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 $P = \{\varphi_i\}$, $i = 1, \ldots, n$, each score $\varphi_i \in [0, 1]$ 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-

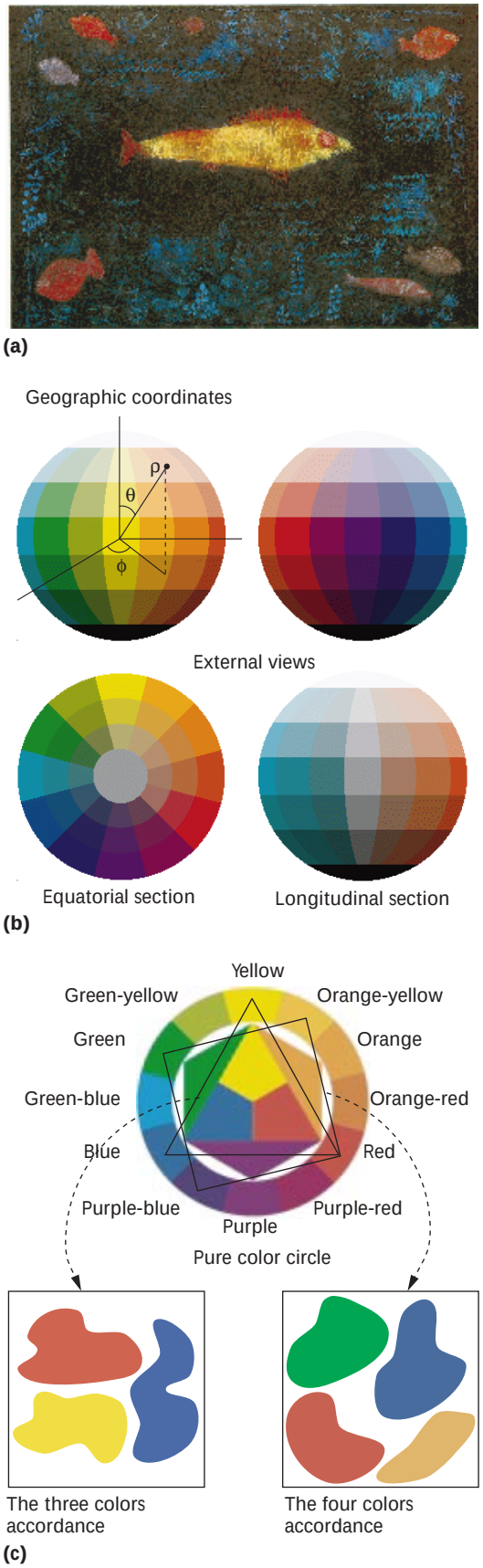


Figure 1. (a) Contrasts and colors in an art image. (b) The Itten sphere. (c) The polygons generating three- and four-chromatic accordance.

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

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

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

of luminance and three levels of saturation. These colors are arranged in a chromatic sphere called the Itten sphere (Figure 1b), such that perceptual contrasting colors have opposite coordinates with respect to the center of the sphere.

Analyzing the polar reference system, we can identify four different types of color contrasts: pure colors, light-dark, warm-cold, and saturated-unsaturated (quality). Psychological studies have suggested that, in western culture, red-orange environments induce a sense of warmth (yellow through red-purple are warm colors). Conversely, green-blue conveys a sensation of cold (yellow-green through purple are cold colors). Cold sensations can be emphasized by the contrast with a warm color or dampened by its coupling with a highly cold tint. The concept of harmonic accordance combines hues and tones to generate a stability effect onto the human eye. Harmony can be achieved by creating color accordances, generated on the sphere by connecting locations through regular polygons (see Figure 1c).

Courtesy of Mediaset

Figure 2. Some frames taken from a "Sergio Tacchini" commerical video. The use of colors and lines in each frame communicates dynamism. In this video, the effect is further enhanced by a high frequency of shots.

Emotion in art images. Color combinations also induce emotional effects. The mapping of low-level color primitives into emotions is quite complex. It must consider theories about the use of colors and cognitive models, and involve cultural and anthropological backgrounds. In other words, people with different cultures will likely share the same expressive level representation but not perceive the same emotions in front of the same color pattern.

Artists use color combinations unconsciously and consciously to produce optical and psychological sensations.[9] Warm colors attract the eye and grab attention of the observer more than cold colors. Cold sensations provided by a large region of cold color can be emphasized by the contrast with a warm color or dampened by its coupling with a highly cold tint. Similarly, small, close cold regions emphasize large, warm regions. Red communicates happiness, dynamism, and power. Orange, the warmest color, resembles the color of fire and thus communicates glory. Green communicates calmness and relaxation, and is the color of hope. Blue, a cold color, improves the dynamism of warm colors, suggesting gentleness, fairness, faithfulness, and virtue. Purple, a melancholy color, sometimes communicates fear. Brown generally serves as the background color for relaxing scenes. Differences in lightness determine a sense of plasticity and the perception of different planes of depth. The absence of contrasting hues and the presence of a single dominant color region inspire a sense of uneasiness strengthened by the presence of dark yellow and purple colors. The presence of regions in harmonic accordance communicates a sense of stability and joy. Calmness and quiet can be conveyed through combining complementary colors. In the presence of two noncomplementary colors, the human eye looks for the complementary of the observed color. This rouses a sense of anguish.

Another basic contribution to the semantics of an image relates to the presence and characteristics of notable lines. In fact, *line slope* is a key feature through which the artist communicates different emotions. For example, an oblique slope communicates dynamism and action, while a flat slope, such as a horizon, communicates calmness and relaxation. Figure 2 shows an example, illustrating how saturated colors can combine with sloped lines to communicate a sensation of dynamism in the observer.

Expression in commercials. Semiotics studies the meaning of symbols and signs. In semiotics, a sign represents anything that conveys meaning according to commonly accepted conventions. Semiotics therefore suggests that signs relate to their meaning according to the specific cultural background. Semioticians usually identify two dis-

tinct steps for the production of meaning:[10]

1. an abstract level formed by *narrative structures*, that is, structures including all those basic signs that create meaning and those values determined by sign combinations, and

2. a concrete level formed by *discourse structures*, describing the way in which the author uses narrative elements to create a story.

