A Conceptual Framework and Empirical Research for Classifying Visual Descriptors

Corinne Jörgensen[2]
Alejandro Jaimes[1]
Ana B. Benitez[1]
Shih-Fu Chang[1]

[1]**Columbia University**
[2]**University at Buffalo, State University of New York**

ABSTRACT

This paper presents exploratory research evaluating a conceptual structure for the description of visual content of images. The structure, which was developed from empirical research in several fields (e.g., Computer Science, Psychology, Information Studies, etc.), classifies visual attributes into a Pyramid containing four syntactic levels (type/technique, global distribution, local structure, composition), and six semantic levels (generic, specific, and abstract object and scene, respectively). Various experiments are presented, which address the Pyramid's ability to achieve several tasks: (1) classification of terms describing image attributes generated in a formal and an informal description task, (2) classification of terms that result from a structured approach to indexing, (3) guidance in the indexing process. Several descriptions, generated by naïve users and indexers are used in experiments that include two image collections: a random web sample, and a set of news images. To test descriptions generated in a structured setting, an Image Indexing Template (developed independently of this project by one of the authors over several years) was also used. The experiments performed suggest that the Pyramid is conceptually robust (i.e., can accommodate a full range of attributes) and that it can be used to organize visual content for retrieval, to guide the indexing process, and to classify descriptions obtained manually and automatically.

Introduction

Technologies for the digitization of analog image collections and the digital production of new images are combining to create vast digital image libraries. The demand for networked access and sharing of these images has created a need for new and more efficient techniques to index their content. Access to these image collections through traditional indexing techniques is problematic for a number of reasons, as existing indexing systems have been created for the needs of limited audiences or targeted for particular types of collections.[1] Newer content-

[1] Two of the more widely used in the United States are the *Thesaurus for Graphic Materials I* (TGM I, Library of Congress, 2000) and the *Art and Architecture Thesaurus* (AAT, Getty Research Institute, 2000). The TGM, although created as a tool for indexing broad, general image collections, was developed to meet the needs of the Prints and Photographs Division and is more appropriate to collections of a historical nature. The AAT is a precise indexing tool which meets the needs of specialized communities of researchers and provides access at a high level of specificity. A review of image indexing systems is provided in Jörgensen, 2000 and Rasmussen, 1997.

based[2] techniques have utility in retrieving subsets of specific visual attributes and may be useful as tools for segmenting large image collections, but currently address only a small portion of the complete range of image attributes of potential interest to users of digital image collections. More recently, there has been an interest among computer scientists in combining these techniques with other more traditional indexing techniques, such as the use of broad ontologies (Chang et al., 1997).

Other recent initiatives focus on metadata structures for image information. Two sets of proposed attributes have been widely disseminated: The *Dublin Core* (Dublin Core, 2000) and the *VRA Core* (VRA Data Standards Committee, 2000). The Art Information Task Force has also proposed the *Categories for the Description of Works of Art* (Getty Information Institute, 1999). Another group addressing metadata standards is the Motion Pictures Experts Group (MPEG). Their latest initiative, known as MPEG-7 [3] is developing standards for the description of multimedia content, which may include any combination of still images, moving images, audio, and text. The research reported herein was completed as part of the MPEG-7 initiative.

The goal of the current research was to test a particular structured representation for the classification of a wide range of image attributes of interest in a retrieval context. The research evaluated and compared the structure in relation to image descriptions resulting from two very different methodologies, conceptual modeling (or a "top-down") approach, and a data-driven (or "bottom-up") approach,

Related Research

Work on issues related to images has been performed by researchers in many different areas. Selective examples follow. Studies in *art* have focused on interpretation and perception (Arnheim, 1984; Buswell, 1935) aesthetics and formal analysis (Barnet, 1997), visual communication (Dondis, 1973), and levels of meaning (Panofsky, 1962). Studies in *cognitive psychology* have dealt with issues such as perception (Hendee & Wells, 1997); visual similarity (Tversky, 1977), mental categories (i.e., concepts) (Armstrong et al., 1983), distinctions between perceptual and conceptual category structure (Burns, 1992; Harnad, 1987), and internal category structure (i.e., levels of categorization) (Morris & Murphy, 1990; Rosch & Mervis, 1975). In the field of *library and information science* (LIS), work has been performed on analyzing the subject of an image (Shatford Layne, 1986; Turner, 1994), indexing (Fidel et al., 1994; Shatford Layne, 1994), the range of attributes that are used to describe images

[2] The phrase "content-based retrieval" comes from the Electrical Engineering/Computer Science community. "Content-based" refers to retrieval of visual information based on what is depicted (color, texture, objects, etc.).
[3] The goal of MPEG-7 is to specify a standard set of descriptors (Ds) for content representation for multimedia information search (indexing and retrieval or "pull" applications), selection and filtering ("push" applications), and management and processing. See http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm for more information.

(Jörgensen, 1998), classification (Lohse et al., 1994), query analysis (Enser, 1993) and indexing schemes (Davis, 1997; Jörgensen, 1996b), among others.

The Conceptual Model

The conceptual model (the "Pyramid") was developed drawing upon this previous body of research. The structure of the Pyramid is briefly outlined below; examples, justification, and further details for each level can be found in Jaimes & Chang, 2000.

The Pyramid (Figure 1) contains ten levels: the first four refer to *syntax*, and the remaining six refer to *semantics*. In addition, levels one to four are directly related to *percept*, and levels five through ten to *visual concept.* While some of these divisions may not be strict, they should be considered because they have a direct impact in understanding *what* the user is searching for and *how* s/he tries to find it. The levels also emphasize the limitations of different indexing techniques (manual and automatic) in terms of the knowledge required.

