Department of Information Science and Media Studies

University of Bergen

Ph. d. Theory Course

Human Computer Visual Communication

Understanding and Interpreting Visual Content

Visual Grammar and Content Based Image Retrieval

Lars-Jacob Hove

Lars-jacob.hove@infomedia.uib.no
www.vq.uib.no

Abstract

This essay presents a discussion on how concepts and tools of visual design and human image understanding can be used to improve automatic description and indexing of digital images.

The main motivation behind the essay is to present a fundament for discussing the issues raised by this question, and present a framework which might provide a basis for further thorough and detailed studies.

The essay consists of 6 main sections:

- Section 1 presents an introduction to the concepts and central questions discussed in the essay.
- Section 2 is a discussion of different levels of image content. A taxonomy and overview of image content is presented.
- Section 3 presents an overview of fundamental concepts and challenges of Content Based Image Retrieval.
- Section 4 presents an overview of some central tools and concepts from visual design. A selection of concepts is discussed in some detail.
- Section 5 presents a suggested approach for implementing concepts from visual design in an image retrieval system.
- Section 6 presents a summary of the issues presented in this essay, as well as some notes concerning the generality and application area of the suggested approach.

1. Introduction - The Challenge of Semantic Image Description

How do we understand and interpret an image? Consider the image in Figure 1, below. One might expect that an observer would recognize the depiction of a scene with a dolphin, two women and two beach balls situated in what appears to be an aqua park. The dolphin appears to be interacting with the woman to the left, while the second woman is watching the pair. Depending on your world view, you might say that the dolphin is joyfully playing with its caretaker. A second caretaker is watching, maybe evaluating the performance of the first caretaker. The trio is quite likely performing before an enthusiastic audience. On the other hand, it might be considered as an exploitation of an unhappy animal. The dolphin is held as a slave by the cynical owners of the aqua park in order to maximize their profits by showcasing the poor animal to a mindless audience. Both these simple narratives might be created from a quick glance at the image.

Figure 1- Image of a dolphin, a ball and two caretakers

Next, consider the image in Figure 2. Although the images differ in terms of composition, framing and colour scheme, we recognize the similarities in *semantic content* between them: a dolphin and at least two people are involved in some kind of interaction, probably in the context of a performance in an aqua park.

Figure 2 – Image of a dolphin interacting with a ball

The above examples illustrate one of the major challenges in digital image management. Until the previous decade, images were prevalently stored as physical objects. Limitations in both hardware and software made computers an ill-suited tool for image collections. However, as the computational power of both hardware and software have increased, the ability to store

more complex data types, such as images, in databases, has been drastically improved. Unfortunately, the tools and methods used to *describe and index* these images have not been proportionally developed.

Consider the task of finding images of a dolphin. A search using *Google Images* on the term "dolphin" returns approximately 2.6 million images[1], indicating the vast number of images available at our fingertips. Even though the human mind is capable of quickly identifying the contents of an image, the sheer volume of images available makes the task of browsing through these images serially practically impossible. Traditional techniques for image description based on textual annotation have proven inadequate for large image collections. While text annotations have high expressive power, it is a difficult and time consuming task to fully annotate large volumes of images and the resulting descriptions are prone to subjectivity and incompleteness (Huang and Rui 1999).

As a response to these challenges, methods based on automatic extraction and comparison of structural image features have been developed (Faloutsos, Barber, Flickner, Hafner and Niblack 1994). These systems, commonly known as Content Based Image Retrieval (CBIR) systems, identify and retrieve images based on a comparison of structural similarity between a seed image and the images in a collection. CBIR systems have shown very promising results for domains where the image sets are homogeneous in nature, such as face or fingerprint identification, or medical image processing. However, automatic retrieval of images based on higher level content, such as their semantic content, has proven difficult (Santini and Jain 1997; Colombo and Del Bimbo 2002; Datta, Joshi, Li and Wang 2007).

This essay is a discussion of the feasibility of improving CBIR systems using knowledge of how humans interpret, understand and create visual structures, as represented by the following question:

> *How can visual design structures be used to improve automatic indexing and description of the semantic content of digital images?*

This question raises three additional questions:

1. What is the "Semantic Content" of digital images?
2. What do we mean by "Automatic Indexing and Description" of images?
3. What are "Visual design structures"?

Each of these questions will be addressed in the following sections. Section 2 presents a classification of different levels of image contents and an overview of how a human observer translates the perceptual structures in the image into semantically meaningful content.

Section 3 presents an overview of fundamental concepts of automatic image processing in Content Based Image Retrieval. A comparison between human and computer-based image interpretation is presented.

Section 4 presents an overview of a set of structures often used in visual composition, based on the work done by Kress and van Leeuwen (2006). An example of how such structures can

[1] http://images.google.com/images?q=dolphin - August 20th 2007

be used by a human to identify and extract semantic content from the image in Figure 1 is presented.

Finally, section 5 is a discussion of what must be done in order to extend CBIR systems with a system for utilizing and interpreting visual design structures.

2. Classification and Interpretation of Image Contents

What do we mean by the phrase *semantic content of digital images*? In ordinary, everyday image use, an observer is commonly interested in the objects or scenes depicted in an image, such as identification of the people present in a family portrait. A history student studying a photograph of a busy street at the end of the last century might be interested in the clothes people wore or the activities depicted in the image. In fields such as cultural studies or art history, a researcher might be interested in the iconic, contextual, cultural or meanings presented by the image, or the stylistic and formal means used in its creation. In technical disciplines such as x-ray imagery, images are regarded as a specific form of signal, where the important content is defined in the syntactical structure of an image, such as colour distribution or abnormalities in particular shapes. These different uses and interpretations of images require different types of knowledge, and are based on different levels of abstraction. Consequently, the term *semantic image content* is ambiguous and needs clarification.