Representing the expressive and emotional level for commercial videos inherits some of the features introduced before for still images and adds new features related to the way frames are concatenated. Semiotics classifies advertising images—and specifically commercials—into four different categories (noted below) that relate to the narrative element.[11] Directors will use the main narrative signs of the video—camera breaks, colors, editing effects, rhythm, shot angles, and lines—in a peculiar way, depending on the specific video typology considered.

- Practical commercials emphasize the qualities of a product according to a common set of values. The advertiser describes the product in a familiar environment so that the audience naturally perceives it as useful. Camera takes are usually frontal, and transitions take place in a smooth and natural way. This implies choosing long dissolves for merging shots, and the prevalence of horizontal and vertical lines—giving the impression of relaxation and solidity.

- Critical commercials introduce a hierarchy of reference values. The product serves as the subject of the story, which focuses on the product's qualities through an apparently objective description of its features. The scene has to appear more realistic than reality itself. For this reason, the commercial has a minimum number of camera breaks. Due to smooth camera motions, the ever-changing colors in the background draw the audience's attention to the (constant) color of the product.

- Utopic commercials provide evidence that the product can succeed in critical tests. Here, the story doesn't follow a realistic plot. Rather, situations appear as in a dream. Mythical scenarios are chosen to present the product, shown to succeed in critical conditions often in a fantastic and unrealistic way. The director creates a movie-like atmosphere, with a set of dominant colors defining a closed-chromatic world and with all the traditional editing effects (cuts, dissolves) possibly taking place.

- Playful commercials emphasize the accordance between users' needs and product qualities. Here a manifest parody of the other typologies of ads takes place. The commercial clearly states to the audience that they're watching advertising material. Situations and places visibly differ from everyday life, and they're deformed in such a caricatural and grotesque fashion that the agreement between product qualities and purchaser's needs is often shown in an ironic way (such as an old woman driving a Ferrari). The director emphasizes the presence of the camera in the ad and uses all possible effects to stimulate the active participation of the audience. Also, everything looks strange and "false"—use of unnatural colors, improbable camera takes, and so on.

Emotion in commercials. The choice of a given combination of narrative signs affects both the expressive and emotional levels of signification. While dissolves communicate calmness and relaxation, cuts increase the video's dynamism (see again Figure 2). Sometimes in commercials, directors emphasize cuts by including a white frame between the end of the first frame and the beginning of the second, thus inducing a sense of shock in the observer. The semantics associated with editing effects thus relates to the use of cuts or dissolves and to the length of the shots in the video. If the video consists of a few long shots, the effect is that of calmness and relaxation. Alternatively, videos with many short shots separated by cuts induce a sense of action and dynamism in the audience.

Content representation and semantics-based retrieval

In the framework of our research, we used the theories presented in the previous section to qualify the content of a visual message at an intermediate semantic level. Specifically, we've exploited Itten's theory and semiotic principles to represent the visual content at the expressive level. Our work supports defining formal rules to qualify the effects of color, geometric, and dynamic feature combinations. We've also exploited psychological principles of visual communication to repre-

sent the visual content at the emotional level. In the next section, we use the semantic descriptors automatically extracted from art images and commercial videos for retrieval purposes.

Content representation for art images

Here we discuss the rules needed to represent the content of art images.

Expressive level. Exploiting Itten's model to qualify an image's chromatic content at an expressive level requires

1. segmenting the image into regions characterized by uniform colors, and
2. representing chromatic and spatial features of image regions.

We segment images by looking for clusters in the color space, then back-projecting cluster centroids in the feature space onto the image. Segmentation occurs through selecting the appropriate color space so that small feature distances correspond to similar colors in the perceptual domain. Adopting the International Commission on Illumination (Commission Internationale de L'Eclairage—CIE) L*u*v* space[12] accomplishes this.

Once segmented into regions, the image's region features can be described in terms of *intra-region* and *inter-region* properties. Intra-region properties include region color, warmth, hue, luminance, saturation, position, and size. Inter-region properties consist of hue, saturation, warmth, luminance contrast, and harmony. To manage the vagueness of chromatic properties, we use a fuzzy representation model to describe a generic property's value. Therefore, we can describe a generic property by introducing n reference values for that property. Then, considering n scores $(\varphi_1, \ldots \varphi_n)$, the i-th score measures the extent to which the region conforms to the i-th reference value. For instance, if we introduce three reference values for the luminance of a region (corresponding to dark, medium, and bright), the descriptor (0.0, 0.1, 0.9) represents a very bright region's luminance.

According to Itten's rules, these abstract properties must be translated into language sentences. In formal language theory notation, region formulas (ϕ) that characterize chromatic and arrangement properties of color patches represent these sentences through

$$\begin{aligned}\phi := \text{region} \mid \text{hue} = \lambda_h \mid \text{lum} = \lambda_l \mid \text{sat} = \lambda_s \mid \\ \mid \text{warmth} = \lambda_w \mid \text{size} = \lambda_S \mid \text{position} = \lambda_p \mid \\ \mid \text{Contrast}_\gamma(\phi_1, \phi_2) \mid \\ \mid \text{Harmony}\ (\phi_1, \ldots, \phi_n) \mid \phi_1 \wedge \phi_2 \mid \phi_1 \vee \phi_2\end{aligned}$$

where λ_γ is a feasible value for the measure γ, with $\gamma \in \{h, l, s, w\}$ denoting the attributes of hue, luminance, saturation, and warmth, respectively. We define semantic clauses in terms of a *fulfillment relation* $\models$ of a generic formula ϕ on a region R. The degree of truth by which ϕ is verified over R is expressed by a value ξ that's computed considering fuzzy descriptors of region properties. We code semantic clauses into a *model-checking* engine. Given a generic formula ϕ and an image I, the engine computes a score representing the degree of truth by which ϕ is verified over I (see the sidebar "Model-Checking Engine" on the next page).[13]

Emotional level. The psychological analysis of effects induced by images suggests that we can use only a limited subset of primary emotions to express the great variability of human emotions (secondary emotions). Explicitly defining the rules mapping expressive and perceptual features onto emotions would require us to build an overcomplicated model taking into account cultural and behavioral issues. Hence, we prefer to determine the relative relevance of each single feature through an adaptation process that fits a specific culture and fashion.