The research on visual information that has been carried out in different fields shows that indexing such information can be particularly complex for several reasons. First, visual content carries information at many different levels (e.g., *syntactic*: the colors in the image; *semantic*: the objects in the image). Second, descriptions of visual content can be highly subjective, varying both across *indexers* and *users*, and for a single user over time. Such descriptions depend on other factors that include, for example, the indexer's knowledge (e.g., art historian), purpose of the database (e.g., education), database content (e.g., fine art images; commercial images), and the task of the user (find a specific image or a "meaningful" image).

Three main factors entered into the construction of the proposed model: (1) range of descriptions; (2) related research in various fields; and (3) generality. In considering the range of descriptions, the focus was only on *visual content* (i.e., any descriptors stimulated by the visual content of the image or video in question; the price of a painting would not be part of *visual content*). Since such content can be described in terms of *syntax* or *semantics* the structure contains a division that groups descriptors based on those two categories. This division is of paramount importance, particularly when we observe research in different fields. Most of the work on content-based retrieval, for example, supports syntactic-level indexing, while work in art places strong emphasis on composition (i.e., relationships between elements) both at the syntactic (i.e.., how colors, lines, and patterns are laid out) and semantic levels (i.e., the meaning of objects and their interactions). Most of the work in information science, on the other hand, focuses on semantics. The structure was developed based on research and existing systems in different fields.

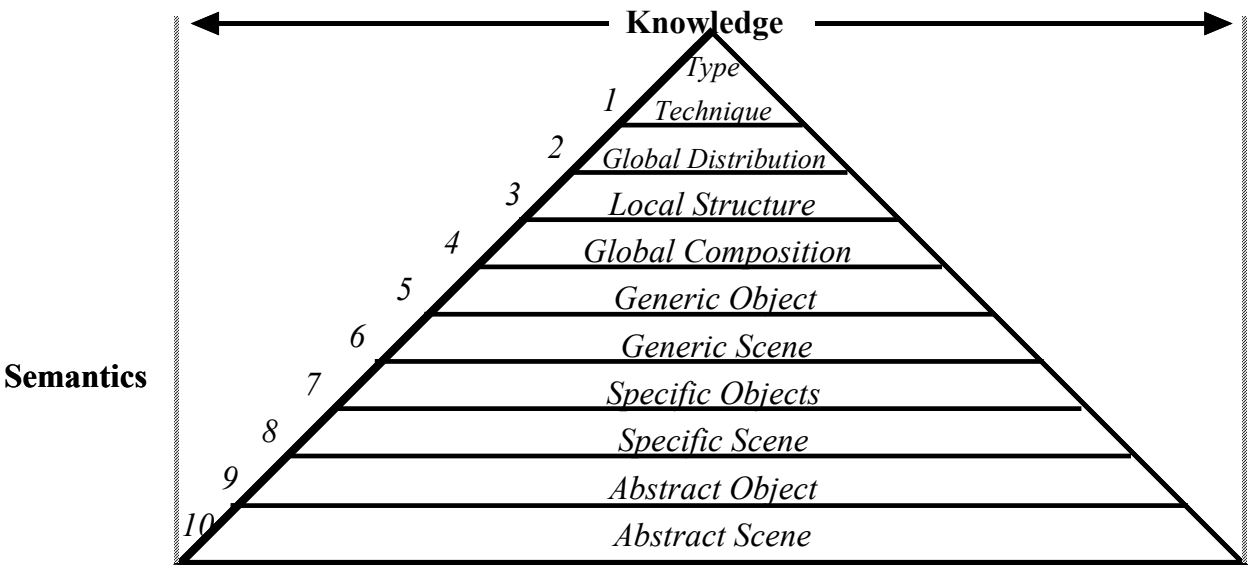

FIG. 1. A "Pyramid" structure for classifying visual content attributes

Syntactic Levels

Type/Technique is the most basic level and describes the general type of image or video sequence or the technique used to produce it (e.g., black and white, color). *Global Distribution*, on the other hand, classifies images or video sequences based on their global content and is measured in terms of low-level perceptual features such as spectral sensitivity (color), and frequency sensitivity (texture). Individual components of the content are not processed at this level (i.e., no "form" is given to these distributions in the sense that the measures are taken globally). Traditionally, global color histogram (Jaimes & Chang, 2000) has been used to index Global Distribution.

In contrast to Global Distribution, the *Local Structure* level is concerned with the extraction and characterization of the individual components of the image. At the most basic level, those components result from low-level processing and include elements such as the Dot, Line, Tone, Color, and Texture. At the *Global Composition* level, the focus is on the specific arrangement or spatial layout of elements in the image. In other words, we analyze the image as a whole, but only use the basic elements described in the previous level (e.g. line and circle) for the analysis. Traditional analysis in art describes composition concepts such as balance, symmetry, center of interest (center of attention or focus), leading line, and viewing angle.

Semantic Levels

At the syntactic levels, no world knowledge is required to perform indexing, so automatic techniques can be used to extract relevant information. Humans, however, mainly use higher level attributes to describe, classify and search for visual material (Jörgensen, 1998). Objects are of particular interest, and they can be placed in categories at different levels- an apple can be classified as a Macintosh apple, as an apple, or as a fruit. When referring to *Generic Objects*[4], we are interested in the basic level categories: the most general level of object description, which can be recognized with everyday knowledge (e.g., apple, man, chair). Similarly, the *Generic Scene* level refers to scenes at their most general level of description. Examples of scene classes include city, landscape, indoor, outdoor, still life, and portrait. It is not necessary to know a specific street or building name in order to determine that it is a city scene, nor is it necessary to know the name of an individual to know that the individual is a woman.