Basic Definitions

Images are visual structures, and interpretation of image contents is highly dependent on the observer's experiences, motivations and needs. Two people observing the same image might have very different interpretations of the content of an image, even though the structural characteristic is unchanged. Consequently, we should distinguish between the physical characteristics of an image, and the interpretations made by a human observer.

Images are representations of light, perceived by our visual senses. Jaimes and Chang (2002) refer to this as the *percept* of the image. Patterns of light are reflected on different materials, and produce the perception of different elements such as texture, colours and shapes. While percept refers to the impressions we perceive through our senses, the *syntax* refers to characteristics of the visual elements themselves and the way in which they are arranged. This is defined as the *Syntactic Image Content*¸ representing the structure of an image: *colour, texture, shapes and the spatial arrangements of these.*

Next, the *concept* of an image refers to a representation, an abstraction or a generic idea generalized from the particular instances of the syntactic image content. As such, it implies the use of background knowledge and an inherent interpretation of what is perceived. This represents the image semantics, and refers to *the meanings* of the syntactic elements and their arrangements (ibid). This is defined as the *Semantic Image Content*, and represents *the meaning of an image, beyond its overt subject matter, including the emotional, intellectual, symbolic, thematic, and narrative connotations.*

The above dichotomy distinguishes between the visual and structural characteristics of an image and interpretations a human observer makes from these. However, it does not provide any insights into the skills and knowledge required to understand and interpret different types of image content. Consequently, a more detailed categorization of semantic image contents is presented in the next section.

Finally, as we are dealing with *digital* images, it is worthwhile to distinguish between an image and the *digital representation* of the image. An image is *a visual representation of an object, scene, person or abstraction, produced on a medium.* A *digital image* is *a set of two-dimensional arrays composed of pixels whose locations hold digital colour and/or brightness information which, when displayed on a suitable interface, form an image.*

Levels of Image Content

Table 1 presents a more detailed classification of image contents. Content is classified as *perceptual structures*, *generic semantic content, specific semantic content* and *abstract content*. The first level is equivalent to the *syntactic image content*. The other levels represent higher degrees of abstraction, and require increased specialized and contextualized knowledge to interpret.

Table 1 - A taxonomy of Levels of Image Content

Content Level	Definition	Interpretation	Examples
Perceptual structures	The basic syntactical structures in an image. Equivalent to the *Syntactic Image Content*	Based on low-level perceptual systems.	Lines, colours, shapes, contours and textures, and the local and global spatial arrangements and distribution of these
Generic semantic content	The basic semantic units. Generic objects, concepts or scenes which share a set of attributes which are common to all, or most of, the members of a particular category.	Based on everyday knowledge, and is presumed to be universal	"Ball", "Dolphin" and "Human" A group of Dolphins at sea.
Specific semantic content	Specific objects, scenes and activities which can be named and identified.	Based on personal knowledge and recognition.	The dolphin "Skippy" Image of the Empire State Building
Abstract content	Meanings that can be derived from specialized or interpretative knowledge about what objects depicted in an image *represents*.	Based on contextual, cultural or technical knowledge of objects, motives and symbols, filtered through individual experience.	Interpretation of X-rays or medical imagery Activities performed Non-visual content A "Smiley" representing happiness (☺)

Perceptual Structures

Perceptual structures refer to the overall syntactical structures of an image. Eakins, Burford and Briggs (2003) distinguish between three categories of perceptual structures. *Perceptual primitives* represent content extracted by low-level perceptual systems, such as colour and some textural descriptions. *Geometric primitives* represent simple two- and three-dimensional non-representational forms, such as lines, arcs, squares and circles. *Visual extensions* are

visual features that do not contain meaning beyond the simple perceptual pattern, such as detection of depth through occlusion or perspective. While these represent different degrees of abstraction of syntactical image contents, they are grouped together as they all represent the perceptual structures of an image.

Generic Semantic Content

The *generic semantic content* refers to what Jaimes and Chang (2002) call the *basic level categories*. This is content that is not derived purely from the perceptual structures. At this level, basic semantic concepts are defined and named. Semantic concepts are generic objects which share a set of attributes which are common to all, or most of, the members of a category. Instances within categories are defined by a set of prototypes, each presenting a subjective indicator of membership in a category. Examples of this are *dolphin*, *human* or *ball.* The two first examples might require a large set of "prototypes" in order to cover the large variance of the shape, while the last example might require a smaller set of prototypes, as there are fever variances between instances of the category. Identification of image content at this level is generally based on everyday knowledge. However, some concepts might require specialist knowledge to identify, and might be subsumed in higher levels (Eakins, Burford et al. 2003).

Specific Semantic Content

The *specific semantic content* refers to particular instances of a concept that can be identified and named. Specific knowledge of the objects in the image is required, and interpretation relies on the factual knowledge of the observer. Examples include individual persons (The dolphin caretaker "Anna" or the dolphin "Skippy"), a particular breed of dog (Chihuahua), or an image of a particular cityscape (An image of the Eiffel tower).

Abstract Content

The *abstract content* refers to image meanings that can be derived from specialized or interpretative knowledge about what the depicted objects *represents*. The generic and specific semantic content primarily concerns what (Jaimes and Chang 2002) refers to as the *visual content* of the image. This is what is directly perceived when an image is observed. The abstract semantic content primarily concerns information that is closely related to the image, but not present. Identification and interpretation of content is based on the observer's knowledge of motives and symbols, and filtered through their individual cultural, technical or emotional experiences.