Explicitly, the definition of the rules mapping expressive and perceptual features onto emotions involves

1. *Identifying a set of primary emotions.* We identified four primary emotions, namely, action, relaxation, joy, and uneasiness, as the most relevant to express the interaction of humans with images. Referring to the general scheme presented above, secondary emotions such as fear and aggressiveness can be expressed as combinations of action and uneasiness.
2. *Identifying for each primary emotion a set of plausible inputs.* In our model, we expect contrasts of warmth and contrasts of hue to generate a sense of action and dynamism. The presence of lines with a high slope reinforces this.
3. *Defining a training set for the model.* Given a database of images, we use some images as templates of primary emotions. This training set adapts weights of perceptual features to

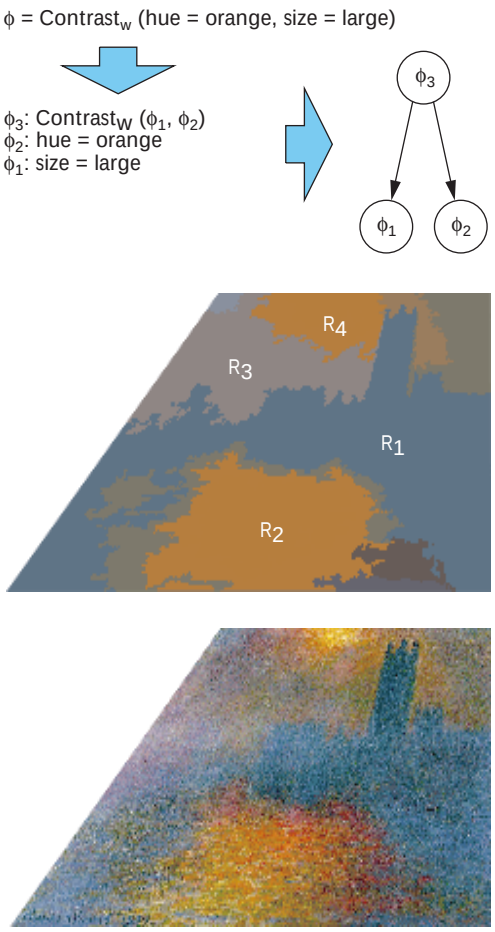


Figure A. The model-checking engine. From top to bottom: formula decomposition, segemented image, and original image.

Model-Checking Engine

A model-checking engine decides the satisfaction of a formula ϕ over an image I with a two-step process. First, the engine recursively decomposes formula ϕ into subformulas in a top-down manner. This allows the representation of ϕ with a tree in which leaf nodes represent intra-region specifications. Afterwards, the engine labels regions in the image description with the subformulas they satisfy in a bottom-up approach. The engine decides the satisfaction of region formulas by directly referring to the chromatic fuzzy descriptors. This first labeling level is then exploited to decide the satisfaction of composition formulas. In this labeling process, the engine first checks if the region contains pixels that can satisfy the subformula. If it can, the formula labels the region with the subformula and with the degree of truth by which the region is satisfied. When the degree of truth falls under a given minimum threshold, the engine drops the image from the candidate list.

In Figure A, the engine computes the degree by which the formula ϕ = Contrast$_w$ (hue = orange, size = large) holds over an image. The engine assumes that the image has just been segmented. For simplicity, the engine considers only four regions, namely R_1, R_2, R_3, and R_4. First, ϕ decomposes into subformulas:

ϕ_3 : Contrast$_w$ (ϕ_1, ϕ_2)
ϕ_2 : size = large
ϕ_1 : hue = orange

Then, according to a bottom-up approach, the engine labels regions with the subformulas they satisfy:

Step 1 Label R_2 and R_4 with ϕ_1
Step 2 Label R_1 with ϕ_2
Step 3 Label R_1, R_2, and R_4 with ϕ_3

At the end of this process, the engine labels three regions with ϕ_3, thus the original formula ϕ is satisfied over the image.

determine the relative relevance of each feature for the emotion induced in subjects of a certain cultural context.

We summarize the dependence between perceptual/expressive features and emotions below. We measure the degree of action communicated by an image as

$$G^I_{action} = g^I_1(F_{warmth-c}, F_{hue-c}, \varphi_{slanted}, w_{a1}, w_{a2}, w_{a3})$$

where $F_{warmth-c}$ and F_{hue-c} represent the expressive features measuring the presence of warmth and hue contrasts, and $\varphi_{slanted}$ the perceptual feature measuring the presence of slanted lines (the w_i's are the feature weights). We can measure the degree of relaxation an image communicates by

$$G^I_{relax} = g^I_2(F_{lum-c}, \varphi_{brown}, \varphi_{green}, w_{r1}, w_{r2}, w_{r3})$$

where F_{lum-c} represents the expressive features measuring the presence of luminance contrasts and φ_{brown} and φ_{green} the perceptual features measuring the presence of brown and green regions. The degree of joy communicated by an image can be measured as

$$G^I_{joy} = g^I_3(F_{harmony}, w_{j1})$$

where $F_{harmony}$ represents the expressive features measuring the presence of regions in harmonic accordance. Finally, we measure the degree of uneasiness an image communicates as

$$G^{I}_{uneas} = g^{I}_{4}(F_{n\text{–}hue\text{–}c}, \varphi_{yellow}, \varphi_{purple}, w_{u1}, w_{u2}, w_{u3})$$

where $F_{n\text{–}hue\text{–}c}$ represents the expressive features measuring the absence of contrasts of hue, and φ_{yellow} and φ_{purple} the perceptual features measuring the presence of yellow and purple regions.

Presently we use a linear mapping to model the g^I_i functions and a linear regression scheme to achieve adaptation of weights w_i based on a set of training examples.

Content representation for commercial videos

Now we'll present the content representation rules for commercial videos.

Expressive level. Semiotic principles permit establishing a link between the set of perceptual features and each of the four semiotic categories of commercial videos (practical, playful, utopic, and critical). This supports organizing a database of commercials on the basis of their semiotic content. Once a video has been segmented into shots, all perceptual features are extracted. Each feature's value is represented according to a fuzzy representation model. A score in [0, 1] is introduced for each feature to qualify the feature's presence in the video. Inter-frame features address

- the presence of cuts (φ_{cuts}) and dissolves ($\varphi_{dissolves}$),
- the presence or absence of colors recurring in many frames ($\varphi_{recurrent}$ ~1 and $\varphi_{recurrent}$ ~0, respectively), and
- the presence or absence of editing effects (φ_{edit} ~1 and φ_{edit} ~0, respectively).

