

# A Conceptual Framework for Indexing Visual Information at Multiple Levels

Alejandro Jaimes and Shih-Fu Chang<sup>1</sup>

Image and Advanced TV Laboratory, Department of Electrical Engineering  
Columbia University, 1312 SW Mudd Building Mailcode 4712 Box F8  
New York, N.Y. 10027

## ABSTRACT

In this paper, we present a conceptual framework for indexing different aspects of visual information. Our framework unifies concepts from the literature in diverse fields such as cognitive psychology, library sciences, art, and the more recent content-based retrieval. We present multiple level structures for visual and non-visual information. The ten-level visual structure presented provides a systematic way of indexing images based on *syntax* (e.g., color, texture, etc.) and *semantics* (e.g., objects, events, etc.), and includes distinctions between *general concept* and *visual concept*. We define different types of relations (e.g., *syntactic*, *semantic*) at different levels of the visual structure, and also use a *semantic information table* to summarize important aspects related to an image. While the focus is on the development of a conceptual indexing structure, our aim is also to bring together the knowledge from various fields, unifying the issues that should be considered when building a digital image library. Our analysis stresses the limitations of state of the art content-based retrieval systems and suggests areas in which improvements are necessary.

**Keywords:** image indexing, image classification, conceptual indexing, content-based retrieval, MPEG-7.

## 1. INTRODUCTION

The recent proliferation of digital images and video has brought new opportunities to end-users that now have a large amount of resources when searching for content. Visual information<sup>2</sup> is widely available on diverse topics, from many different sources, and in many different formats. This is an advantage, but at the same time a challenge since users cannot review large quantities of data when searching such content. It is imperative, therefore, to allow users to efficiently *browse* content or perform *queries* based on their specific needs. In order to provide such functionalities in a digital library, however, it is essential to understand the data, and index it appropriately. This indexing must be structured and it must be based on how users will want to access such information.

In traditional approaches, textual annotations are used for indexing- a cataloguer manually assigns a set of key words or expressions to describe an image. Users can then perform text-based *queries* or *browse* through manually assigned categories. In contrast to text-based approaches, recent techniques in content-based retrieval [38], have focused on indexing<sup>3</sup> images based on their visual content. Users can perform queries by example (e.g., images that look like this one) or user-sketch (e.g., image that looks like this sketch). More recent efforts attempt automatic classification of images based on their content: a system classifies each image, and assigns it a label (e.g., indoor, outdoor, contains a face, etc.).

In both paradigms there are classification issues which are often overlooked, particularly in the content-based retrieval community. The main difficulty in appropriately indexing visual information can be summarized as follows: (1) there is a large amount of information present in a single image (e.g., what to index?), and (2) different levels of description are possible (e.g., how to index?). Consider, for example, a portrait of a man wearing a suit. It would be possible to label the image with the terms “suit” or “man”. The term “man”, in turn, could carry information at multiple levels: *conceptual* (e.g., definition of man in the dictionary), *physical* (size, weight) and *visual* (hair color, clothing), among others. A category label,

---

<sup>1</sup> E-mail: {ajaimes, sfchang}@ee.columbia.edu WWW: <http://www.ee.columbia.edu/~ajaimes, ~sfchang>

<sup>2</sup> In this context, visual information will refer to images (painting, drawing, sketch, etc.), or video. We will use the terms image and visual information interchangeably.

<sup>3</sup> We use the terms indexing and classification interchangeably.

then, implies explicit (e.g., the person in the image is a man, not a woman), and implicit or undefined information (e.g., from that term alone it is not possible to know what the man is wearing).

In this paper, we focus on the problem of multiple levels of description for indexing visual information. We present a novel conceptual framework, which unifies concepts from the literature in diverse fields such as cognitive psychology, library sciences, art, and the more recent content-based retrieval. We make distinctions between *visual* and *non-visual* information and provide the appropriate structures. The ten-level *visual* structure presented provides a systematic way of indexing images based on *syntax* (e.g., color, texture, etc.) and *semantics* (e.g., objects, events, etc.), and includes distinctions between *general concept* and *visual concept*. We define different types of relations (e.g., *syntactic*, *semantic*) at different levels of the *visual* structure, and also use a *semantic information table* to summarize important aspects related to an image (e.g., that appear in the *non-visual* structure).

Our structures place state-of-the art content-based retrieval techniques in perspective, relating them to real user-needs and research in other fields. Using structures such as the ones presented, is beneficial not only in terms of understanding the users and their interests, but also in characterizing the content-based retrieval problem according to the levels of descriptions used to access visual information.

## 1.1 Related work

Work on issues related to images has been performed by researchers in different areas. Studies in *art* have focused on interpretation and perception [1][3], aesthetics and formal analysis [2], visual communication [4], levels of meaning in art [5], etc. Studies in *cognitive psychology* have dealt with issues such as perception [9][10][16], visual similarity [18], mental categories (i.e., concepts) [6], distinctions between perceptual and conceptual category structure [7][8][15], internal category structure (i.e., levels of categorization) [11][12][13][14][17], etc. In the field of *information sciences*, work has been performed in the analysis of the subject of an image [23][32][34][35], issues related to image indexing [22][29][31][33][36], the attributes that can be used to describe images [24][25], classification [26][29], query analysis [20][21], and indexing schemes [27][30], among others. Most work in *content-based retrieval* has focused mainly on using low-level features for automatic classification based global features [56][59], query by example (QBIC, Virage, VisualSEEk) [38], and query by user sketch [42]. More recent work has been performed on object-based classification [44].

There have also been recent efforts related to the organization of multimedia data. Some of that work includes [40][41][46][55], and [48]. In addition, the development of the MPEG-7 standard has triggered a large number of proposals to describe and structure multimedia information [47]. In the October 1999 MPEG-7 draft, descriptions schemes for multimedia data [65] are described: the *visual description scheme* (DS) consists of 6 sub-schemes: *syntactic*, *model*, *semantic*, *summarization*, *creation meta information*, *usage meta information*, *syntactic-semantic links* and *media information*. As will be apparent to the reader in the later sections, each of these DSs could be mapped to different parts of the indexing structures we present: the *syntactic DS* maps to levels 1 through 4 of our *visual structure* (Figure 2), the *semantic DS* to levels 5 through 10 of the same structure, the *summarization* and *meta data DSs* map to the *non-visual structure* (Figure 5), the *media DS* maps to level 1 of our *visual structure* (Figure 2) and also to the physical component of our *non-visual structure* (Figure 5). The work we present differs from the current MPEG-7 DS in various key points: (1) we provide a structured break down of *syntactic* and *semantic* attributes into different levels (e.g., in current MPEG-7 draft, the *object DS* contains an *annotation DS*, but the levels of semantic descriptions that could be used for the object are not considered), (2) we provide a structure to break down relations between elements of the image, based on our visual structures. Based on these differences, part of the work presented in this paper has been proposed to MPEG-7 [61],[63], and some of the components of the work have been included in the recent MPEG-7 drafts [65].

