Image Semantics in the Description and Categorization of Journalistic Photographs

Mari Laine-Hernandez, Media Technology, Helsinki University of Technology, P.O. Box 5500, FI-02015 TKK, Finland, mari.laine-hernandez@tkk.fi

Stina Westman, Media Technology, Helsinki University of Technology, P.O. Box 5500, FI-02015 TKK, Finland, stina.westman@tkk.fi

This paper reports a study on the description and categorization of images. The aim of the study was to evaluate existing indexing frameworks in the context of reportage photographs and to find out how the use of this particular image genre influences the results. The effect of different tasks on image description and categorization was also studied. Subjects performed keywording and free description tasks and the elicited terms were classified using the most extensive one of the reviewed frameworks. Differences were found in the terms used in constrained and unconstrained descriptions. Summarizing terms such as abstract concepts, themes, settings and emotions were used more frequently in keywording than in free description. Free descriptions included more terms referring to locations within the images, people and descriptive terms due to the narrative form the subjects used without prompting. The evaluated framework was found to lack some syntactic and semantic classes present in the data and modifications were suggested. According to the results of this study image categorization is based on high-level interpretive concepts, including affective and abstract themes. The results indicate that image genre influences categorization and keywording modifies and truncates natural image description.

Keywords: image content, free description, keywording, categorization, image categories, multidimensional scaling, hierarchical cluster analysis

Introduction

The digitalization of image collections has increased the availability of pictorial material for both commercial and research use. There exists a growing body of research into image retrieval and description. The nature of visual information, however, creates some special challenges. The range and type of attributes needed for describing image content is still under debate. Several frameworks have been created, yet their match with natural, unconstrained image descriptions formed by users has not been proved. The issue of attribute granularity is also challenging; on how many semantic levels should image access be provided? The meanings carried by images, the specificity of index terms, as well as the queries made to image collections may be of various levels. A query might request a specific item or an instance of a general category. It might also deal with a topical category of images or specify a particular abstract concept or affective response the image should evoke.

The development of content-based image retrieval systems (i.e. systems that use visual image data to perform queries) has been an area of great interest during the last decade, but many challenges still remain. These include defining visual similarity so that it would match the users' mental models of similarity, as well as bridging the semantic gap between the higher-level semantic concepts used by people and the perceptual attributes addressed by the content-based algorithms. Domain and expected users are important in the development of image description and search tools. Systems and description schemes have been developed for both general audiences and domain specialists, as the requirements vary between these two groups. How images are and should be described depends on both the image content and the context. The tasks that give rise to the image search and the images' intended future uses affect the image descriptions and queries.

The indexing of journalistic images has previously been studied by Ornager (1997), but her word association experiment was conducted from a linguistic viewpoint. Markkula and Sormunen (2000) conducted a qualitative study about the current indexing practices of archivists in a newspaper. The purpose of the present study was to review existing frameworks for image descriptions and indexing, and to evaluate their suitability for journalistic photographs. Another goal was to examine the influence of a predetermined image genre, in this case journalistic photographs, on the results. This was achieved by studying how people describe photograph content and what criteria they use for the categorization of photographs. The effect of different tasks on image description and categorization was also studied. Additionally, the possibility of the effect of one particular image-related task to the performance on another was considered.

Past research on image description and categorization

Image semantics in indexing and categorization

Jörgensen (1998) analyzed user behavior in three image description tasks she called viewing, search and memory tasks. The viewing task elicited unconstrained image descriptions from the subjects. The search task approximated a known item search; the subjects saw an image and wrote a query for it in an imaginary, "ideal" retrieval system. The memory task tested the subjects' recall of the images six weeks later. They wrote search descriptions of the images used in the viewing task. The data analysis revealed 12 distinct classes of image attributes used by the subjects. Among these, Jörgensen distinguishes between perceptual (P), interpretive (I) and reactive (R) attributes. The classes with their percentage distributions in the three tasks are shown in Table 1. According to the results, the object class was the prevalent level of image description in all tasks. Jörgensen also stresses the importance of people and associated attributes as well as terms describing the story of the images. Unfortunately, Jörgensen does not define attributes nor report results at the level of individual attributes, only classes. Also the classification of objects as a perceptual class is problematic as their recognition requires semantic interpretation, see e.g. Jaimes and Chang (2000).

Table 1. Distributions of attribute classes (Jörgensen, 1998)

Attribute class	Viewing task	Search task	Memory task
Object (P)	34.3	27.4	26.2
People (P)	8.7	10.3	11.1
Color (P)	9.2	9.7	9.0
Visual elements (P)	7.2	5.4	9.2
Location (P)	8.3	10.7	7.7
Description (P)	6.0	9.0	8.8

Attribute class	Viewing task	Search task	Memory task
People-related attributes (I)	5.2	3.9	2.6
Art historical information (I)	3.8	5.7	7.6
Abstract concepts (I)	3.0	1.5	1.3
Content/story (I)	7.4	10.8	9.4
External relationships (I)	3.3	3.8	4.0
Viewer response (R)	3.7	1.9	3.1

Greisdorf and O'Connor (2002) asked subjects to indicate whether each of ten scenery images matched given query terms pre-selected from basic attribute categories. In a second task the subjects were asked to generate and list all words that they felt could match one or more of the images. Table 2 shows the percentage distribution of the category usage in both tasks. The subjects used low-level attributes (color, shape and texture) more when the query terms were pre-selected. Conceptual and affective terms were more common among the user-supplied terms. The low significance of the action/motion category could be attributed to the image type used in the experiments.

