Content-Based Image Retrieval at the End of the Early Years

Arnold W.M. Smeulders, *Senior Member*, *IEEE*, Marcel Worring, Simone Santini, *Member*, *IEEE*, Amarnath Gupta, *Member*, *IEEE*, and Ramesh Jain, *Fellow*, *IEEE*

Abstract—The paper presents a review of 200 references in content-based image retrieval. The paper starts with discussing the working conditions of content-based retrieval: patterns of use, types of pictures, the role of semantics, and the sensory gap. Subsequent sections discuss computational steps for image retrieval systems. Step one of the review is image processing for retrieval sorted by color, texture, and local geometry. Features for retrieval are discussed next, sorted by: accumulative and global features, salient points, object and shape features, signs, and structural combinations thereof. Similarity of pictures and objects in pictures is reviewed for each of the feature types, in close connection to the types and means of feedback the user of the systems is capable of giving by interaction. We briefly discuss aspects of system engineering: databases, system architecture, and evaluation. In the concluding section, we present our view on: the driving force of the field, the heritage from computer vision, the influence on computer vision, the role of similarity and of interaction, the need for databases, the problem of evaluation, and the role of the semantic gap.

Index Terms—Review, content based, retrieval, semantic gap, sensory gap, narrow domain, broad domain, weak segmentation, accumulative features, salient features, signs, structural features, similarity, semantic interpretation, query space, display space, interactive session, indexing, architecture, evaluation, image databases.

1 Introduction

There is something about Munch's "The Scream" or Constable's "Wivenoe Park" that no words can convey. It has to be seen. The same holds for of a picture of the Kalahari Desert, a dividing cell, or the facial expression of an actor playing King Lear. It is beyond words. Try to imagine an editor taking in pictures without seeing them or a radiologist deciding on a verbal description. Pictures have to be seen and searched as pictures: by objects, by style, by purpose.

Research in content-based image retrieval today is a lively discipline, expanding in breadth. As happens during the maturation process of many a discipline, after early successes in a few applications, research is now concentrating on deeper problems, challenging the hard problems at the crossroads of the discipline from which it was born: computer vision, databases, and information retrieval.

At the current stage of content-based image retrieval research, it is interesting to look back toward the beginning and see which of the original ideas have blossomed, which haven't, and which were made obsolete by the changing landscape of computing. In February 1992, the US National Science Foundation (USNSF) organized a workshop in Redwood, California, to "identify major research areas that should be addressed by researchers for visual information management systems that would be useful in scientific, industrial, medical, environmental, educational, entertainment, and other applications" [81]. In hindsight, the workshop did an excellent job of identifying unsolved problems that researchers should have undertaken. In particular, the workshop correctly stated that "Visual Information Management Systems should not be considered as an application of the existing state of the art (in computer vision and databases) to manage and process images" and that "computer vision researchers should identify features required for *interactive image understanding*, rather than their discipline's current emphasis on automatic techniques" (emphasis added). As possible application fields, the workshop considered mainly Grand Challenge problems, such as weather forecasting, biological modeling, medical images, satellite images, and so on. Undoubtedly, the participants saw enough to justify the use of the large computational and storage capacity necessary for visual databases. This in 1992. The workshop was preceded by many years by the Conference on Database Applications of Pictorial Applications, held in Florence in 1979, probably one of the first conferences of that kind [13]. In the introduction, it was said that: "This has facilitated the advancement of integrated databases [...] on the one hand, of and graphical and image processing (in brief: pictorial) applications on the other." Then, the author proceeds to complain that: "Developments in these two fields have traditionally been unrelated," an observation still very much valid today.

- *A.W.M. Smeulders and M. Worring are with Intelligent Sensory Information Systems, University of Amsterdam, faculty WINS Kruislaan 403, 1098 SJ Amsterdam, The Netherlands. E-mail: {smeulders, worring}@wins.uva.nl.*
- *S. Santini and A. Gupta are with the Department of Electrical and Computer Science Engineering and the San Diego Super Computer Center, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92039-0407. E-mail: ssantini@ece.ucsd.edu.*
- *J. Ramesh is with Praja, Inc., 10455-B Pacific Center Court, San Diego, CA 92121. E-mail: jain@praja.com.*

Manuscript received 23 July 1999; revised 12 May 2000; accepted 12 Sept. 2000.
Recommended for acceptance by K. Bowyer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 110298.

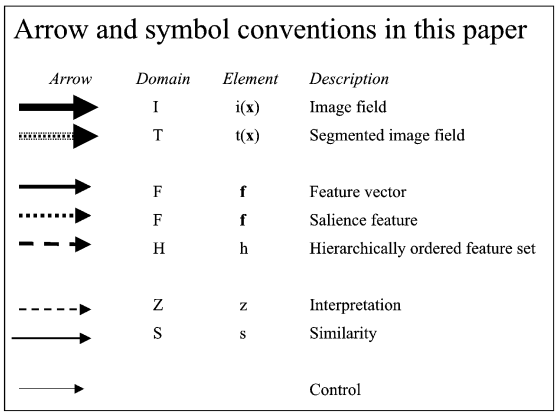


Fig. 1. Data flow and symbol conventions as used in this paper. Different styles of arrows indicate different data structures.

Just after the USNSF workshop, the Mosaic Internet-browser was released, spawning the Web revolution that very quickly changed all cards. In the same era, a host of new digital vision sensors became available. The number of images that the average user could reach increased dramatically in just a few years. Instantly, indexing tools for the Web or digital archives became urgent.

In this paper, we present a view of what we like to call the "early years" of content-based image retrieval. While content based-image retrieval papers published prior to 1990 are rare, often obsolete, and of little direct impact today, the number of papers published since 1997 is just breathtaking. So much, in fact, that compiling a comprehensive review of the state of the art already exceeds the possibility of a paper like this one. A selection was necessary and with it came the need to establish some selection criteria. In addition to the obvious one (completeness of a paper, importance to the field), we have also considered accessibility for the reader. That is to say, we have preferred, whenever possible, to include journal papers over conference papers. We also felt that the field is too young and mobile to make a precise historic account and we have made no attempt in that direction.

We adopt patterns of use and patterns of computation as the leading principles of our review. We follow the data as they flow through the computational process and consider alternative processes with the same position in the flow (Fig. 2). In the data flow diagrams, we use the conventions indicated in Fig. 1. We concentrate on computational methods to arrive at a tool-based overview rather than a system-based overview. The choice implies that references describing complete systems are split, where parts of the method will appear in several sections of the paper. For a system-based review, see [141].

We restrict ourselves to still pictures and leave video databases as a separate topic. Video retrieval could be considered a broader topic than image retrieval as video is built from single images. From another perspective, video retrieval could be considered simpler than image retrieval since video reveals its objects more easily as the points corresponding to one object move together. In still pictures, the author's narrative expression of intention is in frame selection, illumination, and composition. In addition, video has a linear timeline, as important to the narrative structure of video as it is in text. We leave video retrieval for another place, for example, [1], [16].

The paper is organized as indicated in Fig. 2. First we discuss the scope of the content-based retrieval in Section 2. In that section, the characteristics of the domain and sources of knowledge are being discussed. Then, description of content is analyzed in two steps. First, in Section 3, image processing methods by color, texture, and local shape are discussed. They serve as a preprocessing step to the partitioning of the data array and the computation of features, as discussed in Section 4. In Section 5, we discuss the interpretation of a single image and the similarity

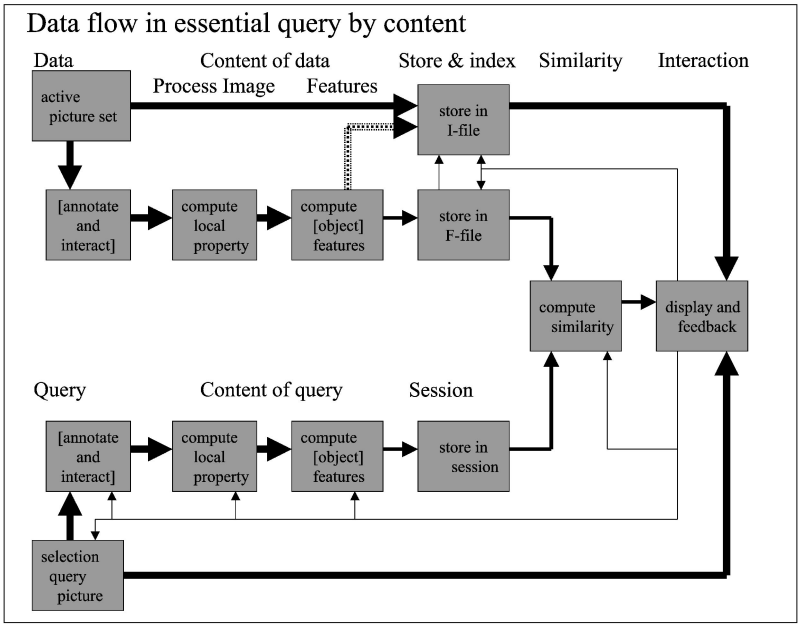


Fig. 2. Basic algorithmic components of query by pictorial example captured in a data-flow scheme while using the conventions of Fig. 1.

Target-, category- and association-search in image retrieval

	Target	*Category*	*Association*
Object goal	1 specific object	an arbitrary object from 1 specific class	not defined at start
Query by example	1 ... N objects	1 ... N objects with class labels	N objects plus association
Similarity	feature-based	class driven	session-specific
Events in F-space	proximity to query	class membership	clusters
Feedback	rank ordered on proximity	likelihood on class membership	relevance feedback on association value
Interactive update:			
of images of query	-	expand query	refine on the way
of features of query	refine on the way	refine on the way	alter on the way
of similarity measure	-	adapt to group	reshape to goal

Fig. 3. Three patterns in the purpose of content-based retrieval systems.

between a pair of images. Query definition, display, and interaction are the topic of Section 6. The paper concludes at the level of systems: indexing, system architecture, and evaluation of performance. Each chapter is concluded by a discussion on the state of the art.

2 Scope

In the literature, a wide variety of content-based retrieval methods and systems may be found. In this section, we discuss patterns in applications, the repertoire of images, the influence of the scene and the role of domain knowledge, and the semantic gap between image features and the user.

2.1 Applications of Content-Based Retrieval

In [31], we see three broad categories of user aims when using the system, see Fig. 3.

- There is a broad class of methods and systems aimed at browsing through a large set of images from unspecified sources. Users of *search by association* at the start have no specific aim other than find interesting things. Search by association often implies iterative refinement of the search, the similarity or the examples with which the search was started. Systems in this category typically are highly interactive, where the specification may by sketch [30] or by example images. The oldest realistic example of such a system is probably [88]. The result of the search can be manipulated interactively by relevance feedback [68], [51]. To support the quest for relevant results, other sources than images are also employed, see for example, [168], [21].
- Another class of users *aims the search* at a specific image. The search may be for a precise copy of the image in mind, as in searching art catalogues, e.g., [48]. Target search may also be for another image of the same object of which the user has an image. This is target search by example. Target search may also be applied when the user has a specific image in mind and the target is interactively specified as similar to a group of given examples, for instance [31]. These systems are suited to search for stamps, art, industrial components, and catalogues, in general.
- The third class of applications, *category search*, aims at retrieving an arbitrary image representative of a specific class. It may be the case that the user has an example and the search is for other elements of the same class. Categories may be derived from labels or emerge from the database [170], [186]. In category search, the user may have available a group of images and the search is for additional images of the same class [28]. A typical application of category search is catalogues of varieties. In [74], [79], systems are designed for classifying trademarks. Systems in this category are usually interactive with a domain specific definition of similarity.

These three types of use are not the whole story [42]. A study [121] of journalists identified five typical patterns of use: searches for one specific image, general browsing to make an interactive choice, searches for a picture to go with a broad story, searches to illustrate a document, and searches for fill-ins only on the esthetic value of the picture. An attempts to formulate a general categorization of user requests for still and moving images are found in [6]. This and similar studies reveal that the range of queries is wider than just retrieving images based on the presence or absence of objects of simple visual characteristics.

2.2 The Image Domain and the Sensory Gap

In the repertoire of images under consideration—the image domain $\mathcal{I}$—there is a gradual distinction between narrow and broad domains [160]. At one end of the spectrum, we have the narrow domain:

> *A narrow domain has a limited and predictable variability in all relevant aspects of its appearance.*

In a narrow domain, one finds a limited variability of the content of the images. Usually, the recording circumstances are also similar over the whole domain. In the narrow domain of lithographs, for instance, the recording is under white light with frontal view and no occlusion. Also, when the object's appearance has limited variability, the semantic description of the image is generally well-defined and, by and large, unique. Another example of a narrow domain is a set of frontal views of faces recorded against a clear background. Although each face is unique and has large variability in the visual details, there are obvious geometrical, physical, and color-related constraints governing the domain. The domain would be wider had the faces been photographed from a crowd or from an outdoor scene. In that case, variations in illumination, clutter in the scene, occlusion, and viewpoint will have a major impact on the analysis.

On the other end of the spectrum, we have the broad domain:

> *A broad domain has an unlimited and unpredictable variability in its appearance even for the same semantic meaning.*

In broad domains, images are polysemic and their semantics are described only partially. It might be the case that there are conspicuous objects in the scene for which the object class is unknown or even that the interpretation of the scene is not unique. A broad class of images can be found in large photo stocks [168] or other photo archives [42]. The

Narrow versus broad domain in image retrieval

	Narrow	*Broad*
Variance of content	low	high
Sources of knowledge	specific	generic
Semantics	homogeneous	heterogeneous
Ground truth	likely	unlikely
Content description	objective	subjective
Scene and sensor	possibly controlled	unknown
Aimed application	specific	generic
Type of application	professional	public
Tools	model-driven, specific invariants	perceptual, cultural, general invariants
Interactivity	limited	pervasive, iterative
Evaluation	quantitative	qualitative
System architecture	tailored database-driven	modular interaction-driven
Size	medium	large to very large
A source of inspiration	object recognition	information retrieval

Fig. 4. Quick reference to narrow versus broad domains.

broadest class available to date is the set of images available on the Internet.