Eakins, Burford et al. (2003) distinguish between four different categories of abstractions. *Contextual abstractions* refer to information which is presumed to be universal, in that it is derived from knowledge of the environment, such as deciding whether a particular image represents day or night. *Cultural abstractions* are presumed to be fairly generalized within the general culture of the viewer. Examples of such abstractions may be activities performed in the image or political, cultural, historical and sporting events. *Technical abstractions* refer to information that requires specific technical expertise to interpret, such as x-ray images. Finally, *emotional abstractions* refer to affective or emotional associations or responses people may have to an image.

Human Interpretation of Image Contents

The classification scheme presented in Table 1 indicates that an increasing level of knowledge is required to interpret the different levels of *semantic* image content. However, little detail is

provided regarding how a human observer perceives and interprets the perceptual structures, or how this is used to recognize the generic semantic content.

Recognizing the generic semantic content is also known as the process of *object recognition*. When we look at images, we are usually unaware of the cognitive processes that allow us to identify the different objects in the image. Yet, these processes are fundamental for our ability to understand and interpret all levels of image contents. If we are unable to successfully identify the basic semantic units in an image, further interpretation of the image is impossible.

According to Messaris (1994), the process of interpreting a perceptual structure into a mental representation involves three successive steps:

1. Detection of outlines and surfaces
2. Assignment of depth and extraction of the third dimension
3. Object recognition.

In the first step, visual information is transmitted from the retina to the brain. Essentially a two-dimensional array of light and colour values, this information is processed by the brain to detect the outlines of objects and the edges of surfaces. The end result is a mental representation than can be thought of as corresponding to an outline drawing of the scene the eyes are looking at (Messaris 1994:11).

In the second step, the brain assigns depth to the various parts of the outlines, i.e. calculates distances between the viewer and each part of the scene. This is a complicated process involving several different kinds of information. When interpreting images, the two most crucial aspects are *texture gradients* and *occlusion. Texture gradients* represent changes within our retinal image of the density of a regular pattern or texture in the scene we are looking at. An example of this would be the way parallel lines stretching away from us appear to converge towards a “vanishing point”. This serves as an indication as to the depth of an image. *Occlusion* represents the blockage of the view of part of one object by another object, helping to infer which objects are in the foreground and which objects are in the background (Messaris 1994, pp51-52).

The final stage of visual perception is *object identification*. While less is known about this than the previous steps, it is assumed that the outlines of objects, whose distance from each other and from the viewer have been determined, are now used to infer the objects’ identities. The outlines are compared to a hierarchical “catalogue” of object structures in the memory of the brain (Messaris 1994, p57). Comparing this to the classification presented by Table 1, we see that this corresponds to identification of the generic semantic content of the image. Identification of further image content is then dependant on contextual, cultural or technical knowledge of objects, motives and symbols, filtered through individual experience.

3. Content Based Image Retrieval

Current available image collections and image databases are, to a large extent, based on keyword annotation for image indexing and retrieval. In systems such as these, images are annotated with descriptive texts or keywords. This is a process which requires a lot of manual work, especially as the number of images grow. Furthermore, while it is trivial to describe the generic semantic content in an image, describing higher levels of semantic content is increasingly complex. Consider again the image in Figure 1. Most observers would easily be

able to describe the basic semantic content using keywords such as “Dolphin”, “Ball”, “Caretaker” and “Aquarium”. However, as we saw in the introduction, different people may perceive the contents of an image differently, and different people’s understanding of the semantics of a certain keyword might wary (Rui, Huang and Mehotra 1998). Consequently, providing a full description of the possible interpretations of the semantic content of an image is a major undertaking, even for a single image. These are known are the problems of *Volume* and *Subjectivity*. During the previous decade, *Content Based Image Retrieval,* or CBIR, emerged from the field of Computer Vision as a possible solution to these problems.

CBIR systems consist of automatic indexing methods and search and retrieval techniques for description and retrieval of images based on their syntactic image content. Current CBIR mechanisms can, to a certain degree, successfully compare and retrieve images based on the syntactical content of an image. However, automatic retrieval of images based on semantic content has proven difficult (Colombo and Del Bimbo 2002). While it is outside the scope of this essay to provide a full overview over CBIR state-of-the art, a basic understanding of CBIR fundamentals and its largest challenges is required for further discussions.

Fundamentals of CBIR

Content Based Image Retrieval is based on extraction and indexing of the perceptual structures of a digital image. When an image is submitted to a CBIR system, the image is analyzed and a set of statistical *feature descriptors* describing the structural features are extracted, e.g. a colour histogram or colour moments. These feature sets form a description of the image structures, giving each image a unique, statistical signature which can be used for image retrieval tasks (Li and Kuo 2002).

Query and retrieval of images in CBIR systems is usually based on a notion of similarity between two images or a single image and images in an image collection. Similarity is determined by a *similarity function*. A similarity function is a mapping between pairs of feature sets and a positive, real world number, which is chosen to be representative of the visual similarity between two images (Li and Kuo 2002).

A simple similarity function might be based on the Euclidian distance between the feature descriptors of two images. If two feature sets are identical, the distance between them would be 0, and the images would be identified as identical. The more dissimilar the feature sets are the larger the distance would be and the perceived similarity would be diminishing.