In addition to the differences with previous work described above, unlike previous efforts, the conceptual structures presented in this paper unify the research in various fields related to image content. This results in more intuitive structures for indexing visual information.

## 1.2 Outline

In section 2 we define some important concepts. In section 3 we present our indexing structures for *visual* and *non-visual* information. In section 4 categorization and similarity issues are discussed, and in section 5 we discuss the image indexing test-bed we are developing based on our framework.

## 2. CONCEPTS AND SEMANTICS

One of the difficulties inherent in the indexing of images is the number of ways in which they can be analyzed. A single image may represent many things, not only because it contains a lot of information, but because what we see in the image can be mapped to a large number of abstract concepts. A distinction between those possible abstract descriptions and more concrete descriptions based only on the visual aspects of the image constitutes an important step in indexing.

In the following sections, we make distinctions between *percept* and *concept*. We then provide definitions for *syntax* and *semantics*, and finally discuss *general concept space* and *visual concept space*. The importance of these definitions in the context of content-based retrieval will be apparent in section 3 when we define our indexing structures.

### 2.1 Percept vs. Concept

Images are multi-dimensional representations of information, but at the most basic level they simply cause a response to light (tonal-light or absence of light) [4]. At the most complex level, however, images represent abstract ideas that largely depend on each individual's knowledge, experience, and even particular mood. We can make distinctions between *percept* and *concept*.

The *percept* refers to what our senses perceive- in the visual system it is light. These patterns of light produce the perception of different elements such as texture and color. No interpretation process takes place when we refer to the *percept*- no knowledge is required.

A *concept*<sup>4</sup>, on the other hand, refers to an abstract or generic idea generalized from particular instances. As such, it implies the use of background knowledge and an inherent interpretation of what is perceived. Concepts can be very abstract in the sense that they depend on an individual's knowledge and interpretation- this tends to be very subjective.

### 2.2 Syntax and Semantics

In a similar way in which percepts require no interpretation, *syntax* refers to the way visual elements are arranged without considering the meaning of such arrangements. *Semantics*, on the other hand, deals with the meaning of those elements and of their arrangements. As will be shown in the discussion that follows, *syntax* can refer to several perceptual levels- from simple global color and texture to local geometric forms such as lines and circles. *Semantics* can also be treated at different levels.

### 2.3 General vs. Visual Concepts

Here we wish to emphasize that *general concepts* and *visual concepts* are different, and that these may vary among individuals.

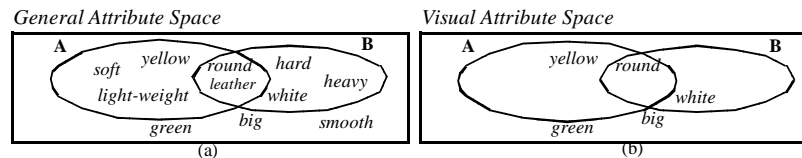
Using a ball as an example, we see that while one possible *general concept* describes a ball as a round mass<sup>4</sup>, different people may have different *general concepts*. A volleyball player may have a different *general concept* of a ball than a baseball player because, as described earlier, a *concept* implies background knowledge and interpretation. It is natural for different individuals to have very different interpretations of ideas (or in this case concrete objects). In Figure 1, we see that the attributes<sup>5</sup> used for the *general* and *visual concepts* of a ball are different (rules could be used to describe concepts, but we use attributes instead to simplify the explanation). Each box represents a universe of attributes, and each circle the set of attributes observers A and B choose to describe a ball. Attributes outside the circles are not chosen by the observers to describe this particular concept. Observer A is a volleyball player, and when asked to give the *general attributes* of a ball, he chooses soft, yellow, round, leather, and light-weight. Observer B is a baseball player, and when asked to give the *general attributes* of a ball, he chooses hard, heavy, white, round, and leather. Note that, naturally, there is also a correlation between some general and visual attributes (e.g., big).

---

<sup>4</sup> Definition from Merriam-Webster dictionary.

<sup>5</sup> In this section, we use the word attribute to refer to a characteristic or quality of an object (e.g, blue, big, heavy). We do not make a distinction between attribute name and attribute type (e.g., color: blue).

These definitions are useful since they point out a very important issue in content-based retrieval: different users have different *concepts* (of even simple objects), and even simple objects can be seen at different conceptual levels. Specifically, there is an important distinction between *general concept* (i.e., helps answer the question: what is it?) and *visual concept* (i.e., helps answer the question: what does it look like?) and this must be considered when designing an image database. We apply these ideas to the construction of our indexing structures. As suggested in [15], conceptual category structure may be based on perceptual structure.



**Figure 1.** We divide attributes into those that are general and those that are visual.

### 3. VISUAL AND NON-VISUAL CONTENT

As noted in the previous section, there are many levels of information present in images, and their multi-dimensionality must be taken into account when organizing them in a digital library. The first step in creating a conceptual indexing structure is to make a distinction between *visual* and *non-visual* content. The *visual content* of an image corresponds to what is directly perceived when the image is observed (i.e., descriptors stimulated directly by the visual content of the image or video in question- the lines, shapes, colors, objects, etc). The *non-visual content* corresponds to information that is closely related to the image, but that is not explicitly given by its appearance. In a painting, for example, the price, current owner, etc. belong to the *non-visual* category. Next we present an indexing structure for the *visual content* of the image and we follow with a structure for *non-visual* information.

#### 3.1 Visual content

Each of the levels of analysis that follows is obtained only from the image. The viewer's knowledge always plays a role, but the general rule here is that information not explicitly obtained from the image does not go into this category (e.g., the price of a painting would not be part of *visual content*). In other words, any descriptors used for visual content, are stimulated by the visual content of the image or video in question