Many problems of practical interest have an image domain in between these extreme ends of the spectrum, see Fig. 4. The notions of broad and narrow domains are helpful in characterizing patterns of use, in selecting features, and in designing systems. In a broad image domain, the gap between the feature description and the semantic interpretation is generally wide. For narrow, specialized image domains, the gap between features and their semantic interpretation is usually smaller, so domain-specific models may help. For faces, many geometric models have been suggested, as well as statistical models [127]. These computational models are not available for broad image domains as the required number of computational variables would be enormous.

For broad image domains in particular, one has to resort to generally valid principles. Is the illumination of the domain white or colored? Does it assume defined and fully visible objects or may the scene contain clutter and occluded objects? Is it a 2D-recording of a 2D-scene or a 2D-recording of a 3D-scene? The given characteristics of illumination, presence or absence of occlusion, clutter, and differences in camera viewpoint determine demands on the retrieval methods.

> *The sensory gap is the gap between the object in the world and the information in a (computational) description derived from a recording of that scene.*

The sensory gap makes the description of objects an ill-posed problem: It yields uncertainty in what is known about the state of the object. The sensory gap is particularly poignant when a precise knowledge of the recording conditions is missing. The 2D-records of different 3D-objects can be identical. Without further knowledge, one has to decide that they *might* represent the same object. Also, a 2D-recording of a 3D-scene contains information accidental for that scene and that sensing but one does not know what part of the information is scene related. The uncertainty due to the sensory gap not only holds for the viewpoint, but also for occlusion (where essential parts telling two objects apart may be out of sight), clutter, and illumination.

Comparing alternative interpretations can attenuate the sensory gap. Content-based image retrieval systems may provide support in this disambiguation through elimination among several potential explanations, much the same as in natural language processing.

2.3 Domain Knowledge

In visual search, explicit representation of the knowledge of the domain is important to alleviate the sensory gap. Among the sources of general knowledge, we mention:

- Laws of syntactic (literal) equality and similarity define the relation between image pixels or image features regardless of its physical or perceptual causes. For instance, considering two images similar because they both exhibit some selected shades of blue in their upper parts is productive in separating outdoor scenes from other images. It is syntactic similarity because the method doesn't make a reference to the reasons by which this similarity exists (in this case, the scattering in the sky) or to the perceptual reasons by which these two images will appear similar to an observer. By the same token, the RGB color space is effective in literal similarity (as it is effective in art [65]) while it does not represent the process of physical color formation or the process of color perception.
- Laws describing the human perception of equality and similarity are important because they define equality on the same basis as the user experiences it. In color, the CIE-Lab and Munsell-spaces were designed to conform to the human perception of color similarity. If the appreciation of a human observer of an object is based on the perception of certain conspicuous items in the image [177], it is natural to direct the computation of broad domain features to these points and regions [157], [138]. Similarly, a biologically plausible architecture [76] of center-surround processing units is likely to select regions which humans would also focus on first.
- Physical laws describing equality and difference of images under differences in sensing and object surface properties. The physics of illumination, surface reflection, and image formation have a general effect on images. The general laws of physics may be employed for large classes of objects. A common example is the law for uniform light reflection off matte objects. These laws are exploited to design color features expressing equality regardless of the pose and viewpoint.
- Geometric and topological rules describe equality and differences of patterns in space. When two objects are geometrically equal, the physical properties of their surfaces or the physical conditions of the sensing may be different. As an example of geometric laws used in retrieval, for all images with depth, local details near the horizon will appear smaller. Also, the horizon is geometrically defined as

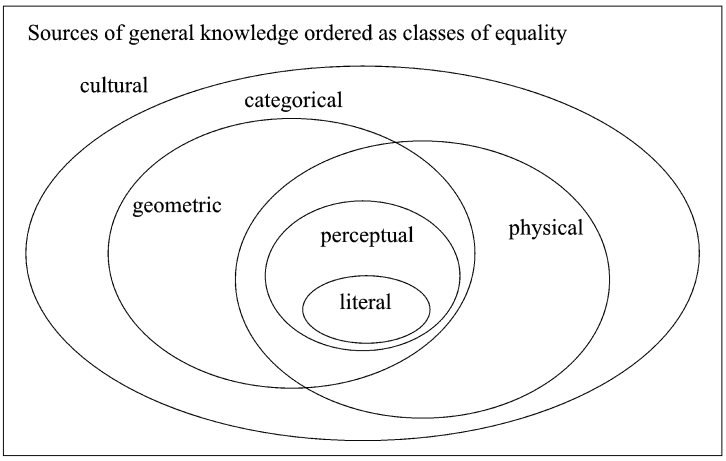


Fig. 5. When searching for a "chair," we may be satisfied with any object under that name, that is, we search for man-defined equality. When we search for all one leg chairs, we add an additional constraint to the general class and restrict the equality class. The same holds when searching for a red chair, adding a condition independent of the geometrical constraint. When we search for a chair perceptually equivalent to a given chair, at least physical and geometrical equality must hold. Finally, when we search for exactly the same image of that chair, literal equality is requested, still ignoring the variations due to noise, of course.

a virtual line containing the focal points. Another example of geometric laws is the expression of spatial [22] or topological relationships [172] between objects.

- Category-based rules encode the characteristics common to class z of the space of all notions $\mathcal{Z}$. If z is the class of all teapots, the characteristics include the presence of a spout. Categories are almost exclusively used in a narrow domains. The domain knowledge may take the form of further constraints to the literal image qualities, additional physical or geometrical laws, or domain-specific man-made customs. When the domain is engineering drawings, detailed geometric knowledge will steer the detection of symbols. In medieval art, color and the relative position of objects have a symbolic meaning [30], generating a set of constraints useful in the search. Each application domain has a private set of constraints.
- Finally, man-made customs or man-related patterns introduce rules of culture-based equality and difference. Under culture, we also assume language. In the search for indoor pictures, one may check for many straight lines and perpendicular corners as a first selection criterion. Utensils have a deterministic size to allow grip. Fashion determines colors [95].

These laws are ordered as indicated in Fig. 5.

2.4 Use and User, the Semantic Gap

We opine that most of the disappointments with early retrieval systems come from the lack of recognizing the existence of the semantic gap and its consequences for system set-up.

> *The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.*

A linguistic description is almost always contextual, whereas an image may live by itself. A linguistic description of an image is a daunting, probably impossible task [146]. A user looks for images containing certain objects or conveying a certain message. Image descriptions, on the other hand, rely on data-driven features and the two may be disconnected. Association of a complete semantic system to image data would entail at least solving the general object recognition problem from a single image. Since this problem is yet unsolved, research is focused on different methods to associate higher level semantics to data-driven observables.

As indicated in Fig. 2, the most immediate means of semantic characterization entail annotation by keywords or captions. This reduces content-based access to information retrieval [135]. Common objections to the practice of labeling are cost and coverage. On the cost side, labeling thousands of images is a cumbersome and expensive job to the degree that the deployment of the economic balance behind the database is likely to decrease. To solve the problem, systems in [21], [142] use a program that explores the Internet, collecting images and inserting them in a predefined taxonomy on the basis of the text surrounding them. A similar approach for digital libraries is taken by [24]. On the coverage side, labeling is seldom complete, context sensitive, and, in any case, there is a significant fraction of requests whose semantics can't be captured by labeling alone [6], [64]. Both methods will cover the semantic gap only in isolated cases.

2.5 Discussion on Scope

The pivotal point in content-based retrieval is that the user seeks semantic similarity, but the database can only provide similarity by data processing. This is what we called the semantic gap. At the same time, the sensory gap between the properties in an image and the properties of the object plays a limiting role in retrieving the content of the image.

We discussed applications of content-based retrieval in three broad types: target search, category search, and search by association. Target search connects with the tradition of pattern matching in computer vision. New challenges in content-based retrieval are the huge amount of objects to search among, the incomplete query specification, the

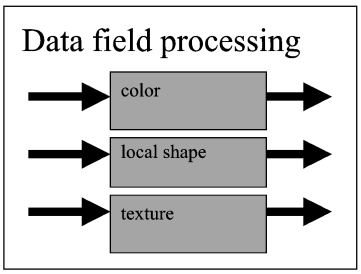


Fig. 6. The data flow diagrams for image processing modules, see Fig. 1 for conventions.

incomplete image description, and the variability of sensing conditions and object states. Category search builds on the object recognition and statistical pattern recognition methods in computer vision. New challenges in content-based retrieval compared to the achievements of object recognition are the interactive manipulation of results, the usually very large number of object classes, and the absence of an explicit training phase for feature and classifier tuning.

Search by association is further removed from most of the computer vision tradition. It is hampered most by the semantic gap. As long as the gap is there, use of content-based retrieval for browsing will not be within the grasp of the general public as humans are accustomed to rely on the immediate semantic imprint the moment they see an image. The aim of content-based retrieval systems must be to provide maximum support in bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics.

We analyze characteristics of the image domain, the domain knowledge, and the types of use as the prime factors determining the functionality of a system. An important distinction is that between broad and narrow domains. The broader the domain, the more browsing or search by association can be the right solution. The narrower the domain, the more likely an application of domain knowledge will succeed. *The challenge for image search engines on a broad domain is to tailor the engine to the narrow domain the user has in mind via specification, examples, and interaction.*

3 Description of Content: Image Processing

It is important to establish that content-based retrieval does not rely on describing the content of the image in its entirety. It may be sufficient that a retrieval system presents similar images, similar in some user-defined sense. The description of content should serve that goal primarily.

We consider the description of content in two steps. First, we discuss image-processing operations that transpose the image data into another spatial data array, see Fig. 6. We divide the methods over local color, the local texture, or local geometry. They may be characterized in general by:

$$f(\mathbf{x}) = g \circ i(\mathbf{x}), \tag{1}$$

where $i(\mathbf{x})$ is the image, element of image space $\mathcal{I}$, g is an operator on images, and the resulting image field is given by $f(\mathbf{x})$. Computational parameters of g may include the size of the neighborhood around $\mathbf{x}$ to compute $f(\mathbf{x})$ or a homogeneity criterion when the size of the patch to compute $f(\mathbf{x})$ depends on the actual data, as in [163], [126], for example.

So, the purpose of image processing in image retrieval must be to enhance aspects in the image data relevant to the query and to reduce the remaining aspects.

One such goal can be met by using invariance as a tool to deal with accidental distortions in the information introduced by the sensory gap. From the above discussion on the sensory gap, it is clear that invariant features may carry more object-specific information than other features as they are insensitive to the accidental conditions of the sensing. The aim of invariant descriptions is to identify objects, no matter from how and where they are observed, at the loss of some of the information content. If two objects t_i (or two appearances of the same object) are equivalent under a group of transformations W, they are in an equivalence class [18]:

$$t_1 \overset{W}{\sim} t_2 \Longleftrightarrow \exists w \in W : t_2 = w \circ t_1. \tag{2}$$

A property f of t is invariant under W if and only if f_t remains the same regardless the unwanted condition expressed by W,

$$t_1 \overset{W}{\sim} t_2 \Longrightarrow f_{t_1} = f_{t_2}. \tag{3}$$

The degree of invariance, that is, the dimensionality of the group W, should be tailored to the recording circumstances. In general, a feature with a very wide class of invariance loses the power to discriminate among essential differences. The size of the class of images considered equivalent grows with the dimensionality of W. In the end, the invariance may be so wide that no discrimination among objects is retained. The aim is to select the tightest set of invariants suited for the expected set of nonconstant conditions. What is needed in image search is a specification of the minimal invariant conditions in the specification of the query discussed in [159]. The minimal set of invariant conditions can only be specified by the user as it is part of his or hers intention. The oldest work on invariance in computer vision has been done in object recognition, as reported, among others, in [117]. Invariant description in image retrieval is relatively new, but quickly gaining ground for a good introduction, see [18], [32]. An alternative to invariant features[1] is to represent the viewing conditions separately from the objects in the scene. This way no information is lost in the reduction to invariant features, while the information is only rearranged. It should be left to the later stages to decide what is important.

3.1 Color Image Processing

Color has been an active area of research in image retrieval, more than in any other branch of computer vision. Color makes the image $i(\mathbf{x})$ take values in a color vector space. The interest in color may be ascribed to the superior discriminating potentiality of a three-dimensional domain compared to the single dimensional domain of gray-level images.

1. As pointed out by one of the referees of this paper.

Two aspects of color return in many of the contributions. One is that the recorded color varies considerably with the orientation of the surface, the viewpoint of the camera, the position of the illumination, the spectrum of the illuminant, and the way the light interacts with the object. This variability should be dealt with in one way or another. Second, the human perception of color is an intricate topic where many attempts have been made to capture perceptual similarity.

Only when there is no variation in the recording or in the perception is the RGB color representation a good choice since that representation was designed to match the input channel of the eye. RGB-representations are in wide-spread use. They describe the image in its literal color properties. An image expressed as $(R(\mathbf{x}), G(\mathbf{x}), B(\mathbf{x}))$ (indices will be omitted from now on) makes most sense when recording in the absence of variance, as is the case, e.g., for art paintings [64], the color composition of photographs [48], and trademarks [79], [39], where two-dimensional images are recorded in frontal view under standard conditions.