In contrast to Generic Object, the *Specific Object* level refers to identified and named objects. Specific knowledge of the objects in the image or the video sequence is required, and such knowledge is usually objective since it relies on known facts. Examples include individual persons or objects (e.g., Bill Clinton, Eiffel Tower). Similarly, *Specific Scene* refers to attributes that carry specific knowledge about the scene (e.g., Paris, Times Square, Central Park, etc.).

Attributes at the *Abstract Object* level, on the other hand, refer to specialized or interpretative knowledge about what the objects represent. This indexing level is the most difficult one in the sense that it is completely subjective, and assessments between different users may vary greatly. For example, a woman in a picture may represent anger to one observer and pensiveness to another. Existing indexing systems (for example, some descriptors in the *TGM I*, Library of Congress, 2000) show the importance and applicability of attributes at this level. Similarly, the *Abstract Scene* level refers to what the image as a whole represents and may be very subjective. Jörgensen (1995b) showed that users sometimes describe images in affective (e.g. emotion) or abstract (e.g. atmosphere, theme) terms. Examples at the abstract scene level include sadness, happiness, power, heaven, and paradise.

The shape of the structure itself reflects, in some sense, the amount of knowledge required to perform indexing, and the different "types" of information identified by different researchers. Analyzing Figure 1 from top to bottom, it is apparent that at the lower levels of the pyramid, more knowledge and information is required. This

[4] Although it is possible to mathematically define specific, generic, and abstract, such definitions would be difficult to apply. Instead, we present intuitive definitions with examples.

assumption, however, may have some exceptions. For example, an average observer may not be able to determine the technique that was used to produce a painting, but an expert in art would be better able to determine exactly what was used. Indexing in this particular case would require more knowledge at the *type/technique* level than at the *generic objects* level (since special knowledge about art techniques would be needed). In most cases, however, the knowledge required for indexing will increase in our structure from top to bottom: more knowledge is necessary to recognize a specific scene (e.g., Central Park in New York City) than to determine the generic scene level (e.g., park); something in the image must indicate that the park is actually Central Park. Similarly, automatically determining the type of content (e.g., color or black and white image), for example, is less expensive than recognizing generic objects (e.g., face detection) and recognizing specific objects (e.g., face recognition).

The model does not depend on any particular database model, visual information type, user type, or purpose; therefore, not every level of the structure would be necessary in every case. Different projects demonstrate that while many researchers are addressing various levels of the "puzzle," there has not heretofore been a conceptual structure which can unite these diverse efforts and demonstrate the relationships among the pieces of this puzzle.

While the focus of the research reported herein is to demonstrate the applicability of the structure to a wide range of attributes, the authors contend that such a structure can also facilitate searching by disambiguating among terms that could appear at several different levels of the structure. Additionally, it makes explicit the generic and specific levels of description which are not well accommodated in some current systems.

Research Questions

The current research tests one possible structure as the basis for attribute classification, the conceptual "Pyramid" discussed above, and asks:

Can the Pyramid classify a full range of visual content descriptors for an image (both syntactic and semantic) in at least one level of its structure?

More specifically, the research question was operationalized by the following four questions:

1. How well can the Pyramid classify terms describing image attributes generated by naïve users, both in a spontaneous informal description and a formal description for retrieval mode?
2. How well can the Pyramid classify the attributes that result from a structured approach to indexing?

3. How well can the Pyramid guide the process of generating and assigning a full range of indexing attributes to images?
4. How well can the Pyramid classify varying conceptual levels of information (specific, generic, and abstract)?

Methodology

For the current research, the definition of attribute follows that in Jörgensen (1995): "any kind of feature, component, or property of a stimulus that can be represented by an information processing system." There are no assumptions about the format of these representations or whether the attributes correspond to innate detectors. In this research, "information processing system" includes systems for organizing image information and image retrieval systems. An image attribute is therefore a feature, component, or property of a stimulus (an image) that can be represented by an information processing system and includes other cognitive, affective, or interpretive responses to the image. These attributes make up what we refer to as the "Visual Structure" of the image, as they are initially stimulated by visual perception. In the structures we present, however, all of the attributes are derived directly from the image – the price of a painting, for example, would not be included in the structure we describe.

To address the research questions, several experiments were designed using the Pyramid (Figure 1) and an Indexing Template (Appendix A) to map existing image descriptions, to generate image descriptions, and to classify image descriptions.

The Image Indexing Template

In contrast to the conceptual modeling described above, another tool, the *Indexing Template*, was developed using a data-driven approach (Jörgensen, 1996) and modified for a larger project to meet the indexing requirements of a large sample of images forming a prototype imagebase (Jörgensen & Srihari, 1999). The first step in our research was to test the comparability of attributes produced by these two different approaches (conceptual modeling and data-driven) by manually mapping attribute types from the Indexing Template to the Pyramid (see Appendix A). Some sections mapped easily on a one-to-one basis (e.g. Visual Elements attributes to Local Structure), while in other cases attributes from several different facets in the Indexing Template were needed to populate certain Pyramid levels. Harmonizing the two schemes was primarily a matter of resolving different levels of analyses. For instance, the Indexing Template distinguishes between "Living Things" and "Objects," while the Pyramid includes both of these as "Objects." The Pyramid describes the generic and specific naming levels explicitly, while the Indexing Template left this component to standard online bibliographic system implementation using "descriptors" (general keywords) and identifiers (proper nouns).