Figure 3 – Different depictions of a dolphin

As an illustration, consider the three images in Figure 3. A comparison between the three images based on colour features, would likely report a higher degree of similarity between the first and the second image, and a low degree of similarity between the second and third image. A comparison based on shape would likely report a high degree of similarity between

the second and third, and a low degree of similarity between the first and the second image. It is unlikely that a retrieval system based on feature descriptors will return all three images, even though they all depict a single dolphin.

Finally, consider the image presented in Figure 4. It is likely that most people would correctly identify it as an image of a banana. However, looking solely at the structural features of the image, it is clearly very similar to the last image in Figure 3. Both are gray-scale images, dominated by a single, shape, with similar salient characteristics.

Figure 4 - A depiction of a banana

The Semantic Gap

The foregoing discussion provides insight into the most challenging problem with the CBIR approach: The system has no semantic understanding of the visual structures it is processing. While retrieval based on the structural features might be very relevant for some homogeneous application domains, such as fingerprint identification or x-ray analysis, retrieval of images from heterogeneous collections of images is more likely to be based on the *semantic* content of the image. Consider the comparison between the tasks of retrieving information about dolphins from a text based source, and retrieving images about dolphins from a digital image collection.

The former task would be relatively easy to perform using a linguistic query such as “Find me all documents which contain the word ‘dolphin’”. The system could also easily expand the search by including terms such as “Cetacean”, or retrieve texts in different languages by replacing the term “dolphin” with other versions of the word, without any input from the user. However, achieving similar results by querying a state of the art, automatically annotated image retrieval system is considerably more difficult. As such systems are based on raw image properties all retrieved images will be structurally similar, not semantically similar. This problem has been dubbed *the semantic gap*. In the context of this essay, the semantic gap can be defined as:

> *The gap between the syntactic, perceptual visual structures in an image, and the semantic associations attributed to these by a human observer.*

This means that while a CBIR system based on the syntactical image content is capable of extracting the syntactical and perceptual structures present in a digital image, it has no understanding of the semantic contents in the image.

Human and Computer Image Interpretation

One possible way to explain the challenges of the semantic gap is to compare human interpretation of image contents with the way CBIR systems processed digital images. As we have seen, human interpretation of perceptual structures consists of three steps, *detection of*

outlines and surfaces, assignment of depth and extraction of the third dimension, and *object recognition*.

A digital image has been defined as a set of two-dimensional arrays composed of pixels whose locations hold digital colour and/or brightness information. We immediately see that is very similar to the visual information being transmitted to the brain through the eyes: a two-dimensional array of light and colour values. There is no denying that the digital representation of an image contains less information than an analogue signal sent to the brain by the perceptual system. However, there is still a high degree of similarity between the raw data available to the two processes.

While the human brain is capable of detecting outlines and contours of object present in an image, this has proven to be very difficult in the case of digital image processing. In image processing, this process is called *image segmentation*, and is defined as *the process by which an image is divided into spatial sub regions.* Reliable image segmentation is especially critical for characterizing shapes and outlines present within an image. However, achieving good and reliable results in image segmentation has been challenging, as the computational processes for has proven to be very complex. Successful object segmentation for broad domains of general images has so far not been achieved.

Generally, it is very difficult to perform precise, automatic object segmentation owing to the complexity of the individual object shape, the existence of noise and occlusion. One of the major challenges is that it has proven difficult to correctly identify the boundary of an object, as defined by *the border between a shape and its environment, represented by the contour*. It is therefore difficult to distinguish between different objects, and separate objects in the foreground from the background (Smeulders, Worring, Santini, Gupta and Jain 2000; Kimia 2002; Datta, Joshi et al. 2007). Consequently, current technology for image processing is unable to follow human image processing beyond the first step. Unless it is possible to distinguish between the different contours present in an image, the task of determining depth and relative size between these objects becomes void.

As a final note it should be noted that there has been some efforts to create software systems that act similar to the third step of human image processing, object recognition through the use of shape prototypes. Some examples of this can be found in (Carson 1997; Sclaroff and Liu 2001; Kimia 2002; Hove 2004). This approach might be promising if the objects in the image are clearly distinguishable from the background and each other, or if the object of interest is the only salient feature of the image, such as in Figure 5 and Figure 3. However, these approaches are most likely to succeed in narrow domain areas. While the human brain apparently has an unlimited capacity to classify different shape prototypes in a hierarchical matter (Messaris 1994), creating a set of such prototypes for even a very narrow domain is a difficult and time-consuming task (Hove 2004).

Community Based Object Identification

Even though the challenge of reliable, automatic object detection remains unresolved, progress is being made towards improving automatic management and retrieval of digital images (Eidenberger 2004; Datta, Joshi et al. 2007). One interesting field of research is the recent initiatives involving community-based indexing methods, such as The ESP Game (von Ahn and Dabbish 2004) and Peekaboom (von Ahn, Ruoran and Blum 2006). This research has shown that it is possible to let a community of users detect and index objects present in an image, as well as the spatial distribution of these objects. A game-like structure is used to

encourage both semantic labelling of image contents as well as spatial tagging of this content. Approaches such as these might provide us with images where the generic semantic concepts are identified and tagged without the use of fully automated object detection.

4. Visual Design – Rules and Grammar

One possible approach to automatic description of image contents is to look to the world of *visual design*. Are there any rules, guidelines, heuristics or techniques that guide the *construction* of visual structures? And if such rules exist, is it possible to use such rules to assist CBIR systems in extracting higher level semantic content from images?