A significant improvement over the RGB-color space (at least for retrieval applications) comes from the use of opponent color representations [169], which uses the opponent color axes $(R-G, 2B-R-G, R+G+B)$. This representation has the advantage of isolating the brightness information on the third axis. With this solution, the first two chromaticity axes can be down-sampled as humans are more sensitive to brightness than they are to chroma. They are invariant to changes in illumination intensity and shadows.

Others approaches use the Munsell or the Lab-spaces because of their relative perceptual uniformity. The Lab-representation is designed so that the Euclidean distance between two colors representations models the human perception of color differences. Care should be taken when digitizing the nonlinear conversion to Lab-space [115].

The HSV-representation is often selected for its invariant properties. The hue is invariant under the orientation of the object with respect to the illumination and camera direction and hence more suited for object retrieval.

A wide variety of tight photometric color invariants for object retrieval were derived in [57] from an analysis of the Schafer model of object reflection. They derive for matte patches under white light the invariant color space

$$\left(\frac{R-G}{R+G}, -\frac{B-R}{B+R}, \frac{G-B}{G+B}\right),$$

only dependent on sensor and surface albedo. For a shiny surface and white illumination, they derive the invariant representation as

$$\frac{|R-G|}{|R-G|+|B-R|+|G-B|}$$

and two more permutations. The color models are robust against major viewpoint distortion.

Color constancy is the capability of humans to perceive the same apparent color in the presence of variations in illumination which change the physical spectrum of the perceived light. In computer vision, color constancy was first considered in [49]. For flat, matte, and uniformly illuminated objects, the paper forms the canonical gamut defined as the convex set of physically feasible normalized RGB, i.e., rgb-responses. The reference then maps all observed rgb-responses in the image into the canonical one. The map explaining all observations determines the color constancy. In [47], this is improved to include specular reflection, shape, and varying illumination. By computing the blue-ratio vector $(\frac{r}{b}, \frac{g}{b}, 1)$, only color orientation is used and intensity is ruled out. In this 2D-space, the color constancy map can again be selected from a canonical gamut of colors and surfaces. In [56], the ratios

$$\left(\frac{R_{x_1}G_{x_2}}{R_{x_2}G_{x_1}}, \frac{G_{x_1}B_{x_2}}{G_{x_2}B_{x_1}}, \frac{B_{x_1}R_{x_2}}{B_{x_2}R_{x_1}}\right)$$

offer more stability to surface geometry variations. Color constancy was applied to retrieval in [54] by using an illumination invariant color representation. The authors index the ratio of neighboring colors. Color constant indexing leads to some loss in discriminating power among objects, but yields illumination independent retrieval instead. The scheme was improved in [158] by using algebraic invariants.

Rather than invariant descriptions, another approach to cope with the inequalities in observation due to surface reflection is to search for clusters in a color histogram of the image. In the RGB-histogram, clusters of pixels reflected off an object form elongated streaks. Hence, in [126], a nonparametric cluster algorithm in RGB-space is used to identify which pixels in the image originate from one uniformly colored object.

3.2 Image Processing for Local Shape

Under the name "local shape," we collect all properties that capture conspicuous geometric details in the image. We prefer the name local shape over differential geometrical properties to express the result rather than the method. The result of local shape evaluation is a dense image data field different from object shape (discussed in Section 4).

Local shape characteristics derived from directional color derivatives—in the paper referred to as texture properties—have been used in [115] to derive perceptually conspicuous details in highly textured patches of diverse materials. A wide, rather unstructured variety of image detectors can be found in [165].

Scale space theory was devised as the complete and unique primary step in preattentive vision, capturing all conspicuous information [187]. It provides the theoretical basis for the detection of conspicuous details on any scale. In [105], a series of Gabor filters of different directions and scale have been used to enhance image properties [137]. Conspicuous shape geometric invariants are presented in [136]. A method employing local shape and intensity information for viewpoint and occlusion invariant object retrieval is given in [148]. The method relies on voting among a complete family of differential geometric invariants. Also, [178] searches for differential affine-invariant descriptors. From surface reflection, in [5], the local sign of the Gaussian curvature is computed while making no assumptions on the albedo or the model of diffuse reflectance.

Combining shape and color both in invariant fashion is a powerful combination, as described by [56], where the colors inside and outside affine curvature maximums in color edges are stored to identify objects.

3.3 Image Texture Processing

In computer vision, texture is defined as all what is left after color and local shape have been considered or it is defined by such terms as structure and randomness. Many common textures are composed of small textons usually too great in number to be perceived as isolated objects. The elements can be placed more or less regularly or randomly. They can be almost identical or subject to large variations in their appearance and pose. In the context of image retrieval, research is mostly directed toward statistical or generative methods for the characterization of patches.

Basic texture properties include the Markovian analysis, dating back to Haralick in 1973, and generalized versions thereof [91], [58]. In retrieval, the property is computed in a sliding mask for localization [99], [59].

Another important texture analysis technique uses multiscale autoregressive MRSAR-models, which consider texture as the outcome of a deterministic dynamic system subject to state and observation noise [174], [106]. Other models exploit statistical regularities in the texture field [9].

Wavelets [34] have received wide attention. They have often been considered for their locality and their compression efficiency. Many wavelet transforms are generated by groups of dilations or dilations and rotations that have been said to have some semantic correspondent. The lowest levels of the wavelet transforms [34], [26] have been applied to texture representation [92], [162] sometimes in conjunction with Markovian analysis [25]. Other transforms have also been explored, most notably fractals [44]. A solid comparative study on texture classification from mostly transform-based properties can be found in [133].

Texture search proved useful in satellite images [98] and images of documents [33]. Textures also served as a support feature for segmentation-based recognition [102], but the texture properties discussed so far offer little semantic referent. They are therefore ill-suited for retrieval applications in which the user wants to use verbal descriptions of the image. Therefore, in retrieval research, in [101], the Wold features of periodicity, directionality, and randomness are used, which agree reasonably well with linguistic descriptions of textures as implemented in [128].

3.4 Discussion on Image Processing

Image processing in content-based retrieval should primarily be engaged in enhancing the image information the query poses, not in describing the content of the image in its entirety.

To enhance the image information, retrieval has set the spotlights on color, as color has a high discriminating power among objects in a scene, much higher than gray levels. The purpose of most image color processing is to reduce the influence of the accidental conditions of the scene and sensing (i.e., the sensory gap). Progress has been made in tailored color space representation for well-described classes of variant conditions. Also, the application of geometric description derived from scale space theory will reveal viewpoint and scene independent salient point sets, thus opening the way to similarity of images on a few most informative regions or points.

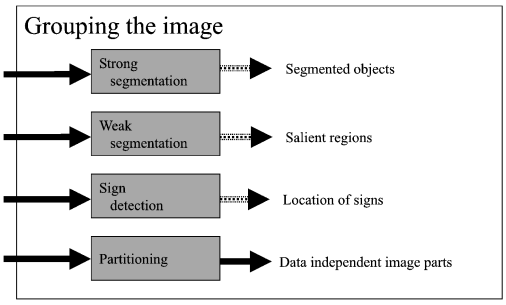


Fig. 7. Symbolic representation of different ways to group image data.

In this section, we have made a separation between color, local geometry, and texture. At this point, it is safe to conclude that the division is artificial. Wavelets say something about the local shape as well as the texture and so do many scale space and local filter strategies. For the purposes of content-based retrieval, an integrated view on color, texture, and local geometry is urgently needed as only an integrated view on local properties can provide the means to distinguish among hundreds of thousands different images. A recent advancement in that direction is the fusion of illumination and scale invariant color and texture information into a consistent set of localized properties [66]. Also, in [20], homogeneous regions are represented as collections of ellipsoids of uniform color or texture, but invariant texture properties deserve more attention, [173] and [185]. Further research is needed in the design of complete sets of image properties with well-described variant conditions which they are capable of handling. Invariance is just one side of the coin, where discriminating power is the other. In content-based image retrieval, the first steps are taken to establish the discriminating power of invariant properties [55]. This is essential as the balance between stability against variations and retained discriminatory power determines the effectiveness of a property.

4 Description of Content: Features

In the first section, we discuss the ultimate form of spatial data by grouping the data into object silhouettes, clusters of points or point-sets. In subsequent sections, we leave the spatial domain to condense the pictorial information into feature values.

4.1 Grouping Data

In content-based image retrieval, the image is often divided in parts before features are computed from each part, see Fig. 7. Partitionings of the image aim at obtaining more selective features by selecting pixels in a trade-off against having more information in features when no subdivision of the image is used at all. We distinguish the following partitionings:

- When searching for an object, it would be most advantageous to do a complete object segmentation first: "Strong segmentation is a division of the image

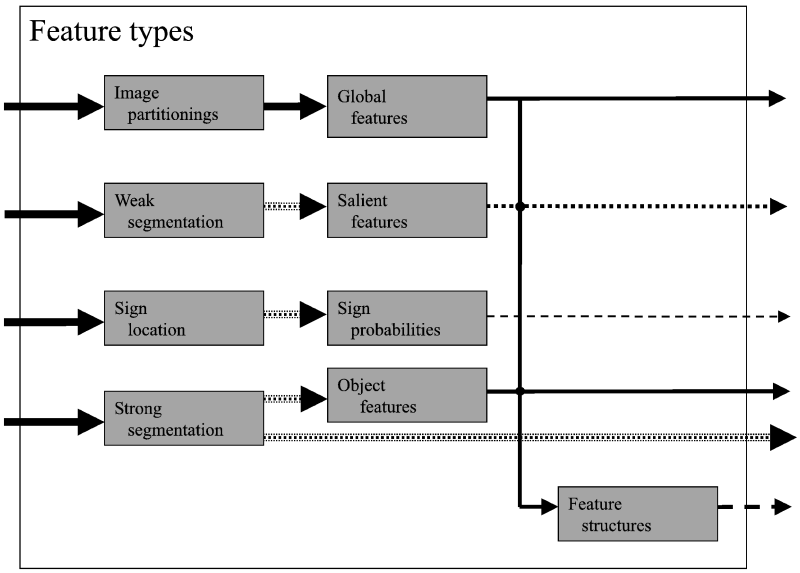


Fig. 8. The different types of features using the data flow conventions of Fig. 1.

data into regions in such a way that region T contains the pixels of the silhouette of object O in the real world and nothing else, specified by: $T = O$."

It should be noted immediately that object segmentation for broad domains of general images is not likely to succeed, with a possible exception for sophisticated techniques in very narrow domains.

- The difficulty of achieving strong segmentation may be circumvented by weak segmentation where grouping is based on data-driven properties: "Weak segmentation is a grouping of the image data in conspicuous regions T internally homogenous according to some criterion, hopefully with $T \subset O$."

The criterion is satisfied if region T is within the bounds of object O, but there is no guarantee that the region covers all of the object's area. When the image contains two nearly identical objects close to each other, the weak segmentation algorithm may falsely observe just one patch. Fortunately, in content-based retrieval, this type of error is rarely obstructive for the goal. In [125], the homogeneity criterion is implemented by requesting that colors be spatially coherent vectors in a region. Color is the criterion in [50], [126]. In [20], [112], the homogeneity criterion is based on color and texture. The limit case of weak segmentation is a set of isolated points [148], [57]. No homogeneity criterion is needed then, but the effectiveness of the isolated points rest on the quality of their selection. When occlusion is present in the image, weak segmentation is the best one can hope for. Weak segmentation is used in many retrieval systems, either as a purpose of its own or as a preprocessing stage for data-driven model-based object segmentation.

- When the object has a (nearly) fixed shape, like a traffic light or an eye, we call it a sign: "Localizing signs is finding an object with a fixed shape and semantic meaning, with $T = \mathbf{x}_{\mathbf{center}}$."

Signs are helpful in content-based retrieval as they deliver an immediate and unique semantic interpretation.

- The weakest form of grouping is partitioning: "A partitioning is a division of the data array regardless of the data, symbolized by: $T \neq O$."

The area T may be the entire image or a conventional partitioning as the central part of the image against the upper, right, left, and lower parts [67]. The feasibility of fixed partitioning comes from the fact that images are created in accordance with certain canons or normative rules, such as placing the horizon about 2/3 up in the picture or keeping the main subject in the central area. This rule is often violated, but this violation in itself has semantic significance. Another possibility of partitioning is to divide the image in tiles of equal size and summarize the dominant feature values in each tile [130].

Each of these four approaches to partitioning leads to a preferred type of features, as summarized in Fig. 8 and illustrated in Fig. 9, where feature hierarchies are used to make a combination on all types.

