Semantics in Visual Information Retrieval

Carlo Colombo, Alberto Del Bimbo, and Pietro Pala
University of Florence, Italy

A compositional approach increases the level of representation that can be automatically extracted and used in a visual information retrieval system. Visual information at the perceptual level is aggregated according to a set of rules. These rules reflect the specific context and transform perceptual words into phrases capturing pictorial content at a higher, and closer to the human, semantic level.

Visual information retrieval systems have entered a new era. First-generation systems allowed access to images and videos through textual data.[1,2] Typical searches for these systems include, for example, "all images of paintings of the Florentine school of the 15th century" or "all images by Cezanne with landscapes." Such systems expressed information through alphanumeric keywords or scripts. They employed representation schemes like relational models, frame models, and object-oriented models. On the other hand, current-generation retrieval systems support full retrieval by visual content.[3,4] Access to visual information is not only performed at a conceptual level, using keywords as in the textual domain, but also at a perceptual level, using objective measurements of visual content. In these systems, image processing, pattern recognition, and computer vision constitute an integral part of the system's architecture and operation. They objectively analyze pixel distribution and extract the content descriptors automatically from raw sensory data. Image content descriptors are commonly represented as feature vectors, whose elements correspond to significant parameters that model image attributes. Therefore, visual attributes are regarded as points in a multidimensional feature space, where point closeness reflects feature similarity.

These advances (for comprehensive reviews of the field, see the "Further Reading" sidebar) have paved the way for third-generation systems, featuring full multimedia data management and networking support. Forthcoming standards such as MPEG-4 and MPEG-7 (see the Nack and Lindsay article in this issue) provide the framework for efficient representation, processing, and retrieval of visual information.

Yet many problems must still be addressed and solved before these technologies can emerge. An important issue is the design of indexing structures for efficient retrieval from large, possibly distributed, multimedia data repositories. To achieve this goal, image and video content descriptors can be internally organized and accessed through multidimensional index structures.[5] A second key problem is to bridge the semantic gap between the system and users. That is, devise representations capturing visual content at high semantic levels especially relevant for retrieval tasks. Specifically, automatically obtaining a representation of high-level visual content remains an open issue. Virtually all the systems based on automatic storage and retrieval of visual information proposed so far use low-level perceptual representations of pictorial data, which have limited semantics.

Building up a representation proves tantamount to defining a model of the world, possibly through a formal description language, whose semantics capture only a few significant aspects of the information content.[6]

Different languages and semantics induce diverse world representations. In text, for example, the meaning of single words is specific yet limited, and an aggregate of several words—a phrase—produces a higher degree of significance and expressivity. Hence, the rules for the syntactic composition of signs in a given language also generate a new world representation, offering richer semantics in the hierarchy of signification.

To avoid equivocation, a retrieval system should embed a semantic level reflecting as much as possible the one humans refer to during interrogation. The most common way to enrich a visual information retrieval system's semantics is to annotate pictorial information manually at storage time through a set of external keywords describing the pictorial content. Unfortunately, textual annotation has several problems:

1. It's too expensive to go through manual annotation with large databases.

2. Annotation is subjective (generally, the anno-

1070-986X/99/$10.00 © 1999 IEEE

tator and the user are different persons).

3. Keywords typically don't support retrieval by similarity.

Automatically increasing the semantic level of representation provides an alternative. Starting from perceptual features—the atomic elements of visual information—some intermediate semantic levels can be extracted using a suitable set of rules. Perceptual features represent the evidence upon which to build the interpretation of visual data. A process of syntactic construction called *compositional semantics* builds the semantic representation.

In this article, we discuss how to extract automatically—from raw image and video data—two distinct semantic levels and how to represent these levels through appropriate language rules. As we organize semantic levels according to a signification hierarchy, the corresponding description languages become stratified, allowing the composition of higher semantic levels according to syntactic rules that combine perceptual features and lower level signs. Since these languages directly depend on objective features, the approach naturally accommodates visual search by example and retrieval by similarity. We'll address two different visual contexts: art paintings and commercial videos. We'll also present retrieval examples showing that compositional semantics improves accordance with human judgment and expectation.

A language-oriented approach

Here we discuss the compositional semantics framework we developed. We also provide background theories in art and advertising.

Compositional semantics framework

The compositional semantics framework involves a bottom-up analysis and processing of a visual message, starting from its perceptual features. For still images, these features are image colors and edges. For videos or image sequences, additional features include the presence of editing effects, the motion of objects within a scene, and so on. Without loss of generality, the perceptual properties of a visual message can be represented through a set of scores $P = \{\phi_i\}$, $i = 1, \ldots, n$, each score $\phi_i \in [0, 1]$ representing the extent to which the i-th feature appears in the message.

We devised two distinct levels of the signification hierarchy, namely the *expressive* and the *emotional* levels, as plausible intermediate steps involved in the construction of meaning.

Further Reading

For a comprehensive introduction to visual information retrieval, see

A. Del Bimbo, *Visual Information Retrieval*, Academic Press, London, 1999.

P. Aigrain, H. Zhang, and D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review," *Multimedia Tools and Applications*, Vol. 3, No. 4, Dec. 1996, pp. 179-202.

A. Gupta and R. Jain, "Visual Information Retrieval," *Comm. of the ACM*, Vol. 40, No. 5, May 1997, pp. 71-79.

A review of the state of the art in visual information processing can be found in

B. Furht, S.W. Smoliar, and H.J. Zhang, *Video and Image Processing in Multimedia Systems*, Kluwer Academic Publishers, Boston, 1996.