Intra-frame features address the presence of

- horizontal and vertical or slanted lines ($\varphi_{hor/vert}$ ~1 and $\varphi_{hor/vert}$ ~0, respectively) and
- saturated or unsaturated colors ($\varphi_{saturated}$ ~1 and $\varphi_{saturated}$ ~0, respectively).

Table 1 summarizes how perceptual features are mapped onto the four semiotic categories. We adopted a linear mapping where table entries indicate expected feature weights and "×" indicates irrelevance.

We represent the video content through sentences qualifying the presence of critical, utopic, practical, and playful features. These sentences are expressed through formulas (Φ) defined as

$$\Phi := F_{practical} \geq k_1 \mid F_{playful} \geq k_2 \mid F_{utopic} \geq k_3 \mid F_{critical} \geq k_4 \mid \Phi_1 \wedge \Phi_2 \mid \Phi_1 \vee \Phi_2$$

where the k_i's represent fixed threshold values.

Table 1. Perceptual mapping onto semiotic categories.

Perceptual Features	Semiotic Categories			
	$F_{practical}$	$F_{playful}$	F_{utopic}	$F_{critical}$
$\varphi_{saturated}$	~0	~1	×	×
$\varphi_{recurrent}$	×	×	~1	~0
$\varphi_{hor/vert}$	~1	~0	×	~1
φ_{cuts}	×	~1	~1	×
$\varphi_{dissolves}$	~1	×	~1	×
φ_{edit}	×	×	×	~0

Emotional level. High-level properties of a commercial relate to the feelings it inspires in the observer. We've organized these characteristics in a hierarchical fashion. A first classification separates commercials that induce action from those that induce quietness. Each class then further splits to specify the kind of action and quietness. Explicitly, action in a commercial can induce two different feelings—suspense and excitement. Similarly, quietness can be further specified as relaxation and happiness.

In the following, we describe the mapping between emotional features and low-level perceptual features reflecting reference principles of visual communication. (See also Table 2 on the next page.)

- *Action*: The degree of action G^V_{action} for a video can be improved by the presence of red and purple. Action videos are often characterized by framings exhibiting high slanting. Short sequences are joined by cuts. The main distinguishing feature of these videos lies in the presence of a high degree of motion in the scene.
- *Excitement*: Videos that communicate excitement (with degree G^V_{excite}) typically feature short sequences joined through cuts.
- *Suspense*: To improve the degree of suspense $G^V_{suspense}$ in a video that communicates action, directors combine both long (φ_{long}) and short (φ_{short}) sequences in the video and join them through frequent cuts.

Table 2. Perceptual mapping onto emotional features.

Perceptual Features	Emotion Categories					
	Action	Excitement	Suspense	Quietness	Relaxation	Happiness
$\varphi_{\text{dissolves}}$	×	×	×	~1	~1	~1
φ_{cuts}	~1	~1	~1	~1	~1	~1
φ_{long}	×	×	~1	~1	~1	~1
φ_{short}	×	~1	~1	~1	~1	~1
φ_{motion}	~1	~1	~1	×	~0	~1
$\varphi_{\text{hor/vert}}$	~0	~0	~0	~1	~1	~1
φ_{red}	~1	~1	~1	×	×	×
φ_{orange}	×	×	×	~1	~1	~1
φ_{green}	×	×	×	~1	~1	~1
φ_{blue}	×	×	×	~1	~1	~1
φ_{purple}	~1	~1	~1	~0	~0	~0
φ_{white}	×	×	×	~1	~1	~1
φ_{black}	×	×	×	~ 0	~ 0	~ 0

"×" indicates irrelevance

- *Quietness*: The degree of quietness G^V_{quiet} for a video can be improved by the presence of blue, orange, green, and white colors, and lowered by the presence of black and purple. Quiet videos feature horizontal framings. A few long sequences might be present, possibly joined through dissolves.

- *Relaxation*: Videos that communicate relaxation (with degree G^V_{relax}) don't show relevant motion components.

- *Happiness*: Videos that communicate happiness (with degree G^V_{happy}) share the same features as quiet videos but also exhibit a relevant motion component.

Presently, we use a linear approximation to model emotional feature mapping, constructed by weight adaptation according to a linear regression scheme. Following the above scheme, the extent to which a video k conforms to one of the six classes is computed through six scores $\{G^V_j(k)\}^6_1$, each representing a weighted combination of low-level video features. To achieve good discrimination between the six classes of commercials requires that for a generic video belonging to category j, the ratio G^V_j/G^V_m must be high, at least for all the categories not a generalization of category j (we call these categories j-*opponent* categories).

Model validation and retrieval results

At the Visual Information Processing Lab of the University of Florence, we've used the compositional semantics framework expounded earlier to develop systems for retrieving still images and videos based on both expressive and emotional content.[13,14] We discuss the method used in these systems for extracting perceptual features from raw visual data in the sidebar "Perceptual Features for Still Images and Videos."

In this section, we present examples of retrieving art images and commercial videos according to the expressive level. You can find retrieval based on emotional features elsewhere.[15] We evaluated retrieval performance in terms of effectiveness, that is, a measure of the agreement between human evaluators and the system in ranking a test set of images according to their similarity to a query. Measuring effectiveness reliably requires a small image test set (typically 20 to 50 images). Given a sample query, we evaluated the agreement as the percentage of human evaluators who rank images in the same (or very close) position as the system does. We define effectiveness as

$$S_j(i) = \sum_{k=P_j(i)-\sigma_j(i)}^{k=P_j(i)+\sigma_j(i)} Q_j(i,k)$$

where i is an image from the test set, j denotes the sample query, and $\sigma_j(i)$ is the width of the window centered in the rank $P_j(i)$ assigned by the system. $Q_j(i, k)$ is the percentage of people who ranked the i-th image in a position between $P_j(i) - \sigma_j(i)$ and $P_j(i) + \sigma_j(i)$.

Perceptual Features for Still Images and Videos

Here we discuss the perceptual features for still images and videos.

Still images

An image's semantics relates to its color content and the presence of elements such as lines that induce dynamism and action.

Colors

Color cluster analysis helps segment images into color regions. We can obtain clustering in 3D space by using an improved version of the standard K-means algorithm, which avoids converging with nonoptimal solutions. This algorithm uses competitive learning as the basic technique for grouping points in the color space. An image's chromatic content can be expressed using a set of eight numbers normalized in [0, 1] denoting one color out of the set {red, orange, yellow, green, blue, purple, white, and black}. Each number quantifies the presence in the image of a region exhibiting the *i*-th color.