4.2 Global and Accumulating Features

In the computational process, features are calculated next. The general class of accumulating features aggregate the spatial information of a partitioning irrespective of the image data. A special type of accumulative features are the global features which are calculated from the entire image. Accumulating features are symbolized by:

$$F_j = \sum_{T_j} h \circ f(\mathbf{x}), \tag{4}$$

where Σ represents an aggregations operation (the sum in this case, but it may be a more complex operator). F_j is the set of accumulative features or a set of accumulative features ranked in a histogram. F_j is part of feature space $\mathcal{F}$. T_j is the partitioning over which the value of F_j is computed, see Fig. 9 for an illustration. In the case of global features, $T_{j=void}$ represents the image, otherwise, T_j

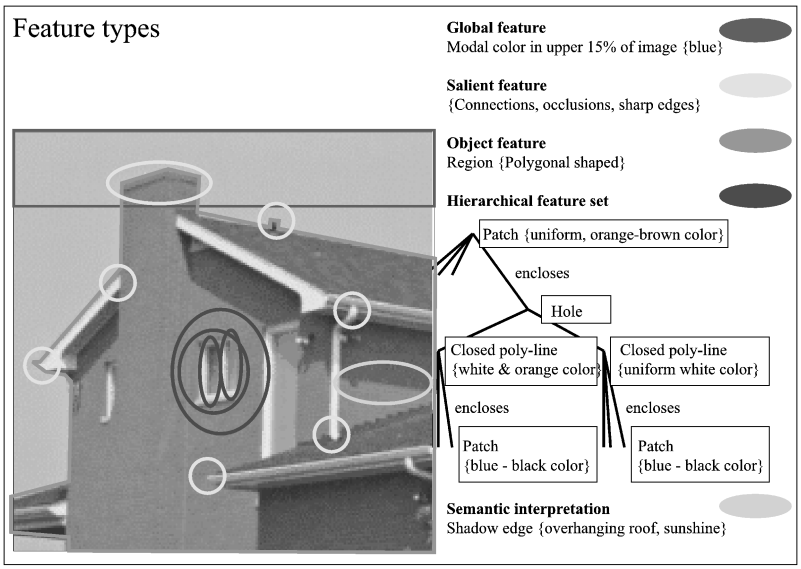


Fig. 9. Illustration of the various feature types as discussed in the paper.

represents a fixed tiling of the image. The operator h may hold relative weights, for example, to compute transform coefficients.

A simple but very effective approach to accumulating features is to use the histogram, that is, the set of features $\mathbf{F}(m)$ ordered by histogram index m. The original idea to use histograms for retrieval comes from Swain and Ballard [169], who realized that the power to identify an object using color is much larger than that of a gray-valued image. As a histogram loses all information about the location of an object in the image, [169], [41] project the histogram back into the image to locate it by searching for best matches. A histogram may be effective for retrieval as long as there is a uniqueness in the color pattern held against the pattern in the rest of the entire data set. In addition, the histogram shows an obvious robustness to translation of the object and rotation about the viewing axis. Swain and Ballard also argue that color histograms change slowly with change in viewpoint and scale and with occlusion.

All of this is in favor of the use of histograms. When very large data sets are at stake, plain histogram comparison will saturate the discrimination. For a 64-bin histogram, experiments show that, for reasonable conditions, the discriminating power among images is limited to 25,000 images [167]. To keep up performance, in [125], a joint histogram is used, providing discrimination among 250,000 images in their database, rendering 80 percent recall among the best 10 for two shots from the same scene using simple features. Other joint histograms add local texture or local shape [61], directed edges [78], and local higher order structures [48].

Another alternative is to add a dimension representing the local distance. This is the correlogram [73], defined as a three-dimensional histogram where the colors of any pair are along the first and second dimension and the spatial distance between them along the third. The autocorrelogram defining the distances between pixels of identical colors is found on the diagonal of the correlogram. A more general version is the geometric histogram [134], with the normal histogram, the correlogram, and several alternatives as special cases. This also includes the histogram of the triangular pixel values, reported to outperform all of the above as it contains more information.

To avoid an explosion of dimensions of the histogram, one could also prefer to reconsider the quality of the information along each of the dimensions. In Section 3, we have considered invariant representations suited to enrich the information on the axes of the histogram as it rules out the accidental influence of sensing and scene conditions.

A different view on accumulative features is to demand that all information (or all relevant information) in the image is preserved in the feature values. When the bit-content of the features is less than the original image, this boils down to compression transforms. Many compression transforms are known, but the quest is for transforms simultaneously suited as retrieval features. As proper querying for similarity is based on a suitable distance function between images, the transform has to be applied on a metric space. The components of the transform have to correspond to semantically meaningful characteristics of the image. Finally, the transform should admit indexing in compressed form yielding a big computational advantage over having the image be untransformed first. Schneier and Abdel-Mottaleb [149] is just one of many where the cosine-based JPEG-coding scheme is used for image retrieval. The JPEG-transform fulfills the first and third requirement, but fails on a lack of semantics. In the MPEG-standard, the possibility of including semantic descriptors in the compression transform is introduced [29]. For an overview of feature indexes in the compressed domain, see [103]. In [92], a wavelet packet was applied to texture images and, for each packet, entropy and energy measures were determined and collected in a feature vector. In [75], vector quantization was applied in the space of coefficients to reduce its dimensionality. This approach was extended to incorporate the metric of the color space in [146]. In [77], a wavelet transform was applied independently to the three channels of a color image and only the sign of the most significant coefficients is retained. In a recent paper [3], a

scheme is offered for a broad spectrum of invariant descriptors suitable for application on Fourier, wavelets, and splines and for geometry and color alike.

Another type of complete feature sets capturing all information in the image is to use moments. Their invariant combinations of moments [72] and [89] have been successfully employed in retrieval of objects in [48], especially when the image contains just the object.

4.3 Salient Features

Another way to avoid the brittleness of strong segmentation is to opt for weak segmentation. This leads to a grouping of the data into homogeneous regions. From the merged regions, a selection is made on their salience. The most conspicuous regions are stored. The limit case of a weak segmentation is the detection of conspicuous points, see Fig. 9. Salient features may be covered by the generic equation:

$$F_j(\mathbf{x}) = \bigwedge_{T_j} h \circ f(\mathbf{x}), \tag{5}$$

where $\wedge$ stands for a local selection operation and operator h maximizes the saliency of the processed image field $f(\mathbf{x})$. The area T_j over which the value of F_j is searched for is usually the whole image, although there would be no objection to concentrating on the center or top part of the image in search for specific events.

As the information of the image is condensed into just a limited number of feature values, the information should be selected with precision for greatest saliency and proven robustness. That is why saliency in [100] is defined as the special points which survive longest when gradually blurring the image in scale space. Also, in [138], lifetime is an important selection criterion for salient points in addition to wiggliness, spatial width, and phase congruency. To enhance the quality of salient descriptions, in [178], invariant and salient features of local patches have been considered. In each case, the image is summarized in a list of conspicuous points. In [148], salient and invariant transitions in gray value images are recorded. Similarly, in [57], [55], photometric invariance is the leading principle in summarizing the image in salient transitions in the image. Salient feature calculations lead to sets of regions or points with known location and feature values capturing their salience.

In [20], first, the most conspicuous homogeneous regions in the image are derived and mapped into feature space. Then, expectation-maximization [37] is used to determine the parameters of a mixture of Gaussians to model the distribution of points into the feature space. The means and covariance matrices of these Gaussians, projected on the image plane, are represented as ellipsoids characterized by their center $\mathbf{x}$, their area, eccentricity, and direction. The average values of the color and texture descriptions inside the ellipse are also stored.

4.4 Signs

When one of the possible interpretations of an image is so preponderant that it can be considered *the* meaning of the image, the image holds a sign, characterized by the probability P on interpretation z:

$$P_z(\mathbf{x}) = P(z|h_z \circ f(\mathbf{x})) \tag{6}$$

with symbols as in (5). The analysis leads to a localization of a sign with its probability. Typical signs are an icon, a character, a traffic light, or a trademark. In the case of maps, the interpretation of map symbols and their spatial relationships provides access to the content of the map [144]. Other systems based on signs are designed with specific application domains in mind, like OCR from an image [200], faces to detect from the image [197], medical images [90], [17], textile [95], art [65], or detecting the constituent components of silhouettes of plants based on a visual lexicon in [180].

For signs, a strong semantic interpretation is within grasp and the undisputed semantic interpretation brings clarity in interpreting the image. That is the attractiveness of using signs, in spite of the fact that the analysis tends to become application-oriented.

4.5 Shape and Object Features

The theoretically best way to enhance object-specific information contained in images is by segmenting the object in the image. But, as discussed above, the brittleness of segmentation algorithms prevents the use of automatic segmentation in broad domains. In fact, in many cases, it is not necessary to know exactly where an object is in the image as long as one can identify the presence of the object by its unique characteristics. When the domain is narrow, a tailored segmentation algorithm may be needed more and, fortunately, also be better feasible. When segmentation is applied, we have:

$$t_j(\mathbf{x}) = s_j \circ f(\mathbf{x}), \tag{7}$$

where $f(\mathbf{x})$ is the data field resulting from the processing above (equal to the image $i(\mathbf{x})$ when g is the identity operator), s_j is the segmentation operator for object j, and $t_j(\mathbf{x})$ indicates the object area T_j. For shape, F_j is a (possibly ordered) set of features from $\mathcal{F}$ for j:

$$F_j = \sum_{T_j} h \circ t_j(\mathbf{x}), \tag{8}$$

where Σ represents an aggregation operation and h is the functional computing shape in this case. Object internal features are computed similar to (4).

The object internal features are largely identical to the accumulative features, now computed over the object area. They need no further discussion here.

An abundant comparison of shape for retrieval can be found in [109], evaluating many features on a 500-element trademark data set. Straightforward features of general applicability include Fourier features and moment invariants of the object this time, sets of consecutive boundary segments, or encoding of contour shapes [43].

For retrieval, we need a shape representation that allows a robust measurement of distances in the presence of considerable deformations. Many sophisticated models widely used in computer vision often prove too brittle for image retrieval. On the other hand, the (interactive) use of retrieval makes some mismatch acceptable and, therefore, precision can be traded for robustness and computational efficiency.

More sophisticated methods include elastic matching and multiresolution representation of shapes. In elastic deformation of image portions [36], [122] or modal matching techniques [150], image patches are deformed to minimize a cost functional that depends on a weighed sum of the mismatch of the two patches and on the deformation energy. The complexity of the optimization problem depends on the number of points on the contour. Hence, the optimization is computationally expensive and this, in spite of the greater precision of these methods, has limited their diffusion in image databases.

Multiscale models of contours have been studied as a representation for image databases in [116]. Contours were extracted from images and progressively smoothed by dividing them into regions of constant sign of the second derivative and progressively reducing the number of such regions. At the final step, every contour is reduced to an ellipsoid which could be characterized by some of the features in [48]. A different view on multiresolution shape is offered in [94], where the contour is sampled by a polygon and then simplified by removing points from the contour until a polygon survives selecting them on perceptual grounds. When computational efficiency is at stake, an approach for the description of the object boundaries is found in [201], where an ordered set of critical points on the boundary are found from curvature extremes. Such sets of selected and ordered contour points are stored in [108] relative to the basis spanned by an arbitrary pair of the points. All point pairs are used as a basis to make the redundant representation geometrically invariant, a technique similar to [192] for unordered point sets.

For retrieval of objects in 2D-images of the 3D-worlds, a viewpoint invariant description of the contour is important. A good review of global shape invariants is given in [140].

4.6 Description of Structure and Lay-Out

When feature calculations are available for different entities in the image, they may be stored with a relationship between them, see Fig. 9 for an illustration. Such a structural feature set may contain feature values plus spatial relationships, a hierarchically ordered set of feature values, or relationships between point sets or object sets. The process is symbolized by:

$$H_{j,k} = \sum_{T_{j,k}} h \circ f(\mathbf{x}), \tag{9}$$

where $T_{j,k}$ indicates the $k\text{th}$ part of the $j\text{th}$ object and $H_{j,k}$ is an (ordered) spatial relationship describing object j in k elements. Structural and layout feature descriptions are captured in a graph, hierarchy, or any other ordered set of feature values and their relationships.

To that end, in [107], [50], lay-out descriptions of an object are discussed in the form of a graph of relations between blobs. A similar lay-out description of an image in terms of a graph representing the spatial relations between the objects of interest was used in [129] for the description of medical images. In [53], a graph is formed of topological relationships of homogenous RGB-regions. When selected features and the topological relationships are viewpoint invariant, the description is viewpoint invariant, but the selection of the RGB-representation as used in the paper will only suit that purpose to a limited degree. The systems in [70], [163] study spatial relationships between regions, each characterized by locations, size, and features. In the later system, matching is based on the 2D-string representation founded by Chang and Hau [22]. For a narrow domain, in [129], [132], the relevant elements of a medical X-ray image are characterized separately and joined together in a graph that encodes their spatial relations.

Starting from a shape description, the authors in [94] decompose an object into its main components, making the matching between images of the same object easier. Automatic identification of salient regions in the image, based on nonparametric clustering followed by decomposition of the shapes found into limbs, is explored in [52].

4.7 Discussion on Features

Also in the description of the image by features, it should be kept in mind that for retrieval a total understanding of the image is rarely needed. Strong segmentation of the scene and complete feature descriptions may not be necessary at all to achieve a similarity ranking. Of course, the deeper one goes into the semantics of the pictures, the deeper the understanding of the picture will also have to be, but the critical point in the advancement of content-based retrieval is the semantic meaning of the image that is rarely self-evident.

The theoretically best approach to a semantic interpretation of an image remains the use of a strong segmentation of the scene. Automatic strong segmentation is, however, hard to achieve, if not impossible for general domains. The brittleness of strong segmentation is a mostly unsurpassable obstacle when describing the content of images by describing the content of its objects. Especially for broad domains and for sensory conditions where clutter and occlusion are to be expected, automatic strong segmentation is hard, if not impossible. In that case, segmentation is to be done by hand when retrieval relies on it.

Narrow domains such as trademark validation, the identification of textiles, and the recognition of fish depend on the shape of the object, assessing similarity on the basis of the silhouettes. The fine-to-coarse decompositions are attractive in their discriminating power and computational efficiency. Again, a major bottleneck is the highly accurate segmentation of the object (as well as a frontal viewpoint of the object). In selected narrow domains, this may be achieved by recording the object against a clear background.