The work brought two structures for image description together (the deductively developed Pyramid and the inductively developed Indexing Template) and compared their results. This mapping demonstrated that the Pyramid does accommodate the wide variety of attributes described in empirical research.

Images

We used two groups of images for the experiments. The first (Set A) consisted of a subset of a random sample of images (approximately 700 images) taken from the World Wide Web by an automated process. It included a variety of images produced for a wide range of purposes, ranging from simple icons to complex images (photographs, cartoons, illustrations, graphics, animations, and so on) produced by different institutions and individuals. The second smaller set of twelve color photographs of current news events in different parts of the world (Set B) was obtained randomly from an Internet news newsgroup. Only the images in set B included textual description (a date and a short descriptive caption).

Image Descriptions

The image descriptions were primarily produced by two groups of individuals: *naïve users* (i.e., no prior training in indexing of visual information) and *indexers* (i.e., trained in indexing visual information). The *naïve users* were all beginning M.L.S. students[5] from a variety of backgrounds who had no previous experience with image indexing or retrieval nor with more general indexing and classification procedures. Using the methodology reported in Jörgensen (1998), forty-one naïve users viewed projected images and each one produced both spontaneous (informal) and retrieval-oriented (formal) image descriptions (i.e., each image was described by the same individual using both methods) for a subset of four images from Set A (web images). The informal descriptions were lists of words or short phrases, whereas the formal descriptions were more complete sentences or long phrases describing an image to be retrieved. These were used as input to several of the experiments. Descriptions generated by six individuals (each described four images using two methods, for a total of 48 image descriptions) were selected randomly providing approximately 500 terms to be mapped with an average of 10.4 terms per image. In addition to these, two authors generated spontaneous descriptions for the twelve news images (Set B).

The *indexers* were twenty-two students that had received training in the theory and practice of indexing images in Dr. Jörgensen's "Indexing and Surrogation"class. These students produced structured indexing records using the Indexing Template described above for the two sets of images (Sets A and B). The indexers used a controlled vocabulary which they developed for the project using selected terms from existing thesauri such as the *AAT* and

the *TGM I*; these terms were selected based on their appropriateness for *visual* description. Indexers were able to add free-text terms when no appropriate thesaurus term was present and after class discussion these terms were added to the thesaurus. Images were indexed online through a web-browser interface with multiple frames displaying the image to be indexed, the Indexing Template, and the controlled vocabulary. Indexers were presented with a random selection from the image collection and were not allowed to "choose" which images they indexed; they indexed a total of approximately 550 images.

Indexers were instructed to index the "visually relevant" and "visually striking" content of the images, rather than attempting to fill each slot of the Indexing Template. Indexers spent an average of ten to twenty minutes indexing each image. Inter-indexer consistency was developed through discussion of the indexing process and comparison of practice indexing records in class, and was assisted by the controlled vocabulary.

In addition to these indexing records, other image descriptions which were used were caption information for the twelve news images (also a form of "indexing"). Additionally, image descriptions were generated by the authors using the Pyramid structure; these are discussed further below.

Experiments

Using the image descriptions from *naïve users* and *indexers* as data, several experiments addressing the research questions were performed. Each experiment and its results are presented in expanded sections below. As the data sets are small for each experiment, this work can be considered preliminary and exploratory. However, the results as aggregated across the experiments suggest that further work in developing a conceptual approach to image indexing (such as that instantiated in the Pyramid) would be beneficial.

Classifying Spontaneous Descriptions

Experiment I addressed Research Question I by determining how well the Pyramid classifies terms from naïve participants' descriptions (described above). The researchers mapped the six participants' spontaneous descriptions (242 mapped terms) and retrieval-oriented descriptions (241 mapped terms) for the same four images from image Set A to the ten levels of the Pyramid.

These descriptions contained attributes at all levels of the Pyramid (Experiment I, Table 1); almost all levels of the Pyramid were used by all participants. The exceptions to this occur at the lowest syntactic levels for *spontaneous* descriptions (Global Distribution and Global Composition); this is in agreement with previous analysis of spontaneous descriptions demonstrating that lower-level descriptors are less-used in spontaneous

[5] Department of Library and Information Studies at the University at Buffalo.

image descriptions (Jörgensen, 1999). However, when naïve participants were asked to describe images more formally in a *retrieval* context, we see that attributes then occur at these lower syntactic levels as well, as with descriptions generated by *indexers*. The results indicate that the Pyramid is able to accommodate a variety of attributes as described by naïve users in both a spontaneous describing task and in a more formal, retrieval-oriented task.[6]

TABLE 1. Mapping of image attributes to the pyramid levels as generated by different methods: experiment I (spontaneous, and retrieval-oriented); II (author, and caption); and III (indexing).

Experiment	I	I	II	II	III
PYRAMID LEVEL	Spontaneous	Retrieval	Author	Caption	Indexing
I. Type/Technique	X	X			X
II. Global		X	X		X
III. Local Structure	X	X	X		X
IV. Global		X	X		X
V. Generic Objects	X	X	X	X	X
VI. Specific Objects	X	X	X	X	X
VII. Abstract Objects	X	X	X	X	X
VIII. Generic Scene	X	X	X	X	X
IX. Specific Scene	X	X	X	X	X
X. Abstract Scene	X	X	X	X	X

Experiment II used data for image Set B (news images). Two of the authors (one from LIS with considerable image indexing experience) and the other from a Computer Science/Electrical Engineering background (no previous image indexing experience) *spontaneously* described (without using the pyramid or template) a set of ten randomly selected images (five unique images each). There were no major differences in the overall range and types of attributes generated between the two authors, with both describing objects, events, people, as well as emotions and themes (see Figure 2 below for sample descriptions).