Lines

Detecting significant line slopes in an image can be accomplished by using the Hough transform to generate a line slope histogram (see Figure B). The feature $\varphi_{hor/vert} \in [0, 1]$ gives the ratio of horizontal and vertical lines with respect to the overall number of lines in the image.

Videos

Video analysis primarily aims to segment video, that is, to identify each shot's start and end points and the video's characterization through its most representative keyframes. Once a video has been fragmented into shots and video editing features have been extracted, each shot's content can be internally described by segmenting each shot's keyframe as described in the previous section of this sidebar.

Cuts

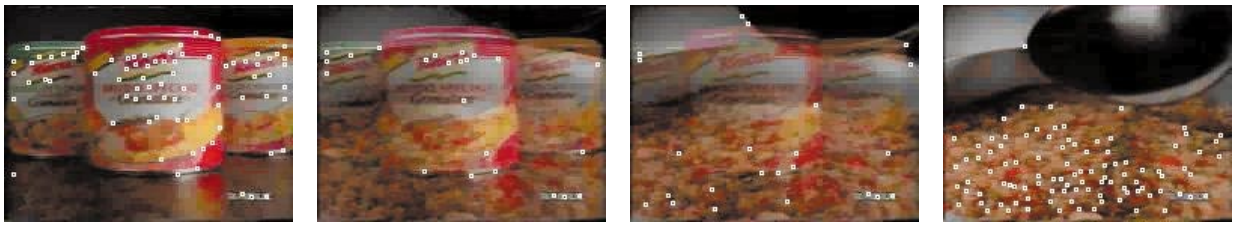
Rapid motion in the scene and sudden changes in lighting yield low correlation between contiguous frames, especially in cases adopting a high temporal subsampling rate. To avoid false cut detection, we studied a metric insensitive to such variations while reliably detecting true cuts. We partition each frame into nine subframes and represent each subframe by considering the color histograms in the hue, saturation, intensity (HSI) color space. We detect cuts by considering the volume of the difference of subframe histograms in two consecutive frames. The presence of a cut at frame *i* can be detected by thresholding the average volume value of for the nine subframes. After repeating the above procedure for all frames $i = 1 \ldots$ #frames, we simply obtain the overall feature related to the presence of cuts in a video as φ_{cuts} = #cuts/#frames, where $\varphi_{cuts} \in [0, 1]$.

Dissolves

The *dissolve* effect merges two sequences by partly overlapping them. Detecting dissolves in commercials proves particularly difficult because dissolves typically occur in a limited number of consecutive frames. Due to this peculiarity, existing approaches to detect dissolves (developed for movies) have shown poor performance. We use instead *corner statistics* to detect dissolves. During the editing effect, corners associated with the first sequence gradually disappear and those associated with the second sequence gradually appear.

continued on p. 48

Figure B. Perceptual features for still images. From left to right: video frame, computed edges, and line slope histogram.



(1) (2) (3) (4)

Figure C. Corners detected during a dissolve effect.

continued from p. 47

This yields a local minimum in the number of corners detected during the dissolve (see Figure C). An image corner is characterized by large and distinct values of the gradient auto-correlation matrix's eigenvalues. We evaluate the feature $\varphi_{\text{dissolves}} \in [0, 1]$ as $\varphi_{\text{dissolves}} = \#\text{dissolves}/\#\text{frames}$.

Motion

We analyze motion by tracking corners in a sequence. For each shot, we compute a feature $\varphi_{\text{motion}} \in [0, 1]$ that represents the average intensity of motion. $\varphi_{\text{motion}} = 0$ means that motion is absent during the sequence. Higher values of φ_{motion} indicate the presence of increasingly relevant motion components.

Inter-shot features

We represent color-related inter-shot features used in a video as $\varphi_{\text{recurrent}} \in [0, 1]$ (expressing the ratio between colors that recur in a high percentage of keyframes and the overall number of significant colors in the video sequence) and $\varphi_{\text{saturated}}$ (expressing the relative presence of saturated colors in the scene). Another relevant inter-shot characteristic—the *rhythm* of a sequence of shots—relates to shot duration and to the use of cuts and dissolves to join shots. We define the rhythm $r(i_1, i_2) \in [0, 1]$ of a video sequence over a frame interval $[i_1, i_2]$ as

$$r(i_2, i_2) = \frac{\#\text{cuts} + \#\text{dissolves}}{i_2 - i_1 + 1}$$

where #cuts and #dissolves are measured in the same interval. A simple feature measuring an entire sequence's internal rhythm is the average rhythm, which relates to the overall number of breaks $\varphi_{\text{edit}} = r(1, \#\text{frames})$.

Courtesy of WebMuseum

Figure 3. Images used to test system effectiveness for the representation of expressive content.

Image retrieval

For art image retrieval, we used a test set of 40 images, including Renaissance to contemporary painters. We measured retrieval effectiveness with reference to four different queries, addressing contrasts of luminance, warmth, saturation, and harmony.

We collected answers from 35 experts in the fine arts field. We asked them to rank database images according to each reference query. Figure 3 shows the test database images. Figure 4 shows the eight sets of top-ranked images the system retrieved in response to the four reference queries. (For more information about these images visit http://www.cineca.it/wm/.)

Figure 5 shows the plots of effectiveness S as a function of rankings. They show the agreement between the people interviewed and the system in ranking images according to the queries. Figure 5 shows only rankings from 1 to 8, since they represent agreement on the most representative images. The plots show a very large agreement

Figure 4. Top-ranked images according to queries for contrast of luminance (a), contrast of saturation (b), contrast of warmth (c), and harmonic accordance (d).

