown

Figure 4.9

Jon Madison

http://www.flickr.com/photos/smartbrother/31720258/

Figure 4.12

Unknown

Figure 4.13

University of Leuven (Belgium) / CWI Amsterdam

A. ACKNOWLEDGEMENTS

I would like to thank many people, but especially these, for assisting me with writing this thesis, helping me with my graduation project and support in other ways:

Ben Schouten, for his assistance on writing my thesis and sharing his vast amount of knowledge regarding Content-Based Image Retrieval and providing me with much useful literature

Arjan Wiegel, my colleague at ArchiVision, for helping me develop Photoindex. I owe him much gratitude for his help on the Macromedia Director parts of Photoindex.

Janine Huizenga and **Tom Demeijer** for their feedback on the ideas during the design-process of Photoindex

Roelof Terpstra, my employer, as well as my other colleagues at ArchiVision for letting me write my thesis at the workplace

Ina van der Brug, for managing the graduation process at the HKU and answering many questions for reassuring me during the graduation process

Willem-Jan Renger, for helping me think out the base concept of Photoindex, during the initial stage of the graduation process

Students at the HKU for their advice, feedback, for being friends and good company: Matthijs Braamhaar, Joost Broersen, Tamas Cselenyi, Bas Derks, Carola Grewelding, Rose Klaver, Hay Kranen, Ton Savenije, Willempje Vrins, and probably more..

Mark Huiskes and Eric Pauwels at the CWI Amsterdam, for discussing the design of Photoindex and sharing knowledge on Content-Based Image Retrieval

The HKU, for providing an excellent work-area with a great ambience with use of computer systems

My father, mother, sister and her friend for support in many kinds of ways