Author 1 terms	**Author 2 terms**
Airport	Interview
Greek policeman	Outdoors
Guns	Three men
Outdoor	Reporters
Duty	Grim expressions
Terrorism	Microphones thrust in face
Protection	
Death	

FIG. 2. Sample spontaneous descriptions of news images.

[6] It should be noted that the terms from the spontaneous and retrieval-oriented terms were not necessarily the same; different terms at different levels of the Pyramid were used in the spontaneous and structured descriptions. Comparative analysis of the two describing tasks data is interesting but not directly relevant to the questions presented here, which focus on whether a *range* of attributes for each task is accommodated at all ten levels of the Pyramid.

Additional data from the captions for the same images was used in the second part of the experiment. The caption for the image described in FIG. 2 is as follows:

US special envoy to the former Yugoslavia Robert Gelbard (R) talks to the press as the leader of Kosovo Albanians, Ibrahim Rugova (L w/ glasses), listens following their talks in Pristina, Yugoslavia March 10. Gelbard came to Pristina March 10 to seek peace between Serbian police and ethnic Albanians.

The spontaneous author descriptions were mapped to the pyramid (by a different author who had not generated the description); terms from data were present in all levels except that of Type/Technique. Additionally, terms from the image captions (of the same 10 images) were also mapped. In contrast, the captions mapping lacked the four syntactic visual information levels (which is to be expected), while more of this information appeared in the authors' descriptions (Table 1, Experiment II). At the higher semantic levels, information appeared on all levels of the Pyramid across both the authors' descriptions and the captions; however, with the caption information there were quite a few more terms which belong in the Specific Object and Specific Event levels (again, a not unexpected result). Specific Object and Specific Event level information depends upon the observer's prior familiarity with a particular object/person/place depicted and may quite often be missing in a spontaneous description. Interestingly, the mapping results for the captions are more closely related to the "spontaneous" descriptions than to descriptions in a retrieval context.

Overall, the experiments with spontaneous and caption descriptions show support for the ten-level conceptual structure as instantiated by the Pyramid. The Pyramid accommodated a full range of attributes gathered in experimental work and all attributes were classified at some level of the Pyramid.

Classifying Structured Descriptions

Experiment III addressed how well the Pyramid accommodates the attributes that result from a structured approach to indexing. Structured indexing implies the use of indexing tools such as some type of metadata structure and a controlled vocabulary, as well as training in the process of indexing. For this experiment, the Indexing Template (Appendix A) and structured image descriptions[7] were used (gathered by the process described above). Thirty-three randomly-selected indexing records for 33 unique images generated by the student indexers for images from Set A were mapped to the Pyramid by the authors (approximately 1,050 terms). Each of the authors performed mapping for a different set of images.

[7] Although the indexers were instructed to index only "salient" aspects of each image, a large number of attributes were generated in each case. There seemed to be an overall tendency to "over-index" the image using the template. Additionally, students' indexing was being graded, prompting them to be very thorough. This, however, produced in-depth indexing for each image.

IMAGE TERM	PYRAMID LEVEL
painting	Type/Technique
oil	Type/Technique
cracked	Global Distribution
red, white	Local Structure
background	Local Structure
rectangle	Local Structure
center	Local Structure
eye level	Global Composition
flag	Generic Object
historical landscape	Generic Scene
patriotism	Abstract Object
pride	Abstract Scene

FIG. 3. Sample image indexing record terms from the Indexing Template mapped to the Pyramid levels.

Results demonstrated that attributes from the Indexing Template were mapped at all levels of the Pyramid, and each indexer used attributes from several levels of the Pyramid (FIG. 3). However, only *one* term occurs at the Specific Scene level across all the indexers, which is to be expected as there was no descriptive text, which would contain such information, attached to these images.[8]

It is interesting to compare the mapping of the indexers' descriptions with the mapping of the spontaneous descriptions of the untrained users (Table 1). Of note is the consistency in descriptor levels in the Pyramid between the naïve respondents' descriptions generated when respondents were asked to describe the images within a retrieval context and the indexers' descriptions. This suggests that when respondents were asked to describe the images as if they wanted to find them (retrieval mode), their descriptions become more "formal" or structured, as was shown previously in Jörgensen (1996). This also suggests that the needs of image searchers may be more closely suited by a structured method (e.g., the Pyramid being tested) to accommodate image descriptions.

This preliminary analysis demonstrates good correspondence between the index terms and the levels of the Pyramid. This suggests that the Pyramid's conceptual structure can accommodate a wide variety of attributes produced as a result of a structured indexing process such as that using the Indexing Template. While mapping of the descriptors to the levels of the Pyramid was straightforward in most cases, some further guidelines would be beneficial for performing such mappings. The Pyramid is designed both to describe an entire image *and* to be used recursively to describe specific areas of an image. In the mapping process, the capability of the Pyramid to be used recursively resolved some of the earlier questions encountered in the mapping. These results suggest that the Pyramid itself may be a candidate for guiding the image indexing process.

[8] Whereas, in the case of the naïve users' descriptions, "accuracy" was not such a concern and they did in fact supply specific terms with no concrete knowledge of the correctness of these terms.

Generation of Indexing Based On The Pyramid

Experiment IV tested how well the Pyramid guides the process of generating and assigning a full range of indexing attributes to images. In this experiment, two image samples were indexed using the Pyramid to suggest attributes which should be indexed.