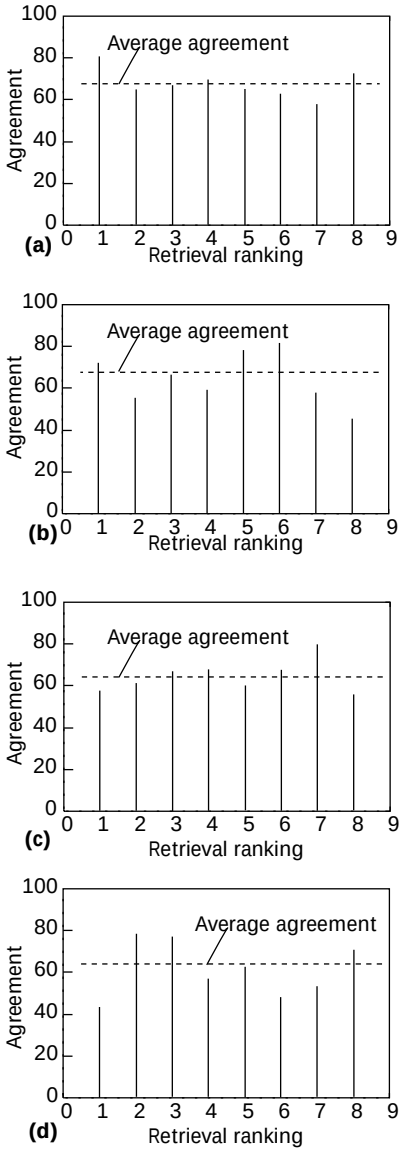


Figure 5. System effectiveness measured against experts for four different sample queries: constrast of luminance (a), constrast of saturation (b), contrast of warmth (c), and harmonic accordance (d).

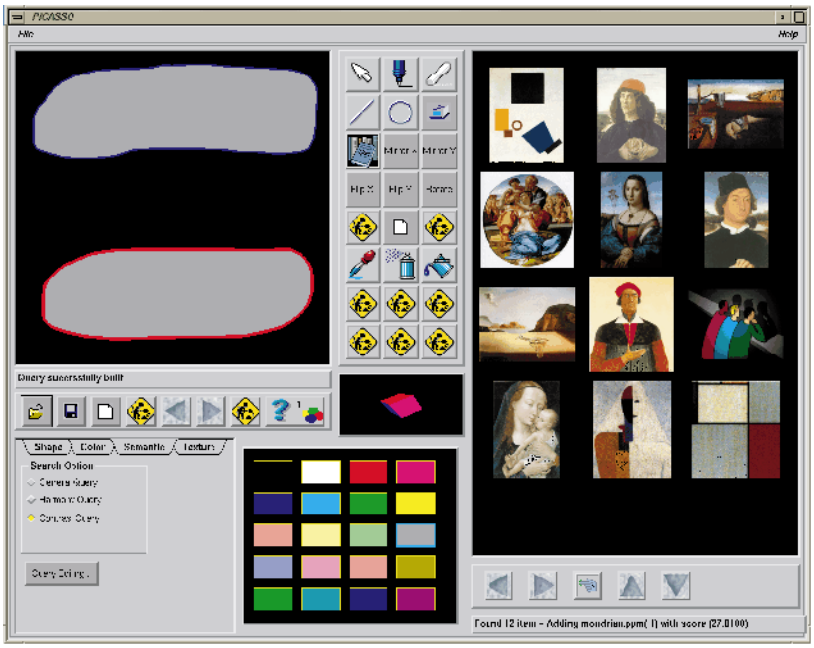


Figure 6. Results of a query for images with two large regions showing constrasting luminance.

between the experts and the system in assigning similarity rankings.

Figure 6 shows an example of retrieval according to the expressive level using a database of 477 images representing paintings from the 15th to the 20th century. Two dialog boxes define properties (hue and dimension) of the two sketched regions of Figure 6. The user may define the degree of truth (60 percent) assigned to the query. Figure 6 (right) shows the retrieved paintings. The 12 best-matched images all feature large regions with contrasting luminance. The top-ranked image represents an outstanding example of luminance contrast, featuring a black region over a white background. Images in the second, third, fifth, sixth, and seventh positions exempli-

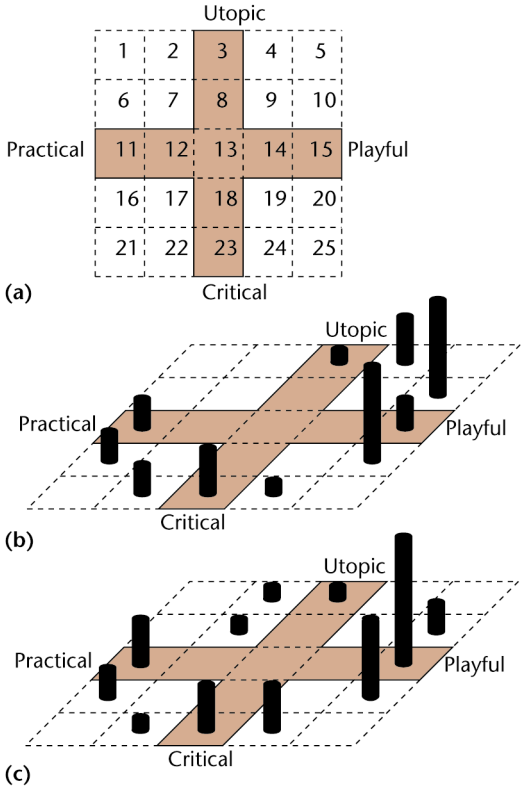


Figure 7. Representation of database commercials' semiotic features (a) along with the experts' (b) and system's (c) classification of them.

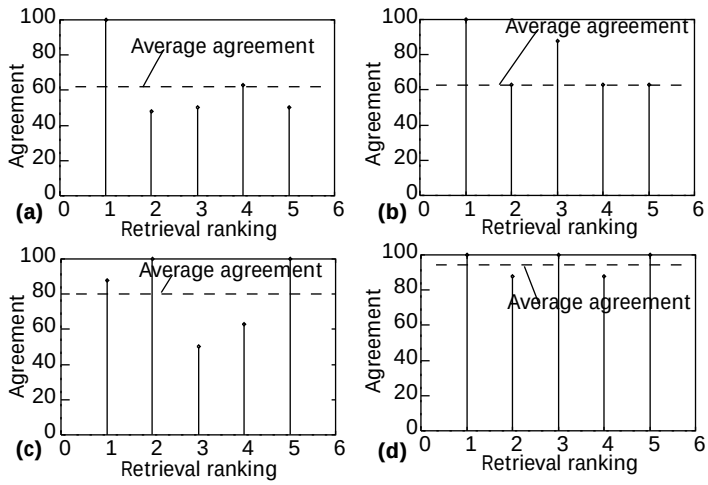


Figure 8. Plots of the agreement between system and experts' classification of commercials with reference to queries for purely practical (a), critical (b), utopic (c) and playful (d) commercials.

fy how contrast of luminance between large regions can be used to convey the perception of different planes of depth.