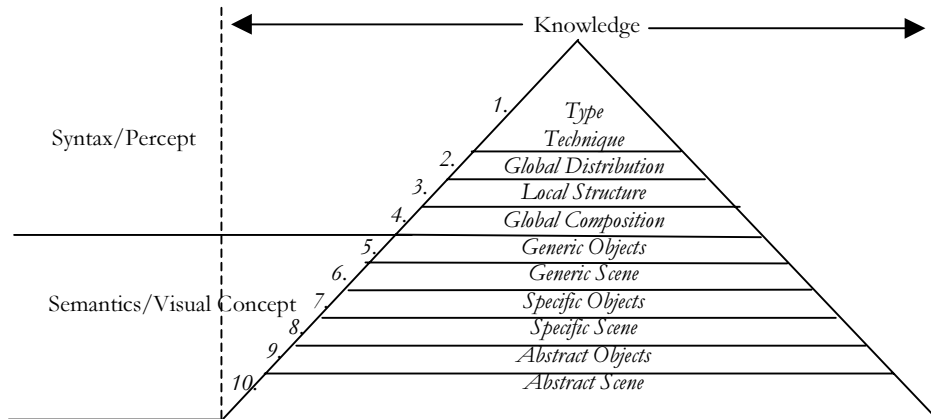
Our visual structure contains ten levels: the first four refer to *syntax*, and the remaining six refer to *semantics*. In addition, levels one to four are directly related to *percept*, and levels five through ten to *visual concept*. While some of these divisions may not be strict, they should be considered because they have a direct impact in understanding *what* the user is searching for and *how* he tries to find it in a database. They also emphasize the limitations of different indexing techniques (manual and automatic) in terms of the knowledge required. An overview of the structure is given in Figure 2. Observing this figure from top to bottom, it is clear that at the lower levels of the pyramid, more knowledge and information is required to perform indexing. The width of each level gives an indication of the amount of knowledge required there- for example, more information is needed to name specific objects in a scene. Each level is explained below and a discussion of the relationship between levels appears in section 3.1.11.

Observing this structure, it will be apparent that most of the efforts in content-based retrieval have focused on *syntax* (i.e., levels one through four). Techniques to perform semantic classification at levels five through ten, however, are highly desirable. The structure we present, helps identify the level of attributes handled by a specific technique, or provided by a given description (e.g., MPEG-7 annotations).

In the discussions that follow, we refer the reader to the examples of Figure 7 in the Appendix.

##### 3.1.1 Type/Technique

At the most basic level, we are interested in the general visual characteristics of the image or the video sequence. Descriptions of the type of image or video sequence or the technique used to produce it are very general, but prove to be of great importance. Images, for example, may be placed in categories such as painting, black and white (b&w), color photograph, and drawing. Related classification schemes at this level have been done conceptually in [40],[46], and automatically in WebSEEK [53].



**Figure 2.** The indexing structure is represented by a pyramid.

In the case of digital photographs, the two main categories could be color and grayscale, with additional categories/descriptions which affect general visual characteristics. These could include number of colors, compression scheme, resolution, etc. We note that some of these may have some overlap with the non-visual indexing aspects described in section 3.2. Figure 7a shows an interesting example.

### 3.1.2 Global Distribution

The *type/technique* in the previous level gives general information about the visual characteristics of the image or the video sequence, but gives little information about the visual content. *Global distribution* aims to classify 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). *Global distribution* features, therefore, may include global color (e.g., dominant color, average, histogram), global texture (e.g., coarseness, directionality, contrast), global shape (e.g. aspect ratio), global motion (e.g. speed, acceleration, and trajectory), camera motion, global deformation (e.g. growing speed), and temporal/spatial dimensions (e.g. spatial area and temporal dimension), among others. The example in Figure 7b shows two images that have similar texture/color. Notice that in this particular case these attributes are quite useful, but they would not be useful if a user were searching for an object

Even though some of these measures are difficult to quantify for a human observer, these global low-level features have been successfully used in various content-based retrieval systems to perform query by example (QBIC, WebSEEk, Virage) and to organize the contents of a database for browsing [38]. An interesting comparison of human and machine assessments of image similarity based on global features at this level can be found in [54].

### 3.1.3 Local Structure

In contrast to *Global Structure*, which does not provide any information about the individual parts of the image or the video sequence, the *Local Structure* level is concerned with the extraction and characterization of the image's components. At the most basic level, those components result from low-level processing and include elements such as the *Dot, Line, Tone, Color,* and *Texture*. In the Visual Literacy literature [4], some of these are referred to as the "basic elements" of visual communication and are regarded as the *basic syntax symbols*. Other examples of local structure attributes are temporal/spatial position (e.g. start time and centroid), local color (e.g. MxN Layout), local motion, local deformation, and local shape/2D geometry (e.g. bounding box). Figure 7c shows images in which attributes of this type may be of importance. In x-rays and microscopic images there is often a strong concern for local details.

Such elements have also been used in content-based retrieval systems, mainly on query by user-sketch interfaces such as those in [42][37], and VisualSEEk [38]. The concern here is not with objects, but rather with the basic elements that represent them and with combinations of such elements- a square, for example, is formed by four lines. In that sense, we can include here some "basic shapes" such as circle, ellipse and polygon. Note that this can be considered a very basic level of "grouping" as performed by humans when perceiving visual information.

### 3.1.4 Global Composition

At this level, we are interested in the specific arrangement of the basic elements given by the local structure, but the focus is on the *Global Composition*. In other words, we analyze the image as a whole, but use the basic elements described above (line, circle, etc.) for the analysis.

*Global Composition* refers to the arrangement or spatial layout of elements in the image. Traditional analysis in art describes composition concepts such as balance, symmetry, center of interest (e.g., center of attention or focus), leading line, viewing angle, etc. [1]. At this level, however, there is no knowledge of specific objects; only *basic elements* (i.e. dot, line, etc.) or groups of basic elements are considered. In that sense, the view of an image is simplified to an image that contains only basic syntax symbols: an image is represented by a structured set of lines, circles, squares, etc. Again, we present images with similar composition in Figure 7d (both images have objects in the center, and the leading line is diagonal). The composition of the images in Figure 7f is also similar, where the leading line is horizontal.

### 3.1.5 Generic Objects

Up to the previous level the emphasis had been on the perceptual aspects of the image. No world knowledge is required to perform indexing at any of the levels above, and automatic techniques rely only on low-level processing. While this is an advantage for automatic indexing and classification, studies have demonstrated that humans mainly use higher level attributes to describe, classify and search for images [24][25][26]. Objects are of particular interest, but they can also 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*, we are interested in what Rosch [14] calls the basic level categories: the most general level of object description. In the study of art, this level corresponds to *pre-Iconography* [5], and in information sciences [34] refers to it as the *generic of* level. The common underlying idea in these concepts and our definition of *Generic Objects* is that only general everyday knowledge is necessary to recognize the objects. A Machintosh apple, for example, would be classified as an apple at this level: that is the most general level of description of that object.

A possible difference between our definition and the definitions in [5][34], lies in the fact that we define *visual objects* as entities that can be seen, sometimes differing from the traditional definition of object. Objects like the sky or the ocean would perhaps not be considered objects under the traditional definition, but correspond to our *visual objects* (as well as the traditional objects like car, house, etc.). Examples of generic the objects "car", and "woman" are shown in Figure 7e. Figure 7g shows a "building", but note that in that figure the name of the building is used, so that particular attribute is a specific object attribute.