General content-based retrieval systems have dealt with segmentation brittleness in a few ways. First, a weaker version of segmentation has been introduced in content-based retrieval. In weak segmentation, the result is a homogeneous region by some criterion, but not necessarily covering the complete object silhouette. It results in a fuzzy, blobby description of objects, rather than a precise segmentation. Salient features of the weak segments capture the essential information of the object in a nutshell. The extreme form of the weak segmentation is the selection of a salient point set as the ultimately efficient data reduction in the representation of an object, very much like the focus-of-attention algorithms for an earlier age. Only points on the

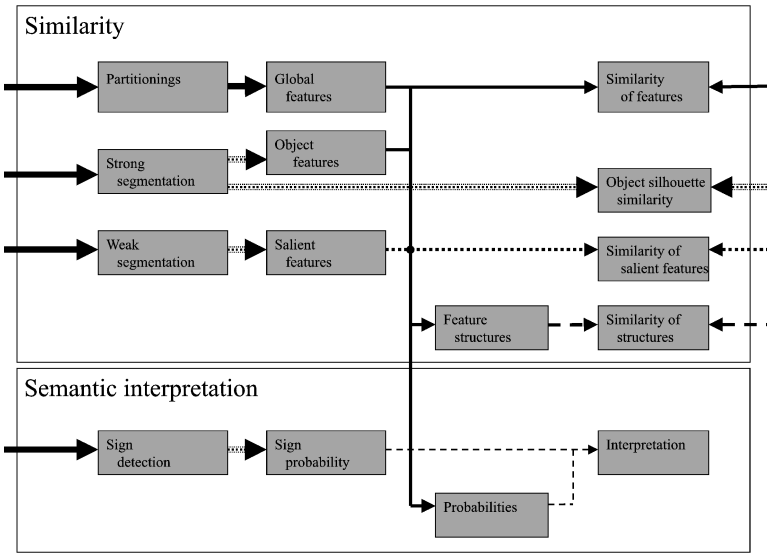


Fig. 10. Data flow diagram of similarity and interpretation.

interior of the object can be used for identifying the object and conspicuous points at the borders of objects have to be ignored. Little work has been done on how to make the selection. Weak segmentation and salient features are a typical innovation of content-based retrieval. It is expected that salience will receive much attention in the further expansion of the field, especially when computational considerations gain in importance.

The alternative is to do no segmentation at all. Content-based retrieval has gained from the use of accumulative features, computed on the global image or partitionings thereof, disregarding the content, the most notable being the histogram. Where most attention has gone to color histograms, histograms of local geometric properties and texture are following. To compensate for the complete loss of spatial information, the geometric histogram was recently defined with an additional dimension for the spatial layout of pixel properties. As it is a superset of the histogram, an improved discriminability for large data sets is anticipated. When accumulative features are calculated from the central part of a photograph may be very effective in telling them apart by topic, but the center does not always reveals the purpose. Likewise, features calculated from the top part of a picture may be effective in telling indoor scenes from outdoor scenes, but again this holds to a limited degree. A danger of accumulative features is their inability to discriminate among different entities and semantic meanings in the image. More work on semantic-driven groupings will increase the power of accumulative descriptors to capture the content of the image.

Structural descriptions match well with weak segmentation, salient regions, and weak semantics. One has to be certain that the structure is within one object and not an accidental combination of patches which have no meaning in the object world. The same brittleness of strong segmentation lurks here. We expect a sharp increase in the research of local, partial, or fuzzy structural descriptors for the purpose of content-based retrieval, especially of broad domains.

5 Interpretation and Similarity

When the information from images is captured in a feature set, there are two possibilities for endowing them with meaning: One derives an unilateral interpretation from the feature set, while the other one compares the feature set with the elements in a given data set on the basis of a similarity function, see Fig. 10.

5.1 Semantic Interpretation

In content-based retrieval, it is useful to push the semantic interpretation of features derived from the image as far as one can.

> *Semantic features aim at encoding interpretations of the image which may be relevant to the application.*

Of course, such interpretations are a subset of the possible interpretations of an image. To that end, consider a feature vector $\mathbf{F}$ derived from an image i. For given semantic interpretations z from the set of all interpretations $\mathcal{Z}$, a learning phase leads to conditional probabilities:

$$P = P(z|\mathbf{F}). \tag{10}$$

A strong semantic feature with interpretation z_j would generate a $P(z|\mathbf{F}) = \delta(z - z_j)$. If the feature carries no semantics, it would generate a distribution $P(z|\mathbf{F}) = P(z)$ independent of the value of the feature. In practice, many feature types will generate a probability distribution that is neither a pulse nor independent of the feature value. This means that the feature value "skews" the interpretation of the image, but does not determine it completely.

Under the umbrella *weak semantics*, we collect the approaches that try to combine features in some semantically meaningful interpretation. Weak semantics aims at encoding, in a simple and approximate way, a subset of the possible interpretations of an image that are of interest in a given application. As an example, the system in [30] uses color features derived from Itten's color theory to encode the semantics associated to color contrast and harmony in art application.

In the MAVIS2-system [84], data are considered at four semantic levels, embodied in four layers called the raw media, the selection, the selection expression, and conceptual layers. Each layer encodes information at an increasingly symbolic level. Agents are trained to create links between features, feature signatures at the selection layer, interrelated signatures at the selection expression layer, and concept (expressed as textual labels) at the conceptual layer. In addition to the vertical connections, the two top layers have intralayer connections that measure the similarity between concepts at that semantic level and contribute to the determination of the similarity between elements at the lower semantic level.

5.2 Similarity between Features

A different road to assigning a meaning to an observed feature set, is to compare a pair of observations by a similarity function. While searching for a query image $i_q(\mathbf{x})$ among the elements of the data set of images, $i_d(\mathbf{x})$, knowledge of the domain will be expressed by formulating a similarity measure $S_{q,d}$ between the images q and d on the basis of some feature set. The similarity measure depends on the type of features, see Fig. 10.

The similarity of two feature vectors $\mathbf{F}$, accumulative or object features alike, is given by:

$$S_{q,d} = s(\mathbf{F}_q, \mathbf{F}_d). \tag{11}$$

At its best use, the similarity measure can be manipulated to represent different semantic contents; images are then grouped by similarity in such a way that close images are similar with respect to use and purpose. There is surprisingly little work dedicated to characterizing similarity measures. A few ideas, however, have emerged. A common assumption is that the similarity between two feature vectors $\mathbf{F}$ can be expressed as:

$$s(\mathbf{F}_q, \mathbf{F}_d) = g \circ d(\mathbf{F}_q, \mathbf{F}_d), \tag{12}$$

where g is a positive, monotonically nonincreasing function and d is a distance function on $\mathcal{F}$. This assumption is consistent with a class of psychological models of human similarity perception [154], [147] and requires that the feature space be metric. If the feature space is a vector space, d often is a simple Euclidean distance, although there is indication that more complex distance measures might be necessary [147]. This similarity model was well-suited for early query by example systems in which images were ordered by similarity with one example.

A different view sees similarity as an essentially probabilistic concept. This view is rooted in the psychological literature [8] and, in the context of content-based retrieval, it has been proposed, for example, in [114]. A general form of such a similarity measure would be

$$s(\mathbf{F}_q, \mathbf{F}_d) = f(P(\hat{\mathbf{F}}_q \approx \mathbf{F}_d)), \tag{13}$$

where $\approx$ means that the two features describe images of the same class and $\hat{\mathbf{F}}_q = \bar{\mathbf{F}}_q + \nu, \bar{\mathbf{F}}_q$ are the "real" stimulus and ν noise due to sensory and measurement conditions.

Measuring the distance between histograms has been an active line of research since the early years of content-based retrieval, where histograms can be seen as a set of ordered features:

$$s(\mathbf{F}^q, \mathbf{F}^d) = g \circ d(\mathbf{F}^q, \mathbf{F}^d), \tag{14}$$

In content-based retrieval, histograms have mostly been used in conjunction with color features, but there is nothing against being used in texture or local geometric properties. Swain and Ballard [169] proposed the use of the intersection distance $d_\cap(\mathbf{F}^q, \mathbf{F}^d) = \sum_{j=1}^{n} \min(\mathbf{F}_j^q, \mathbf{F}_j^d)$, where $\mathbf{F}^q$ and $\mathbf{F}^d$ are two histograms containing n bins each. They also proved that if all images have the same number of pixels, i.e., $\sum_j \mathbf{F}_j$ is the same for all images, then this distance has the same ordinal properties as the L_1 distance. In [62], a different approach is followed. The distance between two histograms is defined in vector form as

$$d_\Sigma(\mathbf{F}^q, \mathbf{F}^d) = \sqrt{(\mathbf{F}^q - \mathbf{F}^d)^t \mathbf{\Sigma} (\mathbf{F}^q - \mathbf{F}^d)},$$

where the matrix $\mathbf{\Sigma}$ expresses the similarity between the $j\text{th}$ and the $k\text{th}$ bins. This has the advantage of considering the similarity between values in the feature space, i.e., of incorporating the metric of the feature space into the similarity measure.

Other commonly used distance functions for color histograms include the Minkowski distances

$$d_r(\mathbf{F}^q, \mathbf{F}^d) = \left[\sum_{j=1}^{n} |\mathbf{F}_j^q - \mathbf{F}_j^d|^r \right]^{\frac{1}{r}}.$$

These measures do not take into account the similarity between different, but related bins of a histogram. In [166], it is observed that this may lead to false negatives. The paper proposes the use of cumulative histograms of the form $\tilde{\mathbf{F}}^q(m) = \sum_{k=0}^{m} \mathbf{F}_k^q$ after ordering the bins by parameter j. Comparisons between cumulative rather than plain histograms show that the former tend to be more forgiving for changes in the bin assignment due to noise. An alternative, also explored in the paper, is to describe the histogram by the first three statistical moments, where 3 is an empirical finding. In [166], the histogram was applied to color images by representing colors in the $HSV\text{-system}$ and computing the moments of the channel separately, resulting in nine parameters, three moments for each of the three color channels. A recent distance measure for color histograms is found in [4], where a hue histogram and a saturation histogram are formed separately with the advantages of saving on memory and the possibility of excluding colors from a query. The reference compares colors on the basis of the angular distance in RGB-space.

The natural measure to compare ordered sets of accumulative features is nonparametric test statistics. They can be applied to the distributions of the coefficients of transforms to determine the likelihood that two samples derive from the same distribution [35], [131]. They can also be applied to compare the equality of two histograms and all variations thereof.

5.3 Similarity of Object Silhouettes

In [183], a good review is given of methods to compare shapes directly after segmentation into a set of object points $t(\mathbf{x})$:

$$S_{q,d} = s(t_q(\mathbf{x}), t_d(\mathbf{x})), \tag{15}$$

without an intermediate description in terms of shape features.

For shape comparison, the authors make a distinction between transforms, moments, deformation matching, scale space matching, and dissimilarity measurement. Difficulties for shape matching based on global transforms are the inexplicability of the result and the brittleness for small deviations. Moments, specifically their invariant combinations, have been frequently used in retrieval [89]. Matching a query and an object in the data file can be done along the ordered set of eigen shapes [150] or with elastic matching [36], [11]. Scale space matching is based on progressively simplifying the contour by smoothing [116]. By comparing the signature of annihilated zero crossings of the curvature, two shapes are matched in a scale and rotation invariant fashion. A discrete analogue can be found in [94], where points are removed from the digitized contour on the basis of perceptually motivated rules. Results on a 2,000-element database are reported to perform better than most of the methods listed above.

When based on a metric, dissimilarity measures will render an ordered range of deviations suited for a predictable interpretation. In [183], an analysis is given for the Hausdorff and related metrics between two shapes on robustness and computational complexity. The directed Hausdorff metric is defined as the maximum distance between a point on query object q and its closest counterpart on d. The partial Hausdorff metric, defined as the kth maximum rather than the absolute maximum, is used in [63] for affine invariant retrieval.

5.4 Similarity of Structural Features

The result of a structural description is a hierarchically ordered set of feature values H, see Fig. 9. In this section, we consider the similarity of

$$S_{q,d} = s(H_q, H_d) \tag{16}$$

between the two structural or layout descriptions.

Many different techniques have been reported for the similarity of feature structures. In [191], [74], a Bayesian framework is developed for the matching of relational attributed graphs by discrete relaxation. This is applied to line patterns from aerial photographs.

A metric for the comparison of two topological arrangements of named parts, applied to medical images, is defined in [172]. The distance is derived from the number of edit-steps needed to nullify the difference in the Voronoi-diagrams of two images.

In [23], 2D-strings describing spatial relationships between objects are discussed and, much later, reviewed in [198]. From such topological relationships of image regions, in [71], a 2D-indexing is built in trees of symbol strings, each representing the projection of a region on the coordinate axis. The distance between H_q and H_d is the weighed number of editing operations required to transform the one tree to the other. In [153], a graph is formed from the image on the basis of symmetry as it appears from the medial axis. Similarity is assessed in two stages via graph-based matching followed by energy-deformation matching.

In [53], hierarchically ordered trees are compared for the purpose of retrieval by rewriting them into strings. A distance-based similarity measure establishes the similarity scores between corresponding leaves in the trees. At the level of trees, the total similarity score of corresponding branches is taken as the measure for (sub)tree-similarity. From a small size experiment, it is concluded that hierarchically ordered feature sets are more efficient than plain feature sets, with projected computational shortcuts for larger data sets.

