RC23353 (W0410-001) October 1, 2004
Computer Science

IBM Research Report

Semantic-Friendly Indexing and Quering of Images Based on the Extraction of the Objective Semantic Cues

Aleksandra Mojsilovic, José Gomes, Bernice Rogowitz
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Semantic-friendly Indexing and Quering of Images Based on the Extraction of the Objective Semantic Cues

Aleksandra Mojsilović, José Gomes and Bernice Rogowitz
IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532
{aleksand@us.ibm.com, josegome@us.ibm.com, rogowitz@us.ibm.com}

Abstract: *Abstract image semantics resists all forms of modeling, very much like any kind of intelligence does. However, in order to develop more satisfying image navigation systems, we need tools to construct a semantic bridge between the user and the database. In this paper we present an image indexing scheme and a query language, which allow the user to introduce cognitive dimension to the search. At an abstract level, this approach consists of: 1) learning the "natural language" that humans speak to communicate their semantic experience of images, 2) understanding the relationships between this language and objective measurable image attributes, and then 3) developing corresponding feature extraction schemes.*

More precisely, we have conducted a number of subjective experiments in which we asked human subjects to group images, and then explain verbally why they did so. The results of this study indicated that a part of the abstraction involved in image interpretation is often driven by semantic categories, which can be broken into more tangible semantic entities, i.e. objective semantic indicators. By analyzing our experimental data, we have identified some candidate semantic categories (i.e. portraits, people, crowds, cityscapes, landscapes, etc.) and their underlying semantic indicators (i.e. skin, sky, water, object, etc.). These experiments also helped us derive important low-level image descriptors, accounting for our perception of these indicators.

We have then used these findings to develop an image feature extraction and indexing scheme. In particular, our feature set has been carefully designed to match the way humans communicate image meaning. This led us to the development of a "semantic-friendly" query language for browsing and searching diverse collections of images.

We have implemented our approach into an Internet search engine, and tested it on a large number of images. The results we obtained are very promising.

Keywords: image semantic categorization, image browsing and retrieval, color naming, perceptual features

Corresponding author:
Aleksandra Mojsilovic
IBM T. J. Watson Research Center
19 Skyline Drive, Room 4S-A10
Hawthorne, NY 10532
phone: (914) 784-7429
fax: (914) 784-7667
email: aleksand@us.im.com

I INTRODUCTION

The number of digital image libraries is growing rapidly, driving the need to develop better, more intuitive tools for searching, navigating and browsing image collections. Organizing the contents of the digital library semantically is an emerging method for achieving this goal. Due to great difficulties in recognizing and classifying images at a general level, not much success has been achieved in identifying high-level semantics for image browsing and retrieval. Therefore, the image retrieval systems have been typically constrained to the use of low-level pictorial properties in a domain specific application. Many such applications have been developed successfully. Examples include: finding "objectionable" images on the Internet [1], [2], face and people detection [3], [4], classification into city versus landscape [5], indoor-outdoor classification [6], retrieval of color patterns for digital catalogs [7], and shape retrieval [8], [56]. As a result of applying fixed combinations of features for image classification the capability of these systems to search general broad-content databases is limited considerably. Many researchers have recognized this problem and proposed semantic-sensitive approaches to the problem of searching general-purpose image databases [9], [16], [17]. Hence, although great progress has been made since the early image retrieval and browsing systems [10]-[14], (for an excellent review of the early years see [15]), none of the existing methods fully captures enough of the semantic-related information to be used as a navigation tool in a general broad-content, image-rich database. Why is that so?

1. 1 Motivation: The "ingredients" of semantics

Our inability to capture "image semantics" comes from the incompatibility of the information computed from the image data directly and our subjective interpretation of the same data. Human vision incorporates numerous mechanisms, whose purpose is to continuously guide our perception through the constant flow of visual information. Each of these mechanisms is a "building block" in our semantic representation. These capabilities of human vision have been studied extensively, starting from the early Gestalt psychologists [18], through the theories of texture perception [19], to the studies in computer vision [20], [21], [30]. Although a representation composed of these perceptual groupings and their pictorial properties alone cannot capture image semantics, the very same representation is indirectly related to semantics. This motivates our work, as we will attempt to design an empirical computational model, which, in a fairly simple manner, respects these mechanisms, follows some of the internal processes we have knowledge of (or at least can estimate to a certain extent), and incorporates perceptually driven features to capture some valuable indicators of image content. In that process we will try to understand what are the important cues in our perception of some elementary semantic templates (for instance, cues like "sky", "water" or "grass" are important in identifying "landscapes"). We will try to understand what attributes and pictorial features are characteristic for certain cues, and develop algorithms to quantify these attributes (as an example, the cue "sky" can be described as a "uniform, upper, blue region", which then implies the use of a segmentation technique, texture features and a color-naming algorithm to capture these properties). In such an approach, these "measurable" cues can be seen as *semantic indicators*, which can be further combined, not to understand images in the way humans do, but to assist retrieval and browsing in an image-rich environment. Also, as we will show in the paper, an image retrieval system that takes an advantage of such an approach, indirectly relies on the cognitive capabilities of the user.

1.2 Our approach to modeling image semantics

A descriptive model for the formation of image semantic information, can be summarized as follows:

Level 1: *Abstract semantics*, which contributes to our interpretation of the scene (e.g. happy person, scary photograph).

Level 2: *Semantic templates* (i.e. semantic categories), which constitute our accumulated semantic knowledge. The examples are infinite, landscape photographs, portraits, cartoons, newspaper adds, still life, objects/catalog images, etc.

Level 3: *Semantic indicators*, or *semantic cues*, (i.e. image elements, their composition and relationships that are characteristic for certain semantic categories). For example, a large human face is the most important feature in our perception of human portraits.

Level 4: *Low-level primitives* and their *perceptual groupings*, i.e. measurable image attributes and their relationships that describe certain semantic indicators. Let us take the "Portrait" semantic template as an example, and the dominant face as its main semantic cue. The foreground/background separation property of human visual system contributes to our perception of the central object (i.e. "there is a dominant object in the scene"). The ability to detect lines, segments, curves, and boundaries, and the ability to extract the built-in geometric prototypes contribute to our shape perception ("the object is oval"). Our color-perception and color categorization mechanism add another piece of information to that representation ("the object has the color of human skin").

In a narrow application domain we can model each of these levels reliably, as we know the users expectations of the database (Level 1), the important classes of images in the particular domain (Level 2) and the important properties of each class (Level 3). However, in a broad-content database, the semantic gap has proven unbridgeable. The closest we can get in capturing the meaning of an image is to understand some fairly general semantic templates and their "cues", and implement feature detection algorithms to capture these properties. Although very far from the direct semantic interpretation, this representation provides a toolkit for breaking the image into the semantic cues (such as "face", "sky", or "red oval object"). The ability to detect and combine these cues facilitates the design of natural similarity metrics, query languages and user interfaces, thus minimizing the gap between the recorded image data and our understanding of the scene.

Based on the analysis above our work focuses on: a) understanding some important semantic categories that drive our visual perception, b) studying human similarity judgments to extract meaningful, discriminating attributes of these semantic categories, and c) designing and implementing feature extraction algorithms, similarity metrics and query language, that provide useful semantic cues for image search and navigation.

Our work had three major phases: 1) experiments and modeling, 2) algorithmic design, and 3) implementation. In the first part we have conducted several subjective experiments (Section II). These experiments helped us identify several candidate semantic categories human observers rely upon in organizing photographic images and measuring their similarity. Since image similarity is such a complex judgment we do not consider these categories as final and immutable set, we see them as a set of reference points in our perception of image content. In the second phase, we have used these findings to develop and implement a set of image features related to the semantics of the categories and cues we had identified (Sections III-VII). In particular, these features have been carefully designed to match the way humans communicate the meaning of an image, and thus led us to the development of a natural "semantic-friendly" language for indexing and querying image databases (Section VIII). To evaluate our approach, in the third phase we have implemented our algorithms into an Internet search engine, and tested it on a large number of images (Section IX). The limitations of the method, additional comments and concluding remarks can be found in Section X.

II EXPERIMENTS

In order to gain insight into how humans categorize images and perceive their semantics, we have conducted a set of subjective experiments. Following the hierarchy of image semantics, described in Section 1.2, the goal of these experiments was to identify some elementary and fairly general semantic categories (templates), to identify semantic cues that contribute to the perception of these categories, and understand how they are combined in interpreting image semantics. Most importantly, the experiments were designed to help us "break down" the semantic cues into the combinations of low-level pictorial features, and learn how to use them operationally to manage diverse collections of images. This section briefly describes the subjective experiments and their findings. More details on these experiments can be found in [22].

2.1 Stimuli, experiments and data analysis

Our work was a continuation of the previous study [23], which describes two experimental procedures designed to judge the similarity of 97 images in a photographic database. The multidimensional scaling analysis (MDS) of the experimental data discovered two important dimensions in the human perception of photographic images: *natural vs. man-made*, and *humans vs. non-human*. In a further qualitative analysis the images appeared to cluster into broad semantic categories such as "portraits," "natural scenes", "fruits and vegetables," etc. The objective of our work was to continue from these findings, enhance our understanding of these perceptual categories, and devise a similarity model where each perceptual category will be described in terms of the low-level features.

We used 196 images, selected to include a wide range of topics (people, nature, buildings, texture, objects, indoor scenes, animals, etc). The first 97 images (*Set 1*) were identical to those used in the previous study [23]. The second set (*Set 2*) contained 99 images especially selected to be as different as possible from the first set [23].

As a starting point in determining the candidate semantic categories, we used the similarity data from [23] and performed hierarchical cluster analysis (HCA). We found that the perceptual distances (i.e. image similarity judgments from the perspective of human observers) between the 97 images were indeed organized into clusters. To confirm the stability of the most important clusters in the HCA solution we split the original data in several ways and performed separate HCAs for each part. The clusters that remained stable for various solutions determined the *preliminary categories* (PC). We have then conducted the following two experiments: Image Similarity Experiment, aimed at developing a set of candidate semantic categories in the domain of photographic images, and Category Naming and Description experiment, aimed at refining these candidate semantic categories, deriving a semantic name for each category, and identifying a set of low-level features that describe each category.

2.1.1 Image Similarity Experiment

The purpose of this experiment was to collect a second set of similarity judgments, which would allow us to examine the perceptual validity and reliability of the initial clusters and further refine them into candidate semantic categories. For this experiment, we printed 97 thumbnails of all the images from Set 1, organized into clusters, and glued them on a tabletop. We also printed thumbnails of the 99 images from Set 2. Twelve subjects participated in the experiment. The subjects were asked to assign each image from Set 2 into one of the preliminary categories, placing them onto the tabletop so that the most similar images were near each other. We provided no instructions concerning the characteristics on which the similarity judgments were to be made, since this was the very information we were trying to discover. The subjects were not allowed to change the initial categories - these images were glued to the tabletop and could not be moved. However, the subjects were allowed to do whatever they liked with the new images. They could change their assignments during the experiment, move images from one category into another, or start their own categories.

The first step in the data analysis was to compute the similarity matrix Δ_{S_2} for the images from Set 2. The matrix entry $\Delta_{S2}(i, j)$ represents the number of times images *i* and *j* occurred in the same category. This matrix was used as an input to the multidimensional scaling algorithm. The next step was to compute the similarity matrix $\Delta_{S2,PC}$ for both the images from Set 2 and preliminary categories. This matrix was used as an input to the HCA to determine a new set of clusters, representing some candidate semantic categories in the human perception of image similarity.

2.1.2 Category Naming and Description Experiment

Testing human perception is an extremely difficult task, and we were aware that the clusters we identified as the candidate semantic categories might have reflected our decisions in designing the experiment. Therefore, to refine the categories and determine whether they were semantically distinct we performed another experiment. In this experiment, we asked the subjects to give names to

the categories identified in the Image Similarity Experiment. To further delineate the categories, and to identify high-level image features that discriminate the categories perceptually, the subjects were also asked to provide written descriptions for each of the categories and list their most important features. This experiment was helpful in many different ways. First, it was used to assess the robustness of the categories and test whether people see them in a consistent manner. For example, if the HCA revealed two separate categories, yet the names and descriptors were indistinguishable, we merged the categories. Conversely, if the HCA merged two sub-categories, yet the observers used distinctly different names and descriptors to identify them, we kept them as two separate categories. And finally, as will be described later, the written explanations were immensely valuable in determining pictorial features that best capture the semantics of each category.

Based on the semantic names the observers assigned to each cluster, we identified the following candidate categories (the categories are shown in Fig. 1, organized as much as possible to resemble the 2D MDS configuration):

C1: Portraits. Portraits and close-ups of people. A common attribute for all images in this group is a dominant human face.
C2a: People "outdoors". Images of people, mainly taken outdoors from a medium viewing distance.
C2b: People "indoors". Images of people, mainly taken indoors from a medium viewing distance.
C3: Outdoor scenes with people. Images of people taken from a large viewing distance. People are shown in an outdoor environment, and are quite small relative to the image
C4: Crowds of people. Images showing large groups of people on a complex background.
C5: Cityscapes. Images of urban life, with typical high spatial frequencies and strong angular patterns.
C6: Outdoor architecture. Images of buildings, bridges, architectural details standing on their own (as opposed to being in a cityscape).
C7: Techno-scenes". Outdoor pictures that include man-made environments and scenes from everyday life, primarily in a cityscape.
C8a: Objects indoors. Images of (dominant) man-made object indoors.
C8b: Indoor scenes with objects. Images of man-made environments.
C9: Objects outdoors. Images of man-made objects outdoors.
C10: Waterscapes with human influence.
C11: Landscapes with human influence.
C12: Waterscapes.
C13: Landscapes with mountains. Images where mountain is the primary feature.
C14: Sky/Clouds. Images with predominant sky/clouds texture.
C15: Winter and snow. Images representing scenes with snow.
C16: Green landscapes and greenery. Natural scenes with an overwhelming impression of green as the primary color.
C17: Landscapes with fields and foliage. Images of trees, fields and woods
C18: Plants, flowers, fruits and vegetables. This category can be further divided into *C18a* representing close-ups of flowers, fruits or vegetables, *C18b* showing plants on a smaller scale, and *C18c* representing images of plants pictured indoors, under artificial illumination, or arranged and organized by humans.
C19:Animals and wildlife.
C20: Textures, patterns and close-ups.