A Framework for Visual Grammar Rules

Kress and van Leeuwen (2006) present a systematic and comprehensive account of a grammar of visual design. It offers a toolkit for understanding and interpreting images. This is primarily *descriptive,* not *normative*, and the concepts and structures presented should be interpreted as tools for *understanding* visual structures, not as rules for how such structures should be.

While it is outside the scope of this essay to provide a complete overview of the topics covered by Kress and van Leeuwen, some of the relevant concepts relevant to this discussion have been adapted and is presented in Table 2. This is clearly not an exhaustive list of relevant structures, but is included as examples of structures that might be useful for automatic image description and indexing. Three main groups of such concepts are presented: *Narrative structures, representation* and *composition*.

Table 2 – A selection of visual structures. Adapted from Kress and van Leeuwen (2006)

Concepts	Description
Narrative	*Narrative structures in an image*
Participants	People, objects, things and places which represent the subject matter of an image
Actor	A participant which is the active part in an interaction between participants
Goal	A participant being the receiving part in an interaction between participants
Interaction vector	An implicit or explicit visual structure representing an interaction between an *actor* and a *goal*
Representation	*Ways of representing image participants*
Framing	Size of the ‘frame’ of an image
Perspective	The point of view in an image.
Composition	*Placement of image participants*
Spatial Distribution	Distribution of participants in various zones of importance in the image.
Salience	The saliency of a particular participant
Framing	The presence of framing devices which connect or disconnect participants
Colour	The presence, or use, of different colours in an image

Narrative Structures

Narrative structures are a set of structures which represent the subject matter of an image. The basic narrative structure is the *participant*. Participants are the peoples, objects, things, places or other structures which represent the interesting elements in an image. It is important to note the emphasis on *interesting* elements. In theory, every perceptual structure in the image might be considered a participant, even though they might not be directly relevant for our interpretation of the image. Consider the scene in Figure 1. The interesting elements of this image might be the caretakers, the dolphin or the beach balls, even though elements such as the individual stones in the walls, or the leaves of the plants might be considered participants. While the latter structures might provide additional cues and context in the image, they are probably not the most important elements for interpreting the image. Consequently, we define participants *as the important visual elements in the image.*

Participants in an image are often involved in some kind of *narrative process*, such as interacting with another participant. Kress and van Leeuwen (2006) presents a variety of such processes describing various forms of narrative structures, such as *action processes, reactional processes, speech processes* and *mental and conversational processes*. However, as all these processes involve some kind of relationship between participants, they have been grouped together in this overview. Consequently, we define a narrative process as an interaction between two participants. Narrative processes commonly consist of an *actor*, representing the active part in the process, and a *goal*, representing the receiving part in the process. Visually, a narrative process between an actor and a goal is represented through an implicit or explicit visual structure called an *interaction vector*. One example of such a vector is the outstretched arm of the leftmost caretaker in Figure 1.

Representation

Representation concerns how participants are represented in an image. Two concepts that are relevant for image description and indexing processing are *framing* and *perspective*.

Framing represents the size of the "frame" used in an image, or the distance between the image observer and the participants in the image. Images are often classified as *close shots*, *medium distance shots* or *long distance shots*. The distance at which a participant is depicted, might be used to indicate the relationship between the observer and the participants, the importance of the participant, or the relative importance of a set of participants. A close-up shot of a person might indicate a personal relationship between the observer and the depicted participant. Medium- or long distance shots might similarly put the observer at a more analytical distance.

Perspective relates to the point of view used in the image, or the placement of the observer and the participants. This can be used to indicate the nature of the relationship between the observer and the participants, or between participants in the image. A 'low shot' might put a participant in a position of power over the observer, while a 'high shot' might indicate the opposite.

Composition

Composition describes the spatial and structural relationships between the participants, and between a participant and the image as a whole. Composition might be used to give further emphasis to the narrative structures in the image, indicate which participants are important, or present the observer with additional information used for interpreting the image. Examples of

compositional concepts are *Spatial distribution, salience, framing structures* and *use of colours*.

Spatial distribution relates to the distribution of participants in various zones of importance of an image, as defined by cultural conventions. For example, in western cultures, a "left" position might represent "something given", while a right position might represent something "new". Furthermore, a "top" position might represent the "ideal" while a "bottom" position might represent the "real".

Salience describes the saliency of a particular participant. In an image, various participants are commonly given saliency according to their importance. Important participants are more likely to be placed in the foreground, have a relatively larger size, or have a relatively higher contrast to the image, than less important participants.

Framing describes the presence of framing devices, or visual structures which connects or disconnects participants. One example of this would be to separate important participants from less important. In Figure 1 the boundary presented by the two stone walls places the first caretaker and the dolphin in one image segment, while the other caretaker is placed in another segment.

Finally, *colour* represents the presence, or use, of different colours in an image. Different colours might have different affordances and associations, determined by context. For example, "Love" or "Anger" might be associated with the colour red depending on the context.

Using the framework – an example

Figure 5 presents an example of the above concepts applied on the image presented in Figure 1. The example shows how these concepts can be used to construct narratives based on the generic semantic content of the image.

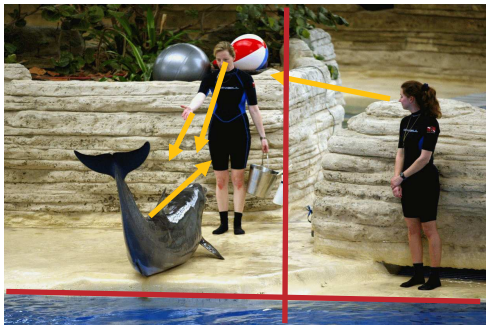

Figure 5 - Aqua Park Image illustrated with visual concepts

First of all, we can identify at least three important, or interesting, elements in the image: the dolphin and the two caretakers. These elements represent the *participants* in the image. The two balls or the stone walls might be considered participants. However, while they contribute to provide contextual information that might help an observer identify that the scene likely takes place in an aquarium, they are not explicitly involved in any activity in the image. Consequently, they are not considered participants in this particular instance.