In [163], images are transformed into homogeneous regions for retrieval based on color layout. The regions are scanned, typically five equally spaced vertical scans, and converted into a string of symbols taken from a visual dictionary. The strings are summarized into region-relative histograms, $\mathbf{F}(i,j)$, indicating how many times a symbol precedes another symbol in one of the scans. During querying, the similarity of q to d is given by $\sum_i (\mathbf{F}^q \mathbf{F}^d)^{-1} \sum_j \mathbf{F}^q \mathbf{F}^d$, that is, the element-by-element correspondence of the region ordered histograms.

5.5 Similarity of Salient Features

Salient features are used to capture the information in the image in a limited number of salient points. Similarity between images can then be checked in several different ways.

In the first place, the color, texture, or local shape characteristics may be used to identify the salient points of the data as identical to the salient points of the query.

$$S_{q,d} = g \circ d(\mathbf{F}_q, \mathbf{F}_d), \tag{17}$$

where $\mathbf{F}_q$ and $\mathbf{F}_d$ are feature vectors of salient properties and g is an optional monotone function. A measure of similarity between the feature values measured of the blobs resulting from weak segmentation consists of a Mahalanobis distance between the feature vector composed of the color, texture, position, area, eccentricity, and direction of the two ellipses [20]. If the features of the ellipse are collected in a vector $\mathbf{F}$, the distance between q and d is given by

$$d_{q,d} = \left[(\mathbf{F}_q - \mathbf{F}_d)^T \mathbf{\Sigma}^{-1} (\mathbf{F}_q - \mathbf{F}_d) \right]^{\frac{1}{2}},$$

where $\mathbf{\Sigma}$ is a diagonal weights matrix set by the user. The similarity between two blobs is defined as $S_{q,d} = \exp(-d_{q,d}/2)$.

In the second place, one can store all salient points from one image in a histogram on the basis of a few characteristics, such as color on the inside versus color on the outside. The similarity is then based on the group-wise presence of enough similar points [57].

$$S_{q,d} = g \circ d(\mathbf{F}^q, \mathbf{F}^d), \tag{18}$$

where $\mathbf{F}^q$ and $\mathbf{F}^d$ are histograms merely indicating the presence of salient points. The metric $d(\mathbf{F}^q, \mathbf{F}^d)$ is now aiming at measuring the presence of the same set of salient points. Comparing sparsely occupied histograms has long been used in text retrieval, where vector space modeling [143] implies the registration in a N-dimensional histogram $\mathbf{F}$ with as many dimensions as there are different words in the dictionary, typically 10,000. In a binary vector space, each dimension is expressing whether that word is present or absent in the text. A text is a point in this high dimensional space. Differences between the text d in the data file and the query q boil down to the intersection distance discussed above: distance

$$d_{\cap}(\mathbf{F}^q, \mathbf{F}^d) = \sum_{1..i..N} \cap_i \mathbf{F}^d(..i..), \mathbf{F}^q(..i..),$$

over all dimensions. The same strategy is used when comparing salient point features derived from different images. The intersection is appropriate when both q and d may be partially occluded in the image or cluttered with the background. When q is neither cluttered or occluded but d may still be, the intersection should be replaced by the <-operation. The model has been used in image retrieval in [158], while keeping access to their location in the image by back-projection [169]. Following the development of the vector space model in text retrieval, a weight per dimension may favor the appearance of some salient features over an other. See also [69] for a comparison with correlograms.

A third alternative for similarity of salient points is to concentrate only on the spatial relationships among the salient points sets $\mathbf{P}_q$ and $\mathbf{P}_d$

$$S_{q,d} = g \circ d(\mathbf{P}_q, \mathbf{P}_d). \tag{19}$$

In point-by-point-based methods for shape comparison, shape similarity is studied in [83], where maximum curvature points on the contour and the length between them are used to characterize the object. To avoid the extensive computations, one can compute the algebraic invariants of point sets, known as the crossratio. Due to their invariant character, these measures tend to have only a limited discriminating power among different objects. A more recent version for the similarity of nameless point-sets is found in geometric hashing [192], where each triplet spans a base for the remaining points of the object. An unknown object is compared on each triplet to see whether enough similarly located points are found. Geometric hashing, though attractive in its concept, is too computationally expensive to be used on the very large data sets of image retrieval due to the anonymity of the points. Similarity of two points sets $\mathbf{P}_q$ and $\mathbf{P}_d$ given in a row-wise matrix is discussed in [188]. A distance is given for similarity invariance by,

$$D^2(\mathbf{P}_q, \mathbf{P}_d) = 1 - \frac{\|\mathbf{P}_q \mathbf{P}_d^T\|^2 + 2\det(\mathbf{P}_q \mathbf{P}_d^T)}{\|\mathbf{P}_d\|^2 \|\mathbf{P}_q\|^2}$$

and, for affine transformations,

$$D^2(\mathbf{P}_q, \mathbf{P}_d) = 2 - tr(\mathbf{P}_d^+ \mathbf{P}_d \cdot \mathbf{P}_q^+ \mathbf{P}_q),$$

where $\mathbf{P}_d^+$ and $\mathbf{P}_q^+$ are the pseudoinverse of $\mathbf{P}_d$ and $\mathbf{P}_q$, respectively.

5.6 Similarity at the Semantic Level

In [70], knowledge-based type abstraction hierarchies are used to access image data based on context and a user profile, generated automatically from cluster analysis of the database. Also in [24], the aim is to create a very large concept-space inspired by the thesaurus-based search from the information retrieval community. In [115], a linguistic description of texture patch visual qualities is given and ordered in a hierarchy of perceptual importance on the basis of extensive psychological experimentation.

A more general concept of similarity is needed for relevance feedback, in which similarity with respect to an ensemble of images is required. To that end, in [45], more complex relationships are presented between similarity and distance functions defining a weighted measure of two simpler similarities

$$S(s, S_1, S_2) = w_1 \exp(-d(S_1, s)) + w_2 \exp(-d(S_2, s)).$$

The purpose of the bireferential measure is to find all regions that are similar to two specified query points, an idea that generalizes to similarity queries given multiple examples. The approach can be extended with the definition of a complete algebra of similarity measures with suitable composition operators [45], [38]. It is then possible to define operators corresponding to the disjunction, conjunction, and negation of similarity measures, much like traditional databases. The algebra is useful for the user to manipulate the similarity directly as a means to express characteristics in specific feature values.

5.7 Learning an Interpretation

As data sets grow large and the available processing power matches that growth, the opportunity arises to learn from experience. Rather than designing, implementing, and testing an algorithm to detect the visual characteristics for each different semantic term, it becomes possible to learn the semantics of objects from their appearance.

For a review on statistical pattern recognition, see [80]. In [182], a variety of techniques is discussed treating retrieval as a classification problem.

One approach is principal component analysis over a stack of images taken from the same class z of objects. This can be done in feature space [118] or at the level of the entire image, for example, faces in [113]. The analysis yields a set of "eigenface" images, capturing the common characteristics of a face without the need of a geometric model.

Effective ways to learn from partially labeled data have recently been introduced in [194], [139], both using the principle of transduction [181]. This saves the effort of labeling the entire data set, unfeasible and unreliable as it grows big.

In [186], preliminary work is reported towards automatic detection of categories on totally unlabeled data sets. They represent objects as probabilistic constellations of features. Recurring salient rigid parts are selected automatically by maximization of the expectation.

In [176], a very large number of precomputed features is considered, of which a small subset is selected by boosting [80] to learn the image class.

An interesting technique to bridge the gap between textual and pictorial descriptions to exploit information at the level of documents is borrowed from information retrieval, called latent semantic indexing [151], [199]. First, a corpus is formed of documents (in this case, images with a caption) from which features are computed. Then, by singular value decomposition, the dictionary covering the captions is correlated with the features derived from the pictures. The search is for hidden correlations of features and captions.

5.8 Discussion on Interpretation and Similarity

Whenever the image itself permits an obvious interpretation, the ideal content-based system should employ that information. A strong semantic interpretation occurs when a sign can be positively identified in the image. This is rarely the case due to the large variety of signs in a broad class of images and the enormity of the task to define a reliable detection algorithm for each of them. Weak semantics rely on inexact categorization induced by similarity measures, preferably online by interaction. The categorization may agree with semantic concepts of the user, but the agreement is, in general, imperfect. Therefore, the use of weak semantics is usually paired with the ability to gear the semantics of the user to his or her needs by interpretation. Tunable semantics is likely to receive more attention in the future, especially when data sets grow big.

Similarity is an interpretation of the image based on the difference with another image. For each of the feature types, a different similarity measure is needed. For similarity between feature sets, special attention has gone to establishing similarity among histograms due to their computational efficiency and retrieval effectiveness.

Similarity of shape has received considerable attention in the context of object-based retrieval. Generally, global shape matching schemes break down when there is occlusion or clutter in the scene. Most global shape comparison methods implicitly require a frontal viewpoint against a clear enough background to achieve a sufficiently precise segmentation. With the recent inclusion of perceptually robust points in the shape of objects, an important step forward has been made.

Similarity of hierarchically ordered descriptions deserves considerable attention as it is one mechanism to circumvent the problems with segmentation while maintaining some of the semantically meaningful relationships in the image. Part of the difficulty here is to provide matching of partial disturbances in the hierarchical order and the influence of sensor-related variances in the description.

Learning computational models for semantics is an interesting and relatively new approach. It gains attention quickly as the data sets and the machine power grow big. Learning opens up the possibility to an interpretation of the image without designing and testing a detector for each new notion. One such approach is appearance-based learning of the common characteristics of stacks of images from the same class. Appearance-based learning is suited for narrow domains. For success of the learning approach, there is a trade-off between standardizing the objects in the data set and the size of the data set. The more standardized the data are the less data will be needed, but, on the other hand, the less broadly applicable the result will be. Interesting approaches to derive semantic classes from captions or a partially labeled or unlabeled data set have been presented recently, see above.

6 Interaction

We turn our attention to the interacting user. Interaction of users with a data set has been studied most thoroughly in categorical information retrieval [123]. The techniques reported there need rethinking when used for image retrieval as the meaning of an image, due to the semantic gap, can only be defined in context. Image retrieval requires active participation of the user to a much higher degree than required by categorized querying. In content-based image retrieval, interaction is a complex interplay between the user, the images, and their semantic interpretations.

6.1 Query Space: Definition and Initialization

To structure the description of methods, we first define *query space*. The first component of query space is the selection of images $\mathcal{I}_Q$ from the large image archive $\mathcal{I}$. Typically, the choice is based on factual descriptions like the name of the archive, the owner, date of creation, or Web site address. Any standard retrieval technique can be used for the selection. The second component is a selection of the features $\mathcal{F}_Q \subset \mathcal{F}$ derived from the images in $\mathcal{I}_Q$. In practice, the user is not always capable of selecting the features most fit to reach the goal. For example, how should a general user decide between shape description using moments or Fourier coefficients? Under all circumstances, however, the user should be capable of indicating the class of features relevant for the task, like shape, texture, or both. In addition to feature class, [57] has the user indicate the requested invariance. The user can, for example, specify an interest in features robust against varying viewpoint, while the expected illumination is specified as white light in all cases. The appropriate features can then be automatically selected by the system. As concerns the third component of query space, the user should also select a similarity function, $\mathcal{S}_Q$. To adapt to different data sets and goals, $\mathcal{S}_Q$ should be a parameterized function. Commonly, the parameters are weights for the different features. The fourth component of query space is a set of labels $\mathcal{Z}_Q \subset \mathcal{Z}$ to capture goal-dependent semantics. Given the above, we define an abstract query space:

> *The query space Q is the goal dependent 4-tuple* $\{\mathcal{I}_Q, \mathcal{F}_Q, \mathcal{S}_Q, \mathcal{Z}_Q\}$.

To start a query session, an instantiation $Q = \{I_Q, F_Q, S_Q, Z_Q\}$ of the abstract query space is created. When no knowledge about past or anticipated use of the system is available, the initial query space Q^0 should not be biased toward specific

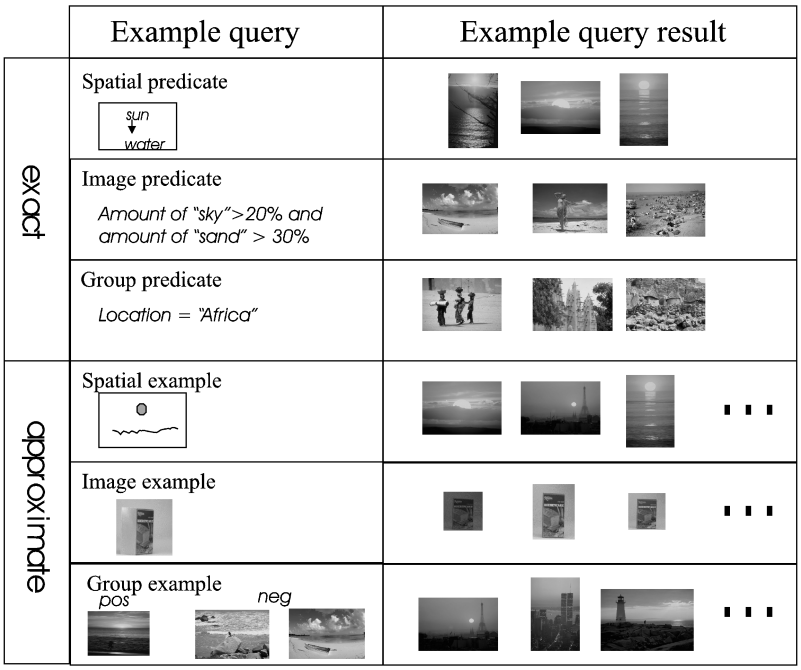


Fig. 11. Example queries for each of the six different query types and possible results from the Corel image database.

images or make some image pairs a priori more similar than others. The active set of images I_Q is therefore equal to all of $\mathcal{I}_\mathcal{Q}$. Furthermore, the features of F_Q are normalized based on the distribution of the feature values over I_Q e.g., [48], [142]. To make S_Q unbiased over F_Q, the parameters should be tuned, arriving at a *natural distance measure*. Such a measure can be obtained by normalization of the similarity between individual features to a fixed range [184], [142]. For the instantiation of a semantic label, the semantic gap prevents attachment to an image with full certainty. Therefore, in the ideal case, the instantiation Z_Q of $\mathcal{Z}_\mathcal{Q}$ assigns, for each $i \in I_Q$ and each $z \in Z_Q$, a probability $P_i(z)$, rather than a strict label.