Since image similarity is such a complex judgment, these categories should not be considered as the final and immutable set, we see them only as a set of reference points in our perception of image content. As the experimental decisions represented only a partial (and subjective) view into the image world, and since the experimental set included only 200 images, we are not making an attempt in claiming the generality of this result. However, as we will demonstrate later in this work, the categories we had identified have proven to be a useful starting point in capturing and modeling some elementary aspects of image semantics.

2.1.3 The Analysis of the verbal descriptors: Identifying "semantic ingredients" for our set of "semantic templates"

Although the individual subjects used different verbal descriptors to characterize different categories, there were many consistent trends. We observed that certain image elements have dominating influence in identifying some categories. For example, in the "nature" categories, "water", "sky/clouds", "snow", "grass" and "mountains" emerged as very important cues. The same held true for the images with people - our subjects were extremely sensitive to the presence of people in the image, even if the image was dominated by a landscape, object, man-made structure. We also discovered that color composition and color features played an important role in comparing natural scenes, but were seldom used to describe images of people, man-made objects and environments. Within these

categories, spatial organization, spatial frequency and shape features mainly influence similarity judgments. Strong hues (such as bright red, yellow, lime green, pink, etc.) in combination with spatial properties, shape features, straight lines, straight boundaries, geometry and overall color composition were used to describe man-made objects in the picture. On the other hand, regions in the landscape and nature images were usually categorized as having rigid boundaries and "random" edges.

Fig. 1: The set of candidate semantic categories.

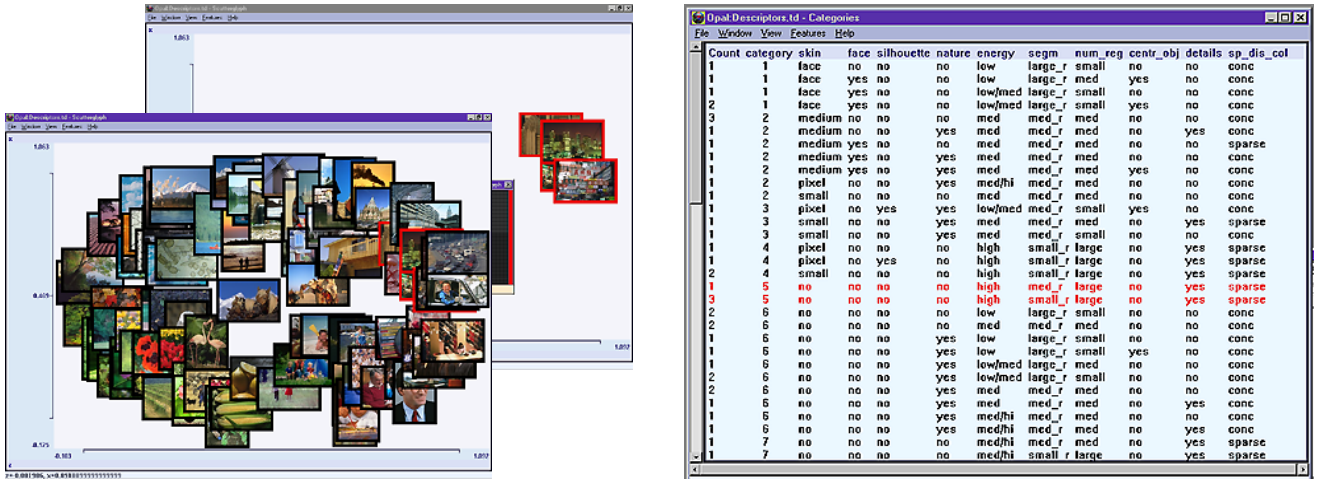
2.2 The connection to the low-level primitives: Deriving the pictorial features that describe some important "semantic cues"

Having identified some elementary candidate semantic categories -- *semantic templates* -- and discovered important semantic cues in the perception of these categories – *semantic ingredients* -- the next step is to model these cues so they can be used operationally in image retrieval and browsing applications. Unlike the previous work, which uses the low-level primitives (e.g. color, color layout, texture and shape) to represent information about the semantic meaning [7], [23], [24], [25], [26], we focused on the descriptions provided by the subjects in our experiments, and analyzed them with the following question in mind: *Is it possible to find a set of low-level features and their organization capable of modeling these semantic cues, and ultimately the semantic templates?*

As a starting point we used the written descriptions of the categories (gathered in the second experiment) and devised a list of verbal descriptors people found crucial in distinguishing the categories. We have then translated these descriptors into calculable image-processing features. For example, "image consisting primarily of a human face, with little or no background scene", used to describe category "Portraits", in the "image-processing language" corresponds to a "dominant, large, skin colored region". Or, "busy scene", used to describe category "Crowded scenes with people", in the "image-processing language" corresponds to "high spatial frequencies". We have further expanded the list by adding some features we considered useful, producing a list of over 40 image-processing features, called *the complete feature set* (CFS). As an illustration, here we list some features from the CFS: "number of regions after segmentation" (large, medium, small, one region), "image energy" (high, medium, low frequencies), "regularity" (regular, irregular), "existence of the central object" (yes, no), "edge distribution" (regular/directional, regular/nondirectional, irregular/directional, etc.), "color composition" (bright, dark, saturated, pale, gray overtones, etc.), "objects with bright color" (yes, no), "blobs of bright color" (yes, no), "spatial distribution of dominant colors" (sparse, concentrated), "presence of geometric structures" (yes, no), "number of edges" (large, medium, small, no edges), "corners" (yes, no), "straight lines" (occasional, defining an object, no straight lines), etc.

To find which of these features correlate with the semantics of each category, we used the Opal visualization package, which integrates numerous, linked views of tabular data, with automatic color brushing between the visualizations [27]. We used Opal to mine the experimental data and compare it to the image-processing descriptors for a set of 100 images. Specifically, for each category we were interested in finding a feature combination that discriminates that category against all the other images. Fig. 2 illustrates one visual query from Opal. The left view in Fig. 2 represents a *scatter glyph plot* of the image set, with images from the "Cityscapes" category selected with red. The coloring is linked to all the other views in the application. The values of the features in the complete feature set, for all the images in the scatter glyph are shown in the *category table window* (the right view in Fig. 2). This view shows that the feature combination colored in red is unique for our selection and that the following rule discriminates the category "Cityscapes" :

Skin = no skin, *Face* = no face, *Nature* = no nature, *Number of edges* = large, *Details* = yes,*Energy* = high, *Number of regions* = large, *Region size* = small, *Central object* = no central object,



Count	category	skin	face	silhouette	nature	energy	segm	num_reg	centr_obj	details	sp_dis_col
1	1	face	no	no	no	low	large_r	small	no	no	conc
1	1	face	yes	no	no	low	large_r	med	yes	no	conc
1	1	face	yes	no	no	low/med	large_r	small	no	no	conc
2	1	face	yes	no	no	low/med	large_r	small	yes	no	conc
3	2	medium	no	no	no	med	med_r	med	no	no	conc
1	2	medium	no	no	yes	med	med_r	med	no	yes	conc
1	2	medium	yes	no	no	med	med_r	med	no	no	sparse
1	2	medium	yes	no	yes	med	med_r	med	no	no	conc
1	2	medium	yes	no	yes	med	med_r	med	yes	no	conc
1	2	pixel	no	no	yes	med/hi	med_r	med	no	no	conc
1	2	small	no	no	no	med	med_r	med	no	no	conc
1	3	pixel	no	yes	yes	low/med	med_r	small	yes	no	conc
1	3	small	no	no	yes	med	med_r	med	no	yes	sparse
1	3	small	no	no	yes	med	med_r	small	no	no	conc
1	4	pixel	no	no	no	high	small_r	large	no	yes	sparse
1	4	pixel	no	yes	no	high	small_r	large	no	yes	sparse
2	4	small	no	no	no	high	small_r	large	no	yes	sparse
1	5	no	no	no	no	high	med_r	large	no	yes	sparse
3	5	no	no	no	no	high	small_r	large	no	yes	sparse
2	6	no	no	no	no	low	large_r	small	no	no	conc
2	6	no	no	no	no	med	med_r	med	no	no	conc
1	6	no	no	no	yes	low	large_r	small	no	no	conc
1	6	no	no	no	yes	low	large_r	small	yes	no	conc
1	6	no	no	no	yes	low/med	large_r	med	no	no	conc
2	6	no	no	no	yes	low/med	large_r	small	no	no	conc
2	6	no	no	no	yes	med	med_r	med	no	no	conc
1	6	no	no	no	yes	med	med_r	med	no	yes	conc
1	6	no	no	no	yes	med/hi	med_r	med	no	no	conc
1	6	no	no	no	yes	med/hi	med_r	med	no	yes	conc
1	7	no	no	no	no	med/hi	med_r	med	no	yes	sparse
1	7	no	no	no	no	med/hi	small_r	large	no	yes	sparse

Fig. 2: Opal visualization.

Similar analysis is performed for all other categories. We have also discovered that within a certain category, not all the features are equally important. For example, all images in the "Cityscapes" category have high spatial frequencies, many details, dominant brown/gray overtones, and segmentation yields large number of small regions. These features are thus considered as the *required features* for the "Cityscapes" category. On the other hand, most of the images from this category (but not all of them) have straight lines or regions with regular geometry, originating from the man-made objects in the scene. Or, although the dominant colors are primarily brow/gray/dark, many images have blobs of saturated colors, again because of the man-made objects in the scene. Therefore, straight lines, geometry and blobs of saturated color are considered as the *frequently occurring features* for the "Citycsapes" category.

III FEATURE EXTRACTION: THE HIGH-LEVEL VIEW

By now we had conducted subjective experiments that helped us identify several elementary semantic templates involved in our perception of image content. By analyzing the subjective data for each template, we have gained some understanding of the visual cues involved in the perception of these templates, and identified the basic primitives and perceptual mechanisms that "translate" these primitives into the determining characteristics of each cue. To model the elementary cues we have then encoded these processes into the calculable image features and the relationships among them, thus providing an indirect link to image content.

In this section we give a high-level overview of the entire feature extraction scheme. The scheme is quite elaborate, as it relies upon numerous image processing and analysis operations to compute a large number of local and global features. Our implementation includes many algorithms that are well studied in the computer vision and image-processing communities – these methods will be documented

through their original references. However, some descriptors from our representation have not been addressed before (color names, color composition), or provide a novel point of view (line and skin detection) – these implementations and methods will be described in detail in Sections IV, V, VI, and VII.

The complete feature extraction scheme is illustrated in Fig. 3. The input image is first rescaled, to account for different viewing conditions, and preprocessed, to account for the spatial averaging and color constancy mechanisms in human visual system.

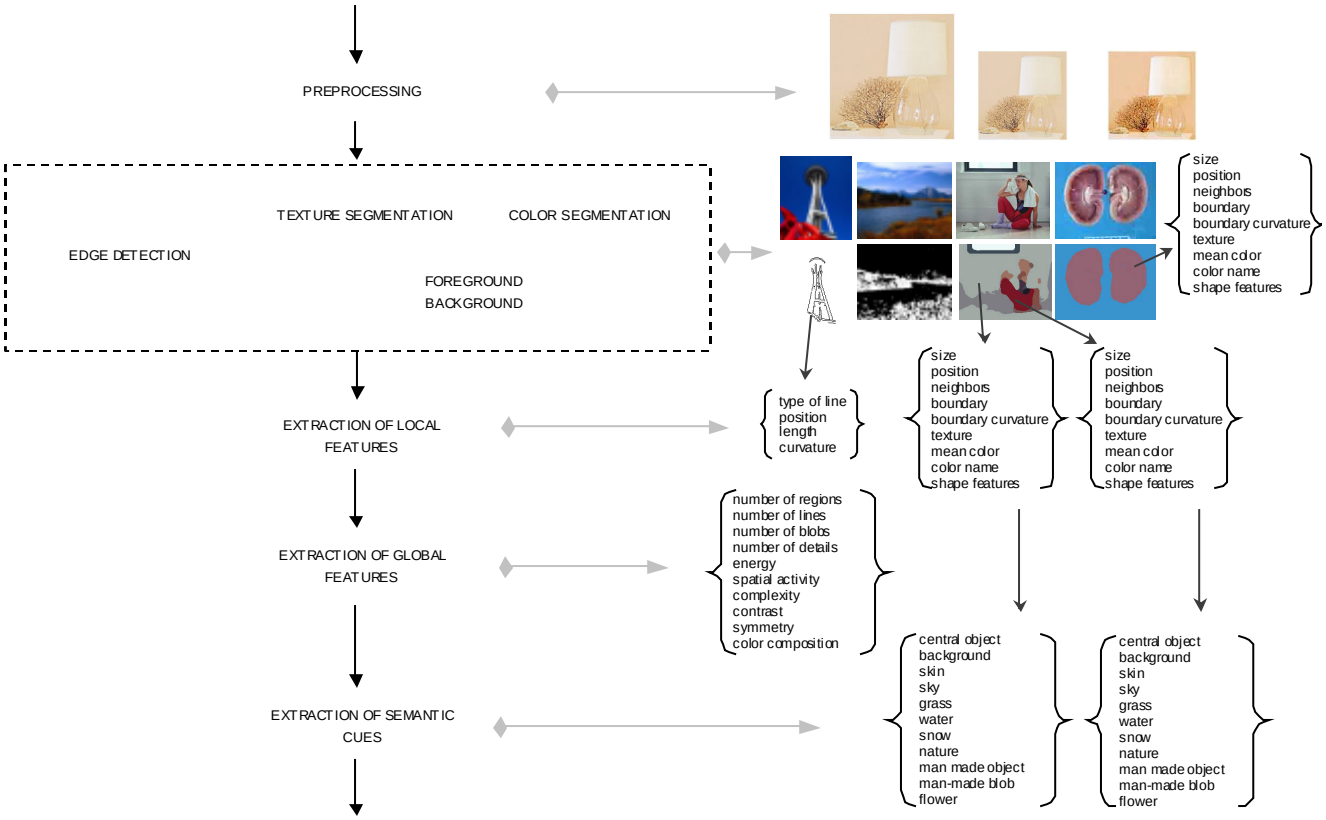


Fig. 3: The overview of the feature extraction process.