Table 2. Distribution of category usage (Greisdorf & O'Connor, 2002)

Category	Pre-selected terms used	User-supplied terms used
Color, shape, texture	34.8	13.0
Object	13.1	24.2
Action/motion	15.5	4.3

Category	Pre-selected terms used	User-supplied terms used
Place/location	9.1	19.8
Affect/emotion	27.5	38.6

Hollink et al. (2004) had their subjects perform an illustration task. They were asked to read the text and form an image in their mind that could be an illustration for the text. They wrote a free description of the image and searched for the image using a textual search. Hollink et al. classified the descriptions and the queries using their framework for the classification of image descriptions. The framework divides visual descriptions into perceptual and conceptual descriptions (general, specific and abstract concepts). The most frequently used level in both tasks was the general conceptual level (63.4% of terms in the description task and 67.5% in the querying task). At other levels some differences existed between the tasks. The free descriptions contained less specific and more abstract and perceptual attributes than queries (8.2% vs. 20.4%, 11.1% vs. 5.2 and 17.3% vs. 7.0%), but there was no statistical difference between the attribute distributions of the two tasks.

Jörgensen's results indicate that people mostly use attributes of objects, people and story as well as other semantic level terms when freely describing image content. The importance of the conceptual level in the results of Hollink et al. (2004) supports this. The finding has also been supported by studies on image searchers' needs, most recently by a survey conducted by Eakins et al. (2004). Their results indicate that semantic image content (semantic terms, cultural and technical abstractions) is the most important type of image content for searches. The user studies by Greisdorf and O'Connor (2002) and Hollink et al. (2004) seem to suggest that there are differences between free image descriptions and using imposed terms or adapting descriptions to a query system. Query tasks result in more specific terms and less abstract and perceptual terms than unconstrained image description tasks. Free description tasks lead to the use of more semantic attributes (objects, location, emotion) compared to tasks requiring the use of imposed terms. Hollink et al. (2004) compared free description and querying of the same (mental) image, but to the best of the authors' knowledge no studies have compared free descriptions and constrained annotations.

Image indexing frameworks

For the purpose of this study several image indexing frameworks and taxonomies of image attributes or image content were reviewed. These include the classifications used by Hollink et al. (2004) and Jörgensen (1998) in their user studies. In addition, two theoretical frameworks were reviewed. Shatford (1986) has extended Erwin Panofsky's theory of meaning in art images to apply to all images. Based on Panofsky's three levels (pre-iconographical, iconographical and iconological), Shatford categorized the subjects of images as generic of, specific of and about. Ofness refers to the factual content of the image ("what is the image of"). Aboutness on the other hand refers to the expressional content ("what is the image about?"). Shatford also added four facets to each of the three levels: Who? What? Where? and When? The Panofsky/Shatford facet matrix has become a widespread model for describing image content and it has been used widely in research. Jaimes and Chang (2000) have developed a conceptual pyramid model for describing visual content based on previous research. The pyramid contains ten levels: the first four describe the syntax of an image and the remaining six refer to its semantics. The model relies in its classification on the amount of knowledge required to identify and index attributes on each level. The higher the level, the more knowledge is needed to formulate a description. The first four levels are so-called perceptual levels, on which no world knowledge is needed. The six remaining levels are conceptual levels. General, specific or abstract knowledge is required to formulate descriptions on these levels.

The models are summarized in Table 3. The "syntax/semantics" column indicates a rough division of low-level content descriptors and high-level semantic keywords. Some of the frameworks focus more on the semantic image content whereas others include both syntactic and semantic levels. Although the pyramid by Jaimes and Chang contains the most levels, Jörgensen's framework is the most extensive since it contains also class-specific attributes. It also covers subjective reactions to images.

Table 3. Image content levels in various theories and frameworks summarized

<table>
<tr><th></th><th>Jaimes & Chang (2000)</th><th>Panofsky / Shatford (1986)</th><th>Hollink et al. (2004)</th><th colspan="2">Jörgensen (1998)</th></tr>
<tr><td rowspan="5">SYNTAX</td><td>Type/technique</td><td rowspan="5"></td><td rowspan="5">Perceptual</td><td>Interpretive</td><td>Art historical information</td></tr>
<tr><td>Global distribution</td><td rowspan="4">Perceptual</td><td>Color, visual elements</td></tr>
<tr><td>Local structure</td><td>Color, visual elements</td></tr>
<tr><td>Global composition</td><td>Color, visual elements, location</td></tr>
<tr><td rowspan="6">SEMANTICS</td><td>Generic objects</td><td rowspan="2">Pre-iconography / generic "of"</td><td rowspan="2">General conceptual</td><td>Perceptual / Interpretive</td><td>Objects, people</td></tr>
<tr><td>Generic scene</td><td>Interpretive</td><td>Content/ story</td></tr>
<tr><td>Specific objects</td><td rowspan="2">Iconography / specific "of"</td><td rowspan="2">Specific conceptual</td><td rowspan="2">Interpretive</td><td rowspan="2">Content/ story</td></tr>
<tr><td>Specific scene</td></tr>
<tr><td>Abstract objects</td><td rowspan="2">Iconology / "about"</td><td rowspan="2">Abstract conceptual</td><td rowspan="2">Interpretive</td><td rowspan="2">Abstract, people-related and reactive attributes</td></tr>
<tr><td>Abstract scene</td></tr>
</table>

Image similarity and categorization

Sormunen et al. (1999) asked photojournalists to evaluate the similarity of photographs in order to build a test collection for evaluating content-based image retrieval. They found that journalists evaluated the similarity of images based on the following criteria: shooting distance and angle, colors, composition, cropping, photo direction, background, direction of movement, objects in the image, number of people in the image, action, facial expressions and gestures, and abstract theme. Rogowitz et al. (1998; Mojsilović & Rogowitz, 2001b) conducted psychophysical experiments in which subjects organized natural images according to perceived similarity. The results obtained using multidimensional scaling (MDS) revealed clearly interpretable axes along which the subjects evaluated similarity: natural vs. man-made axis (ranging from natural to man-made objects and scenes) and human vs. non-human axis (ranging from close-ups of people to images of animals and inanimate scenes). Teeselink et al. (2000) conducted a free categorization task and a foreground-background categorization task on images of natural scenes. They were looking for a generalized relation between perceived categories of image content and perceived foreground-background separations. The number of categories was predefined: the subjects were asked to categorize the photos first into two groups, then to three, four and five groups. The results indicated that the most important basis for categorization was the presence of people in the photographs. The second most important feature was whether the photographs depicted a landscape or objects/buildings. Based on their recent study on free categorization of images, Rorissa and Hastings (2004) concluded that interpretive attributes are better candidates than perceptual attributes for indexing categories of images. In their experiment the main categories formed by the subjects were: exercising, single men/women, working/busy, couples, poses, entertainment/fun, costume and facial expression. The test image material Rorissa and Hastings used consisted of photographs of people with the backgrounds removed.