### 3.1.6 Generic Scene

Just like an image can be indexed according to the individual objects that appear in it, it is possible to index the image as a whole based on the set of all of the objects it contains and their arrangement. Examples of scene classes include city, landscape, indoor, outdoor, still life, portrait, etc. Some work in automatic scene classification has been performed by [56][59], and studies in basic scene categories include [17][11].

The guideline for this level is that only general knowledge is required. 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 it is a portrait. Figure 7f shows two images whose attributes correspond to generic scene. Other examples for the same images may include "mountain scene", "beach", etc.

### 3.1.7 Specific Objects

In contrast to the previous level, *Specific Objects* refers to objects that can identified and named. Shatford refers to this level as *specific of* [34]. Specific knowledge of the objects in the image is required, and such knowledge is usually objective since it relies on known facts. Examples include individual persons (e.g., Bill Clinton in Figure 6), and objects (e.g., also "Alex" and "Crysler building" in Figure 7g).

### 3.1.8 Specific Scene

This level is analogous to *General Scene* with the difference that here there is specific knowledge about the scene. While different objects in the image may contribute in different ways to determine that the image depicts a specific scene, a single object is sometimes enough. A picture that clearly shows the Eiffel Tower, for example, can be classified as a scene of Paris, based only on that object (see Figure 7h). The other image in the same figure shows a similar example for Washington D.C.

### 3.1.9 Abstract Objects

At this level, specialized or interpretative knowledge about what the objects *represent* is used. This is referred to as *Iconology* (interpretation) in art [5], or the *about* level in [34]. This indexing level is the most difficult one in the sense that it is completely subjective and assessments between different users vary greatly. The importance of this level was shown in experiments by [25], where viewers used abstract attributes to describe images. For example, a woman in a picture may represent anger to one observer, or perhaps pensiveness to another observer. Other examples of abstract object descriptions appear in Figure 7i as “arts” and “law”.

### 3.1.10 Abstract Scene

The *Abstract Scene* level refers to what the image as a whole represents. It may be very subjective. It was shown in [25], for example, that users sometimes describe images in affective (e.g. emotion) or abstract (e.g. atmosphere, theme) terms. Other examples at the abstract scene level include sadness, happiness, power, heaven, and paradise. Examples of abstract scene for the images in Figure 7j are “Agreement”, and “Industry”.

### 3.1.11 Relationships across levels

We have chosen a pyramid representation because it directly reflects several important issues inherent in our structure. Analyzing Figure 2 from top to bottom, it is apparent that at the lower levels of the pyramid, more knowledge and information is required to perform the indexing. This knowledge is represented by the width of each level. It is important to point out, however, that this assumption may have some exceptions. An average observer, for example, may not be able to determine the technique that was used to produce a painting- but an expert in art would be 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).

Although inter-level dependencies exist, each level can be seen as an independent perspective or dimension when observing an image and the way each level is treated will depend on the nature of the database, users and purpose.

### 3.1.12 Visual Content Relationships

In this section, we briefly present a representation for relations between image elements<sup>8</sup>. As shown in Figure 3, this structure accommodates relations at different levels and is based on the *visual structure* presented earlier (see Figure 2). We note that relations at some levels (e.g., 1, 6, 8, and 10 in Figure 3) are most useful when applied between entities to which the structure is applied (e.g., scenes from different images may be compared). Elements within each level are related according to two types of relations: *syntactic* and *semantic* (only for levels 5 through 10). For example: two circles (*local structure*) can be related *spatially* (e.g., next to), *temporally* (e.g., before) and/or *visually* (e.g., darker than). Elements at the *semantic* levels (e.g., objects) can have *syntactic* and *semantic* relations- (e.g., two people are next to each other, and they are friends). In addition, each relation can be described at different levels (*generic*, *specific*, and *abstract*). We note that relations between levels 1,6,8, and 10 can be most useful between entities represented by the structure (e.g., between images, between parts of images, scenes, etc.)

The *visual structure* in Figure 2 is divided into *syntax/percept* (levels 1 to 4) and *visual concept/semantics* (levels 5 to 10). To represent relations, we observe such division and take into consideration the following [64]: (1) Knowledge of an object embodies knowledge of the object’s spatial dimensions, that is, of the gradable characteristics of its typical, possible or actual, extension in space; (2) knowledge of space implies the availability of some system of axes which determine the designation of certain dimensions of, and distances, between objects in space. We use this to argue that relations that take place in the *syntactic* levels of the *visual structure* can only occur in 2D space<sup>6</sup> since no knowledge of the objects exist (i.e., relationships in 3D space cannot be determined). At the *local structure* level, for example, only the basic elements of visual literacy are considered, so relations at that level are only described between such elements (i.e., which do not include 3D information). Relations between elements of levels 5 through 10, however, can be described in terms of 2D or 3D.

In a similar way, the *relations* themselves are divided into the classes *syntactic* (i.e., related to perception) and *semantic* (i.e. related to meaning). *Syntactic* relations can occur between elements<sup>7</sup> at any of the levels shown in Figure 3, but *semantic*

<sup>6</sup> In some cases, we could have depth information associated with each pixel, without having knowledge of the objects. Here we make the assumption that depth information is not available.

<sup>7</sup> We use the word element here since it may refer to any image component (e.g., dot, line, object, etc.), depending on the level of analysis used.

relations occur only between elements of levels 5 through 10. *Semantic* relationships between different colors in a painting, for example, could be determined (e.g., the combination of colors is warm), but we do not include these at that level of our model.

Following the work of [62], we divide *spatial relationships* into the following classes: (1) topological (i.e., how the boundaries of elements relate) and (2) orientation (i.e., where the elements are placed relative to each other). Topological relations include near, far, touching, etc. and orientation relations include diagonal to, in front of, etc.

*Temporal relations* refer to those that connect elements with respect to time (e.g., in video these include before, after, between, etc.), and *visual relations* refer only to visual features (e.g., bluer, darker, etc.). *Semantic* relations are associated with meaning (e.g., owner of, friend of, etc.). A more detailed explanation of relations is provided in [61].

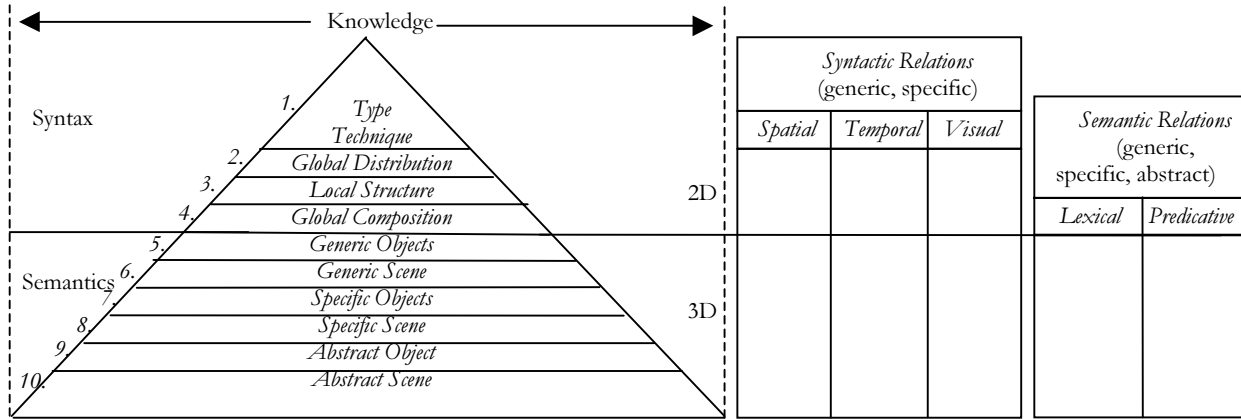


Figure 3. Relationships are based on the visual structure.