The query space forms the basis for specifying queries, display of query results, and for interaction.

6.2 Query Specification

For specifying a query q in Q, many different interaction methodologies have been proposed. A query falls in one of two major categories: *exact query*, where the query answer set $A(q)$ equals the images in I_Q, satisfying a set of given criteria, and an *approximate query*, where $A(q)$ is a ranking of the images in I_Q with respect to the query, based on S_Q. Within each of the two categories, three subclasses can be defined depending on whether the query relates to the spatial content of the image, to the global image information, or to groups of images. An overview of the initialization and specification of queries is shown in Fig. 11.

For exact queries, the three subclasses are based on different predicates the result should satisfy:

- *Exact query by spatial predicate* is based on the location of silhouettes, homogeneous regions, or signs. Query on silhouette location is applicable in narrow domains only. Typically, the user queries using an interpretation $z \in \mathcal{Z}_\mathcal{Q}$. To answer the query, the system then selects an appropriate algorithm for segmenting the image and extracting the domain-dependent features. In [156], the user interactively indicates semantically salient regions to provide a starting point. The user also provides sufficient context to derive a measure for the probability of z. Implicit spatial relations between regions sketched by the user in [163] yield a pictorial predicate. Other systems let the user explicitly define the predicate on relations between homogeneous regions [20]. In both cases, to be added to the query result, the homogenous regions as extracted from the image must comply with the predicate. A Web search system in which the user places icons representing categories like human, sky, and water in the requested spatial order is presented in [97]. In [144], users pose spatial-predicate queries on geographical signs located in maps based on their absolute or relative positions.
- *Exact query by image predicate* is a specification of predicates on global image descriptions, often in the form of range predicates. Due to the semantic gap, range predicates on features are seldom used in a direct way. In [120], ranges on color values are predefined in predicates like "MostlyBlue" and "SomeYellow." Learning from user annotations of a partitioning of the image allows for feature range queries like: "amount of sky > 50 percent and amount of sand > 30 percent" [130].
- *Exact query by group predicate* is a query using an element $z \in \mathcal{Z}_\mathcal{Q}$, where $\mathcal{Z}_\mathcal{Q}$ is a set of categories that partitions $\mathcal{I}_\mathcal{Q}$. Both in [21] and [179], the user queries on a hierarchical taxonomy of categories. The difference is that the categories are based on contextual information in [21] while they are interpretations of the content in [179].

In the approximate types of query specifications, the user specifies a single feature vector or one particular spatial configuration in $\mathcal{F}_\mathcal{Q}$, where it is anticipated that no image will satisfy the query exactly.

- *Approximate query by spatial example* results in an image or spatial structure corresponding to literal image values and their spatial relationships. Pictorial specification of a spatial example requires a feature space such that feature values can be selected or sketched by the user. Low-level feature selectors use color pickers or selections from shape and texture examples [48], [61]. Kato et al. [88] were the first to let users create a sketch of the global image composition which was then matched to the edges in $\mathcal{I}_\mathcal{Q}$. Sketched outlines of objects in [93] are first normalized to remove irrelevant detail from the query object before matching it to objects segmented from the image. When specification is by parameterized template [36], [150], each image in $\mathcal{I}_\mathcal{Q}$ is processed to find the best match with edges of the images. The segmentation result is improved if the user may annotate the template with salient details like color corners and specific textures. Preidentification of all salient details in images in $\mathcal{I}_\mathcal{Q}$ can then be employed to speed up the search process [161]. When weak segmentation of the query image and all images in $\mathcal{I}_\mathcal{Q}$ is performed, the user can specify the query by indicating example regions [20], [163].
- *Approximate query by image example* feeds the system a complete array of pixels and queries for the most similar images, in effect asking for the *k*-nearest-neighbors in feature space. Most of the current systems have relied upon this form of querying [48], [61]. The general approach is to use an S_Q based on global image features. Query by example queries are subclassified [184] into *query by external image example*, if the query image is not in the database, versus *query by internal image example*. The difference in external and internal example is minor for the user, but affects the computational support as, for internal examples, all relations between images can be precomputed. Query by image example is suited for applications where the target is an image of the same object or set of objects under different viewing conditions [57]. In other cases, the use of one image cannot provide sufficient context for the query to select one of its many interpretations [146].
- *Approximate image query by group example* is specification through a selection of images which ensemble defines the goal. The rationale is to put the image in its proper semantic context to make one of the possible interpretations $z \in \mathcal{Z}_\mathcal{Q}$ preponderant. One option is that the user selects $m > 1$ images from a palette of images presented to find images best matching the common characteristics of the m images [31]. An *m*-query set is capable of defining the target more precisely. At the same time, the *m*-query set defines relevant feature value variations and nullifies irrelevant variations in the query. Group properties are amplified further by adding negative examples. This is achieved in [10] by constructing a query q best describing positive and negative examples indicated by the user. When, for each group in the database, a small set of representative images can be found, they are stored in a visual dictionary from which the user creates the query [146].

  Of course, the above queries can always be combined into more complex queries. For example, both [20], [163] compare the similarity of regions using features. In addition, they encode spatial relations between the regions in predicates. Even with such complex queries, a single query q is rarely sufficient to make $A(q)$ the user desired answer set. For most image queries, the user must engage in active interaction with the system on the basis of the query results as displayed.

6.3 Query Space Display

There are several ways to display the query result to the user. In addition, *system feedback* can be given to help the user in understanding the result. We define:

> *The visualization operator V maps the query space Q into the display space D having perceived dimension d.*

Note that d is the intrinsic dimensionality of the query result or d is induced by the projection function in V if the query result is of too high a dimension to visualize directly. In both cases, d is not necessarily equal to the two dimensions of the screen, so an additional projection operator might be required to map D onto the screen.

When the query is exact, the result of the query is a set of images fulfilling the predicate. As an image either fulfills the predicate or not, there is no intrinsic order in the query result and $d = 0$ is sufficient.

For approximate queries, the images in I_Q are given a similarity ranking based on S_Q with respect to the query. In many systems, the role of V is limited to bounding the number of images displayed, which are then displayed in a 2D rectangular grid [48], [21]. Note, however, that we should have $d = 1$. If the user refines its query using query by example, the images displayed do not have to be the images closest to the query. In [184], images are selected that together provide a representative overview of the whole active database. An alternative display model displays the image set minimizing the expected number of total iterations [31].

The space spanned by the features in F_Q is a high-dimensional space. When images are described by feature vectors, every image has an associated position in this space. In [146], [175], [68], the operator V maps the high-dimensional feature space onto a display space with $d = 3$. Images are placed in such a way that distances between images in $\mathcal{D}$ reflect S_Q. A simulated fisheye lens is used to induce perception of depth in [175]. In the reference, the set of images to display depends on how well the user selections conform to selections made in the community of users. To improve the user's comprehension of the information space, [68] provides the user with a dynamic view on F_Q through continuous variation of the active feature set. The display in [86] combines exact and

approximate query results. First, the images in I_Q are organized in 2D-layers according to labels in Z_Q. Then, in each layer, images are positioned based on S_Q.

In exact queries based on accumulative features, back-projection [169], [41] can be used to give system feedback, indicating which parts of the image fulfill the criteria. For example, in [130], each tile in the partition of the image shows the semantic label, like sky, building, or grass, the tile received. For approximate queries, in addition to mere rank ordering, in [20], system feedback is given by highlighting the subparts of the images contributing most to the ranking result.

6.4 Interacting with Query Space

In early systems, the process of query specification and display is iterated, where, in each step, the user revises the query. Updating the query often is still appropriate for *exact queries*. For *approximate queries*, however, the interactive session should be considered in its entirety. During the session, the system updates the query space, attempting to learn the goals from the user's feedback. As for the user, the result is what is visualized in display space, we define:

An interactive query session is a sequence of query spaces $\{Q^0, Q^1, ..., Q^{n-1}, Q^n\}$ *with* $A^n(q) = V(Q^n)$.

In a truly successful session, $A^n(q)$ is the user's search goal. The interaction process is schematically indicated in Fig. 12. The interaction of the user yields a relevance feedback RF_i in every iteration i of the session. The transition from Q^i to Q^{i+1} materializes the feedback of the user. For target search, category search, or associative search, various ways of user feedback have been considered. All are balancing between obtaining as much information from the user as possible and keeping the burden on the user minimal. The simplest form of feedback is to indicate which images are relevant [31]. In [28], [110], the user in addition explicitly indicates nonrelevant images. The system in [142] considers five levels of significance, which gives more information to the system, but makes the process more difficult for the user. When $d \geq 2$, the user can manipulate the projected distances between images, putting away nonrelevant images and bringing relevant images closer to each other [146]. The user can also explicitly bring in semantic information by annotating individual images, groups of images, or regions inside images [111] with a semantic label.

In general, user feedback leads to an update of query space:

$$\{I_Q^i, F_Q^i, S_Q^i, Z_Q^i\} \xrightarrow{RF_i} \{I_Q^{i+1}, F_Q^{i+1}, S_Q^{i+1}, Z_Q^{i+1}\}. \quad (20)$$

Different ways of updating Q are open. In [184], the images displayed correspond to a partitioning of I_Q. By selecting an image, one of the sets in the partition is selected and the set I_Q is reduced. Thus, the user zooms in on a *target* or a *category*.

The methods follows the pattern:

$$I_Q^i \xrightarrow{RF_i} I_Q^{i+1}. \quad (21)$$

In current systems, the feature vectors in F_Q corresponding to images in I_Q are assumed fixed. This has great advantages in terms of efficiency. When features are parameterized, however, feedback from the user could lead to optimization of the parameters. For example, in parameterized detection of objects based on salient contour details, the user can manipulate the segmentation result to have the system select a more appropriate salient detail based on the image evidence [161]. The general pattern is:

$$F_Q^i \xrightarrow{RF_i} F_Q^{i+1}. \quad (22)$$

For *associative search*, users typically interact to learn the system the right associations. Hence, the system updates the similarity function:

$$S_Q^i \xrightarrow{RF_i} S_Q^{i+1}. \quad (23)$$

In [28], [142], $\mathcal{S}_Q$ is parameterized by a weight vector on the distances between individual features. The weights in [28] are updated by comparing the variance of a feature in the set of positive examples to the variance in the union of positive and negative examples. If the variance for the positive examples is significantly smaller, it is likely that the feature is important to the user. The system in [142] first updates the weight of different feature classes. The ranking of images according to the overall similarity function is compared to the rankings corresponding to each individual feature class. Both positive and negative examples are used to compute the weight of the feature, computed as the inverse of the variance over the positive examples. The feedback RF_i in [146] leads to an update of the user-desired distances between pairs of images in I_Q. The parameters of the continuous similarity function should be updated to match the new distances. A regularization term is introduced, limiting the deviation from the initial natural distance function.

The final set of methods follow the pattern:

$$Z_Q^i \xrightarrow{RF_i} Z_Q^{i+1}. \quad (24)$$

The system in [111] precomputes a hierarchical grouping of partitionings (or images for that matter) based on the similarity for each individual feature. The feedback from the user is employed to create compound groupings

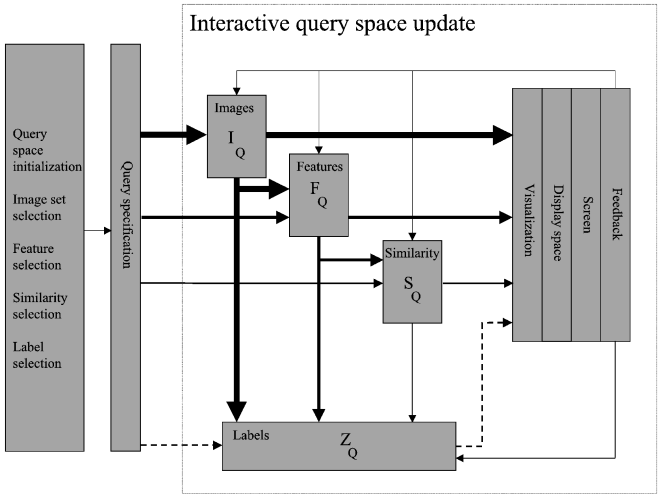


Fig. 12. The framework for interaction in content-based image retrieval.

corresponding to a user given $z \in \mathcal{Z}_Q$. The compound groupings are such that they include all of the positive and none of the negative examples. Unlabeled images in the compound group receive label z. The update of probabilities P is based on different partitionings of I_Q. For *category* and *target* search, a system may refine the likelihood of a particular interpretation, updating the label based on feature values or on similarity values. The method in [110] falls in this class. It considers *category search*, where $\mathcal{Z}_Q$ is {relevant, nonrelevant}. In the limit case for only one relevant image, the method boils down to *target search*. All images indicated by the user as relevant or nonrelevant in current or previous iterations are collected and a Parzen estimator is constructed incrementally to find optimal separation between the two classes. The generic pattern, which uses similarity in updating probabilities, is the form used in [31] for target search with $\mathcal{Z}_Q$ = {target}. In the reference, an elaborate Bayesian framework is derived to compute the likelihood of any image in the database being the target, given the history of actions RF_i. In each iteration, the user selects examples from the set of images displayed. Image pairs are formed by taking one selected and one displayed, but nonselected, image. The probability of being the target for an image in I_Q is increased or decreased depending on the similarity to the selected and the nonselected example in the pair.