Methodology

An empirical study on image description and categorization was conducted to find the answers to the following research questions: Which image levels from Table 3 are prevalent in the description, keywording and categorization of journalistic photographs? How does the usage of image levels differ between the keywording and free description of image content? Does categorization of the images depend on the nature of the earlier description task (keywording/free description)? Do reviewed results regarding image description and categorization apply to journalistic photographs?

Material

The test material consisted of 40 reportage-type photographs from two online image collections by image journalists and amateur photographers. Of the images, 31 included people, ranging from single person close-ups to images of masses. People were depicted in various activities, situations and environments. The remaining 9 photographs depicted inanimate objects, animals or scenery. Several criteria were used as basis for the selection of photographs: broad range of color distribution, colorfulness (calculated according to Hasler and Süsstrunk 2003) and lightness levels; strong visual elements (texture, shape); various distances of the object to the viewer; wide range of topics and semantic content; and emotional content. The photographs were selected so that they could be categorized in various ways. This was assumed because each of the selected photographs had common features with several other selected photographs. Some assumptions of possible image groupings were made during the selection to verify these multilevel linkages.

Procedure and participants

A total of 20 subjects (12 male, 8 female) participated in the study. The participants were students of technology and university employees. All participants were native Finnish speakers and the experiments were conducted in Finnish. The experiment contained three tasks: keywording, free description and categorization. The subjects were divided into two groups at random and maintaining gender balance. One group performed the keywording task and the other performed the free description task. All subjects participated in the categorization task after the first task. The test photographs in the keywording and free description tasks were displayed one at a time on a CRT computer screen from a normal (unrestricted) viewing distance. The maximum width/length of a photograph was about 14 cm. The photographs were displayed in random order without time

limitations. While looking at the photograph the subject wrote down their description in a text field next to the photograph. In the keywording task the subjects were asked to write down the first five words to come to their mind that best describe the photograph they see. This task reflects the nature of keyword annotation frequently used in image databases at photo agencies and newspapers. In the free description task the subjects were asked to write down a free description of the photograph as they would when describing its content to another person. This task is a completely unrestricted image description aimed to reveal the natural way people describe and communicate image content. The categorization task took place after the keywording or free description task. The subjects were presented with print versions of the photographs from the first task glued to pieces of grey cardboard. They were asked to organize the photographs into categories according to similarity. The number of categories was not restricted, and also single photographs could form categories. The subjects did not receive any further instructions regarding how they should judge the similarity of the photographs. After the completion of the task the subject was asked to explain and name the categories.

Data analysis for the keywording and free description tasks

The data from the keywording task contained individual words or multiword terms such as full names as was requested by the instructions. An average of 202 keywords was elicited per subject. This is slightly higher than the requested 5 terms per photograph and is due to some multiword descriptions being categorized as separate terms. An example of an answer (translated from Finnish) in the keywording task contained the following words: *cows, chub, close-up, early spring, pasture.* In the free description task, an unconstrained description was called for. The subjects mostly wrote complete or near-complete sentences from which meaningful words were extracted so that the results from the two tasks could be compared. An average of 315 words per subject was extracted with considerable variation in length; the number of words per subject ranged from 139 to 428 with a standard deviation of 98. One subject's description of one photograph yielded in average 8 extracted words. An example of an answer in the free description task is: *"Two cows peeking out from between barbed wires. One of the cows is white and the other one has black spots. The ground is covered in snow and the sky is bright."* The resulting words were categorized according to Jörgensen's (1998) framework. This framework was selected because it was the most detailed and extensive one. This assured that the categorization of the terms would be as detailed as possible.

Data analysis for the categorization task

The categorization data was analyzed using multidimensional scaling and hierarchical cluster analysis in Matlab. The data was first converted to an aggregate dissimilarity matrix. The percent overlap S_{ij} for each pair of photographs i and j was calculated as the ratio of the number of subjects who placed both i and j in the same category to the total number of subjects. The percent overlap gives a measure of similarity, which was then converted to a measure of dissimilarity: $\delta_{ij} = 1 - S_{ij}$. Rorissa and Hastings (2004) also used this procedure in their study of free sorting of images. The hierarchical cluster analysis was done using the complete-linkage (farthest-neighbor) method, applied in similar studies also by Lohse et al. (1994), Vailaya et al. (1998) and Teeselink et al. (2000). Two-dimensional non-metric multidimensional scaling was performed on the data. Multidimensional scaling was used by Lohse et al. (1990) to confirm clustering results.

Results

Keywording and free description

The percentage distributions of classes and attributes in this study are listed together with Jörgensen's (1998) results (average of three tasks from Table 1) in Table 4. The viewing task from Jörgensen is the one most closely related to the free description task and the search task to the keywording task. However, no systematic consistency was found either between these specific task pair results or the differences between the two different task types in the two studies. Thus the unweighted average of Jörgensen's tasks was used as a basis for comparison. The attribute classes are written in capital letters and the attributes belonging to each class are listed under the class name.