The yellow arrows represent four potentially interesting *interaction vectors*, which might indicate interesting narrative processes. First of all, a substantial part of the dolphin's body is pointing towards the leftmost caretaker. The dolphin is identified as an *actor*, the caretaker as the *goal.* This can signify that the dolphin is focused on the caretaker, and is presumably involved in some sort of interaction, or transaction, with her. Similarly, the same caretaker's outstretched hand is signalling something to the dolphin, and her eyes are looking directly at the dolphin, possibly representing another transaction – the caretaker is expecting something from the dolphin, maybe passing some sort of instructions. Finally, the rightmost caretaker is watching the pair, possibly focusing on the actions and behaviour of the other caretaker. As seen in the introduction to this essay, combining these vectors with our understanding of the basic semantic units in the image (Dolphin, human, ball, aqua park) allows us to construct several possible narratives from this photograph.

The *framing* of the image is a long distance shot, and the *perspective* (point-of-view) of the shot is slightly from above. The observer is placed at an analytical distance to the participants, which might lead to the two different narratives described in the introduction to the essay. Another choice of perspective might lead to different narratives. For example, placing the point of view directly behind the dolphin, placing the caretaker in a "position of power" above the dolphin might induce stronger feelings towards the "cynical aqua park" narrative than the "human and dolphins playing" narrative, or indicate that the dolphin sees humans as a kind of "god".

Finally, the structural composition of the background might induce even further narratives. First of all, the two distinct stone walls present a natural *framing* (represented by the red, vertical line) grouping the dolphin and the leftmost caretaker together, excluding the second caretaker from their interaction. Finally, the line formed by the water on the bottom firmly places the dolphin out of its natural habitat, into the human domain, perhaps inducing thoughts of human dominance over the dolphin.

5. Automatic Image Indexing and Visual Grammar

Let us return to our initial question – *How can visual design structures be used to improve automatic indexing and description of the semantic content of digital images?* A natural starting point for this discussion is to compare the classification of semantic image content to the set of visual grammar structures presented in section 4.

The *generic semantic content* represents the basic semantic units in an image, while the *specific semantic content* represents such units that have been individually identified and named. In images such as Figure 1 this includes every little detail in the image, from the individual leaves of the trees, the stones in the walls or the eyes of the caretakers. However, in most cases it is likely that some of the generic or specific content is more important to an observer than others, such as the two caretakers, the dolphin and possibly the beach balls. This is very similar to the definition of image *participants*: the important visual elements in an image. Consequently, it is argued that the participants of the image represent at least some of the important generic and specific content in an image.

The other structures presented in section 4 are, to a certain degree, defined by perceptual structures in an image. *Interaction vectors* and roles such as *actor* and *goal* are defined by the syntactical structure of the individual participants and through their spatial relationships, e.g. the outstretched hand of the caretaker or the shape of the dolphin's body in Figure 1.

Similarly, *representation* and *composition* are primarily defined by the spatial arrangements of the participants and their syntactical features.

Finally, the *abstract content* of an image represents the meanings that can be derived from the general and specific contents based on the experiences and knowledge of the viewer. The example in section 4 illustrated how the visual grammar structures could be used to construct at least some basic narratives based on the image in Figure 1. Based on this, it can be argued that it is possible for a human to identify some abstract content by applying knowledge of visual grammar structures on an image. However, two important questions remain:

- Which requirements must be met in order to use this knowledge in an image retrieval system?
- Which steps must be taken in order to implement this knowledge in an image retrieval system?

Requirements for Utilizing Visual Grammar in Image Retrieval

The idea of using a set of rules based on visual grammar as an addition to CBIR systems seems enticing. However, there is a major challenge which needs to be addressed first: *object recognition*.

The visual grammar concepts presented in this text have the *participant* as the fundamental unit. The other concepts are based on various operations, comparisons and descriptions of the participants in an image, with a possible exception of *colour*. Without positive identification of the participants, the value of the other concepts is greatly diminished. Identification of participants becomes a fundamental requirement for using this approach. Consequently, in order to make use of this concept, an image retrieval system needs access to images where the participants are already identified. The three main approaches to this are: *automatic indexing, manual indexing* and *semi-automatic* indexing.

Automatic indexing is based on using technology from the field of CBIR for automatic indexing and description of participants. However, we have seen that current technology is very limited in its ability to automatically identify and describe the generic semantic content in an image. It is unlikely that a system using current technology will be able to automatically provide correct identification of participants. As such, automatic indexing of image participants for a general domain is not currently within reach. However, in certain types of content, such as face recognition, object recognition and identification is possible.

Manual indexing is based on having one or more individuals manually identify and tag the semantic content of an image. This approach is likely to provide results of high quality. However, significant time and manual effort is required to describe images, and it is unlikely that this approach would be suitable for a large number of images.

Finally, *semi-automatic indexing* represents the recent initiatives in community-based indexing methods. Rather than using individuals for indexing and describing images, the task is performed by a potentially large community of users. While this approach might result in a varied quality in the results, it can potentially process a much larger volume of images.