The image is then subjected to the edge-detection and segmentation procedures. Edge-detection includes the line/curve detection algorithm that isolates the "significant regular edges", i.e. the edges that are most likely to "influence" our perception of image content (this "interesting spin" on the edge-detection techniques is described in Section IV in more detail). Image segmentation includes texture segmentation and pixel labeling, color segmentation, and foreground/background segmentation. (The details of these algorithms are given in Section V). Ideally, these operations provide a strong segmentation (the division into regions, which represent the boundaries of the perceived objects, and nothing else). This is achievable only in a narrow application domain; in a general-purpose application these segmentations provide a partitioning into "likely-to-be" meaningful regions. We will therefore assume, that although these are not necessarily constituent objects, the pictorial properties of the identified regions still provide a valuable link to image semantics.

The next step is the computation of the local features, where to all *relevant regions* and *relevant objects* from the segmented image we assign a structure with the information about the size, position, neighbors, boundary, boundary curvature, texture, elementary shape descriptors (boundary, eccentricity, moments and symmetry features), mean color and color name (e.g. "red", "light pink", or "dark gray"; this novel algorithm is fully described in Section VI).

Next, the local features are merged to provide the global features, such as: the number of regions, the number of blobs, the number of regions with specific color, the number of details, the number of lines, energy, spatial activity, measures of contrast and image

complexity. The color names from all relevant regions are also combined into a *color name histogram* to determine the color appearance of the image. The color name histogram generates descriptions such as *grayish*, *brownish*, *vivid colors*, *graphic/artificial colors*, *dark image*, *green regions*, etc.

In the last step, the features are analyzed and combined to discover the presence of certain semantic cues. Specifically, to each relevant region and each relevant object we assign a structure of *semantic flags* - each flag representing one of the semantic cues (central object, background, skin, sky, grass, water, snow, man-made object, man-made blob, nature, flower, etc.). More details regarding combining features and modeling the semantic cues are given in Section VII. As a part of the modeling task we have developed a new algorithm for the detection of skin regions – Section VII also gives the details of this algorithm).

IV FEATURE EXTRACTION: LINE AND CURVE DETECTION

Lines and curves that carry strong semantic information are of special interest in the image modeling applications. For example, several perfectly straight or highly regular lines are often related to man-made objects (in which case the length of these lines also indicates the viewing distance). Another example is the horizon, a long horizontal line typically found in traditional landscape photographs. This section describes an interesting approach for detecting such lines and curves. Although the scheme we propose has its foundation entirely within the well-known theory of edge detection, it presents a novel point of view and an efficient implementation.

4.1 Motivation

Let us consider a "photometrically" and geometrically smooth edge in the image, a line, for instance, because it is as regular as a curve can be, and let us assume that the image is not noisy in the neighborhood of this line (i.e. the line is a straight isophote). No assumption is made about the level of contrast across the line. Then, observe that, in general, the edge-preserving restoration techniques do not keep this line invariant. In an edge-preserving technique, like the one introduced by Perona and Malik in [31], the local anisotropy is controlled by the local magnitude of the image derivative. Since there are edges with different levels of contrast, the tangential smoothing along the edges is more important than the normal one. However but the contribution of the latter is by no means negligible, which has an undesirable effect – namely, the line we have considered will be displaced by an amount that depends upon the profile of the intensity discontinuity. Therefore, we propose that such a "perfect" line should be kept perfectly invariant, regardless of the intensity profile in the neighborhood of the line. This natural idea is the foundation of our method.

4.2 Line and curve detection: The curvature flow and vanishing set of the Laplacian operator

The fact that we propose not to smooth in the normal direction at all makes an important difference with respect to the edge-preserving restoration techniques. From the curve evolution theory, a way to preserve the "perfect" line is to smooth the image by the curvature flow, so that all its isophotes undergo the curvature motion [33]. Since the curvature flow corresponds only to the tangential smoothing, one may wonder how this method handles noise. An answer to this question is in the work of Grayson [32], who proved that under the curvature motion any planar closed simple curve would shrink to a point (disappear) in a finite time, and small looped isophotes arising from the noise are naturally eliminated. Of course, this also implies that any closed curve containing a straight portion would eventually vanish too. (In effect, only infinite lines are invariant under the curvature motion, not the segments of lines. The latter are transformed in the following way: they become shorter and vanish in a finite time. Still, until a portion of the line disappears, it remains strictly invariant.)

Once the image has been smoothed via the curvature motion [33], it remains to extract the vanishing set of the Laplacian operator. We do that by using the *marching lines algorithm*, i.e. bilinear interpolation [34]. Practically, it consists of the following procedure. Consider four adjacent pixels, seen as the four corners of a square. The image intensity function is bilinearly interpolated inside the square and the vanishing set of this function is a segment (or two segments) that runs from a border of the square to another one. The ends of these

segments have real coordinates, not integer ones. It is easy to see, by considering similar adjacent squares, that the segment will inevitably be continued. This proves that the extracted curve will be closed (or will end at the border of the image), and it will be simple (i.e. not self-intersecting). It is important to emphasize that these closed curves do not necessarily enclose homogeneous regions. However, although their global topology does not necessarily form a satisfying segmentation, these curves represent the *significant edges*, i.e. edges that are often very important to the perception and understanding of the scene.

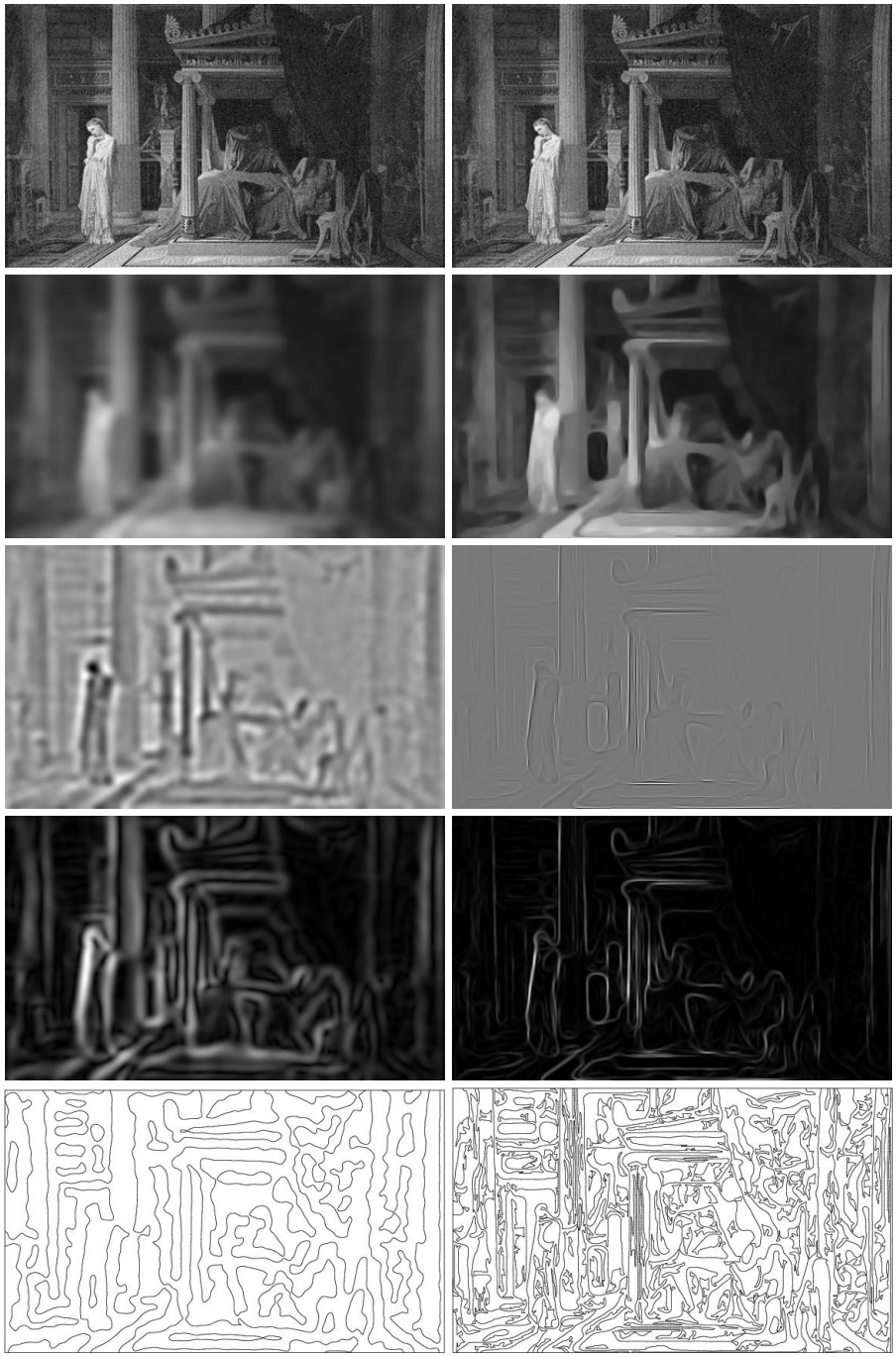


Fig. 4: The comparison between the isotropic algorithm (left column) and the proposed approach (right column). The top row shows the input image, the second, third, fourth and fifth rows show the smoothed image, its Laplacian, the magnitude of its first derivative, and the extracted edges. All the extracted edges are not shown (there are too many of them), only one tenth of the randomly selected lines appear.

A natural extension of this method is to use a purely geometric approach to find significant portions of the detected curves (lines, circles, etc.). It is important to point out that the extracted curves are smooth because their nodes have real coordinates. Furthermore, normal vectors and curvatures at these nodes may be computed very accurately by bilinear interpolation, directly from the first and second derivatives of the smoothed image. Also, our approach detects lines, or regular curves on the basis of their sole geometric regularity, independently of the photometric aspects. (We found the justification for doing this in the fact that contrast-dependent methods cannot detect low contrast edges even if they are prominent due to evident geometric regularity. However, in many cases, geometry does not entirely account for the importance of edges and we need to evaluate the photometric prominence of the extracted curves. This is simply done by computing the magnitude of the smoothed image derivative.).

We illustrate the algorithm on a noisy gray-scale copy of the painting Antiochus and Stratonice from Ingres because it contains numerous lines of different lengths and different contrast levels (Fig. 4). By comparing the two flows we can observe the negative effect of normal smoothing on edge detection (albeit its positive effect on image segmentation). The input noisy image has been obtained by adding the Gaussian noise of standard deviation 15/255 to the ground truth image. The amount of smoothing is identical for both methods. The reader may observe that our method produces many more curves than the other one, since the absence of normal smoothing prevents several isophotes to merge into one (which is exactly the effect we want to achieve). Observe also how the straight edges (independently on their contrast level) are much more likely to be detected by our method than by the isotropic algorithm.

V FEATURE EXTRACTION: PREPROCESSING AND SEGMENTATION

In its early stages the human visual system performs a significant amount of spatial averaging, which accounts for the way we extract dominant objects, interpret color information and perceive images at the global level [35]. However, the amount of averaging depends on the spatial frequencies, spatial relationships among the colors, size of the observed objects and global context. For example, the capability of the human visual system to distinguish different colors drops rapidly for high spatial frequencies -- consequently we describe textured areas with a single color, since only spatial averages of the microvariations are perceived. On the other hand, we do not average isolated edges, as they represent object and region boundaries. Furthermore, the very notion of region or object color has it roots in the fact that humans perceive and interpret color information independently of illumination conditions, the phenomenon known as the *color constancy* [28]. Therefore, before proceeding to image segmentation and scene understanding, we need to account for this “mechanism” as well.

Based on these observations we have implemented the segmentation part as an adaptive low-pass filtering operation. This relatively simple approach exploits some basic facts about human perception and follows them in few processing steps, with a goal of constructing a simplified representation consistent with the “brain image” we use when addressing image content. The algorithm is illustrated in Fig. 5. As the first step, we use a variant of the “white world approach” to estimate the scene illuminant and apply a simple chromatic-adaptation transform to account for different illumination conditions [29]. This is followed with an edge-detection/pixel-labeling scheme. Based on the edge density in the neighborhood of each pixel we estimate if the pixel is *perceptually important* (uniform region and texture) or *perceptually unimportant* (noise and image elements whose color content we typically do not process or care about, such as boundary regions, edges, contours, etc.). The pixel labeling operation also produces the *texture map*. After the labeling procedure, all perceptually important pixels are subjected to the adaptive Gaussian low-pass filtering operation. The support of the filter is determined based on the pixel labels, so that it averages more within uniform regions, less in fine textures, even less in coarse textures, and ignores transition regions and edges. The smoothed image is then subjected to the mean-shift color segmentation [37], to produce the color-segmented image and to identify all *relevant regions*. Finally, texture and color information are combined to achieve the foreground-background segmentation and identify *relevant objects* and *background regions*. An example image from our database and the corresponding processing steps are illustrated in Fig. 6.

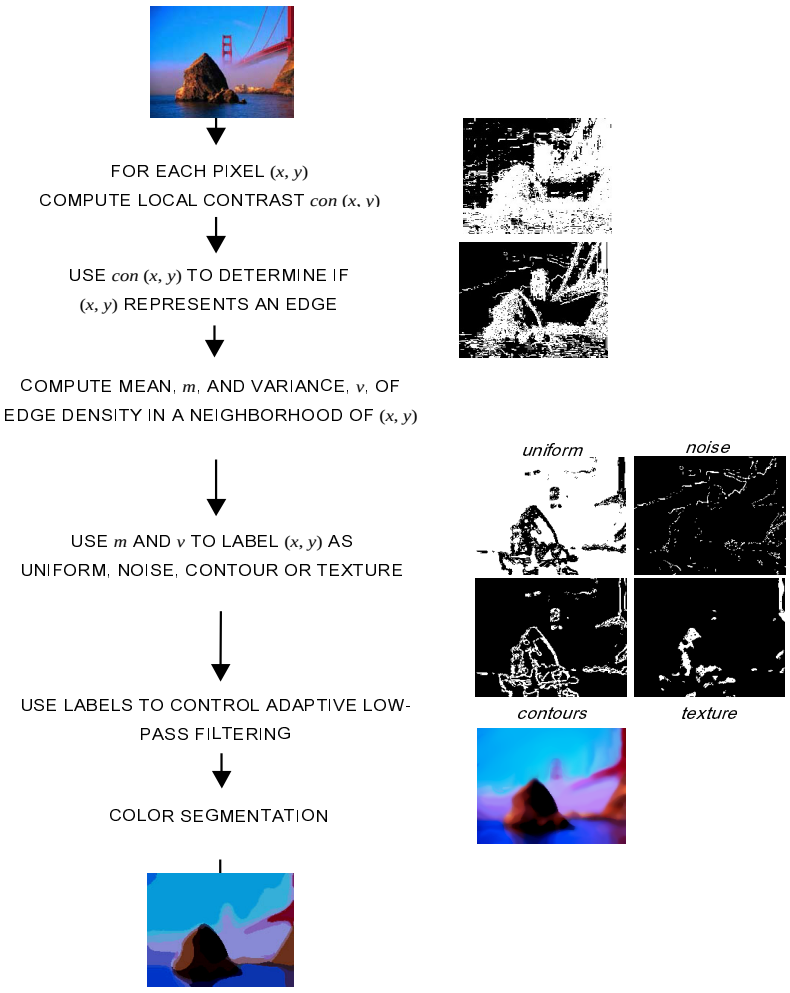


Fig. 5: The block diagram of the image segmentation part.