Interpretational levels were the most prevalent description levels used in this study. They accounted for 51.4% of all the terms in the keywording task and 28.1% of all the terms in the free description task. The classes CONTENT/STORY and PEOPLE-RELATED ATTRIBUTES accounted for roughly one

third of the terms used by the subjects: 40.1% in the keywording task and 26.1% in the free description task. The most used attribute class in both tasks was OBJECTS (26.3% in the keywording task and 29.1% in the free description task). The usage of image levels differed somewhat between the keywording and free description tasks, the most notable difference occurring in the use of the class LOCATION. References to either the general or specific location of elements within the image were nearly absent in the keywording task (0.3% of all terms) but both location attributes were used quite frequently in the free description task (10.2%). Number in the DESCRIPTION class was another attribute nearly missing in the keywording task (0.4% of all terms) but used in the free description task, where numeration of people and objects accounted for 3.4% of terms. Also attributes such as body parts and people (e.g., "woman") were more common in the free description task. The use of ABSTRACT CONCEPTS and CONTENT/STORY attributes differed between the tasks. Both of these attribute classes were used more in keywording than in free description (10.8% vs. 1.7% and 28.2% vs. 17.4%). The most salient differences appeared regarding attributes concerning theme, event and setting. Attributes depicting emotion or relationship were also used more in the keywording task. Also worth noting are attributes of conjecture and uncertainty, referring to words such as "maybe" or "apparently", depicting the subject's own, sometimes uncertain, interpretations of the photograph's content. These VIEWER RESPONSE attributes were used more in free description than in keywording (3.6% vs. 0.9%).

Table 4. Percentage distribution of classes and attributes from Jörgensen (1998) (J) in the keywording (Key) and free description (Free) tasks

Class / attribute	J	Key	Free
OBJECTS (P)	**29.3**	**26.3**	**29.1**
object		22.8	19.5
text		0.0	0.3
body part		2.0	6.0
clothing		1.5	3.4
PEOPLE (P)	**10.0**	**4.1**	**7.0**
people		4.1	7.0
COLOR (P)	**9.3**	**3.0**	**6.2**
color		2.5	4.4
color value		0.5	1.8
VISUAL ELEMENTS (P)	**7.2**	**6.4**	**4.0**
composition		0.4	0.6
focal point		0.4	0.2
motion		1.9	0.4
orientation		0.6	0.3
perspective		1.7	1.3
shape		0.6	0.6
texture		0.1	0.1
visual component		0.9	0.4
LOCATION (P)	**8.9**	**0.3**	**10.2**
general		0.2	4.4
specific		0.1	5.8
DESCRIPTION (P)	**8.0**	**7.4**	**12.0**
description		7.1	8.5
number		0.4	3.4
PEOPLE-RELATED ATTRIBUTES (I)	**3.9**	**12.1**	**8.7**
relationship		1.3	0.5
social status		8.6	7.8
emotion		2.2	0.4
ART-HISTORICAL INFORMATION (I)	**5.7**	**0.0**	**0.0**
ABSTRACT (I) CONCEPTS	**2.0**	**10.8**	**1.7**
abstract		3.9	1.0
atmosphere		1.0	0.2
state		0.8	0.5
symbolic aspect		0.0	0.0
theme		5.2	0.0
CONTENT/ STORY (I)	**9.2**	**28.2**	**17.4**
activity		8.9	9.6
category		0.0	0.2
event		6.7	1.8
setting		10.5	5.0
time aspect		2.1	0.8
EXTERNAL RELATION (I)	**3.7**	**0.4**	**0.3**
comparison		0.0	0.1
similarity		0.1	0.1
reference		0.3	0.1
VIEWER RESPONSE (R)	**2.9**	**0.9**	**3.6**
personal reaction		0.7	0.5
conjecture		0.2	2.2
drawing		0.0	0.0
uncertainty		0.0	0.8

Categorization

The number of categories formed by subjects in the categorization task varied between 6 and 24, and the number of photographs per category between 1 and 10. On average the subjects in both groups

(keywording and free description) created 15 categories with an average of 3 photographs per category. Examples of individual thematic image categories that occurred frequently are: religion (6.8% of all categories named), animals (5.7%), politics (5.7%), scenery (5.4%), sports (5.4%) and music (5.0%). Categories were also formed based on the following unifying concepts: activity/event (16.1%), cultural references such as cultural background or country (13.6%), terms describing emotions and/or atmosphere (11.1%), and visual elements such as shape, color or perspective (5.7%). References to people (e.g. children, soldiers, Prince Charles) were present in 24.7% of category names. Category names also often combined two concept types, e.g. Indian woman, children playing, colorful scenery. The results of the hierarchical cluster analysis are presented in Figure 1. The cluster labels have been extracted from the explanations and names provided by the subjects. Some of the eight top-level clusters have clearly distinguishable sub-clusters, totaling fifteen clusters, equaling the mean number of groups into which the subjects categorized the photographs. The quality of the solution was evaluated by calculating the cophenetic correlation coefficient. The magnitude of the coefficient should be close to 1 for a high-quality solution. It was 0.91 for the solution obtained.