For the first part of the experiment, two of the authors (those without image indexing experience) indexed a subset (thirty-one images across both authors) of image Set A (web images), producing 287 image index terms. In contrast to the indexing performed by the student indexers, no controlled vocabulary was used for this indexing. The indexing of this set of images was performed by the authors on randomly-selected *unseen* images (not used by the same author in previous mapping work or seen in any other context).

Sample image descriptions for this work using the Pyramid as a guideline are depicted in Figure 4. The major conclusion from this experiment is that descriptions for each level of the Pyramid were easily generated.[9] It should be noted that the descriptions generated here are shorter than the descriptions generated by student indexers using the Indexing Template (9.3 for the authors versus 10.4 for the student indexers), perhaps as a result of a lack of a controlled vocabulary (see also Note 7). However, the goal here was not to demonstrate the completeness of indexing done using the Pyramid but to demonstrate that the levels of the Pyramid can suggest adequate conceptual guidance for the process of indexing and that all Pyramid levels are relevant to the visual content of an image. Further detailed analysis of this data should demonstrate whether a full range of attributes, as represented in the Indexing Template, is present.

Pyramid Level	Image 1 terms	Image 2 terms
Type/Technique	color photograph	color photograph
Global Distribution	white, brown	clear
Local Structure	curves	curves, lines
Global Composition	centered, eye level, close-up	leading line
Generic Object	person, man, head, neck, shirt	ducks, lake, mountain, bridge, vegetation
Generic Scene	portrait, indoors	outdoor, daytime, landscape
Specific Object		
Specific Scene		
Abstract Object	Efficiency	family
Abstract Scene		vacation dream

FIG. 4 - Sample image descriptions generated using the Pyramid as a guideline

The second part of Experiment IV followed the procedures for the web image indexing using the Pyramid. Two authors (one of whom also did the web image indexing) each indexed five images from Set B (news images), using the Pyramid again to guide the indexing (135 terms or 13.5 terms per image). Results of the mapping were

[9] Although the examples do not contain Specific Object and Specific Scene, these were populated as well based upon the authors' general knowledge.

identical to the mapping of spontaneous descriptions for the previous Set A, with information lacking only in the Specific Object and Specific Scene Levels, for similar reasons. When captions are included, these levels are populated as well (e.g., Mirage 2000 Jet Fighter; Aden, Yemen). The Pyramid's variety of levels suggested a wider variety of information is assigned to these images than happens with spontaneous description (similar to results reported in Jörgensen, 1996), increasing the images' utility and their access points for retrieval. Therefore, the Pyramid is capable of generating attributes in the same areas as covered by the image Indexing Template.

Levels of Indexing

The fourth research question concerns how well the Pyramid structure can accommodate varying levels of information. The results from the news image indexing using the Pyramid are most instructive for this question, as the web image sample had little specific descriptive information associated with the images. The generic and specific levels of data are handled well by this conceptual structure, although we did find important open issues. One of the significant questions was the level at which a description should be placed. The choice between generic and specific, for example, was sometimes difficult since annotations at these levels may depend on the specific application or context (e.g., generic object: cat; specific object: Siamese cat, or generic object: Siamese cat; specific object: Felix). The distinction between object and scene, at the abstract level, also proved to be challenging for the indexers in some cases, since the same abstract term could apply to a single object or the entire scene (e.g., flag or scene representing patriotism). Although it is perfectly valid to have the same term at two different levels, we found that indexers were sometimes confused. As for the type of terms to use, the Indexing Template seemed to provide a more comfortable fit in some cases since it elicits terms in more detailed categories (e.g., abstract theme, symbol, and emotion at the abstract levels). For example, the Pyramid does not make a distinction between object and event, which is made by the template. Lastly, we found that syntactic level descriptions (e.g., global distribution) were easier to generate for some images than others (e.g., a texture image from the web- global color blue vs. a news image), although the pyramid usefully accommodates automatic techniques at all of the syntactic levels (Jaimes & Chang 2000).

The fourth research question concerns how well the Pyramid structure can accommodate varying levels of information (specific, generic, and abstract). The results from the news image indexing using the Pyramid are most instructive for this question, as the web image sample had little specific descriptive information associated with the images. The generic and specific levels of data are handled well by this conceptual structure. Abstract qualities were slightly more problematic for indexers to distinguish and in some cases these were not felt to be particularly intuitive. The Pyramid structure defines an Abstract Object as a localized entity with abstract meaning

(e.g. flag = patriotism); if the abstraction cannot be localized it becomes an Abstract Scene (e.g. democracy). The Indexing Template seemed to provide a more comfortable "fit" for some of these more abstract terms. For instance, term such as "democracy," "patriotism," and "respect," are perhaps more easily characterized by the more finely distinguished theme, symbol, and emotion of the Indexing Template than abstract object or scene. It may be that at the "abstract" level the object/scene distinction is less useful than a finer-grained analysis, or perhaps than a unitary approach (a *single* "abstract" level). Additionally, the Pyramid does not make a distinction between object and event, which is a natural part of descriptive language; this distinction is usefully made by the template. Both of these are open issues and bear further consideration and testing.

Discussion and Conclusions

This paper has presented preliminary results from exploratory research evaluating a conceptual structure, the "Pyramid," for the description of visual content of images. A variety of techniques were used to evaluate the conceptual structure. While the research demonstrated that the distribution of attributes among the levels of the Pyramid varies depending upon who generated them (indexers, researchers, naïve participants) and upon the task (describing, indexing, retrieval), the researchers found no instances where an attribute could not be accommodated by a level of the Pyramid. In addition, the Pyramid provides guidance to indexers by making explicit specific, generic, and abstract levels of description. Especially useful is the recursive nature of the Pyramid, which permits associations among objects and attributes (e.g., can be applied to a scene, object, section of an image, etc.).