Video retrieval

We evaluated video retrieval effectiveness using a test set of 20 commercial videos. A team of five experts in the semiotic and marketing fields classified each video in terms of the four semiotic categories. We used the common representation of commercials categories based on the *semiotic square*. This instrument, first introduced by Greimas,[10] combines pairs of four semiotic objects with the same semantic level, according to three basic relationships: opposition, completion, and contradiction. Objects placed at opposite sides of the square are complementary. Figure 7a shows the practical-playful and critical-utopic diagonals of the semiotic square as coordinate axes. We asked the experts to classify the commercials by associating each commercial with a position on the square. To ease the classification task, 25 rectangular regions were identified in the square through a regular partition, which supports the definition of a video's three different degrees of conformity to a generic category. For instance, all videos classified in region 4 of Figure 7a feature a high degree of conformity to the utopic category and a medium conformity to the playful one.

Figure 7 shows how the experts (Figure 7b) and the system (Figure 7c) classified database commercials. Each region (cylinders of zero height aren't shown) in the square has a vertical cylinder, whose height is proportional to the percentage of commercials located in that region.

Figure 8 shows a more accurate measure of system effectiveness by displaying the agreement between the system and experts with reference to queries for purely practical, critical, utopic, and playful commercials. Each query considered the five top-ranked commercials. For a generic commercial, we measured the agreement between the system and experts as $A = 1 - d/d_{max}$, where d is the "city block" distance between the two blocks in the semiotic square where the system and the experts located the commercial. d_{max} is the maximum value of d, here 8. The best average agreement corresponds to the query for playful commercials, evidencing the effectiveness of the features used to model this category. The worst performance corresponds to queries for practical commercials. In this case, the system classified many commercials as practical, while the experts rated them as critical. Such a classification mismatch originates from the recurrent presence in critical commercials of foreground views of the promoted product. The experts can easily detect and recognize a foreground view, but the system can't detect this feature.

Figures 9 through 12 show examples of video retrieval according to the expressive level using a

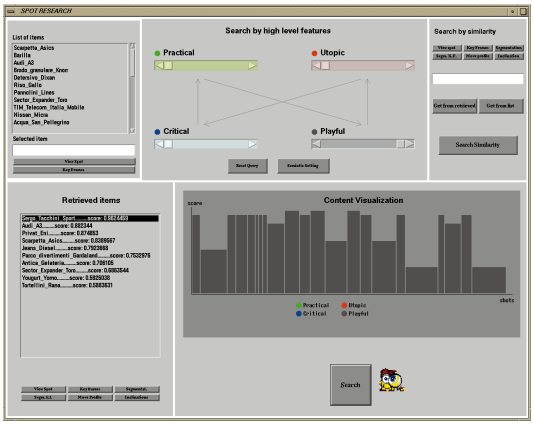
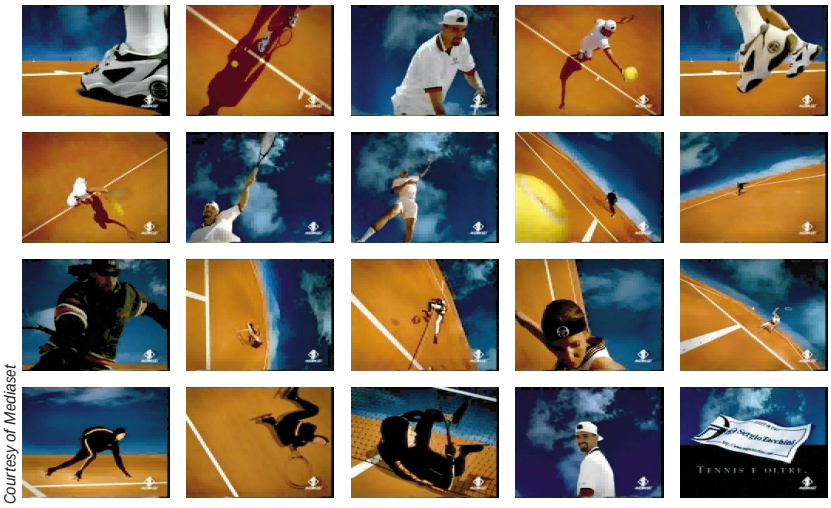


Figure 9. Retrieval of playful commercials.



Courtesy of Mediaset

Figure 10. Some of the keyframes for the first-ranked spot in Figure 9 ("Sergio Tacchini").

database of 131 commercials. Users can specify queries by defining the degree by which the retrieved commercials have to conform to the four semiotic categories. Figure 9 shows the output of the retrieval system in response to a query for purely playful commercials. At the right of the list of retrieved items, a bar graph displays the degree by which each shot of the top-ranked video—the white, thin vertical lines represent cuts and dissolves—belongs to the playful category. The top-ranked videos in this category all advertise products for "young and smart" people (sportswear, sport watches, blue jeans, and so on). These results aren't surprising, since they reflect the common marketing practice of targeting a commercial to a specific audience.

Figures 10 and 11 report some of the keyframes for the two top-ranked spots in Figure 9. The first best-ranked spot (advertising sportswear) presents all the typical features of a playful commercial, including a very fast rhythm, unorthodox camera takes, situations—like skating in a tennis court—reflecting semantic issues at a higher level than that of our computer analysis, and very saturated colors. Similar features appear in the second-ranked commercial. In Figure 11 notice the presence of quasi-identical keyframes (the close-ups of the man) typical of a nonlinear story. In this spot the camera frenetically switches between close-up views of the man and his dogs, almost never alternating details and global views like a utopic spot would do.

Figure 12 shows the output of the retrieval system in response to a query for purely critical commercials. In contrast to the previous example, the top-ranked videos obtained in response to the query all advertise typical home products. Figures

Courtesy of RAI

Figure 11. Some of the keyframes for the second-ranked spot in Figure 9 ("Audi A3").