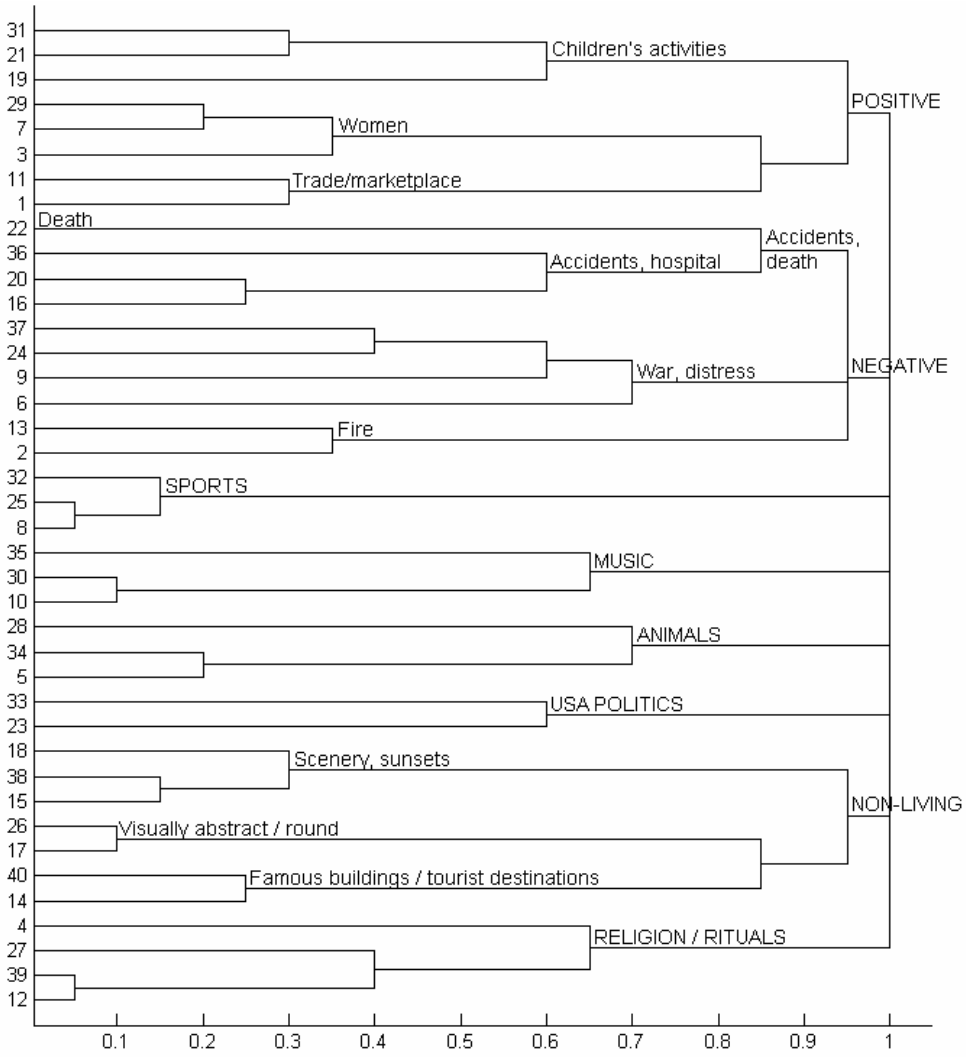

Figure 1. A dendrogram for the 40 test photographs obtained with the complete-linkage method

The results of two-dimensional non-metric multidimensional scaling are shown in Figure 2. The MDS did not reveal any main coordinates in the organization of the photographs. It could be concluded that hierarchical clustering was a more appropriate method, since the task concerned categorizing (or clustering) photographs. It is possible that the small number of photographs with large variation in

semantic content prevented the emergence of axes. Rogowitz et al. (1998) also comment that “very low dimensional spaces cannot represent the full complexity of perceptual similarity judgments”. The stress value in the non-metric scaling was 0.14 which is considered fair.