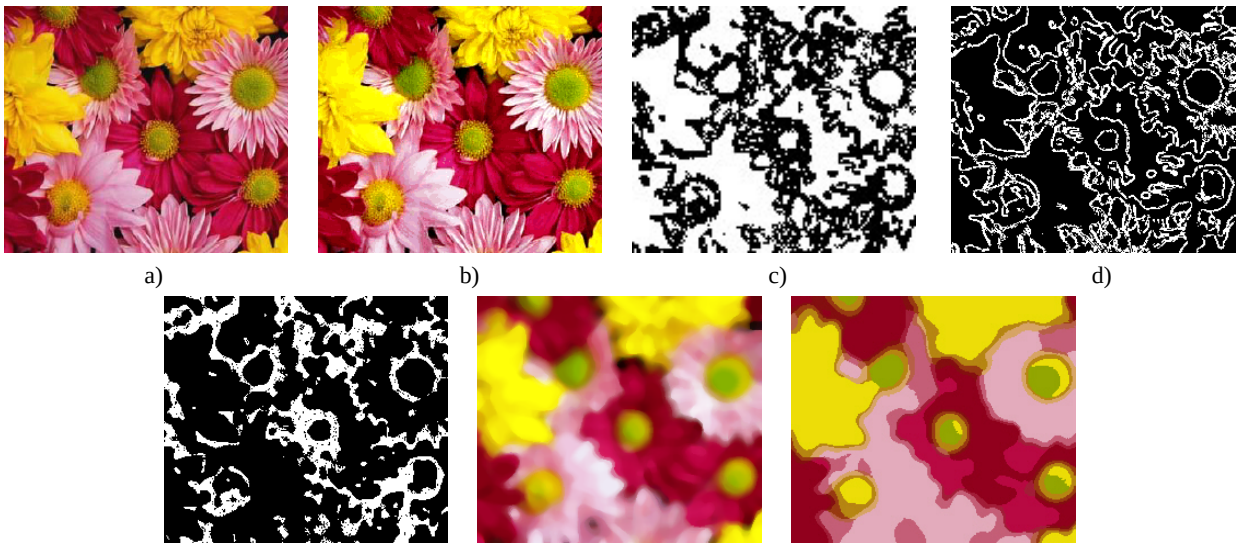


Fig. 6: The image segmentation result: a) original image, b) preprocessed image, c) pixels labeled as uniform (white), d) pixels labeled as color and texture edges (white), e) pixels labeled as texture (white), f) smoothed image, and g) color segmented image.

VI Feature Extraction: Color names

Although color spaces allow us to describe colors and measure their similarity, in everyday life we mainly identify colors by their names. For that reason, the color names we use to describe certain objects, image regions or the overall color scheme provide strong links to image content. For example, we always describe sky as light blue, although its spans a wide range of different blues. Similarly, we often rely upon descriptors such as “bright”, “brownish”, “grayish” and “dark” to convey the impression of the atmosphere in the scene. Therefore, when combined with image segmentation, color names can be used to select objects, identify important semantic cues, describe the appearance of the image, and ultimately generate semantic annotations. In our feature extraction scheme the color naming ability is needed both locally, to provide the local features (i.e. color names for the specific objects and regions), and at the global level to provide the description of the overall color composition. Both tasks require a flexible computational model for color categorization. In this section we make an attempt towards developing a computational color-naming model based on human behavior in color categorization.

6.1 The basics of color naming and the overview of our approach

The issue of color names and color categories has been debated for a long time. However all studies agree that color perception is indeed divided into categories, as all human languages have a number of color terms to name some of these categories [57]. Although the mechanism of color naming is still not entirely understood, there are several studies that influenced the research in this area. One of the most widely used models originated from the work by Berlin and Key, who studied color naming patterns across a variety of languages and discovered remarkable regularities in the shape of the basic color vocabulary [53]. Berlin and Kay introduced a concept of *basic color terms* and identified 11 basic terms in English. Numerous later studies demonstrated that the basic color categories (*prototypes* or *focal colors*) play a crucial role in color naming, as we seem to represent all other colors relative to these prototypes [29]. Although researchers still do not agree on the issues such as the universality of color names (whether they are universal or culture-specific), or the formation of the categories (whether they are genetically determined or learned), they all accept that color categories exist and that there is a special relationship between the categories and language. From the perspective of designing computer models of color naming, the most important facts are: 1) regardless of language or culture, each color category has a focal point in the perceptual space and spans the volume around this point with fuzzy boundaries, and 2) there is a large consensus in the language community about what the focal point for a particular word is (even though there is less of a consensus about the boundaries of its color region) [57].

Based on the elements discussed so far, here we analyze the basic ingredients of a fairly general color-naming algorithm for applications in digital image libraries. Since prototypical colors are the crucial part in the perception of color names, the first step in our work is to determine a well-represented set of color prototypes, *vocabulary*, and the corresponding color naming *syntax*. The next step requires a *color naming mechanism*, which for an arbitrary input color determines the category membership. Finally, we will extend this approach to name color regions and provide the description of overall color composition for an arbitrary input image. In order to respect the studies of color categorization and develop a perceptual color naming method, we need to make sure that our method satisfies the following properties. It should account for the basic color terms and use a systematic syntax to combine them. The names attached to different colors should reflect perceived color differences among them. Segmenting the color space into the color categories should produce smooth regions. Finally, we need a representation that everyone can understand and use, and an implementation that produces results in agreement with human judgments.

6.2 Developing the vocabulary of color names

The selection of the color naming vocabulary and syntax is the central part of our design, as we need to make sure that the selected prototypes are basic, yet general enough to allow for a variety of descriptions we make (“the sky is blue, but the sea is dark blue”, “the

image is vivid", or "the whole scene is grayish and the colors are moderate"). There were several attempts towards designing a standard method for choosing color names [48], [54], [58]. One solution, which seems as an optimal choice for a general color naming application is the *ISCC-NBS dictionary* developed by the National Bureau of Standards [54]. This dictionary gives names to 267 regions in the Munsell color space, and employs English terms to describe colors along the dimensions of hue, lightness and saturation. In the ISCC-NBS dictionary there are five discrete values for lightness (*very dark*, *dark*, *medium*, *light* and *very light*), four discrete values for saturation (*grayish*, *moderate*, *strong* and *vivid*), three terms that address both lightness and saturation (*brilliant*, *pale* and *deep*), and 28 names for hues constructed from the basic set (*red*, *orange*, *yellow*, *green*, *blue*, *violet*, *purple*, *pink*, *brown*, *olive*, *black*, *white* and *gray*).

There are several advantages for implementing the ISCC-NBS dictionary in a color-naming algorithm: it is developed using controlled perceptual experiments, it includes the basic color terms (with the modifiers that allow for additional flexibility), and each prototype is represented with its *centroid color*, thus preserving the notion of color foci. There are, however some irregularities in the representation, such as the lack of systematic syntax and the problem of several artificially constructed names (*blackish red*, *reddish gray*), which were introduced to preserve the overall structure of the vocabulary and are usually not understood by non-trained observers. To address these problems, we had previously performed several subjective experiments and constructed a slightly modified version of the vocabulary. The details of these experiments can be found in [36]. It is important to emphasize that the goal of our experiments was only to "correct" the syntax of the color names from the ISCC-NBS dictionary, while preserving the color values of the corresponding prototypes. Consequently, our vocabulary can be viewed as "Renamed ISCC-NBS", since it operates on the same set of prototypes as the ISCC-NBS model. The difference between them is due to the fact that: 1) some of the color prototypes that were not perceived consistently by our subjects have been removed from the model, and 2) some of the ISCC-NBS names have been changed to reflect the majority of experimental decisions. Our findings support the observations from the previous studies [54], [55]. The findings were generalized through the following syntax (note that symbol : denotes "is defined as", and symbol | denotes meta-or):

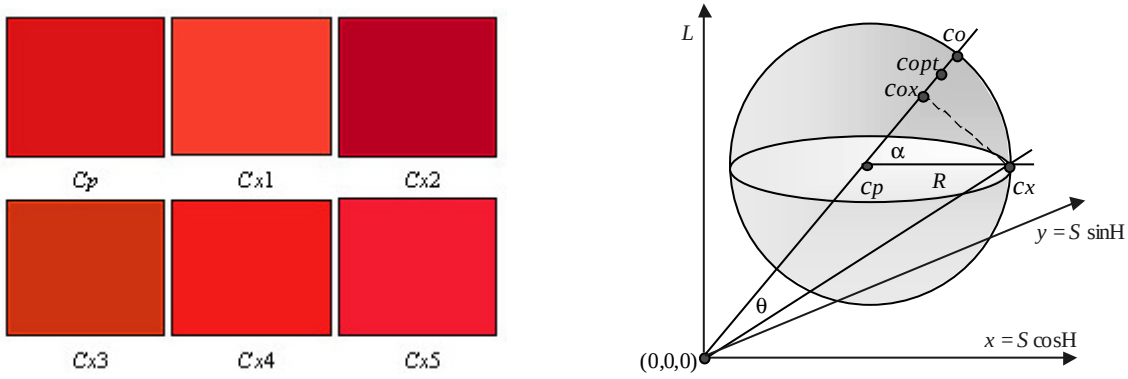
<color name> : <chromatic name> | <achromatic name>
<chromatic name> : <lightness> <saturation> <hue> |
<saturation><lightness> <hue>
<achromatic name> : <lightness> <achromatic term>
<lightness> : *blackish* | *very dark* | *dark* | *medium* | *light* | *very light* | *whitish*
<saturation> : *grayish* | *moderate* | *medium* | *strong* | *vivid*
<achromatic term> : <generic achromatic term> | <ish form> <generic achromatic term>
<generic achromatic term>: *gray* | *black* | *white*

<hue> : <generic hue> | <halfway hue> | <quarterway hue>
<generic hue> : *red* | *orange* | *brown* | *yellow* | *green* | *blue* | *purple* | *pink* | *beige* | *olive*
<halfway hue> : <generic hue> - <genereic hue>
<quarterway hue> : <ish form> <generic hue>
<ish form> : *reddish* | *brownish* | *yellowish* | *greenish* | *bluish* | *purplish* | *pinkish*

6.3 Developing the color naming metric

Since the color naming process should address the graded nature of category membership and take into account the universality of color foci, we decided to perform color categorization through the *color naming metric*. Assuming a well-represented set of prototypes (foci), the metric computes the distance between the input color and all the prototypes, thus providing a membership value for each categorical judgment. Although commonly used as the measure of color similarity, the Euclidean distance in the *CIE Lab* color space has several serious drawbacks for the application in our method. The first problem is related to the sparse sampling of the color space. It is well known that the uniformity of the *Lab* suffers from defects, so that its "nice" perceptual properties remain in effect only within a radius of few *just-noticeable differences* [28]. Since there are only 267 points in our vocabulary, the distances between the colors may be large and the metric only partially reflects the degree of color similarity. The other, more serious problem is related to our perception of color names and their similarity. Let us assume an arbitrary color represented by a point c_p in the *Lab* space, and a set of neighboring colors $\{c_{xi}\}$, on a circle with the radius L in that space. Although all the pairs (c_p, c_{xi}) are equally distant, we do not perceive them as

equally similar. This is illustrated in Fig. 7a, where the color c_p is compared to the colors c_{x1} - c_{x5}, all with $D_{Lab}(c_p, c_{xi}) = 10$, demonstrating (within the printer gamut of course) that the perceptual differences between these colors are not equal.



a) b)

Fig. 7: a) An example of the equidistant pairs of colors. The color data is given in Table I. b) The derivation of the color naming metric.

Table I: Color data for Fig. 7a. $D_{Lab}(c_p, c_{xi})$ and $D_{HLS}(c_p, c_{xi})$ are the color distances between the points c_p and c_{xi} in the *Lab* and *HSL* color spaces, respectively. $\theta_{Lab}(c_p, c_{xi})$ and $\theta_{HLS}(c_p, c_{xi})$ are the spatial angles between the points c_p and c_{xi} in the *Lab* and *HSL* color spaces.

c_p: (*L*,*a*,*b*) = (47,70,55) (*h*,*s*,*l*) = (0,90,85)						
	(*L*,*a*,*b*)	(*h*,*s*,*l*)	D_{Lab}	D_{HLS}	θ_{Lab}	θ_{HLS}
c_{x1}	(57,70,55)	(5,82,98)	10.0	17.0	4.8	7.5
c_{x2}	(37,70,55)	(349,100,73)	10.0	24.0	5.3	11.1
c_{x3}	(47,60,55)	(11,91,81)	10.0	17.8	4.4	8.3
c_{x4}	(53,76,60)	(0,89,96)	9.9	11.0	0.90	3.8
c_{x5}	(53,76,50)	(354,89,95)	9.9	13.7	4.8	5.5

6.3.1 Testing the hypothesis: Color similarity experiment

To test the relationship between the perceptual similarity, color distances and angles in the *Lab* and *HSL* color spaces, we conducted a subjective experiment. Four subjects participated in the experiment. The subjects were given ten sets of color samples. Each set consisted of the "prototype" color, c_p, and five colors, $\{c_{xi}\}_{i=1,..5}$, with the same *Lab* distance to the prototype ($D_{Lab}(c_p, c_{xi}) = const$). An example of such a set is given in Fig. 7a. The distance between the prototype and each color c_{xi} for the first set was $D_{Lab}(c_p, c_{xi}) = 6$. The distances for the Sets 2-10 where: 6, 8, 10, 10, 14, 16, 18, 24, and 30, respectively. For each set the subjects were asked to rank the samples c_{xi} according to the perceived similarity to the set prototype, with 1 being the most similar and 5 being the least similar. The sets were displayed on a computer monitor. For each set we have also computed distances, $D_{HSL}(c_p, c_{xi})$, and angles, $\theta_{HLS}(c_p, c_{xi})$, in the *HSL* color space, and correlated them with the subjects' rankings. By analyzing these correlations we have made an interesting observation; the correlation between the rankings and θ_{HSL} was 97%, and the correlation between the rankings and D_{HSL} was 95%, indicating that θ_{HSL} and D_{HSL} may be useful indicators of perceptual similarity between the equidistant colors, although none of the two distances alone represents an accurate color naming metric.