The rest of this discussion assumes that it is possible to create a set of images where the participants have been identified. It is likely that improvements will be made for both

automatic and semi-automatic indexing methods, and manual indexing is already a possibility. While current methods might not ideal for this purpose, it is argued that it might indeed be possible to create a set of images with participants identified. Furthermore, even if future improvements in either automatic or semiautomatic indexing improve to the point where they can provide reliable identification of the generic semantic content in an image, it will not automatically allow users of the system to query the collection based on complex, abstract or contextual content. This will still requires some sort of additional description. It is believed that an approach based on visual grammar be a promising approach in this regard.

The Suggested Approach

The discussion above has indicated that concepts from visual design might provide a potentially interesting approach towards improving image retrieval systems. As a conclusion to this discussion, an approach for implementing this in an image retrieval system is suggested. Two vital tasks are identified and discussed:

1. Developing an indexing scheme for describing the conceptual structures in an image, e.g. participants and interaction vectors
2. Developing a set of rules based on concepts from visual design, and translating these rules into a set of formal rules which can be executed by a software

Figure 4 presents a schematic overview of a system for indexing and describing images based on this approach. While the figure is very simplified, it presents an overview of the interaction and relationships between the components defined above and a set of images.

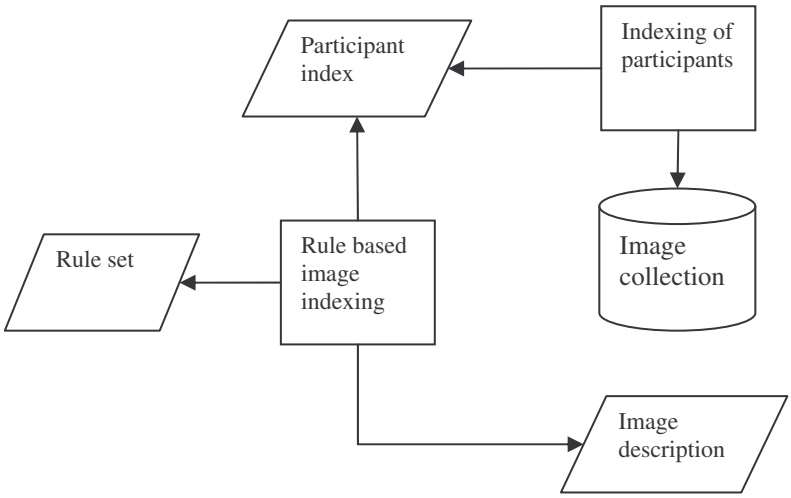

Figure 6 - Simplified view of the relationship between images, participant indexes and rule set

Developing a participant indexing scheme

The fundamental requirement for this approach is that the image system has *a priori* information about the participants present in the images being processed. Consequently, the first step must be the development of an indexing scheme for the describing participants.

A structure for this scheme and the details for implementing it must be evaluated through a separate study. However, the minimum requirement of the indexing scheme is that the size

and overall contour of each participant are represented with coordinates within the image. This will allow the rules to identify the spatial distribution and salience of the participant, and should be relatively easy to obtain using any of the object recognition methods described above.

Next, the level of detail of the descriptions should be balanced against the cost of obtaining these. The more detailed this description is, the higher the potential benefits are. For example, in order to identify and determine potential interaction vectors, the contour should include the most salient features and major axis of the participant, e.g. the caretaker's extended hand and the shape of the dolphin in Figure 1. However, more detailed descriptions will require either more effort by the indexers, or better indexing algorithms.

Developing a set of rules for visual Grammar

In order for an image retrieval system to use utilize the visual design concepts presented in section 4 these must be translated into a set of clearly defined, unambiguous rules that can be evaluated using computational processes. As with the indexing scheme, the actual development of these rules must be evaluated in a separate, more detailed study. Two main tasks are required in order to do this.

First, the presentation in Table 2 is a qualitative, linguistic description. As such, it is not well-suited for direct translation into computational rules. A more thorough analysis of visual design and grammar should be performed in order to identify and detail a set of formal rules. The rules must have set of conditions for evaluating whether it is satisfied or not.

Next, these rules must be specified in a matter that can be evaluated using software, such as predicate logic, fuzzy logic or methods based in artificial intelligence. The rules should be matched against the level of detail of the description of the participants.

As an illustration of this, consider the task of creating a set of rules for compositional concepts. Given that the system has *a priori* knowledge of the presence of participants, along with their relative spatial distribution, a set of *compositional rules* can be created. In section 2 we saw that an image can be divided into various zones of importance. Based on the size and placement of a particular participant, it is easy to determine whether the participant is within a particular zone. Next, relative importance of the various participants might be determined based on a comparison of relative size and saliency. Finally, it is possible to detect the presence of sharp edges or distinct boundaries in an image using methods from automatic image processing. This can be used to determine the presence of *framing devices*, and the relative placement of participants according to these might provide additional information about relationships between participants.

6. Conclusion

The main motivation behind this paper has been to discuss how we can use knowledge of visual design to improve automatic indexing and description of the semantic content of digital images. Interpretation and understanding of higher level image semantics cannot be inferred from the syntactical and perceptual image structures, even if the generic semantic content are identified through manual or semi-automatic indexing methods. Interpretation and understanding of higher level image semantics is dependent on cultural, technical, emotional and contextual knowledge.

While it has been outside the scope of this paper to fully explore the questions raised by this discussion, it is believed that introducing concepts and tools from visual design might prove a fruitful approach towards enhancing image indexing- and retrieval systems.