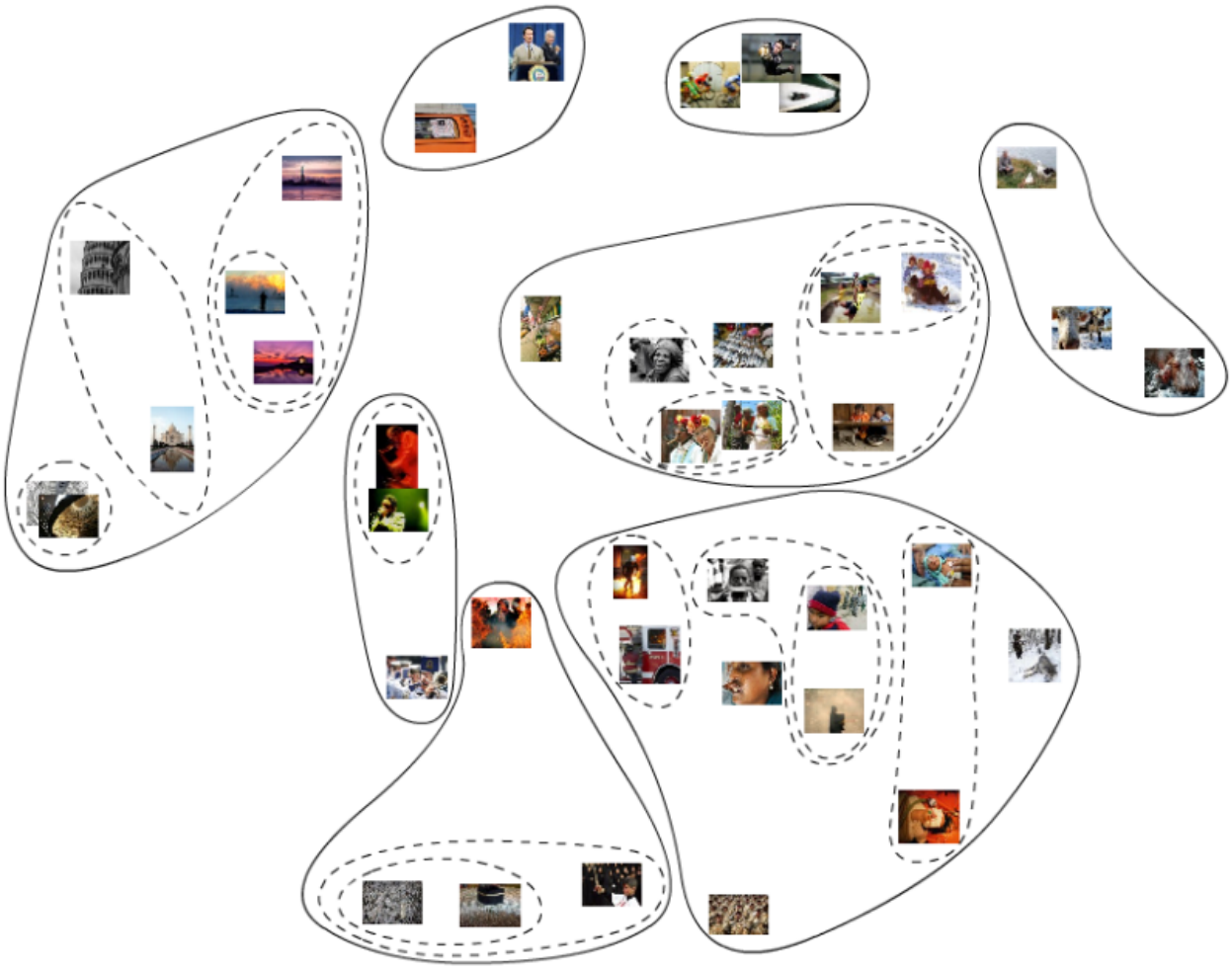

Figure 2. Results of the 2D MDS with the main clusters given by the hierarchical cluster analysis visualized with solid circles and some of the sub-clusters visualized with dashed circles

Influence of earlier task on categorization task

One of the aims of this study was to find out if there are differences in how people categorize photographs depending on the task they have performed before the categorization (in this case keywording vs. free description). For this reason the hierarchical cluster analysis was carried out separately for the two groups of subjects and Jaccard’s coefficient was used to measure if the clustering results differed between the groups. It was also used by Lohse et al. (1994) and Rorissa and Hastings (2004) to test the consistency of subjects in sorting tasks. The calculated Jaccard’s coefficient value for the two subject groups of this study was 0.788. It may be concluded that the nature of the first task did not influence the way in which the subjects later categorized the photographs.

Discussion

The distributions of the terms elicited in the free description task are very similar to Jörgensen’s (1998) results regarding the objects, location, abstract concepts and viewer response classes. The largest difference occurs in the class CONTENT/STORY. This is probably due to the fact that the subjects used a lot of verbs to describe the activities in the photographs, resulting in 9.6 % (vs. 5.3 % in Jörgensen’s

study) of all the words belonging to the attribute “activity” in the class CONTENT/STORY. The results on keywording differ more from those of Jörgensen in that CONTENT/STORY attributes are even more pronounced, and ABSTRACT CONCEPTS and PEOPLE-RELATED ATTRIBUTES are more common, whereas COLOR and LOCATION attributes are used less. Jörgensen names OBJECTS as a perceptual class but the attributes within the class are most often clearly semantic in nature. If objects are taken to belong to the interpretational levels, the joint share of the OBJECTS, CONTENT/STORY and PEOPLE-RELATED ATTRIBUTES covers more than half of all elicited terms (66.7% in keywording and 55.2% in free description). The grouping of the PEOPLE and DESCRIPTION as perceptual should be reconsidered, as they often contain semantic attributes. The review of the classification system was hindered by the fact that Jörgensen (1998) does not define the attribute classes thoroughly.

Based on the results, most of the terms used in the description and keywording of photographs were interpretational semantic. This is consistent with the results from several earlier studies (Fidel, 1997; Greisdorf & O’Connor, 2002; Jörgensen, 1998; Jörgensen, 1999; Mojsilovic & Rogowitz, 2001a; Rorissa & Hastings, 2004). High-level and affective concepts were also used in image categorization while some but relatively few categories were based on similar visual elements. All the semantic levels in the framework by Jaimes and Chang (2000), i.e. generic, specific and abstract, were used in the description and categorization of photographs. It can be concluded from these results and reviewed studies that perceptual (or syntactic) attributes alone are not suitable for labeling and indexing image groups or categories. Within a single semantic content category, on the other hand, perceptual attributes can be used e.g. to aid browsing.

There were differences between the terms elicited in the two description tasks. Attributes referring to theme, event and setting, atmosphere and emotion were used more in the keywording task, which seems to reflect the subjects’ need to limit and summarize their interpretations. The use of attributes such as numbers, colors and descriptions was less common in the keywording task, also likely due to the need to express the content concisely. Some visual elements such as motion and composition were more notably mentioned in the keywording task. This seems to be the result of describing the whole photograph at once, rather than recounting its content part by part. Unconstrained description resulted in several mentions of the location of objects within the photograph and in the enumeration of people and objects due to the narrative form the subjects used without prompting. Regular mentions of uncertainty and conjecture led to more frequent use of reactive attributes in the free description task.

The photographs were often categorized based on the presence of people as well as content/story-related and abstract concepts. These are among the criteria photojournalists used in their evaluations of image similarity in the study by Sormunen et al (1999). Salient themes and activities were further used to categorize photographs depicting people. Scenery and architecture photographs were frequently categorized together, and the main clusters show a division between photographs including people and photographs portraying landscapes or buildings. This finding is consistent with Teeselink et al.’s (2000) results. The clusters show that emotional content (i.e. positive vs. negative photographs) was one of the bases for similarity evaluations. The affective responses to images and the atmosphere of the image itself have been deemed important in the selection process of journalistic photographs (Kobré, 2000; Markkula & Sormunen, 1998) and based on these categorization results they also seem to be a natural way for people to organize photographs. The influence of the photographs’ emotional tone on categorization has not been discussed much in previous studies, although the image categories reported by Rorissa and Hastings (2004) included several affective themes. Other studies on free categorization (Vailaya et al., 1998; Teeselink et al., 2000) seem to concentrate purely on the generic semantic level of image content (justified because of their interest in, for example, automatic image classification), whereas several of the journalistic photographs used in this study contained affective subject matter. The results seem to indicate that the genre of the test photographs (in this case photojournalistic) does influence the categorization. Therefore the results of earlier studies on image categorization cannot entirely be generalized to apply to journalistic photographs. When investigating how people categorize images, affective and symbolic aspects of image content should be taken into account.