The limitations of the research are the small number of images used, the use of student indexers, and the use of the researchers themselves to produce some of the data. However, the limited number of images still produced a large number of terms which were mapped in the experiment and were more than adequate to demonstrate that the Pyramid levels accommodate a wide range of terms. The student indexers produced high-quality indexing records, and the researcher data did not differ from data gathered by other methods. Additional data from professional indexers would certainly further substantiate the results.

As these various experiments produced consistent and positive results, the authors feel that the Pyramid is a robust conceptualization of visual image content and encourage further work both developing and using this structure for image representation and retrieval. We suggest that the results support the use of the Pyramid both as a method of organizing visual content information for retrieval and as a method for stimulating additional

attributes which could be added to existing records. Additionally, the Pyramid's ten-level structure could be used to represent attributes generated by both manual and automatic techniques.

Current and Future Research

Our current work is focusing on testing the Pyramid with audio and video data, and upon the disambiguation of image attributes during retrieval using the Pyramid. For instance, we have shown good results with the Pyramid in distinguishing among images with a "blue ball" as opposed to an image with an overall blue color.

Future research includes testing the Pyramid more widely, using additional material and more experienced indexers to generate descriptors using it, as well as exploring indexer training in using the Pyramid. Other important work should focus on determining whether some combination of the two structures (Indexing Template and Pyramid) would be useful, and the circumstances under which each may be a more appropriate choice to guide indexing. The goal of the current project was not to test one against the other but rather to substantiate that the range of attributes addressed by the Pyramid is adequate. The experimental work pointed up some differences, as discussed earlier, between the deductively developed Pyramid and the inductively developed Indexing Template. Differences that could fruitfully be explored between the two structures concern the range of attributes produced by each, the differences among the attribute types generated and the levels populated, and the number of attributes produced by each, as well as the communities that would find these structures most useful. One very interesting question to investigate is whether the Pyramid can serve as an entry point in providing access to images within a specific domain. While we tested the range of attributes classifiable within a generalized domain (the web images), levels of the Pyramid may be populated differentially across different domains (news image captions). As image collections become even more diverse and accessible, refinements to target specific types of images, video, and audio will become even more important. The current research has produced data that would aid in exploring these questions as well.

APPENDIX A. IMAGE INDEXING TEMPLATE ATTRIBUTES MAPPED TO PYRAMID LEVELS

INDEXING TEMPLATE (GROUP & ATTRIBUTE TYPE)	INDEXING TEMPLATE (ATTRIBUTE)	EXPLANATION/ EXAMPLE	PYRAMID LEVEL
"EXTERNAL" INFO.			
	> Image ID		NA
	> Creator/Author		NA
	> Title		NA
	> Publisher		NA
	> Date		NA
	> Image Type	color, X-ray, graphics,	TYPE/TECHNIQUE
	> Access Conditions		NA
	> Technical	resolution, file size, etc.	NA
"INFERRED" INFO.			
> "Environment"	>> When - time *in image*	Middle Ages, summer	GENERIC SCENE
	>> Where - General	city, rural, indoor, office	GENERIC SCENE
	>> Where - Specific	Paris, Chrysler Building	SPECIFIC SCENE
> Subject/Topic		overall subject/theme:	ABSTRACT OBJECT/SCENE
> Medium		oil, watercolor, digital	TYPE/TECHNIQUE
> Symbolism		Garden of Eden, afterlife	ABSTRACT OBJECT/SCENE
> "Why"	>> Emotions/Mental	sadness, laughter	ABSTRACT OBJECT/SCENE
	>> Relationships	brothers, romance	ABSTRACT OBJECT/SCENE
> "Miscellaneous"	>> Point of	bird's-eye, close-up	GLOBAL COMPOSITION
	>> Style	abstract, realism, etc.	ABSTRACT SCENE
	>> Genre	landscape, portrait	GENERIC SCENE
	>> Atmosphere/overall	gloomy, mysterious	ABSTRACT OBJECT/SCENE
VISUAL ELEMENTS			
> Color	>> Color	Red, blue	GLOBAL DIST/LOCAL STRC.
	>> Color Quality	dark, bright	GLOBAL DIST/LOCAL STRC.
	>> Placement	center, overall,	LOCAL STRUCTURE
> Shape	(>> Placement)	square, elongated,	GLOBAL DIST/LOCAL STRC.
> Texture	(>> Placement)	smooth, shiny, fuzzy	GLOBAL DIST/LOCAL STRC.
LITERAL OBJECTS			
> Category - General		What group: tool	GENERIC/SPECIFIC
> Type - Specific	(>> Placement)	What it is - hammer	GENERIC/SPECIFIC
	>> Shape		
	>> Texture		
	>> Size		
	>> Number		
	>> Color		
LIVING THINGS			
	> Type	human or what animal	GENERIC OBJECT
	(>> Placement)		
	>> Size	large, very small	
	>> Gender	male, female,	SPECIFIC OBJECTS
	>> Age		SPECIFIC OBJECTS
	>> Number		
	>> Pose	seated, standing, lying	GENERIC/SPECIFIC SCENE
	>> Name	Ghandi	SPECIFIC OBJECT/SCENE
	>> Physical Action/Event	running, talking, football	GENERIC/SPECIFIC SCENE
	>> Status	occupation, social status	ABSTRACT OBJECT/SCENE
COLLATERAL INFO.	> Caption		
	> Related Text		
	> Voice Annotations		

Armstrong, S.L, Gleitman, L.R., & Gleitman, H. (1983). What some concepts might not be. Cognition 13, 263-308.