6.3.2 Designing and testing the color naming metric

Following these observations we have designed a new empirical metric for color-naming applications, which attempts to correct the before-mentioned problems. Let us assume a prototype c_p and an arbitrary input color c_x. As discussed previously, for the given $D_{Lab}(c_p,c_x)$, the combination between $D_{HSL}(c_p,c_x)$ and $\theta_{HSL}(c_p,c_x)$ reflects the "reliability" of the *Lab* distance as the measure of similarity in the color-name domain. Thus, we will use this relationship to modify D_{Lab} in the following manner. We first compute the distances between c_p and c_x in the *Lab* and *HSL* spaces:

$$D_{Lab}(c_p,c_x) = L = \sqrt{(l_p - l_x)^2 + (a_p - a_x)^2 + (b_p - b_x)^2}\ ,\ D_{HLS}(c_p,c_x) = R = \sqrt{s_p^2 + s_x^2 - 2s_p s_x \cos(h_p - h_x) + (l_p - l_x)^2}\ .$$

Given the radius *R*, we then find the color $c_o : (h_o, s_o, l_o)$, so that:

$$D_{HLS}(c_p,c_o) = R\ ,\ \theta_{HSL}(c_p,c_o) = \frac{s_p s_o + l_p l_o}{\sqrt{(s_p^2 + l_p^2)(s_o^2 + l_o^2)}} = 0\ . \tag{1}$$

In solving (1) we chose a point that satisfies $\theta_{HSL}(c_x,c_o) < \pi$. This is illustrated in Fig. 7b. According to our hypothesis, given the distance *L*, the optimal perceptual match corresponds to the direction $\theta_{HSL}(c_p,c_o) = 0$. Assuming a small increment ΔR, we then update the initial solution c_o as follows: $R_o = D_{HSL}(c_p,c_o)$, $s_o = s_o(1 \pm \Delta R / R_o)$, and $l_o = l_o(1 \pm \Delta R / R_o)$, until $D_{Lab}(c_p,c_o) \approx D$. At this point, c_o represents the optimal perceptual match to c_p, for the given *Lab* distance. We denote this solution c_{opt}. As an estimate of perceptual dissimilarity between c_x and c_{opt}, we compute the relative difference between c_{opt}, and the projection $c_x \perp c_{opt}$:

$$\Delta d(c_p,c_x) = \frac{d(c_p,c_{opt}) - d(c_p,c_{ox})}{d(c_p,c_{opt})} = \frac{R_o - R\cos\alpha}{R_o} = 1 - \frac{s_p s_x \cos(h_p - h_x) + l_p l_x - s_p^2 - l_p^2}{R_o \sqrt{s_p^2 + l_p{}^2}} \tag{2}$$

As required by our model, in predicting the amount of perceptual similarity this formula takes into account both the distance and the angle in the *HSL* space. Therefore, we use this value to modify the *Lab* distance between the colors c_p and c_x as follows:

$$D(c_p,c_x) = D_{Lab}(c_p,c_x)[1 + k(D_{Lab}(c_p,c_x))\Delta d(c_p,c_x)] \tag{3}$$

i.e. the *Lab* distance is increased proportionally to the amount of dissimilarity Δd. The factor $k(L)$ is introduced to avoid modifying distances between close points, $k(L) = 0$ if $L < 7$, and to limit the amount of increase, $k(L) = const$ if $L > 30$.

To test the stability of the method we have used the metric to name different color regions in the *RGB* and *HSV* color spaces. In both color spaces the color names changed smoothly and the know color regions were identified accurately. To test the agreement with the human observers we asked four subjects to review the color names assigned by our method to 100 randomly selected colors. Each subject received a different set of colors. The subjects agreed with the assigned color name in 91% of cases (362/400).

6.4 Attaching the color names to image regions and pixels

As already mentioned, in our feature extraction scheme the automatic color naming algorithm is used to provide both the local features (i.e. color names for the specific objects and regions, and the global features (i.e. overall color composition).

6.4.1 Local features

At the lower level, the color naming algorithm is used to assign a color name to all regions from the color segmentation and to all objects obtained through the foreground-background segmentation. The color-name representation is also used to improve the segmentation result, by merging the neighboring regions that share the same color name.

6.4.2 Global features

At the global level, the color-naming algorithm is applied to obtain the description of the overall color composition. To obtain such a description, we start from the color-segmented image, and via (3) assign the color name to all perceptually significant pixels. In the next step, we compute the histogram of color names and use it to generate the description of color composition. The structure and syntax of our color vocabulary allow us to describe color composition at different accuracy levels. Depending on how the color names and modifiers are combined, we can formulate descriptions at the *fundamental*, *coarse*, *medium* or *minute* level. At the *fundamental* level, the color names are expressed as <generic hue> or <generic achromatic term> from the syntax described in Section 6.2 (e.g. *brown, gray*). At the *gross* level, the color names are expressed as <luminance> <generic hue>, or <luminance> <generic achromatic term> (e.g. *dark brown, light gray*). At the *medium* level, the color names are obtained by adding <saturation> to the gross descriptions (e.g. *strong dark brown*). At the *minute* level, the complete <color name> as specified in the syntax is used (e.g. *strong dark reddish brown*). These different precision levels follow experimental observations related to the different color naming patterns in human behavior [55], [36]. For example, we typically use fundamental level when referring to well-know objects, or when color-related information is not considered important. According to our experiments, the overall description of photographic images is usually formulated with the gross or medium precision, while the minute level color names are typically used in describing specific objects and regions, or emphasizing the differences among them [36]. Table II illustrates the description of color composition for image in Fig. 6, for all four levels. The table also includes a comparison to the descriptions provided by three different observers in the subjective testing of the method.

Table II: Color composition extracted from the image in Fig. 6 and the comparison to three typical descriptions our subjects provided in the subjective testing of the method.

Minute	Medium	Gross	Fundamental	Subjects' descriptions		
strong purplish red (29%)	*strong red* (29%)	*red* (29%)	*red* (29%)	*purple*	*red*	*yellow*
vivid yellow (21%)	*vivid yellow* (21%)	*yellow* (21%)	*yellow* (21%)	*yellow*	*yellow*	*pink*
moderate pink (22%)	*moderate pink* (22%)	*pink* (22%)	*pink* (31%)	*light green*	*pink*	*green*
moderate dark pink (9%)	*moderate dark pink* (9%)	*dark pink* (9%)	*green* (4%)	*pink*		*purple*
strong light yellowish green (4%)	*light green* (4%)	*light green* (4%)		*dark pink*		

6.4.3 Relationships with semantics

The descriptors derived from the color names are among the most valuable in linking the low-level features to the important semantic cues, as well as in modeling some basic semantic categories. For example, as illustrated in Fig. 8a, upper regions labeled as *vivid blue* or *vivid purplish blue* may represent sky on a bright sunny day, regions with regular boundaries/geometry and bright saturated colors are likely to be man-made objects. Similarly, for the image in Fig. 8b, the relationships between the curvature and color of the bright purple region, and the neighboring green regions indicate that this is a "good" candidate for the category "Flowers". Overall color composition often captures the atmosphere in the scene. By combining the linguistic terms from our syntax it is easy to see why the scene shown in Fig. 8c can be described as "brownish", and easily related to a man-made environment. Linguistic descriptions, which can be constructed using our algorithm are very broad, and include terms as, "bright", "dark", "monochromatic", "grayish", "lot of green", "computer generated", etc.

a) b) c)

Fig. 8: a), b) Color names attached to neighboring regions carry the information about important semantic cues. c) Example of the "brownish" scene -- the color composition for this image at the minute level is: *brownish black* (29%), *dark brown* (6%), *dark yellowish brown* (7%), *dark grayish brown* (2%), *strong yellowish brown* (8%), *light grayish brown* (3%), *moderate brown* (17%), *yellowish white* (3%), *beige* (9%), *light beige* (6%).

VII COMBINING THE FEATURES TO DERIVE SEMANTIC INDICATORS (CUES)

Semantic cues are built from pictorial features and may be defined as special relationships between the latter. For the sake of simplicity, let us assume that these relationships can be described via "ad-hoc formulas". For instance, the semantic indicator "sky" can be defined as "upper, uniform, blue region". Another example is "flower", which is defined as "bright colored region with curved or rigid boundaries, either in the neighborhood of irregular green regions, or on a textured green background".

In many cases, these formulas can be easily derived from our "everyday" language based on the verbal expressions that describe scenes and objects. One such an example is our set of perceptual experiments (see Section II), where we used our subjective data to model some of the elementary semantic categories, the underlying semantic cues, and the visual attributes that contribute to their "semantic appearance". Most semantic indicators, however, resist empirical modeling and require the use of more formal tools. Such a formal modeling can be significantly simplified if we adopt the following hypothesis. Each semantic indicator is built from a small number of image features (say, less than six or seven). Of course, the relationships between these features may be very rich. This hypothesis simplifies the problem because the number of features that contribute to our perception is enormous (even our basic view into the "semantic world" involves more than 40 features), and the complexity of the resulting model depends mostly on the number of considered variables. Also, even for very simple semantic cues (such as sky, water or grass) the choice of underlying image features may be subjective. Nevertheless, we believe that a formal semantic modeling is an achievable task.

One example of such a modeling in our implementation is the semantic indicator "skin''. It involves the presence of regions with human skin colors (yet, not necessarily faces, for the latter are not semantic indicators but real semantic objects). Therefore, the presence of human skin may be detected by the local color and its first-order variation. Of course, the particular relationships between the color components and their first order derivatives over the range of human skin colors are very complex. In this section we propose a variational approach to model the "skin'' semantic indicator. The same algorithm could very well apply to other semantic cues, as long as they satisfy the aforementioned hypothesis.

7.1 Learning the semantic indicators through a variational approach

7.1.1The formal statement of the problem

Consider a set P containing N points of a n-dimensional manifold Ω. In order to illustrate the ideas, we may think of P as a data set of N measures performed on a stochastic (or deterministic) process whose state may be partially (or totally) described using n numeric values. Consequently, each point in P is a sample, each coordinate is a parameter of the process under consideration, and Ω is the set of possible values these variables may take *a priori*.

In the field of image libraries, this situation arises when considering a set of image features (such as spectral, color or texture measures) for the classification purposes. If n such numeric features can be extracted for each image, one may then associate a set of N images with a cloud of N points in R^n, each point representing one image through its feature vector. If within a certain class of images complex relationships between different features exist, one may ideally expect the corresponding points to form a submanifold of R^n. A geometric representation of this submanifold can then be used to distinguish this particular class from all the other image points. Here, $\Omega = R^n$, P is the set of features vectors and M is the submanifold of R^n, hopefully of low dimension, modeling the relations between the features. Another example is stereo vision, as pixel matching algorithms often output a cloud of points in R^3, which belong to the surface of the pictured objects, and it remains to find a surface "passing": through these points. Here, $\Omega = R^3$, P is the set of points and the searched surface is M.

In this section, we propose a theoretical and practical framework which is relevant when the measured variables are expected to have complex distributions, strong relationships, and, more specifically, in a case of functionally related variables. Formally, this restriction is equivalent to saying that the points of P are not distributed arbitrarily in Ω but, instead, due to some noise, belong to a submanifold of Ω of dimension smaller than n. Analyzing the data set is then equivalent to recovering this submanifold.

7.1.2The state of the art in modeling relationships between variables

The case of variables well represented by their mean and variance is covered extensively in statistics [41], [42]. In this section, we are interested exclusively in data sets with more complex distributions where such assumptions are not relevant. For example, when higher order statistics are essential for the analysis, one approach is to use *Pearson* or *Johnson curves* [43], [44], though, neither is appropriate when both the distributions and the relations are expected to be highly complex. Another approach, *exploratory data analysis*, attempts to identify relations between several variables by searching for systematic patterns and clusters [45]. When no distributional assumption is available, neural networks are often used for their flexibility, although they relate very closely to standard statistical regressions [46]. In addition, despite being very effective in many applications, the underlying mechanism of neural networks is difficult to interpret. Therefore, we suggest another formalism, founded on the variational calculus.

7.1.3 Intuition about the method

Suppose that P is a sparse cloud of points in the plane $\Omega = R^2$, and that these points are roughly distributed along a smooth curve. The problem consists in recovering "the smooth curve" M that "passes through" P. We simply develop the idea that P, seen as a subset of Ω, can be transformed continuously into the curve M. This is achieved through an iterative process in which each point of P spreads itself in the direction of its neighbouring points. Little by little, each point in P transforms itself into a short piece of curve oriented toward other points of P, and grows toward them. Eventually, all the pieces will connect producing a smooth and simply connected curve. At the same time, outliers are eliminated and the shape is regularized. This "spreading" process transforms P continuously into the smooth manifold M. In the next section, we design an objective cost, or energy, associated with this spreading shape. As in any variational method, the final shape of M will correspond to the minimum energy, which leads us to two main issues. The first one being "How to define a proper energy ?", and the second one being "How to represent the evolving shape?".