The nature of the earlier description task did not have a statistically significant influence on the categorization of the photographs. Also the number of categories was similar between the two groups. This suggests that the subjects evaluated the photograph content similarly but the different tasks led them to describe the content differently. This has repercussions, among others, for the design of interfaces for

image archival. Keyword annotation, where a word limit may be imposed and individual terms (instead of a narrative) are requested may hinder and narrow down natural image description.

Suggestions for improvement

The framework by Jörgensen (1998) used to classify the terms generated in the tasks was found to lack some important syntactic and semantic classes. It was thus decided to extend the framework. This is part of the evaluation of the suitability of the framework for describing the semantic content of photographs. Three new classes are suggested as extensions: VISUAL QUALITY (attributes sharpness and distortion), ANIMAL (attribute animal), and WEATHER (attribute weather). Some existing classes of the framework were also found incomplete for the purpose of exhaustively classifying the terms elicited in the tasks. Because of this, additions to four classes are proposed. The class names and attributes which are suggested to be added are listed in Table 5.

Table 5. Suggestions of new attributes into existing classes of Jörgensen (1998)

Class	Attributes	Examples
PEOPLE-RELATED ATTRIBUTES	named identity	Prince Charles
	Pose	standing, sitting, crouching
	facial expression	smiling
	Nationality	Indian
VISUAL ELEMENTS	Direction	downwards
DESCRIPTION	Size	large, small
	Dimension	low, high
	Quantity	several, little
CONTENT/STORY	general setting	desert
	specific setting	Sahara
	general event	accident
	specific event	September 11th

In the CONTENT/STORY class, attributes setting and event were replaced by the more specific general setting, specific setting, general event and specific event. All other additions were made on top of the existing framework. These additions cover 7.5 % (24.9% including new content/story attributes) of the elicited terms in the keywording task and 10.2% (17 % including content/story attributes) in the free description task. The class ART-HISTORICAL INFORMATION was left out because no attributes pertaining to that class were found in the tasks. A final data analysis was done using the modified framework. The percentage distributions of the classes and attributes in the tasks are shown in Table 6. The added classes and attributes are indicated with italics.

Table 6. Percentage distribution of classes and attributes according to the modified framework

Class / attribute	Key	Free
OBJECTS (P)	**23.0**	**26.3**
object	19.7	17.1
text	0.0	0.3
body part	1.7	5.6
clothing	1.5	3.4
PEOPLE (P)	**4.1**	**7.0**
people	4.1	7.0
ANIMAL (P)	***2.7***	***2.4***
animal	*2.7*	*2.4*
COLOR (P)	**3.0**	**6.2**
color	2.5	4.4
color value	0.5	1.8

Class / attribute	Key	Free
VISUAL ELEMENTS (P)	**6.0**	**3.8**
composition	0.4	0.6
focal point	0.1	0.0
motion	1.9	0.3
orientation	0.6	0.3
perspective	1.7	1.3
shape	0.6	0.6
texture	0.1	0.1
visual component	0.7	0.4
direction	*0.1*	*0.3*
VISUAL QUALITY (P)	***0.5***	***0.5***
sharpness	*0.4*	*0.5*
distortion	*0.1*	*0.0*

Class / attribute	Key	Free
LOCATION (P)	**0.3**	**10.1**
general	0.2	4.3
specific	0.1	5.7
DESCRIPTION (P)	**6.4**	**10.8**
description	5.6	4.9
number	0.4	3.4
size	*0.4*	*1.6*
dimension	*0.2*	*0.4*
quantity	*0.0*	*0.5*
PEOPLE-RELATED ATTRIBUTES (I)	**12.7**	**10.4**
relationship	1.3	0.5
social status	6.9	5.6
emotion	2.2	0.4
named identity	*1.2*	*0.7*
pose	*0.2*	*1.3*
facial expression	*0.7*	*0.4*
nationality	*0.5*	*1.6*
ABSTRACT CONCEPTS (I)	**10.8**	**1.5**
abstract	3.8	0.9
atmosphere	1.0	0.1
state	0.8	0.5
symbolic aspect	0.1	0.0
theme	5.2	0.0

Class / attribute	Key	Free
CONTENT/STORY (I)	**27.6**	**16.7**
activity	8.7	8.9
category	0.0	0.2
general event	*6.6*	*1.8*
specific event	*0.2*	*0.0*
general setting	*4.2*	*3.2*
specific setting	*6.4*	*1.8*
time aspect	2.1	0.8
WEATHER (I)	*1.0*	*0.5*
weather	*1.0*	*0.5*
EXTERNAL RELATION (R)	**0.4**	**0.3**
comparison	0.0	0.1
similarity	0.1	0.1
reference	0.3	0.1
VIEWER RESPONSE (R)	**0.9**	**3.5**
personal reaction	0.7	0.5
conjecture	0.2	2.2
drawing	0.0	0.0
uncertainty	0.0	0.8

Conclusions

The most prevalent photograph description level in both tasks was the interpretational level including general, specific and abstract semantic concepts. However, constrained keywording resulted in more terms depicting story, setting and theme than the free description task, which led the subjects to enumerate individual objects and describe their locations. Free description also resulted in narrative-type descriptions without prompting and sometimes included the subjects' conjectures and estimates. The nature of the earlier description task had no significant effect on the categorization of the photographs. The groups that had previously performed the keywording and the free description tasks appeared to interpret photograph content similarly. However, the two description tasks resulted in different attributes being used to describe photograph content. This suggests that the limitations imposed by image annotation (separate terms, limited number of terms) may truncate natural image descriptions.