In a similar way in which the elements of the visual structure have different levels (*generic, specific, abstract*), relations can be defined at different levels. *Syntactic* relations can be *generic* (e.g., near) or *specific* (e.g., a numerical distance measure). *Semantic* relationships can be *generic, specific, or abstract* (see Figure 4).

As an example, *spatial global distribution* could be represented by a distance histogram, *local structure* by relations between local components (e.g., distance between visual literacy elements), and *global composition* by global relations between visual literacy elements.

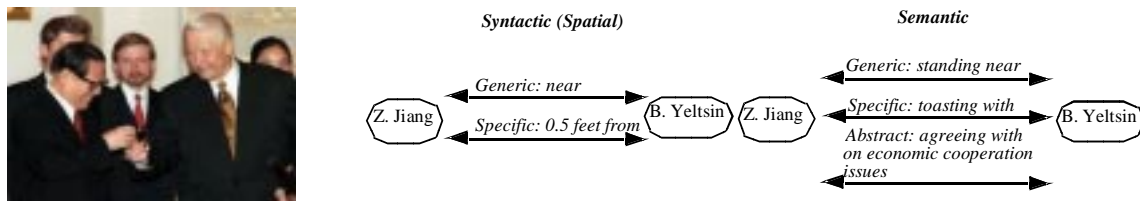
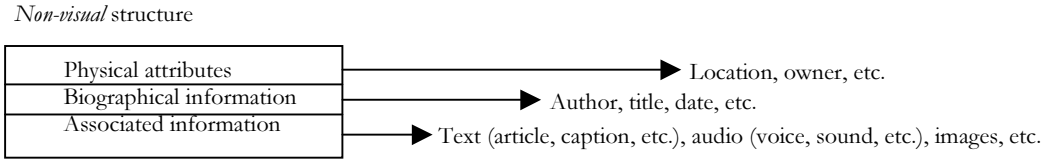


Figure 4. Syntactic (spatial) and semantic relations that could be used for this image.

### 3.2 Non-visual information

As explained at the beginning of this section, *non-visual information* refers to information that is not directly part of the image, but is rather associated with it in some way. Shatford in [33] divides attributes into biographical and relationship attributes. While it is possible for non-visual information to consist of sound, text, hyperlinked text, etc., our goal here is to present a simple structure that gives general guidelines for indexing. We will focus briefly on text information only. Figure 5 gives an overview of this structure.





**Figure 5.** Non-visual information.

### 3.2.1 Biographical Information

The source for the actual image may be direct (e.g., a photograph of a natural scene) or indirect (e.g., image of a sculpture, painting, building, drawing). In either case, there may be *Biographical Information* associated with the image. This information can repeat itself for several objects in the image (e.g., an image of the ceiling of the Sistine chapel may have information about the painting and the chapel itself), exist for the image only, or not exist at all. In most cases, *Biographical Information* is not directly related to the subject of the image, but rather to the image as a whole. Examples include the author, date, title, material, technique, etc.

### 3.2.2 Associated information

The second class of *non-visual information* is directly linked to the image in some way. *Associated Information* may include a caption, article, a sound recording, etc.


As discussed in section 3.3, in many cases this information helps perform some of the indexing in the visual structure, since it may contain specific information about what is depicted in the image (i.e., the subject). In that context, it is usually very helpful at the semantic levels since they require more knowledge that is often not present in the image alone. In some cases, however, the information is not directly related to the subject of the image, but it is associated to the image in some way. A sound recording accompanying a portrait, for example, may include sounds that have nothing to do with the person being depicted- they are associated with the image though, and could be indexed if desired.

### 3.2.3 Physical attributes

*Physical Attributes* simply refer to those that have to do with the image as a physical object. This may include location of the image, location of the original source, storage (e.g., size, compression), etc.

## 3.3 Relationships between indexing structures

Following the work of [34], we define a *Semantic Information Table* to gather high level information about the image (Figure 6). The table can be used for individual objects, groups of objects, the entire scene, or parts of the image. In most cases *visual* and *non-visual* information contribute in filling in the table- simple scene classes such as indoor/outdoor may not be easily determined from the visual content alone; location may not be apparent from the image, etc. Individual objects can be classified and named based on the *non-visual information*, contributing to the mapping between *visual object* and *conceptual object*.



	Specific	Generic	Abstract
Who	B. Clinton, L. Zhaoxing	Men	Intl politics
What action	Signing condolence book	Meeting	Apology
What object	Condolence book #3	Book	Condolences
Where	Oval office	Indoors	Government
When	May 13, 1999	Daytime	War
Why	Bombing	Embassy bombing	Mistake

**Figure 6.** *Visual* and *non-visual* information can be used to semantically characterize an image or its parts. The way in which these two modalities contribute to answer the questions in the *semantic table* may vary depending on the content. The *table* helps answer questions such as: *What is the subject (person/object, etc.)?*, *What is the subject doing?* *Where is the subject?* *When?* *How?* *Why?* The table can be applied to individual objects, groups of objects, the entire scene, or parts of the image.

The relationship between this structure and the visual structure is apparent when applying the table at each level beginning with level 5. We also note that while the table provides a compact representation for some information related to the image, it does not replace the indexing structures presented. The group of structures provides the most complete description.

Having the appropriate indexing structures, we can focus on how the contents of a digital library may be organized. In the next section, we analyze issues that play a crucial role in the organization and retrieval of images.

## 4. FEATURES, SIMILARITY, AND CATEGORIZATION

In order to be successful at building an image digital library, it is not only important to understand the data, but also the human issues related to *classification*. In this section we discuss issues of importance in this respect, and explain how we apply the concepts in building our image indexing test bed. First, we discuss categories. Then, we discuss levels and structure in categorization. Finally, we present some of the issues related to attributes and similarity.