As far as the energy is concerned, it has to reflect the relevant properties of the desired final shape. In the variational methods, the energy is often a weighted sum of several energies, each favoring or penalizing certain shape properties. In our problem, we need at least two energy terms. The first one is the *data attachment term*, which penalizes shapes containing a lot of points that do not belong to P and prevents the spreading process from adding too many points, thus making M too "fat". Shapes that do not contain all the points of P are penalized as well, since the final result should not miss the parts of P. However, the data attachment term alone is useless because its minimization results in nothing but $M = P$. Thus, we introduce the *regularization term*, which favors better-connected and smoother shapes. It will, of course, be in competition with the data attachment term, since P is not smooth, nor well connected. One way to obtain

well-connected shapes is to favor convexity. So, the regularization term favors convexity. Once again, the regularization term alone is useless, because minimizing it would simply connect all the points of *P*, producing the convex hull of *P* as the final result. To summarize, the total energy will only favor spreading toward other neighboring points of *P* as this is the only way the two terms may actually reach an agreement.

The representation of the evolving shape is an important issue as well. In effect, *a priori*, no hypothesis is made neither about the dimension of the final shape *M*, nor its topology. This is a domain where implicit representations are usually superior to shape explicit parametrizations. In solving this issue, we were mostly inspired by [47], [49] where a curve in R^3 is represented by a one-parameter family of concentric tubes of increasing radi. The represented curve is the medial axis of the tubes. If the radius *r* of the tubes is the parameter of the family then the tubes converge toward the curve when *r* tends to 0. This is actually a very general approach, which is valid regardless of the dimension and topology. The key is to consider neighborhoods (or approximations) of the represented object with increasing tolerance. Note, that those neighborhoods are always hyper-surfaces of the ambient spaces, i.e. manifolds of dimension *n* - 1. For instance, in R^3, both concentric tubes and concentric spheres are bi-dimensional although their limits are curves and points. Finally, those hypersurfaces are conveniently encoded as the iso-hypersurfaces of a scalar function defined on Ω. This implicit representation allows us to formulate the problem as a variational one.

7.1.4 Formalization

Let $u : \Omega \to [0,1]$ be a smooth function to be constructed so that the family of hypersurfaces $S_\alpha = u^{-1}(\alpha), \quad 0 < \alpha < 1$, tends to a submanifold M of Ω when α tends to 0. In that sense, u can be interpreted as a weak implicit representation since

$$M = \lim_{\alpha \to 0} u^{-1}(\alpha).$$

In particular, far from *P* the value of u is 1, and tends to 0 as one approaches *P*. We will define u as a solution to a variational problem that is naturally related to the reconstruction one. To construct an objective function, we will first present some useful integral criteria and their Euler-Lagrange equations. We will then see how to combine them and finally, since the variational methods are iterative, we will describe how to initialize u. We start with the case $\Omega = R^n$ and consider independently the four non-negative integrals

$$\int_P u^2, \quad \int_{\Omega \setminus P} (u-1)^2, \quad \int_\Omega \nabla^2 u, \quad \int_\Omega \nabla u \cdot Q_u \nabla u,$$

where Q_u is the projector onto the sub tangent space of Ω corresponding to negative eigenvalues of the Hessian of u. In other words, if $D^2 u = P^T D P$, with $D = Diag(\lambda_i)$, and $P^T P = I$, then $Q_u = P^T G P$ with $G = Diag(\nu(\lambda_i))$, where $\nu(R^-) = \{1\}$, and $\nu(R^{*+}) = \{0\}$. Note that Q_u is symmetric positive.

The motivation for considering the first three integrals is obvious. The minimization of the first one enforces M to pass through *P*. In effect it is null if and only if $u(P) = \{0\}$. Minimizing the second one prevents M from passing through the points that don't belong to *P*. In effect, it is null if and only if $u(M \setminus P) = \{1\}$ a.e. in $\Omega \setminus P$. Finally, minimizing the third integral enforces the smoothness of u and, to some extent (actually to the extent that α is not a singular value of u), that of S_α. The Euler-Lagrange equations of the three first integrals are respectively

$$2u, \quad 2(u-1), \quad -2\Delta u$$

on the domains where the corresponding integrals are defined and **0**, i.e. the null function, elsewhere in Ω. As for the fourth integral, it is minimized by the non-concave functions, the point we will develop further. Note that the motivation for considering convex u's is that it implies the convexity of its iso-hypersurfaces S_α (cf. previous section, which is how the connection between neighboring points is favored). Here is how we minimize this integral. Although Q_u depends upon $D^2 u$, we consider only the first order term in the Euler-Lagrange equation of the integral. It is equal to

$$-2\Delta^- u,$$

where $\Delta^- u$ is the sum of the negative eigenvalues of the symmetric matrix $D^2 u$, i.e.

$$\Delta^- u = \sum_{i=1,\ldots,n} \nu(\lambda_i)\lambda_i .$$

The proof can be found in [40].

In order to achieve the right balance between each effect, it is important to combine these integrals properly. We form a weighted sum of the four integrals depending on only one parameter, the scale σ. We define the scale as the critical distance between two just distinguishable points of *P*, thus "calibrating" our linear combination of contributions. The partial differential equation to solve is

$$\frac{\partial u}{\partial t} = \frac{\beta^2}{\sigma^2}(-1_P u + (1-1_P)(1-u)) + \varepsilon\Delta u + \Delta^- u,$$

where $\beta = \ln(7+\sqrt{48})$, $\varepsilon << 1$ and 1_P is the indicatrix function of *P* in Ω (see [40] for the details of the computation). This PDE is solved using the standard finite differences numerical schemes, and the manifold *M* is extracted using the Marching Cubes algorithm.

7.1.5 Application to modeling skin color

Algorithms for detecting skin in color images typically proceed in two steps: 1) the candidate skin pixels are detected locally, and 2) the geometry of the detected regions is regularized using morphological operators. This section concerns only the first step and deals with the local detection of skin color using the approach described in this section. To implement the "skin model" 200 images with people have been segmented manually, the set *P* is defined as the set of the *Lab* colors values of all the pixels from the skin regions, Ω is the set of existing colors in the *Lab* color system, and *M* is supposed to approximate the set of colors corresponding only to human skin. After reconstructing *M* (see Fig. 9), it is then possible to test whether the condition $u(L,a,b) < u_o$ is satisfied, where (L,a,b) is the color value of the tested pixel and u_0 is the threshold related to the probability of the color being "on" the manifold *M*. Consequently, this test can be used as the *local skin color detector*.

Fig. 10 shows a comparison of our detection method with two other methods on both images containing skin and images without skin. One of the strengths of the presented method is that it can be applied with more than three variables, which can further enhance the detection of skin regions. For example, in Fig. 11, we added a fourth variable being the variance of $\|(L,a,b)-(L_o,a_o,b_o)\|$ in a small neighborhood of each pixel, where (L_o,a_o,b_o) is the color of the considered pixel, and $\|\cdot\|$ denotes the Euclidean distance. This measure, denoted as $v((L_o,a_o,b_o))$ accounts for the textural properties of skin regions, and therefore significantly reduces the false acceptance in detecting the skin candidates. It also needs to be pointed out that the value of the threshold u_o seems to be rather arbitrary, with no direct physical interpretation (although it should not be difficult to relate it to a probability).

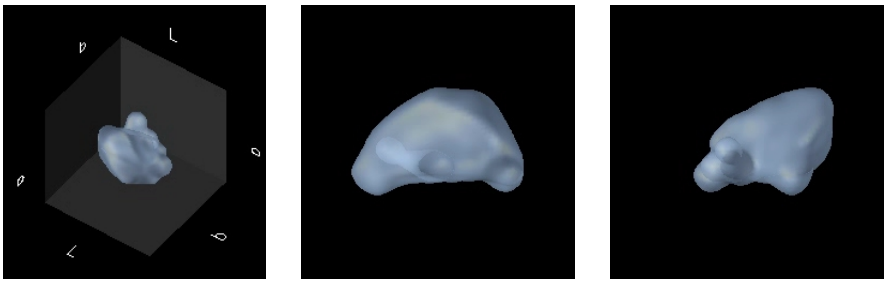


Fig. 9: The manifold of the set of human skin colors in the *CIE Lab* system from three different viewpoints.

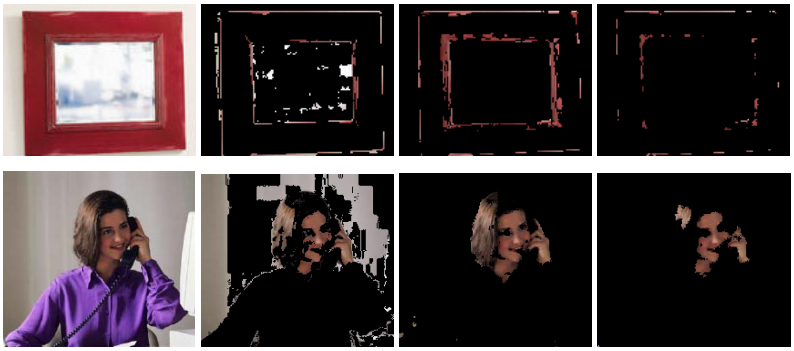


Fig. 10: The examples of skin color detection using (from the left to the right) the methods proposed in [50], [51] and in this section with the variables (L, a, b).

Fig. 11: The examples of skin color detection using (from the left to the right) the methods proposed in [50], in [51] and in this section with the variables (L, a, b) and, in the last column, with the variables (L, a, b, v(L, a, b)). Notice how the additional variable, related to texture properties, enhances the classification in images that do not contain skin.

VIII "NATURAL" QUERYING LANGUAGE

This section presents a natural language for querying image databases. Following the findings from our experiments, the vocabulary of the query language is based on the concept of semantic cues, while the syntax captures the basic patterns in the human perception of semantic templates and categories. The language we propose is simple but expressive. Both its vocabulary and its syntax are elementary. In effect, the words of the language are almost limited to the names of the semantic indicators. Being "elementary" visual cues, semantic indicators are often described with a single word (e.g. people, snow, mountain, object). These words are combined to construct sentences and express an assertion about the image, such as "the number of skin regions is greater than 5, scene is busy and dark". All the images in the database are tested against the query, and only those that satisfy the assertion are selected. Performing such a comparison requires

only the addition of a few words for expressing logical relationships (e.g. “and”), comparison operators (e.g. “greater than”, “equal to”, “similar to”), and several other words (e.g. “the”, “is”) that are common in any natural language. Although semantic cues alone do not carry the level of abstraction involved in our perception of image semantics, as our experiments have shown, humans use these cues when describing images, and communicating their own interpretation of the image content. It is certainly questionable whether this conscious self-introspection of the human subjects reflects what actually happens in their brain. Yet, humans use naturally these very same words when interacting with or designing an image querying system. Therefore, such queries may be qualified as “semantic”, as they carry a cognitive dimension introduced by the user. Consequently, the entire “system” we have developed can be qualified as “semantic”, as it is empowered by the brain of its user. In the next paragraphs, we present the language in details and demonstrate the semantic expressiveness of the querying system.

8.1 Vocabulary and syntax.

According to our findings, our goal is to develop the language with the following important “words”, and the associated syntax rules:

1. The names of the regions in the image: `skin`, `sky`, `grass`, `water`, `nature`, `background`, etc., are always used with the prefix `the area occupied by`, as in the sentence “the area occupied by sky is greater than 30%”.
2. Counters: `lines`, `details`, `objects`, `blobs`, etc. are always used with the prefix `the number of`, as in the sentence “the number of objects is 2”.
3. Dimensions: `width`, `height`, `length`, etc. are always used with the prefix `the`, as in the sentence “the width is equal to the height” (i.e. the width of the image of course).
4. Real and integer numbers are also words of the language, as 30 and 2 in the two previous examples.
5. Comparison relationships: `is greater than`, `is smaller than`, `is equal to`, `is similar to` can be used to form assertions as in the two previous examples.
6. Logical and grouping relationships: `or`, `and`, `( )`, `not` can be used to relate assertions and form other assertions as in the definition of the formal portrait “(the width is greater than the height) and (the background is greater than 20%) and (the number of skin regions is 1) and (the amount of skin is greater then 10%)”. The availability of logical relationships is, of course, recommended in any natural language - in our case they are introduced to support the notions of “required” and “frequently occurring” features, as described in Section 2.2. According to the verbal descriptions recorded in our experiments, assertions involving required features occur in logical conjunctions, as in the sentence “background is large **and** there are two objects”. On the other hand, assertions involving frequently occurring features occur in logical unions with other assertions of the same type, as in the sentence “contrast is low **or** color composition is pale”. Therefore, it was important to support this pattern in the language. The latter, however, is not restricted to these patterns only and allows for additional flexibility in composing the queries.

We have generalized these descriptions into the following formal grammar.

query::= <sentence>
sentence::= <assertion>
assertion::= (<assertion>) | <assertion> or <assertion> | <assertion> and <assertion> | not <assertion> | <regions_size> <comparison> <number> |
<counter> <comparison> <number> | <length> <comparison> <number> | <global> <comparison> <number> | the image is <composition>
length::= the size of the <length_name>
length_name::= width | height | background_boundary
regions_size::= the percentage of <regions_names>
regions_names::= skin | sky | grass | water | nature | central object | background | color object | color blob | green region
counter::= the number of <counter_name>
counter_name::= lines | details | regions | objects | skin regions | color blobs | color regions | colors
global::= the <global_name>
global_name::= contrast | x-symmetry | y-symmetry
composition::= greyish | pinkish | brownish | monochromatic | computer generated | bright | dark | dominant green
comparison::= is greater than | is smaller than | is equal to | is similar to

This grammar allows us to formulate queries in a concise form, which is equivalent to the verbal descriptions we use in everyday life. For instance, using the grammar, the previous examples can be written as:

$$\texttt{sky} > 0.3$$
$$\texttt{objects} = 20$$
$$\texttt{width} = \texttt{height}$$
$$\texttt{width} > \texttt{height} \ \wedge \ \texttt{background} > 20 \wedge \texttt{skin\_region} == 1 \ \wedge \ \texttt{skin} > 10$$
$$\texttt{background} > B_{high} \wedge \texttt{objects} = 2$$
$$\texttt{contrast} < C_{low} \vee \texttt{composition} = \text{"pale"}$$

where B_{high} and C_{low} are be preselected thresholds.

IX Experiments and Results

To evaluate our method, we have implemented an image-driven Internet search engine, ISEE (Image Search and Exploration Engine) and applied it to a large number of images collected from the World Wide Web. ISEE has two parts: an *image web robot*, and a *visual browser*, which provides a user interface for searching and browsing the image repository using the computed features and the querying scheme. Some screenshots from the browser are given in Fig. 12.