Arnheim, R. (1984). Art and visual perception: A psychology of the creative eye. Berkeley, CA: University of California Press.

Barnet, S. (1997). A short guide to writing about art. 5th Edition. New York: Longman.

Burns, B. (1992). Perceived similarity in perceptual and conceptual development: the influence of category information on perceptual organization. In B. Burns (Ed.), Percepts, concepts and categories: the representation and processing of information (pp. 175-228). New York: Elsevier Academic Publishers.

Buswell, G.T. (1935).How people look at pictures: a study of the psychology of perception in art. Chicago, IL: University of Chicago Press.

Chang, S-F., Smith, J.R.,. Beigi, M., & Benitez, A.B. (1997). Visual information retrieval from large distributed on-line repositories. Communications of the ACM 40(12), 63-71.

Davis, E. T. (1997). A Prototype item-level index to the civil war photographic collection of the Ohio Historical Society. Master of Library Science thesis, Kent State University, August, 1997.

Dondis, D.A. (1973). A primer of visual literacy. Cambridge, MA: MIT Press.

Dublin Core v. 1.1 (February 2000). Available: http://purl.oclc.org/metadata/dublin_core

Enser, P.G.B. (1993). Query analysis in a visual information retrieval context, Journal of Document and Text Management 1(1), 25-52.

Fidel, R., Hahn, T.B, Rasmussen, E.M, & Smith, P.J, (Eds.) (1994). Challenges in indexing electronic text and images. ASIS Monograph series. Medford N.J: Learned Information, Inc.

Getty Information Institute. Categories for the Description of Works of Art (December 1999). Available: http://www.getty.edu/gri/standard/cdwa/

Getty Research Institute. Getty Vocabulary Program. Art & Architecture Thesaurus (Februrary 2000). Available: http://shiva.pub.getty.edu/aat_browser/

Harnad, S. (Ed). (1987). Categorical Perception: The Groundwork of Cognition. New York: Cambridge University Press.

Hendee, W.R & Wells, P.N.T. (Eds). (1997). The Perception of Visual Information. Second edition. New York: Springer Verlag.

Jaimes, A. & Chang, S.-F. (1999). Model-based classification of visual information for content-based retrieval. In Minerva M. Yeung; Boon-Lock Yeo; Charles A. Bouman (Eds.), Proceedings SPIE (Vol. 3656 Storage and Retrieval for Image and Video Databases VII)(pp. 402-414). San Jose, CA: International Society for Optical Engineering.

Jaimes, A. & Chang, S.-F. (2000). A conceptual framework for indexing visual information at multiple levels. Internet Imaging 2000IS&T/SPIE, San Jose, CA January 2000, forthcoming.

Jörgensen, C, (1995b). Classifying images: criteria for grouping as revealed in a sorting task. In Raymond Schwartz (Ed.), Advances in Classification Research 6 (Proceedings of the 6th ASIS SIG/CR Classification Research Workshop)(pp. 45-64), Medford NJ: Information Today.

Jörgensen, C. & Srihari, R. (1999). Creating a web-based image database for benchmarking image retrieval systems: a progress report. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas (Eds), Proceedings SPIE (Vol. 3644 Human Vision and Electronic Imaging IV)(pp. 534-541), San Jose, CA: International Society for Optical Engineering.

Jörgensen, C. (1996). Testing an image description template. In Steve Hardin (Ed.), Proceedings of the 59th Annual Meeting of the American Society for Information Science (ASIS '96) (pp. 209-213), Medford NJ: Information Today.

Jörgensen, C. (1998). Image attributes in describing tasks: an investigation. Information Processing & Management 34(2/3), 161-174.

Jörgensen, C. (1999). Retrieving the unretrievable: art, aesthetics, and emotion in image retrieval systems. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas (Eds.), Proceedings SPIE (Vol. 3644 Human Vision and Electronic Imaging IV)(pp. 348-355), San Jose, CA: International Society for Optical Engineering.

Jörgensen, C. (2000). Image indexing: an analysis of selected classification systems in relation to image attributes named by naive users. Final Report to the Office of Sponsored Research, Online Computer Library Center.

Jörgensen, C., (1995a). Image attributes. Ph.D. thesis, Syracuse University.

Library of Congress. Thesaurus for Graphic Materials I: Subject Terms (February 2000). Available: http://lcweb.loc.gov/rr/print/tgm1

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

Morris, M.W. & Murphy, G.L. (1990). Converging operations on a basic level in event taxonomies. Memory and Cognition 18(4), 407-418.

Panofski, E. (1962). Studies in iconology. New York: Harper & Row.

Rasmussen, E. M. (1997). Indexing images. Annual Review of Information Science and Technology 32 (ARIST), (pp. 169-196). Washington DC: American Society for Information Science.

Rosch, E & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology 7, 573-605.

Shatford Layne, S. (1986). Analyzing the subject of a picture: a theoretical approach. Cataloguing and Classification Quarterly, 6(3), 39 - 62.

Shatford Layne, S. (1994). Some issues in the indexing of images, Journal of the American Society for Information Science 45(8) 583-588.

Turner, J. (1994). Determining the Subject content of still and moving image documents for storage and retrieval: an experimental investigation, Ph.D. thesis, University of Toronto.

Tversky, A. (1977). Features of similarity. Psychological Review 84(4), 327-352.

VRA Data Standards Committee. Visual Resources Association. (January 2000). The VRA core categories for visual resources, version 2.0. Available: http://www.vra.oberlin.edu/vc.html