### 4.1 Categories and classification

*Categorization* can be defined as treating a group of entities as equivalent. A *category* is any of several fundamental and distinct classes to which entities or concepts belong- entities within categories appear more similar and entities between categories appear less similar [7]. Before categorization can be undertaken, however, it is essential to have an understanding of the nature of the data being categorized. Since this was done in section 3, we can now focus on the types of categories that could be used. In the literature of classification, researchers have identified two kinds of categories [9]: (1) *Sensory Perception* categories (e.g., texture, color or speech sounds -/e/), and (2) *Generic Knowledge (GK)* categories (e.g., natural kinds- birds, artifacts- cars and events -eating).

In our structure we can identify *Sensory Perception* categories such as color and texture. *GK* categories, however, play a very important role since users are mainly interested in the objects that appear in the images and what those objects may represent. Some theories in cognitive psychology [7] express that classification in *GK categories* is done as follows:

- *Rules*: attribute values of the entity are used (e.g., rule: an image in the people category should have a person in it).
- *Prototypes*: a prototype of the category contains the characteristic attributes of its category's exemplars. These are attributes that are highly probable across category members, but are neither necessary nor sufficient for category membership. A new image is classified according to how similar it is to the category's prototype (e.g., a prototype for the landscape class could be simple sketch of a sunset).
- *Exemplars*: an instance is classified according to its most similar exemplar's category (e.g., instead of having a rule for the people category, we could have a set of example images in that class and use those for classification).

This evidence is helpful in terms organizing images in a database because we can use these techniques to perform classification and to present results to the user. These concepts are being used in the development of our image indexing test bed.

### 4.2 Category Structure

Category structure is a crucial factor in a digital library and brings about several issues of importance which we briefly discuss here. The following issues should be considered: relationships between categories (e.g., hierarchical or entity-relation), the levels of abstraction at which classification should be performed (e.g., studies by Rosch [13] suggest the existence of a basic level and subordinate/superordinate level categories), horizontal category structure (i.e., how each category should be organized and the degrees of membership of elements within each category- these can be fuzzy or binary), etc.

In addition to considering different levels of analysis when indexing visual information, the way in which similarity is measured is of great importance. Issues related to measurements of similarity include the level of consideration (e.g., part vs. whole), the attributes examined, the types of attributes (e.g., levels of our structures), whether the dimensions are separable or not, etc.

## 5. THE IMAGE INDEXING TEST BED

We are developing an image indexing test bed that incorporates the concepts presented, using different techniques to index images based on the structure of Figure 2. In particular, for *type/technique* we are using discriminant analysis as in [53]. For *global distribution*, we use global color histograms and Tamura texture measures [57]. At the *local structure* level, we allow sketch queries as in VideoQ [37], by using automatic segmentation and also multi-scale phase-curvature histograms of coherent edge-maps and projection histograms [50]. *Global composition* is obtained by performing automatic segmentation and merging of generated regions to yield iconic representations of the images.

*Generic objects* are being automatically detected using the *Visual Apprentice* [43]. In the *Visual Apprentice*, visual object detectors are built by defining an *object definition hierarchy* (i.e., specifying the model of an object and its parts) and providing the system with examples. Multiple classifiers are learned automatically by the system at different levels of the hierarchy (*region, perceptual, object-part, and object*), and the best classifiers are automatically selected [45] and combined [44] when performing automatic classification. We also use the AMOS system [60] to perform manual annotation of objects and object search.

At the *generic scene* level we perform city vs. landscape and indoor vs. outdoor classification. This is done automatically using the OF\*IIF technique [49] in which clustering and classification of image regions is performed in conjunction with textual features (e.g., from the image caption), if available, and specialized object detectors (e.g., face or sky detector).

Information about *specific objects and scenes* is obtained from the associated information using the doextract system [66], which extracts names of people, places, etc. Annotations at the *abstract levels*, when performed, are being done manually.

In addition to using the automatic techniques above, we implement some of the concepts presented in section 4. Given that a combination of sensory perception categories and general knowledge categories also takes place in our framework, we use rules, prototypes and exemplars to perform manual classification and also to present the elements of the digital library to the user.

## 6. CONCLUSIONS AND FUTURE WORK

We have presented a conceptual framework for indexing visual information at multiple levels. Our structures are suitable for *syntactic/semantic* as well as *perceptual/conceptual* distinctions. We have separated *visual concepts* and *general concepts*, and presented a structure (the *semantic information table*) that represents *semantic* information from *visual* and *non-visual* data. Our structures allow indexing of visual information at multiple levels and include description of relations at multiple levels. We have also discussed several of the important issues related to the indexing of visual information, as identified by several researchers in different fields.

In addition to providing structures that unify research in different areas, we have discussed the image indexing test-bed we are developing which is based on the concepts presented. We are currently working on validating our framework in the context of MPEG-7, through experiments similar to those performed in [27]. Future work includes expanding our structure to audio, modifying it so that it may include video structure information (e.g., scene transitions), and working on the development of an evaluation scheme for the test-bed we are constructing.

## REFERENCES

The following references have been grouped only to aid the reader. Some of them may belong in several categories.

### Art

- [1] R. Arnheim. *Art and Visual Perception: A psychology of the Creative Eye*. University of California Press. Berkeley, California, 1984.
- [2] S. Barnet. *A Short Guide to Writing About Art*. 5th Edition, Logman, New York, 1997.
- [3] G.T. Buswell. *How People Look at Pictures: A Study of the Psychology of Perception in Art*, University of Chicago press, 1935.
- [4] D. A. Dondis. *A primer of visual literacy*. Cambridge, Mass., MIT Press, 1973.

- [5] E. Panofski. *Studies in Iconology*, Harper & Row, New York, 1962.

### **Cognitive Psychology**

- [6] S. L. Armstrong, L. R. Gleitman and H. Gleitman, "What Some Concepts Might Not Be," *Cognition* 13, pp. 263-308, 1983.
- [7] B. Burns, editor. *Percepts, Concepts and Categories: the Representation and Processing of Information*, Elsevier Academic Publishers, New York, 1992.
- [8] B. Burns, "Perceived Similarity in Perceptual and Conceptual Development: The Influence of Category Information on Perceptual Organization," *Percepts, Concepts and categories*. B. Burns editor, 1992.
- [9] S. Harnad, editor. *Categorical Perception: the Groundwork of Cognition*, Cambridge University Press, New York, 1987.
- [10] W. R. Hendee, P. N.T. Wells, editors. *The Perception of Visual Information*, Second Edition. Springer Verlag, New York, 1997.
- [11] M. W. Morris and G. L. Murphy, "Converging Operations on a Basic Level in Event Taxonomies," *Memory and Cognition*, Vol. 18, No. 4, pp. 407-418, 1990.
- [12] A. Rifkin, "Evidence for a Basic Level in Event Taxonomies," *Memory and Cognition*, Vol. 13 No. 6, pp. 538-556, 1985.
- [13] E. Rosch and C. B. Mervis, "Family Resemblances: Studies in the Internal Structure of Categories," *Cognitive Psychology* 7, pp. 573-605, 1975.
- [14] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson and P. Boyes-Braem, "Basic Objects in Natural Categories," *Cognitive Psychology* 8, pp. 382-439, 1976.
- [15] L. B. Smith and D. Heise. "Perceptual Similarity and Conceptual Structure," *Percepts, Concepts and Categories*. B. Burns, Editor, 1992.
- [16] E.R. Tufte, *Visual Explanations: Images and Qualities, Evidence and Narrative*, Graphics press, Cheshire, Conn., 1997.
- [17] B. Tversky and K. Hemenway, "Categories of Environmental Scenes," *Cognitive Psychology* 15, pp. 121-149 1983.
- [18] A. Tversky, "Features of Similarity," *Psychological Review*. Vol. 84 No. 4, July 1977.