Semantic levels: expression and emotion. At the expressive level, perceptual data are organized into a group of new features—the expressive features—taking into account both spatial and temporal distributions of perceptual features. Expressive features reflect concepts that humans embody at a higher level of abstraction to achieve a more compact visual representation.

Combination rules are modeled as functions F acting over the perceptual feature set P and returning a score expressing the *degree of truth* by which the rule F holds. Hence, a rule F_j can be defined as

$$F_j : [0, 1]^n \rightarrow [0, 1]$$

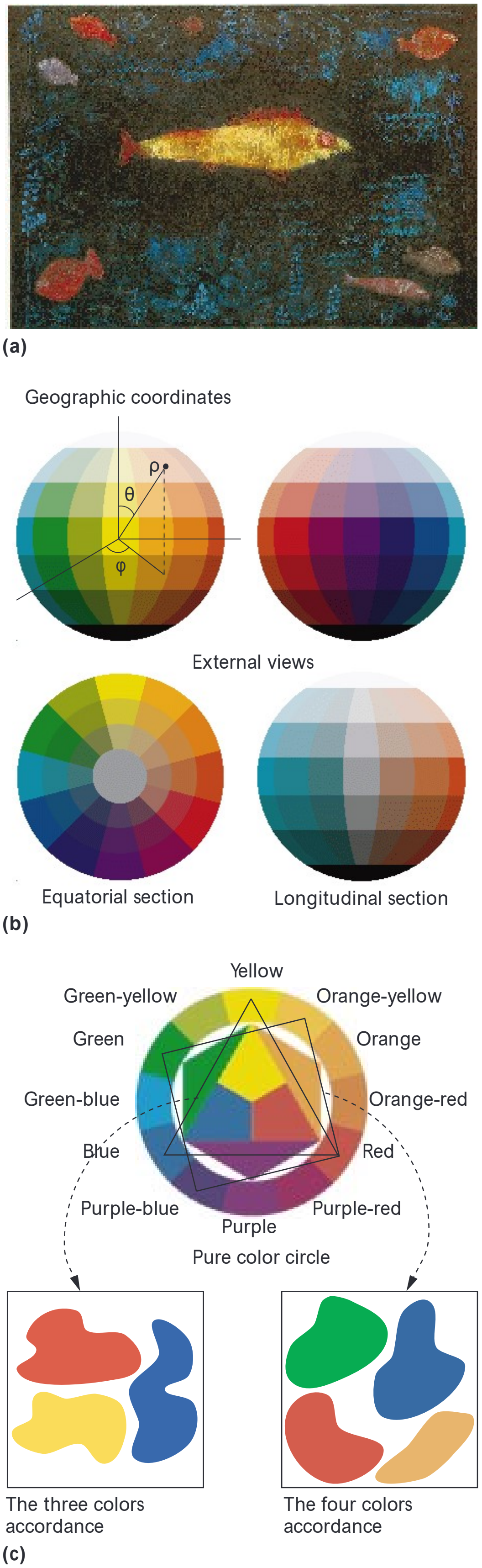
Operators of logical composition between rules extend the signification of the representation. We define these operators as

$$F_1 \wedge F_2 = \min(F_1, F_2) \qquad F_1 \vee F_2 = \max(F_1, F_2)$$

The expressive feature set $F = \{F_1, \ldots, F_m\}$ qualifies the content of the visual message at the expressive level.

A musical example lets us capture the distinction between the expressive and emotional levels. Assume that the laws of harmony and counterpoint characterize music composition at the expressive level. However, following these rules doesn't guarantee a pleasant result for the audience. In other words, musical fruition and understanding involves adopting aesthetic criteria that go beyond expression—that is, the syntax of meaning—and reach emotion as the ultimate semantics of meaning. (The art of J.S. Bach provides a remarkable example of how to reach musi-

Figure 1. (a) Contrasts and colors in an art image. (b) The Itten sphere. (c) The polygons generating three- and four-chromatic accordance.



cal beauty by a skillful adherence to the formal rules of 18th century music.) We formally construct the emotional level, at the top of our signification hierarchy, from the levels below—namely, the expressive and perceptual levels. With a notation similar to the one above, rules at the emotional level are represented through functions G acting over the set of perceptual and expressive features $P \times F$ and returning a score that expresses the degree of truth by which the rule G holds. Hence, a rule G_k can be defined as

$$G_k : [0, 1]^{n+m} \to [0, 1]$$

Operators of logical composition between rules can extend the representation's semantics. The set $G = \{G_k\}$ qualifies the content of the visual message at the emotional level.

Perceptual, expressive, and emotional features qualify the meaning of a visual message at different levels of signification. For all these levels, construction rules depend on the specific data domain to which they refer (such as movies, commercials, and TV news for videos; paintings, photographs, and trademarks for still images). Specifically, the expressive level features objective properties that generally depend on collective cultural backgrounds. The emotional level, on the contrary, relies on subjective elements such as individual cultural background and psychological moods.

Background theories

Here we present theories that provide a reference framework for developing expressive and emotional rules in the domains of art images and commercial videos.

Expression in art images. Among the many authors who recently addressed the psychology of art images, Arnheim discussed the relationships between artistic form and perceptive processes,[7] and Itten[8] formulated a theory about use of color in art and about the semantics it induces. Itten observed that color combinations induce effects such as harmony, disharmony, calmness, and excitement that artists consciously exploit in their paintings. Most of these effects relate to high-level chromatic patterns rather than to physical properties of single points of color (see, for example, Figure 1a). Such rules describe art paintings at the expressive level. The theory characterizes colors according to the categories of hue, luminance, and saturation. Twelve fundamental hues are chosen and each of them is varied through five levels