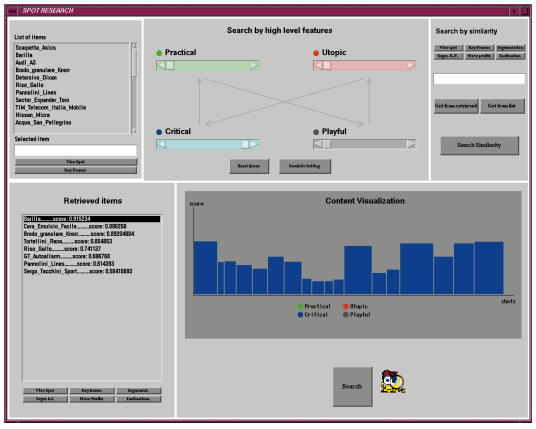


Figure 12. Retrieval of critical commercials.

Courtesy of RAI

Figure 13. Some of the keyframes for the first-ranked spot in Figure 12 ("Pasta Barilla").

Courtesy of Mediaset

Figure 14. Some of the keyframes for the second-ranked spot in Figure 12 ("Cera Emulsio").

13 and 14 show some relevant frames for the two top-ranked commercials in Figure 12.

Conclusions and future work

If, as the adage says, "an image is worth a thousand words," then every designer of multimedia retrieval systems is well aware that the converse also holds true. That is, many times a word can stand for a thousand images, since it represents a *class of equivalence* of objects, thus reflecting a higher semantic level than that of the objects themselves. We believe that future multimedia retrieval systems will have to support access to information at different semantic levels to reflect diverse application needs and user queries.

The research presented in this article attempts to introduce different levels of signification by a layered representation of visual knowledge. The most relevant insight we gained is that defining rules that capture visual meaning can be difficult, especially when dealing with general visual domains. As a good design practice, such rules should derive from specific domain characterizations, and possibly be refined and tailored to specific classes of users.

Future work will address experimenting with different application scenarios (such as TV news and movies), and complementing visual features and language sentences with textual and audio data. MM

Acknowledgments

We thank Bruno Bertelli and Laura Lombardi for useful discussions during the development of this work. We'd also like to acknowledge all those experts who provided their support in the performance tests.

References

1. T. Joseph and A. Cardenas, "PicQuery: A High-Level Query Language for Pictorial Database Management," *IEEE Trans. on Software Engineering*, Vol. 14, No. 5, May 1988, pp. 630-638.
2. N. Roussopolous, C. Faloutsos, and T. Sellis, "An Efficient Pictorial Database System for Pictorial Structured Query Language (PSQL)," *IEEE Trans. on Software Engineering*, Vol. 14, No. 5, May 1988, pp. 639-650.
3. M. Flickner et al., "Query by Image and Video Content: The QBIC System," *Computer*, Vol. 28, No. 9, Sept. 1995, pp. 310-315.
4. J.R. Smith and S.F. Chang, "VisualSeek: A Fully Automated Content-Based Image Query System," *Proc. ACM Multimedia 96*, ACM Press, New York, Nov. 1996.
5. D.A. White and Ramesh Jain, "Similarity Indexing with the SS-tree," *Proc. of the 12th Int'l Conf. on Data Engineering*, IEEE Computer Society Press, Los Alamitos, Calif., Feb. 1996, pp. 516-523.
6. R.J. Brachman and H.J. Levesque, eds., *Readings in Knowledge Representation,* Morgan Kaufmann, Los Altos, Calif., 1985.
7. R. Arnheim, *Art and Visual Perception: A Psychology of the Creative Eye*, Regents of the University of California, Palo Alto, Calif., 1954.
8. J. Itten, *Art of Color* (*Kunst der Farbe*), Otto Maier Verlag, Ravensburg, Germany, 1961 (in German).
9. C.R. Haas, *Advertising Practice (Pratique de la Publicité)*, Bordas, Paris, 1988 (in French).

10. A.J. Greimas, *Structural Semantics (Sémantique Structurale)*, Larousse, Paris, 1966 (in French).
11. J.-M. Floch, *Semiotics, Marketing, and Communication: Below the Signs, the Strategies (Sémiotique, Marketing et Communication: Sous les signes, les stratégies)*, University of France Press, Paris, 1990 (in French).
12. R.C. Carter and E.C. Carter, "CIELUV Color Difference Equations for Self-Luminous Displays," *Color Research and Applications*, Vol. 8, No. 4, 1983, pp. 252-553.
13. J.M. Corridoni, A. Del Bimbo, and P. Pala, "Sensations and Psychological Effects in Color Image Database," *ACM Multimedia Systems*, Vol. 7, No. 3, May 1999, pp. 175-183.
14. M. Caliani et al., "Commercial Video Retrieval by Induced Semantics," *Proc. IEEE Int'l Work. on Content-Based Access of Images and Video Databases*, IEEE CS Press, Los Alamitos, Calif., Jan. 1998, pp. 72-80.
15. C. Colombo, A. Del Bimbo, and P. Pala "Retrieval of Commercials by Video Semantics," *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR 98)*, IEEE CS Press, Los Alamitos, Calif., June 1998, pp. 572-577.

Carlo Colombo is an assistant professor in the Department of Systems and Informatics at the University of Florence, Italy. His main research activities are in the field of computer vision, with specific interests in image and video analysis, human-machine interfaces, robotics, and multimedia. He holds an MS in electronic engineering from the University of Florence, Italy (1992) and a PhD in robotics from the Sant'Anna School of University Studies and Doctoral Research, Pisa, Italy (1996). He is a member of IEEE and the International Association for Pattern Recognition (IAPR), and presently serves as secretary to the IAPR Italian Chapter.

Alberto Del Bimbo is a professor in the Department of Systems and Informatics at the University of Florence, Italy. His scientific interests cover image sequence analysis, shape and object recognition, image databases and multimedia, visual languages, and advanced man-machine interaction. He earned his MS in electronic engineering at the University of Florence, Italy in 1977. He is presently an associate editor of *Pattern Recognition*, *Journal of Visual Languages and Computing*, and *IEEE Transactions on Multimedia*. He is a member of IEEE and IAPR. He presently serves as the Chairman of the Italian Chapter of IAPR. He is the general chair of IEEE Int'l. Conference on Multimedia Computing and Systems (ICMCS 99).

Pietro Pala received his MS in electronic engineering at the University of Florence, Italy, in 1994. In 1998, he received his PhD in information science from the same university. Currently he is a research scientist at the Department of Systems and Informatics at the University of Florence. His current research interests include recognition, image databases, neural networks, and related applications.

Contact the authors at the Department of Systems and Informatics, University of Florence, Via Santa Marta 3, I-50139, Florence, Italy, e-mail {columbus,delbimbo, pala}@dsi.unifi.it.