### **Information Sciences**

- [19] E. T. Davis, "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.
- [20] J. P. Eakins, "Design Criteria for a Shape Retrieval System," *Computers in Industry* Vol. 21, No. 2, pp. 167-184, 1993.
- [21] P.G.B. Enser, "Query Analysis in a Visual Information Retrieval Context," *Journal of Document and Text Management*. Vol 1 No.1, 1993.
- [22] R. Fidel, T. B. Hahn, E. M. Rasmussen, and P. J. Smith, editors. *Challenges in indexing electronic text and images*. *ASIS Monograph series*, 1994.
- [23] B.J., Jones, "Variability and Universality in Human Image Processing," *Understanding Images: Finding Meaning in Digital Imagery*, Francis T. Marchese, editor, TELOS, Santa Clara, CA, 1995.
- [24] C. Jorgensen, "Image Attributes," *Ph.D. thesis*, Syracuse University, 1995.
- [25] C. Jorgensen, "Image Attributes in Describing Tasks: an Investigation", *Information Processing & Management* 34 (2/3): 161-174, 1998.
- [26] C. Jorgensen, "Classifying Images: Criteria for Grouping as Revealed in a Sorting Task," *Advances in Classification Research Vol. 6, Proceedings of the 6th ASIS SIG/CR Classification Research Workshop*, Raymond Schwartz, editor, pp 45-64, 1995.
- [27] C. Jorgensen, "Indexing Images: Testing an Image Description Template", *ASIS 1996 Annual Conference Proceedings*. October 19-24, 1996.
- [28] K. Markey, "Computer Assisted Construction of a Thematic Catalog of Primary and Secondary Subject Matter," *Visual Resources*, Vol 3, pp. 16-49, Gordon and Breach, Science Publishers, 1983.
- [29] B. Orbach, "So That Others May See: Tools for Cataloguing Still Images," *Describing Archival Materials: the Use of the MARC AMC Format*, Haworth Press, 1990.
- [30] E.B. Parker, *LC Thesaurus for Graphic Materials: Topical Terms for Subject Access*. Library of Congress, Washington DC, 1987.

- [31] E.M. Rasmussen, "Indexing Images," *Annual Review of Information Science and Technology (ARIST)*, Vol. 32 p. 169-196, 1997.
- [32] H. Roberts, "Do You Have Any Pictures Of...? Subject Access to Works of Art In Visual Collections and Book Reproductions," *Art Documentation*, Fall, 1988.
- [33] S. Shatford Layne, "Some Issues in the Indexing of Images," *Journal of the American Society for Information Science* Vol. 45, No. 8, pp. 583-588, 1994.
- [34] S. Shatford Layne, "Analyzing the Subject of a Picture: A Theoretical Approach," *Cataloguing and Classification Quarterly*, Vol 6 No. 3, The Haworth Press, 1986.
- [35] J. Turner, "Determining the Subject content of still and moving image documents for storage and retrieval: an experimental investigation", Ph.D. thesis, University of Toronto, 1994.
- [36] J. Turner, "Cross-Language Transfer of Indexing Concepts for Storage and Retrieval of Moving Images: Preliminary Results," *ASIS 1996 Annual Conference Proceedings*. October 19-24, 1996.

### **Content-Based Retrieval**

- [37] S.-F. Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong, "A Fully Automatic Content-Based Video Search Engine Supporting Multi-Object Spatio-temporal Queries," *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image and Video Processing for Interactive Multimedia*, Vol. 8, No. 5, pp. 602-615, Sept. 1998.
- [38] S.-F. Chang, J.R. Smith, M. Beigi and A. Benitez, "Visual Information Retrieval from Large Distributed On-line Repositories", *Communications of the ACM*, December, 1997.
- [39] D. Healey, "Preattentive Processing in Visualization," <http://www.cs.berkeley.edu/~healey/PP/PP.shtml>
- [40] R. S. Heller and C. D. Martin, "A Media Taxonomy," *IEEE Multimedia Magazine*, Vol 5. No. 1, 1995.
- [41] N. Hirzalla, B. Falchuk, and A. Karmouch, "A Temporal Model for Interactive Multimedia Scenarios," *IEEE Multimedia Magazine*, Winter 1995.
- [42] C. E. Jacobs, A. Finkelstein, D. H. Salesin, "Fast Multiresolution Image Querying," *Proceedings of SIGGRAPH 95, in Computer Graphics Proceedings, Annual Conference Series*, pp 277-286, August 1995.
- [43] A. Jaimes and S.-F. Chang, "Model-Based Classification of Visual Information for Content-Based Retrieval", *Storage and Retrieval for Image and Video Databases VII, IS&T/SPIE*, San Jose, CA, January 1999.
- [44] A. Jaimes and S.-F. Chang, "Integrating Multiple Classifiers in Visual Object Detectors Learned from User Input", Invited paper, session on Image and Video Databases, *4th Asian Conference on Computer Vision (ACCV 2000)*, Taipei, Taiwan, January 8-11, 2000.
- [45] A. Jaimes and S.-F. Chang, "Automatic Selection of Visual Features and Classifiers", *Storage and Retrieval for Image and Video Databases VIII, IS&T/SPIE*, San Jose, CA, January 2000.
- [46] G. L. Lohse, K. Biolsi, N. Walker and H. H. Rueter, "A Classification of Visual Representation", *Communications of the ACM*. Vol. 37, No. 12, December 1994.
- [47] MPEG-7 website: <http://drogo.cselt.stet.it>
- [48] H. Purchase, "Defining Multimedia," *IEEE Multimedia Magazine*, Vol. 5 No. 1, 1998.
- [49] S. Paek, C. L. Sable, V. Hatzivassiloglou, A. Jaimes, B. H. Schiffman, S.-F. Chang, K. R. McKeown, "Integration of Visual and Text based Approaches for the Content Labeling and Classification of Photographs," *ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval*. Berkeley, CA. August, 1999.
- [50] R.K. Rajendran and S.F. Chang, "Visual Search Tools and their Application in K-12 Education", *ADVENT project technical report*, Columbia University, May 1999.
- [51] S. Santini and R. Jain, "Gabor Space and the Development of Preattentive Similarity," *Proceedings of ICPR 96, International Conference on Pattern Recognition*, Vienna, Autumn 1996.
- [52] S. Santini, R. Jain, "Beyond Query by Example," *ACM Multimedia 98*, Bristol, England, September 1998.
- [53] J. R. Smith and S.-F. Chang, "An Image and Video Search Engine for the World-Wide Web", *Storage & Retrieval for Image and Video Databases V*, San Jose, CA, February 1997.
- [54] D. McG. Squire, T. Pun, "A Comparison of Human and Machine Assessments of Image Similarity for the Organization of Image Databases," *Scandinavian conference on Image Analysis*, June 9-11, Lappeenranta, Finland, 1997.
- [55] U. Srinivasan, C. Lindley, and B. Simpson-Young, "A Multi-Model Framework for Video Information Systems", in *Database Semantics: Issues in Multimedia Systems*. Kluwer Academic Publishers, pp. 85-108, January 1999.
- [56] M. Szummer and R.W. Picard, "Indoor-Outdoor Image Classification", *IEEE International Workshop on Content-based Access of Image and Video Databases*, in conjunction with ICCV'98. Bombay, India, 1998.
- [57] H. Tamura, S. Mori, and T. Yamawaki, "Textural Features Corresponding to Visual Perception", *IEEE Transactions*