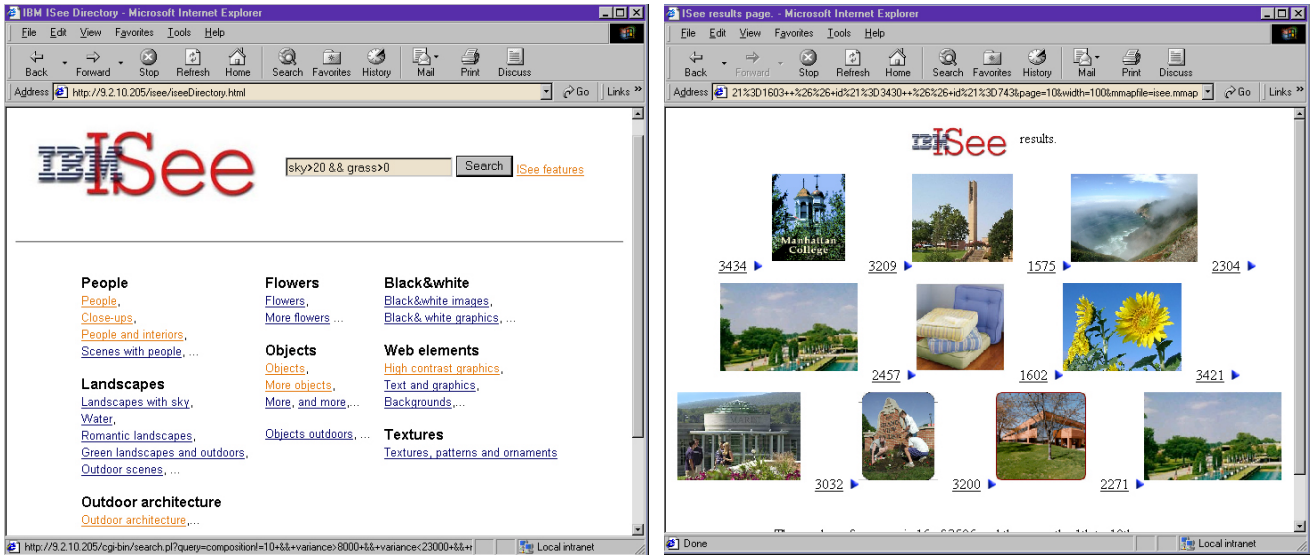


Fig. 12: The image on the left shows the main page from ISee. The user can type queries in the query window, or use the *Web directory* to search for images from certain category. The right image shows the result for the query "sky>20 && grass>0".

Using this browser we have tested the method with respect to the following four tasks: 1) modeling the categories identified through our experiments, 2) expressing new semantic categories in the same domain (i.e. photographic images), 3) expressing semantic categories in a new domain (e.g. medical images), and 4) using the language to measure image similarity (i.e. image retrieval).

9.1 Modeling the semantic categories

The results in this section, demonstrate that our features and language indeed "capture" the semantics of the basic categories identified in our experiments. Fig. 13a shows the result of the query that models the category "People":

```
nature<10 && contrast>8000 && contrast<23000 && regions>140 && skin_regions>=3 &&
skin_regions<10 && skin<15 && lines<4 && color_regions<15.
```

Similar queries can be written for other categories. Fig. 13b, 13c, and 13d show the results for categories “Outdoor Architecture”, “Flowers”, and “Objects Indoors”.

9.2 Expressing new semantic categories

The categories we have modeled are not rigid and deterministic – they depend on the task, user or environment. Hence, rather than the strict set of rules, the proposed model should be viewed as a dynamic system used to navigate through the world of images. Of course, with semantic indicators being the constructing bricks, a user may use her imagination for expressing new semantic categories in a natural manner.

For example, let us assume a new semantic template “High contrast graphics” (note that this template does not correspond to any of the categories from our experiments). The following query captures the semantics of this template:

```
(color_regions>35 && lines>0) || contrast>28000 || (colors<6 && (color_regions+color_blobs)>20)
```

The results for this query are shown in Fig. 14a.

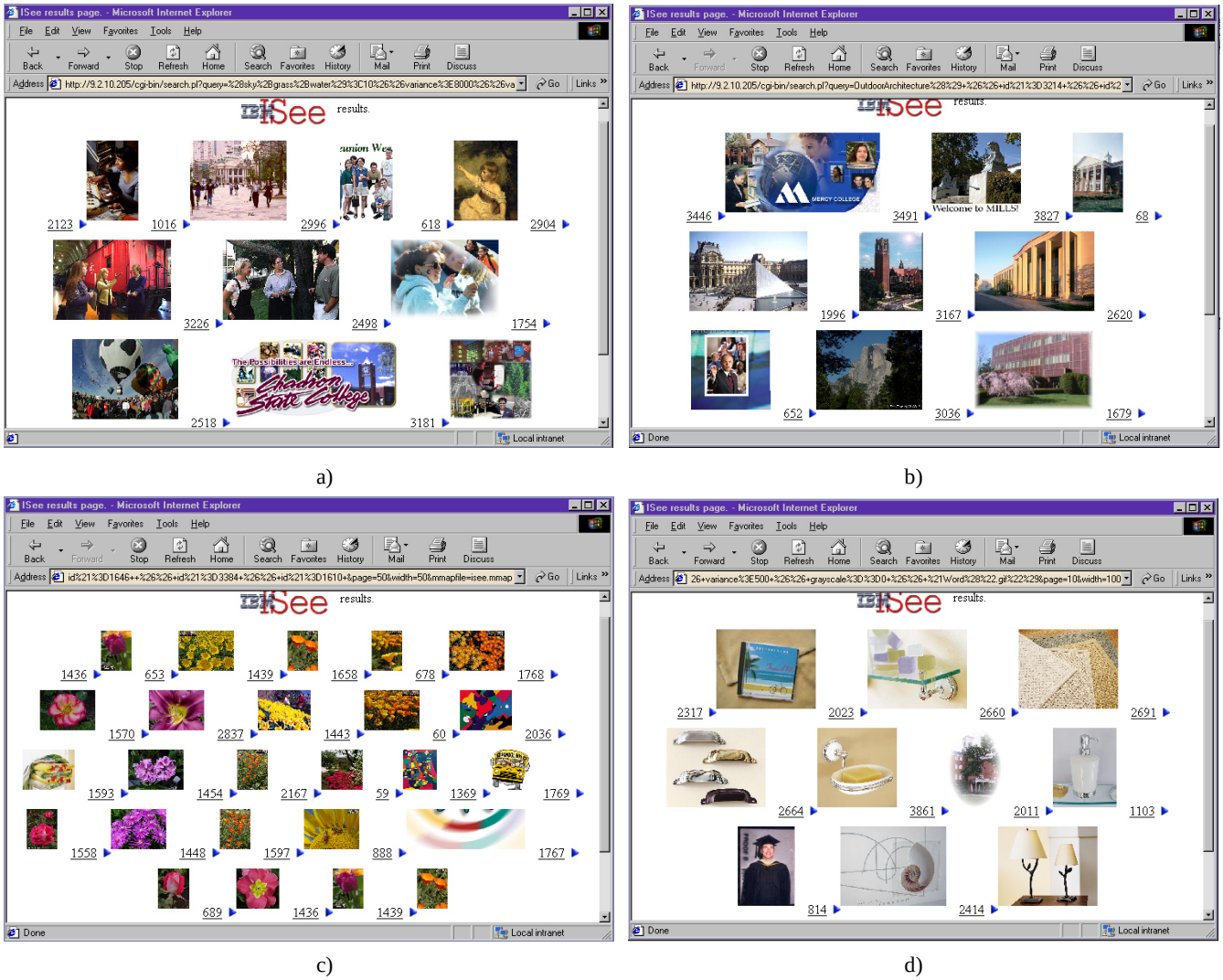


Fig. 13: The first page of the search results for categories “People”, “Outdoor architecture”, “Flowers”, and “Objects Indoors”.

9.3 Searching for similar images

So far, the queries like “an image similar to this example image” are not supported in the query language. However, it is natural to extend the language in order to include the information about known images. This was implemented by allowing the previously described assertions to be attached to the selected (reference) image. For instance, assuming the reference image, "ref", one can query “the contrast is the same as contrast in ref”, which, in its concise form writes `contrast == ref.contrast`. All the images in the database are then tested against this assertion and only those that satisfy it are selected.

Of course, a more realistic query would be “the contrast is similar to the contrast in ref”, which can be written as:

```
contrast> ref.contrast-ref.contrast/10 && contrast< ref.contrast+ref.contrast/10
```

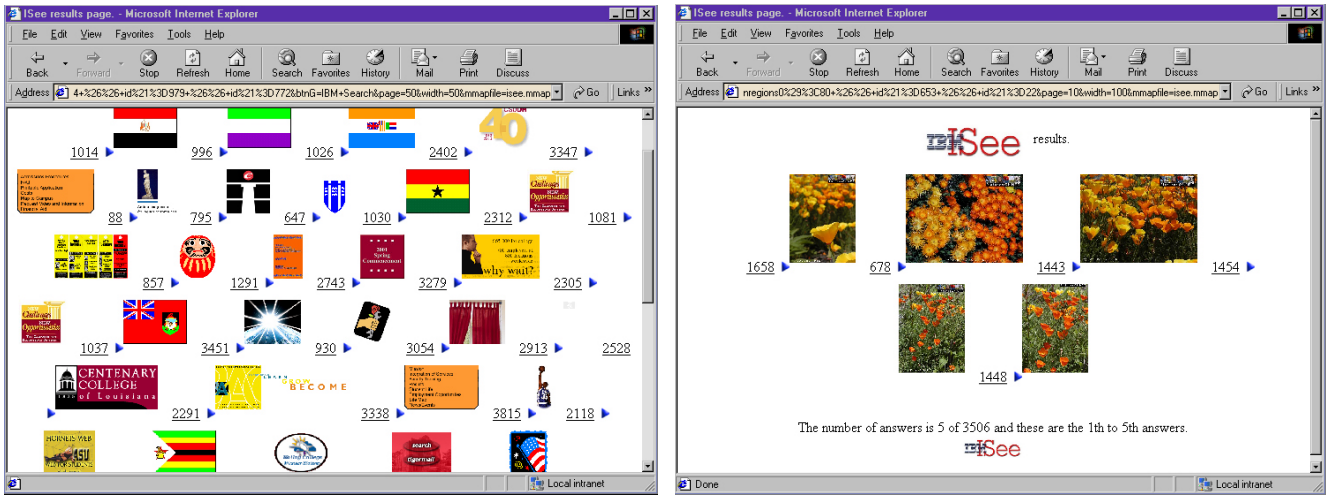
Or by adding a new similarity operator:

```
similar(contrast, ref.contrast)<ref.contrast/10
```

where the construction `similar(.)` represents a simple addition to the language. This extension to the language gives the user the opportunity to query images whose visual attributes are similar to the visual attributes of the selected images. Of course, as in all other cases, the interpretation of similarity is in the user’s mind. This search for similar images can be demonstrated through the following example. In order to find images with vivid-colored flowers in the field, similar to the image number 653 (second image in the top row in Fig. 13c) the user can use the following query:

```
Flowers() && similar(regions, Image(653).regions)<50 && composition==Image(653).composition
```

where in the feature domain *regions* is a measure of image activity (busy image with many details), and *composition* is the measure of color composition. The results for this query are given in Fig. 14b.



a) b)

Fig. 14: a) The first page of the search results for newly introduced category “High-contrast graphics”. b) An image retrieval example.

9.3 Working in a different domain

So far we have focused on semantics within the broad domain of photographic images. The semantic models we are relying upon in capturing the meaning of an image, the underlying cues, and their features have been develop and tested on images from that domain. To go one step further, we were interested in testing the ability of our algorithms to “generalize”, and work in a new domain, a domain with entirely different interpretation of semantics, different templates, cues and perception of image similarity. To test whether our algorithm is robust for such an application we have redesigned our Internet portal, to search and browse medical databases registered on the Internet.

By observing the important characteristics and visual properties for different modalities used in medical-imaging (for example X-rays, histologies, stainings, micrographs, MRI, pathologies, etc.) we have design the queries and applied them to categorize images according to these modalities. For example, the following query can be used to look for images from "Stainings – Papanicolau" category:

```
objects>4 && composition="pinkish" && vertical_lines=1
```

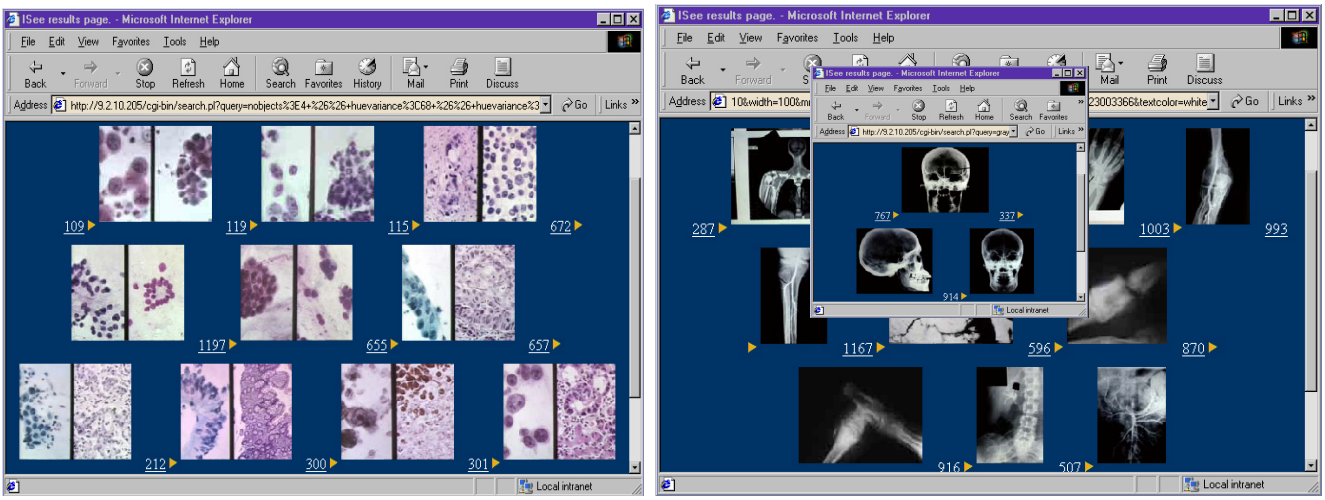
Similarly, to search for X-ray images, the user may use the following query:

```
grayscale2==1 && nregions0<100 && variance>6000 && background_boundary>70
```

or to further narrow the search to the X-ray images of skull:

```
Xray() && nobjects==1 && (excentricity1<2.6 && ysymmetry>0.4) && (background2>50 && background2<75)
```

where the query X-ray() was defined as above. The results for these queries are shown in Fig. 14.



a) b)

Fig. 14: a) The first page of the search results for category "Stainings - Papanicolau".b) a) The first page of the search results for category "X-ray", and "X-ray – Skull".