Two main tasks towards this goal have been identified. First, a system for indexing and describing the important participants must be developed. Next, a set of rules based on concepts from visual design must be developed, and translated into a set of rules which can be evaluated using computational processes.

Application Area and Generality

As a conclusion, some notes on the application area and generality of the proposed are presented. At first glance, application of rules based on visual design concepts might seem restricted to images created using a similar set of rules. For example, one cannot take for granted that every image is taken by a photographer with a firm grasp of the use and meanings of visual concepts such as framing, composition and perspective. Furthermore, in the case of art photos, digitized paintings and similar works of arts, it is quite possible that the creator deliberately violates one or more of these guidelines order to create a particular effect, or evoke certain reactions or emotions in the viewer. Accordingly, this approach might prove more useful for images created primarily to communicate a particular message, or adhering to the conventions of a particular genre, such as portraits, art images or advertisements.

However, it is possible to imagine that such rules *could be used* on images that were not created within a visual design framework. First of all, some of the rules might be applicable even though they weren't consciously applied by the image creator. The example in Figure 1 presented us with several possible narratives constructed using the visual concepts presented in Table 2. However, we do not know whether this image was created with any regard to the described concepts. The image was retrieved from a personal web page[2]. The photographer stated that he captured the image during a performance in an aqua park, and presented the image as a lucky snap-shot of the performance.

Despite of this, we were able to construct several narratives based on the image. Even though the image might not have been *created* within the framework of a visual grammar, it is up to the discretion of the viewer how he or she chooses to read and interpret the image. The image itself does not carry any other information than is presented by its structural syntax. It is left to the discretion of the viewer to create meanings in the image, based on its composition, framing and perspective, filtered through the viewers own experiences. Even though the results might not be in accordance with the creator's intentions, they might still be valid *interpretations* of the image and used for indexing, description and classification of the image.

7. Acknowledgements

This paper is written as a part of a course in Human and Computer Visual Communication, at the department of Information Science and Media Studies at University of Bergen. I would like to thank Joan Nordbotten and Jens Kjeldsen for valuable discussions and insights, as well as their contributions during the work on this essay.

[2] No longer available online

8. References and Bibliography

Carson, C. (1997). Region Based Image Query. Proceedings of IEEE CVPR, Santa Barbara, California, IEEE.

Colombo, C. and A. Del Bimbo (2002). Visible Image Retrieval. Image Databases: Search and Retrieval of Digital Imagery. V. Castelli and L. D. Bergman, John Wiley & Sons**:** 11-31.

Datta, R., D. Joshi, et al. (2007). "Image Retrieval: Ideas, Influences, and Trends of the New Age." ACM Computing Surveys **39 (To appear)**.

Eakins, J. P., B. Burford, et al. (2003). "A taxonomy of the image: on the classification of content for image retrieval." Visual Communication **2**(2): 123-161.

Eidenberger, H. (2004). A new perspective on visual information retrieval. SPIE Electronic Imaging Symposium, San Jose, SPIE.

Faloutsos, C., R. Barber, et al. (1994). "Efficient and effective querying by image content." Journal of Intelligent Information Systems **3**: 231-262.

Hove, L.-J. (2004). Extending Image Retrieval Systems with a Thesaurus for Shapes. The Norwegian Information Technology Conference, Stavanger, Norway, NIK.

Hove, L.-J. (2004). Improving Image Retrieval with a Thesaurus for Shapes. Department of Information Science and Media Studies. Bergen, University of Bergen.

Huang, T. S. and Y. Rui (1999). "Image Retrieval: Current Techniques, Promising Directions And Open Issues." Journal of Visual Communication and Image Representation **10**(4): 39-62.

Jaimes, A. and S.-F. Chang (2002). Concepts and Techniques for Indexing Visual Semantics. Image Databases: Search and Retrieval of Digital Imagery. V. Castelli and L. D. Bergman, John Wiley & Sons, Inc.**:** 497-565.

Kimia, B. B. (2002). Shape Representation for Image Retrieval. Image Databases. Search and Retrieval of Digital Imagery. V. Castelli and L. D. Bergman. New York, John Wiley & Sons, Inc.**:** 345-372.

Kress, G. and T. van Leeuwen (2006). Reading Images - The Grammar of Visual Design. Oxon, Routledge.

Li, Y. and C. C. J. Kuo (2002). Introduction to Content-Based Image Retrieval - Key Techniques. Image Databases: Search and Retrieval of Digital Imagery. V. Castelli and L. D. Bergman. New York, John Wiley & Sons**:** 261-284.

Messaris, P. (1994). Visual Literacy - Image, Mind & Reality. Oxford, Westview Press.

Rui, Y., T. S. Huang, et al. (1998). Relevance Feedback Techniques in Interactive Content-Based Image Retrieval. Storage and Retrieval for Image and Video Databases**:** 25-36.

Santini, S. and R. Jain (1997). "The Graphical Specification of Similarity Queries." Journal of Visual Languages and Computing **7**(17).

Sclaroff, S. and L. Liu (2001). "Deformable Shape Detection and Description via Model-Based Region Grouping." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(5): 475-489.

Smeulders, A., M. Worring, et al. (2000). "Content Based Image Retrieval at the End of the Early Years." IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12): 1349-1380.

von Ahn, L. and L. Dabbish (2004). Labelling Images with a Computer Game. ACM Conference on Human Factors in Computing Systems, Vienna, Austria, ACM.

von Ahn, L., L. Ruoran, et al. (2006). Peekaboom: A Game for Locating Objects in Images. ACM Conference on Human Factors in Computing Systems, Montréal, Québec, Canada, ACM.