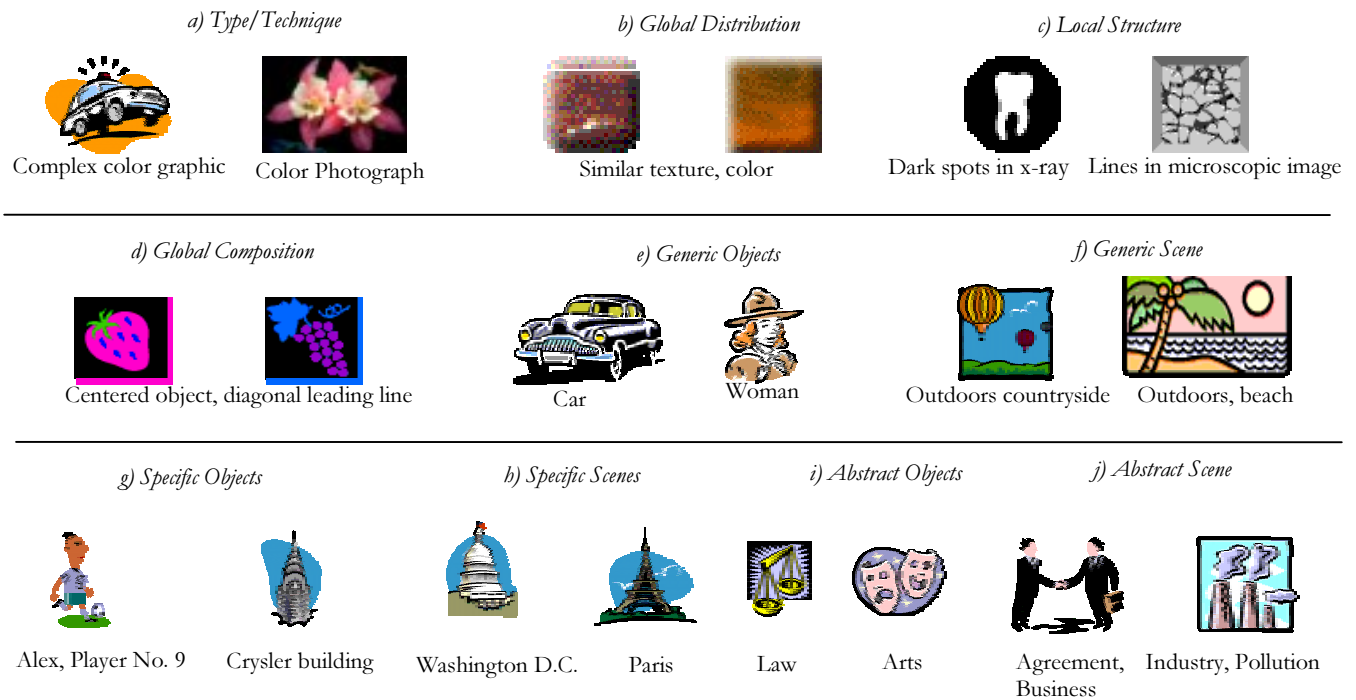
on Systems, Man, and Cybernetics, Vol. SMC-8, No. 6, June 1978.

- [58] A. Triesman, "Preattentive Processing in Vision," *Computer Vision, Graphics, and Image Processing* 31, pp. 156-177, 1985
- [59] A. Vailaya, A. Jain and H.J. Zhang, "On Image Classification: City vs. Landscape", *IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara, California, June 21, 1998.
- [60] D. Zhong and S.-F. Chang, "AMOS: An Active System for Mpeg-4 Video Object Segmentation", *1998 International Conference on Image Processing*, October 4-7, 1998, Chicago, Illinois, USA

**Others**

- [61] A. B. Benitez, A. Jaimes, S.-F. Chang, J. R. Smith, and C.-S. Li, "Fundamental Entity-Relationship Models for the Generic Audio Visual DS", Contribution to ISO/IEC JTC1/SC29/WG11 MPEG99/M4754, Vancouver, Canada, July 1999.
- [62] D. Hernandez, *Qualitative Representation of Spatial Knowledge*, Lecture Notes in Artificial Intelligence, 804. Springer-Verlag, Berlin, 1994.
- [63] A. Jaimes, C. Jorgensen, A. B. Benitez, and S.-F. Chang, "Multiple Level Classification of Visual Descriptors in the Generic AV DS", Contribution to ISO/IEC JTC1/SC29/WG11 MPEG99/M5251, Melbourne, Australia, October 1999.
- [64] E. Lang, K.-U. Carstensen, and G. Simmons. *Modelling Spatial Knowledge on a Linguistic Basis*, Lecture Notes in Artificial Intelligence 481, Springer-Verlag, Berlin, 1991.
- [65] MMDS Group, "MPEG-7 Generic AV Description Schemes (V0.7)", *Doc. ISO/IEC JTC1/SC29/WG11 MPEG99/N2966*, Melbourne, Australia, October 1999.
- [66] N. Wacholder, Y. Ravin, and M. Choi, "Disambiguation of Proper Names in Text", *5th Conference on Applied Natural Language Processing*, Washington D.C., April 1997.

**APPENDIX**



**Figure 7.** Example images for each level of the *visual* structure presented.