Photograph categorization was mainly conducted based on content/story-related and abstract concepts. The main clusters show emotional content and the presence or absence of people being used as categorization criteria. Further studies should be conducted regarding affective content as basis for image categorization. Careful attention should be paid to the selection of test images. Most past research reviewed for this report included narrow image content matter, e.g. scenery images or extreme close-ups of people. The material in this study included photographs depicting emotions (including negative ones), and the categorization results reflected that in the form of top-level clusters. Image categorization seems to depend on semantic content of various levels and be influenced by the image genre. For the purpose of generalizing results from image indexing research the selection of test images is a key issue.

The results of this study have application potential in various areas. Knowledge on image categorization may be used in image retrieval applications. Predicted image categories can serve as input in image search and selection tools as well as image analysis for content-based image retrieval. Knowledge regarding unconstrained image descriptions is useful in the creation of image indexing models. Furthermore, the results may be useful in the design of image retrieval experiments in future studies.

Acknowledgements

The authors wish to acknowledge the support of The National Technology Agency of Finland for this research project. We would also like to thank Professor Pirkko Oittinen for her constructive comments.

REFERENCES

Eakins, J.P., Briggs, P. & Burford, B. (2004). Image Retrieval Interfaces: A User Perspective. *Lecture Notes in Computer Science, 3115*, 628-637.

Fidel, R. (1997). The image retrieval task: implications for the design and evaluation of image databases. *The New Review of Hypermedia and Multimedia, 3*, 181-199.

Greisdorf, H. & O'Connor, B. (2002). Modelling what users see when they look at images: a cognitive viewpoint. *Journal of Documentation, 58(1),* 6-29.

Hasler, D. & Süsstrunk, S.E. (2003). Measuring colourfulness in natural images. In B. Rogowitz & T. Pappas (EDS) *IS&T/SPIE Human Vision and Electronic Imaging VIII, SPIE vol. 5007*, 87-95. Santa Clara, CA.

Hollink, L., Schreiber, G., Wielinga, B. & Worring, M. (2004). Classification of User Image Descriptions. *International Journal of Human-Computer Studies, 61(5),* 601-626.

Jaimes, A. & Chang, S-F. (2000). A Conceptual Framework for Indexing Visual Information at Multiple Levels. In G. Beretta & R. Schettini (EDS) *IS&T/SPIE Internet Imaging I, SPIE vol. 3964,* 2-15. San Jose, CA.

Jörgensen, C. (1998). Attributes of Images in Describing Tasks*. Information Processing & Management, 34(2),* 161-174.

Jörgensen, C. (1999). Retrieving the Unretrievable: Art, Aesthetics, and Emotion in Image Retrieval Systems. In B. Rogowitz & T. Pappas (ED) *IS&T/SPIE Human Vision and Electronic Imaging IV*, *SPIE vol. 3644*, 348-355. San Jose, CA.

Kobré, K. (2000). *Photojournalism: The Professionals' Approach,* 4th ed. Boston, Focal Press. 384 p.

Lohse, G.L., Biolsi, K., Walker, N. & Rueter, H.H. (1994). A classification of visual representations. *Communications of the ACM, 37(12),* 36-49.

Lohse, J., Rueter, H., Biolsi, K. & Walker, N. (1990). Classifying visual knowledge representations: a foundation for visualization research. In A. Kaufman (ED) *IEEE Visualization '90*, 131-138. San Francisco, CA.

Markkula, M. & Sormunen, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval, 1(4)*, 259-285.

Markkula, M. & Sormunen, E. (1998). Searching for Photos - Journalists' Practices in Pictorial IR. In J. Eakins, D. Harper & J. Jose (EDS) The Challenge of Image Retrieval. Newcastle upon Tyne, UK.

Mojsilovic, A. & Rogowitz, B. (2001a). A Psychophysical approach to modeling image semantics. In B. Rogowitz & T. Pappas (EDS) *IST&T/SPIE Human Vision and Electronic Imaging VI, SPIE vol. 4299,* 470-477. San Jose, CA.

Mojsilovic, A. & Rogowitz, B. (2001b). Capturing Image Semantics with Low-level Descriptors. In (EDS) *IEEE International Conference on Image Processing, ICIP 2001, vol. 1*, 18-21. Thessaloniki, Greece.

Ornager, S. (1997). Image retrieval: Theoretical analysis and empirical user studies on accessing information in images. In C. Schwartz & M. Rorvig (EDS) *Proceedings of the 60th Annual Meeting of the American Society for Information Science, 34*, 202-211.

Rogowitz, B.E., Frese, T., Smith, J.R., Bouman, C.A. & Kalin, E. (1998). Perceptual image similarity experiments. In B. Rogowitz & T. Pappas (EDS) *IS&T/SPIE Human Vision and Electronic Imaging III*, *SPIE vol. 3299*, 576-590. San Jose, CA.

Rorissa, A. & Hastings, S.K. (2004). Free sorting of images: Attributes used for categorization. In L. Schamber & C. Barry (EDS) *Annual Meeting of the American Society for Information Science and Technology,* Providence, RI.

Sormunen, E., Markkula, M. & Järvelin, K. (1999). The Perceived Similarity of Photos - A Test-Collection Based Evaluation Framework for the Content-Based Photo Retrieval Algorithms. In: Draper S.W., Dunlop M.D., Ruthven I., van Rijsbergen C.J. (Eds.) *Mira 99: Evaluating interactive information retrieval*.

Teeselink, I.K., Blommaert, F. & de Ridder, H. (2000). Image Categorization*. Journal of Imaging Science and Technology, 44(6),* 552-559.

Vailaya, A., Jain, A. & Zhang, H.J. (1998). On Image Classification: City Images vs. Landscapes*. Pattern Recognition, 31(12),* 1921-1935.