X DISCUSSIONS AND APPLICATIONS OF THE METHOD

10.1 More on the experiments, some limitations and selection of semantic categories

Through our experiments, we have identified candidate semantic templates and semantic cues human observers use to organize photographic images, and have modeled them using the combinations of pictorial features. Since image similarity is such a complex judgment, it is not realistic to believe that we have uncovered the final immutable set of categories. We consider these to be a useful starting point, not a divinely inspired set. As the experimental decisions represent only a partial (and subjective) view of the image world, we are not making an attempt in claiming the generality of the solution.

As we mentioned before, testing the human perception is an extremely difficult task, and we are aware that the experimental setup poses many limitations to our conclusions. The size of the data set, incremental nature of the procedure, and the techniques used may have biased the outcome of the experiments and the derivation of the semantic categories. Having said this, we would like to add that the system we propose does not require that the categories be perfect. It only requires that the categories capture enough of the information in the similarity judgments to help organize the images semantically. We believe that the most useful applications for image libraries will not be those that retrieve exact image matches, but those that provide a meaningful context for browsing and navigation. The current approach certainly provides a differentiation in such tasks.

10.2 More on features, image categorization and retrieval result

We have shown that our system is fairly general and can satisfy queries for which it had not been explicitly designed. However, this level of generality has its price, and there are limitations, which need to be discussed.

We have indeed proposed several new algorithms, and used some of the “well-known” features in a novel way. However, these are only the first-steps, and many of our features can be further improved. Observe, for instance, the results of the query "Flowers()" in Fig. 13c. In addition to many nice guesses, the system has also retrieved an image of towels, a school buss and the rainbow, mainly because these images also have these typical “flower-looking” colors. And we do not find it surprising to discover that a perceptual system suffers from "illusions", just as humans do. These errors are categorized as “false positives” and are acceptable as long as their number is relatively low (we found that many users simply ignore them). But, as a careful reader may have realized, this drawback is just the tip of the iceberg. The system produces "false negatives" as well, i.e. images of beautiful flowers that did not show up because they did not satisfy the query. These errors are a more important concern, since they remain invisible. The only way for the user to know what he or she is missing is to traverse the entire database! So, our system has an intrinsic bias toward showing good results (and hiding the bad ones). This is a “very pleasant bias” for the user who interacts with the system, but it is a limitation. Many of the “false negatives” can be eliminated by further improving the feature extraction schemes. Furthermore, the querying language being natural and the parameters intuitive, the user may use her understanding to correct these errors through multiple tries. However, except by traversing the entire database, there is no exact solution for the problem of “missing images”, and yet it would not be “exact” because two different users may not always agree whether a certain image is a flower or not! An interactive and tractable approach to apprehend "what you are missing" is to relax the query progressively, (make it less and less restrictive). By observing how, and how fast the ratio between correct answers and “false positives” changes, one may apprehend the importance of the missing information. Of course, such a learning scheme would be an interesting future research direction. Another important aspect for the future work is integrating the visual features with the semantic aspects contained in text descriptors, see for example methods proposed by Santini [52], [53].

10.3 Conclusions

The world of digital image libraries is growing rapidly, driving the need to develop better, more intuitive tools for searching, navigating and browsing image collections. In this paper, we have proposed a method for semantic-based image retrieval, categorization and browsing, based on low-level image descriptors derived from perceptual experiments. Although our study was limited to the elementary semantic templates and cues, we believe that our results are promising, as our approach seemed robust when tested with images especially selected to challenge our model. This was demonstrated while testing the method on the large set of images collected on the Internet. Our prototype implementation demonstrates the potential value of the natural query language in image navigation. We believe that our method lays the groundwork for enhancing current image/video retrieval methods, for the better organization of large image databases, and for the development of more intuitive navigation schemes, browsing methods and user interfaces.

REFERENCES

[1] M. Fleck, D. A. Forsyth, and C. Bregler, “Finding naked people”, *Proc. European Conf. Computer Vision*, vol. 2, pp. 593-602, 1996.

[2] J. Z. Wang, J. Li, G. Wiederhold, and O. Firschein, “System for screening objectionable images”, *Computer Comm*., vol. 21, no. 15, pp. 1355-1360, 1998.

[3] J. R. Bach, S. Paul, and R. Jain, “A visual information management system for the interactive retrieval of faces”, *KnowData*, vol. 5, no. 4, pp. 619-628, August 1993.

[4] Y. Li, B. Tao, S. Kei, W. Wolf, “Semantic image retrieval through human subject segmentation and characterization”, *Proc. SPIE Storage and Retrieval for Image and Video Databases*, vol. 3022, pp.340-351, 1997.

[5] A. Vailaya, A. Jain, and H. J. Zhang, “On image classification: city versus landscape”, *Proc. IEEE Workshop Content-based Access of Image and Video Libraries*, pp. 3-8, June 1988.

[6] M. Szummer and R. Picard, “Indoor-outdoor classification”, *Proc. Int. Workshop on Content-Based Access of Image and Video Databases*, pp. 42-51, Bombay, India, Jan. 1998.

[7] A. Mojsilovic, J. Kovacevic, J. Hu, R. J. Safranek, K. Ganapathy, "Matching and retrieval based on the vocabulary and grammar of color patterns", *IEEE Trans. on Image Proc.*, vol. 9, no. 1, pp. 38-54, January 2000.

[8] W. Zhu, and T. Syeda-Mahmood, “Image organization and retrieval using a flexible shape model”, *Proc. Int. Workshop on Content Based Access of Image and Video Databases*, pp..31-39, Bombay, India, Jan. 1998.

[9] J. Z. Wang, J. Li, and G. Wiederhold, “SIMPLIcity: Semantics-sensitive integrated matching for picture libraries”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, Sept. 2001

[10] W. Niblack, R. Berber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, and P.Yanker, “The QBIC project: Quering images by content using color, texture and shape”. in *Proc. SPIE Storage and Retrieval for Image and Video Data Bases*, pp. 172-187, 1994.

[11] A. Pentland, R. W. Picard, and S. Sclaroff, “Photobook: Content-based manipulation of image databases”, *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233-254, 1996.

[12] J. R. Smith, and S. Chang, “VisualSeek: A fully automated content-based query system”, in *Proc. ACM Multimedia*, pp. 87-98, 1996.

[13] Y. Rui, T. S. Huang, and S. Mehrotra, “Content-based image retrieval with relevance feed-back in Mars”, in *Proc. IEEE Conf. on Image Processing*, vol. II, pp. 815-818, 1997.

[14] W. Y. Ma, and B. S. Manjunath, “Netra: A toolbox for navigating large image databases”. in *Proc. IEEE Int. Conf. on Image Processing*, vol. I, pp. 568-571, 1997.

[15] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval: the end of the early years”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349 -1380, December 2000.

[16] J. M. Corridoni, A. Del Bimbo, P. Pala, "Image Retrieval by Color Semantics", *ACM Multimedia Systems*, Vol. 7, no. 7, 1999.

[17] C. Colombo, A. Del Bimbo, P. Pala "Semantics in Visual Information Retrieval",*IEEE Multimedia*, Vol. 6, no. 3, 1999.

[18] M. Wertheimer, “Untersuchungen zur Lehe von der Gestalt II”, *Psychol. Forsch.*, 4, 1923. Translated as “Principles of perceptual organization”, in *Readings in Perception*, D. Beardslee and M. Wertheimer, Eds., (Princeton, N.J.: Van Nostrand, 1958), 115-135.

[19] S. Zucker, “Computational and psychophysical experiments in grouping: Early orientation selection”, in Human and Machine Vision, Beck, Hope & Rosenfeld, Eds. (New York: Academic Press, 1983), 545-567.

[20] D. Marr, *Vision*, San Francisco: W.H. Freeman and Co., 1982.

[21]A. P. Witkin, and J. M. Tenenbaum, “On the role of structure in vision”, in Human and Machine Vision, Beck, Hope & Rosenfeld, Eds., New York: Academic Press, 1983, 481-543.

[22] A. Mojsilovic, and B. Rogowitz, “Capturing image semantics with low-level descriptors”, In *Proc. IEEE International Conference on Image Processing*, ICIP 2001, Thessaloniki, Greece, October 2001.

[23] B. Rogowitz, T. Frese, J. Smith, C. A. Bouman, and E. Kalin, “Perceptual image similarity experiments”, in *Proc. of SPIE*, 1997.

[24] M. Turk and A. Pentland, “Eigenfaces for recognition”, *J. Cogn. Neurosci*., vol. 3, pp. 71-86, 1991.

[25] D. Forsyth and M. Fleck, “Body plans”, in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, San Juan, 1997, pp. 678-683.

[26] A. Mojsilovic, J. Kovacevic, D. Kall, R. J. Safranek, and K. Ganapathy, "Vocabulary and grammar of color patterns", *IEEE Trans. on Image Processing*, vol. 9, no. 3, pp. 417-431, March 2000.

[27] D. Rabenhorst, *Opal: Users manual*, IBM Research Internal Document. (See also www.research.ibm.com/visualanalysis/Opal.html)

[28] G. Wyszecki, W. Stiles, *Color science: Concepts and methods, quantitative data and formulae*, John Wiley&Sons, New York, 1982.

[29] J. M. Lammens, “A computational model of color perception and color naming”, *Ph.D. Thesis*, Univ. of Buffalo, June 1994.

[30] J. Koenderinck, “The structure of images”, *Biol. Cybern.*, vol. 50, pp. 363--370, 84.

[31] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629--639, July 1990.

[32] M. Grayson,“The heat equation shrinks embedded plane curves to round points”, *J. Differential Geom.*, vol. 26, pp. 285-314, 1987.

[33] J. A. Sethian, *Level Set Methods*, Cambridge University Press, 1996.

[34] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit: an Object-Oriented Approach to 3D Graphics*, Prentice Hall, 1996.

[35] E. R. Kandel, J. H. Schwartz, and T. M. Jessel, *Principles of neural science*, Appleton and Lange, New York, 1991.

[36] A. Mojsilovic, "A method for color naming and description of color composition in images", *Proc. IEEE Int. Conf. Image Processing*, Rochester, New York, Sept. 2002. [37] D.Comaniciu, P.Meer, “Mean shift analysis and applications”, *IEEE Int. Conf. Comp. Vis.*, pp.1197-1203, Greece, 1999.

[38] B. Berlin and P. Kay, *Basic Color Terms: Their Universality and Evolution*, Berkeley: University of California, 1969.

[39] K. Kelly, and D. Judd, “The ISCC-NBS color names dictionary and the universal color language (The ISCC-NBS method of designating colors and a dictionary of color names)”, *NBS Circular* 553, November 1, 1955.

[40] J. Gomes, and A. Mojsilovic, “Variational approach to recovering a manifold from sample points”, to appear in *Proc. of the Seventh European Conference on Computer Vision*, May 2002.

[41] S. Kachigan, *Statistical Analysis*, Radius Press, 1986.

[42] S. Hays, W. Statistics, and Y. Holt, *Statistics*, Holt, Rinehart and Winston, New York, 1981.

[43] L. R. Lamberson, G. F. Gruska, and K. Mirkhani, *Non-Normal data Analysis*, Garden City, MI: Multiface Publishing., 1967.

[44] G. Hahn and S. Shapiro, *Statistical Models in Engineering*, John Wiley & Sons, 1967.

[45] StatSoft Inc., *Electronic Statistics Textbook*, volume http://www.statsoftinc.com/ textbook/ stathome.html.Tulsa, StatSoft, 2001.

[46] C. M. Bishop. *Neural Networks for Pattern Recognition*.Oxford University Press, Oxford, 1995.

[47] Luigi Ambrosio and Halil M. Soner, “Level set approach to mean curvature flow in arbitrary codimension”, *J. of Diff. Geom.*, vol. 43, pp. 693-737, 1996.

[48] A. Maerz, and M. R. Paul, *A dictionary of color*, McGraw-Hill Book Co., Inc., New York, NY 1930.

[49] L. Lorigo, O. Faugeras, W.E.L. Grimson, R. Keriven, R. Kikinis, and C-F. Westin, “Co-dimension 2 geodesic active contours for MRI segmentation”, *Proc. of the International Conference on Information Processing in Medical Imaging*, pages 126--139, June 1999.

[50] C. Garcia, G. Tziritas, "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis", *IEEE Transactions on Multimedia*, vol. 1, no. 3, pp.264-277, September 1999.

[51] K. Sobottka and I. Pitas, "Extraction of facial regions and features using color and shape information", *International Conference on Pattern Recognition (ICPR'96)*, Vienna, Austria, vol. III, pp. C421-C425, 25-29 August 1996.

[52] S. Santini ,“Mixed media search in image databases using text and visual similarity”, *ICME 2001, IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, August 2001.

[53] S. Santini, “A query paradigm to discover the relation between text and images”, *Proceedings of SPIE*, Vol. 4315, Storage and Retrieval for Media Databases 2001, San Jose, Jan. 2001

[54] S. E. Palmer, *Vision Science: Photons to phenomenology*, MIT Press, April 1999.

[55] S. Tominaga, “A colour-naming method for computer color vision”, *Proc. of the 1985 IEEE Int. Conf. on Cybernetics and Society*, pp. 573-577, Tucson, Arizona.

[56] A. K. Jain, A. Vailaya, “Shape--Based Retrieval: A case study with trademark images”, *Pattern Recognition*, vol. 31, No 9, pp. 1369-1390, 1998.

[57] T. Belpaeme, *Factors influencing the origins of colour categories*, Ph.D Thesis, Vrije Universiteit Brussel, 2002.

[58] H. Munsell, *A Color Notation*, Munsell Color Company, Baltimore MD, 1946.