6.5 Discussion on Interaction

Any information the user can provide in the search process should be employed to provide the rich context required in establishing the meaning of a picture. The interaction should form an integral component in any modern image retrieval system, rather than a last resort when the automatic methods fail. Already, at the start, interaction can play an important role. Most of the current systems perform query space initialization irrespective of whether a target search, a category search, or an associative search is requested. But, the fact of the matter is that the set of appropriate features and the similarity function depend on the user goal. Asking the user for the required invariance yields a solution for a specific form of target search. For category search and associative search, the user-driven initialization of query space is still an open issue.

For image retrieval, we have identified six query classes, formed by the Cartesian product of the result type {exact, approximate} and the level of granularity of the descriptions {spatial content, image, image groups}. The queries based on spatial content require segmentation of the image. For large data sets, such queries are only feasible when some form of weak segmentation can be applied to all images or when signs are selected from a predefined legend. A balance has to be found between flexibility on the user side and scalability on the system side. Query by image example has been researched most thoroughly, but a single image is only suited when another image of the same object(s) is the aim of the search. In other cases, there is simply not sufficient context. Queries based on groups, as well as techniques for prior identification of groups in data sets, are promising lines of research. Such group-based approaches have the potential to partially bridge the semantic gap while leaving room for efficient solutions.

The query result has an inherent display dimension which is often ignored. Most methods simply display images in a 2D grid. Enhancing the visualization of the query result is, however, a valuable tool in helping the user-navigating query space. As apparent from the query space framework, there is an abundance of information available for display. New visualization tools are urged to allow for user- and goal-dependent choices on what to display.

Through manipulation of the visualized result, the user gives feedback to the system. The interaction patterns as enumerated here reveal that, in current methods, feedback leads to an update of just one of the components of query space. There is no inherent reason why this should be the case. In fact, joint updates could indeed be effective and well worth researching. For example, the pattern which updates category membership based on a dynamic similarity function would combine the advantages of browsing with category and target search.

One final word about the impact of interactivity on the architecture of the system. The interacting user brings about many new challenges for the response time of the system. Content-based image retrieval is only scalable to large data sets when the database is able to anticipate what interactive queries will be made. A frequent assumption is that the image set, the features, and the similarity function are known in advance. In a truly interactive session, the assumptions are no longer valid. A change from static to dynamic indexing is required.

7 System Aspects

7.1 Storage and Indexing

We have been concerned with the content of the image eventually leading to a feature vector $\mathbf{F}$ or a hierarchically ordered set H containing the information of the image. Repetition over all images in the data set yields a file of feature vectors, the data file. In the previous section, we discussed the request as translated into the query image vector, $\mathbf{F}_q$, to be compared with the elements $\mathbf{F}_d$ of data file on the basis of the similarity function.

Scientifically and practically, the most interesting applications of retrieval are on large data sets, where there is statistically sound coverage of the image spectrum and learning general laws from the data sets makes sense. For large image sets, computational performance cannot be ignored as an issue. When storing the feature vectors in a standard, linear file with one record to each feature vector, we are bound to scan through all feature vectors. In that case, we have to perform N fetches of a record plus subsequent calculations to find the data vector most similar to the query feature vector. The response time of the system is $O(N)$ and, so, it is the number of operations when inserting a new element in the database and updating the mutual distance matrix among the elements. Linear scanning the feature vector file puts interactive response times out of reach, most certainly for data sets of 10,000 images and more.

In addition to the number of images, the dimension of the image vector can also be considerable. In [176], over 10,000 features are computed from the image each

describing a local pattern. In the example of a wavelet histogram for texture-based retrieval [162], an image has a nine-dimensional vector for each pixel compressed to a 512-bin histogram to a total of 512^2 histograms of 512 bins per image. The shape indexing technique [153] represents an image vector by a hierarchically ordered set of six types of nodes and three types of links, each encoding a number of image descriptors. Indexing in high dimensional spaces is difficult by the curse of dimensionality, a phenomenon by which indexing techniques become inefficient as the dimensionality of the feature space grows. The performance of R^*-trees degrades by a factor of 12 as the number of dimensions increases from 5 to 10 [190].

We focus on three classes of indexing methods that are in use on large image databases, substantiated by performance figures: *space partitioning*, *data partitioning*, and *distance-based* techniques. In space-partitioning index techniques, the feature space is organized like a tree as discussed in [15]. A node in this tree stands for a region in this space. When the number of points in a region exceeds a prescribed amount, the region is split into subregions which become the children of the node containing the original region. The best known index in this class is the $k-d$ tree. A $k-d$ tree is a data structure that generalizes binary trees in k dimensions, hence the name. It splits an overfilled node of the tree along its k dimensions, an splits the data points in the node along the median of the data values [7]. In the implementation of the reference, there is an approximation bound ϵ. In [152], a 15-fold decrease in response time is reported using the $k-d$ tree for a 20-nearest neighbors query over $N = 500{,}000$ images with $M = 78$ dimensions in feature space and $\epsilon = 0.1$. A further improvement is the $k-d$ B-tree as a multidimensional generalization of standard B-tree with splitting capacity of the $k-d$ tree. It is a balanced data structure with equal depth from the root, with $O(logN)$ performance [196].

Data partitioning index techniques associate, with each point in feature space, a region that represents the neighborhood of that vector. An *R*-tree is such a data partitioning structure to index hyperrectangular regions in *M*-dimensional space. The leaf nodes of an *R*-tree represent the minimum bounding rectangles of sets of feature vectors. An internal node is a rectangle encompassing the rectangles of all its children. An R^+-tree is a variant which does not allow the minimum bounding rectangles in a node to overlap. In R^*-tree, the minimum bounding rectangles may overlap. The ordinary *R*-tree, family has not been very successful for vectors with dimension somewhere over 10. The VAM-split *R*-tree splits along the dimension of maximum variance [190], hence the name, and was shown to have a better performance than standard R^*-tree. Since the splitting criterion is dependent on the spread of the data rather than the number of data in a bucket, it proves to be very effective, even for clustered data and for correlated dimensions. A VAM-split *R*-tree can be constructed two orders of magnitude faster than an R^*-tree, taking less than a second for a 20-nearest-neighbors query for $M = 11$ on an $N = 100{,}000$ image database. The SS-tree [189] and its further improvement, the *SR*-tree [87], use the intersection of the minimum bounding hypersphere and minimum bounding hyperrectangle as the bounding region of a data element. As the dimension grows, it combines, for one bounding region, the advantage of the small volume of the hyperrectangle with the small diameter of the hypersphere. It has been shown that the *SR*-tree is efficient in low and high feature vector sizes.

Distance-based index structures are example-based space-partitioning techniques, and, hence, very suited for query by example when feature space is metric. The primary idea is to pick an example point and divide the rest of the feature space into M' groups in concentric rings around the example. This results in a distance-based index, the vantage point tree first proposed in [195]. In [14], the VP-tree was generalized for high dimensional feature vectors. The MVP-tree is a static data structure that uses multiple example (vantage) points at each node. The first vantage point is used to create M' partitions. And in each partition, a second vantage point creates M' more divisions. MVP-trees, with $M' = 3$, $M_{maxpernode} = 13$, using the L_2 metric found to perform fewer distance calculations during vector comparison for a nearest-neighbor search than competing techniques. The M-tree proposed in [27] is a more robust and scalable indexing strategy that uses the triangle-inequality of metric spaces, but, at the same time, retains the data partitioning properties of spatial access methods such as the *R*-tree and the *SS*-tree. *M*-trees are similar to MVP-trees in the sense that they both split the space based on spherical volumes around some reference vectors. However, *M*-trees are locally adapting, balanced indexes. Hence, they are less affected by the distribution of vectors in feature space.

Although the M-tree cannot guarantee worst case performance, it has been shown to work well in practice, with distance computation costs of $O(\log M)$ in most cases. It also has the advantage of dealing directly with features that can be represented in a metric space but not in a vector space, unlike techniques like FastMap [46] or multidimensional scaling which approximate the feature space with a vector space that maintains approximately the same metric.

It is regrettable that not too much work across the division between the disciplines of vision and databases has been done yet, with a few exceptions in commercial systems [48], [61] and research [46], [85], [171]. Recent work from the vision side is found in [170], where the database organizes itself for narrow domains in clustering hierarchies of the most expressive features, and in [2], where clusters of feature values are sought in a graph-theoretical representation.

7.2 System Architectures

Many systems like Photobook [128], PictoSeek [57], and [116] are rooted in computer vision. In such systems, the data and features are typically stored in files addressed by name. From an architectural point of view, this approach is likely to run into data integrity and performance problems when trying to scale up to a large database and a large number of users.

The large number of elements is clearly an issue in [21], [40], [168] and any other of the numerous Web search engines, where the emphasis is on filling the database using the World Wide Web as a logical repository. Architectural

issues focus on modules for searching the Web. In such architectures, a clear distinction can be made between off-line indexing and online readout for retrieval. The separation simplifies database integrity greatly. Hence, with simple, static index structures one can obtain a good performance during retrieval.

One step beyond the read-only databases is the use of a standard database management system extended for image retrieval. Groundbreaking examples are QBIC [48] and Virage [61]. In [120], an extended relational database system was used. Reducing image retrieval as a plug-in module in an existing database solves the integrity problem for image content and allows dynamic updates. It also provides natural integration with features derived from other sources. Standard databases maintain a narrow data-exchange channel between the search engine and the data and, hence, performance is rather poor. In the references, visualization and knowledge management is not being addressed as part of the integrated system.

For more complete systems, detailed architectures are to be considered. One of the early contributions to do so can be found in the CORE-system [193], which has been the basis for many different applications. The architecture is centered around a general Data Base Management System on top of which modules for analysis, indexing, training, and retrieval have been resolved in the database parlance. In such an approach, the database structure dominates knowledge management, feature calculation, and visualization tools, which severely hampers ease of expression in these areas.

The generic architecture described in [60] is based on a detailed model of the various information types. Each information type holds its data in a separate repository rather than the unified database, specifically separating data-driven and semantic, information bearing features. The advantage is that processing units, one for each new data type, can be put together quickly. The drawback is that the distinction between information-bearing features and data-driven features varies with the level of knowledge in the system and, hence, will be obstructing knowledge-based analysis.

The Infoscopes architecture [82] follows a similar distinction of features. It adds a new dimension to the architecture of content-based systems by making explicit the knowledge for the various parts of the system. The Piction system [155] proposes an architecture for collaborative use of image information and related textual information, while making knowledge explicit.

The systems discussed so far put their main emphasis on data and, later, on knowledge processing. In contrast, systems reviewed in [141] and, particularly, the MARS-system [142], are based on the information retrieval paradigm. The interaction with the user is considered crucial for a successful system. They propose an architecture for future systems which has a sequential processing structure from features to user interaction. This architecture ignores the role of data organization and explicit domain knowledge.

Integration of database research and visualization [96] has brought about techniques for visualizing the structure and content of an image database. Content-based similarity between images is not exploited here. The systems in [68], [184] have pursued integration furthest by using content-based similarity, interaction, and visualization, as well as database techniques for retrieving relevant images. Hence, they use techniques from three of the research fields identified earlier. The El Niño database system [146] proposes an architecture for the integration of several, possibly remote, engines through a mediator. The interaction is viewed as an outside source of knowledge infusion; semantics are expected to emerge in the course of the interaction. The addition of a knowledge component would lead to truly integrated content-based image retrieval systems.

7.3 System Evaluation

The evaluation of image retrieval is a difficult yet essential topic for the successful deployment of systems and their usefulness in practical applications. The initial impetus for the evaluation of image databases comes from the neighboring discipline of information retrieval, in which user-based evaluation techniques have reached a considerable degree of sophistication [143]. The main tool that image retrieval research borrowed from information retrieval are the precision and recall measures. Suppose a data set D and a query q are given. Through the use of human subjects, the data set can be divided into two sets: the set of images relevant for the query q, $R(q)$ and its complement, the set of irrelevant images $\bar{R}(q)$. Suppose that the query q is given to a data set and that it returns a set of images $A(q)$ as the answer. The precision of the answer is the fraction of the returned images that is indeed relevant for the query:

$$p = \frac{|A(q) \cap R(q)|}{|A(q)|}, \tag{25}$$

while the recall is the fraction of relevant images that is returned by the query:

$$r = \frac{|A(q) \cap R(q)|}{|R(q)|} \tag{26}$$

While precision and recall are a useful tool in information retrieval, in the context of image databases, they are not sufficient for two reasons. First, the selection of a relevant set in an image database is much more problematic than in a text database because of the more problematic definition of the meaning of an image. In the case of a document, the human judgment of whether a certain document is relevant to a certain query is relatively stable and there is a strong connection between this judgment and the statistical characteristics of the presence of certain elementary units (words). In the case of an image, relevance is much less stable because of the larger number of interpretation, of a image separated from a linguistic context. Moreover, no analysis in terms of semiotically stable constitutive elements can be done therefore, the correlation between image relevance and low-level feature is much more vague.

The second reason is that image databases usually do not return an undifferentiated set of "relevant" results, but a ranked list of results or some more complex configuration that shows the relation between the results of the query.

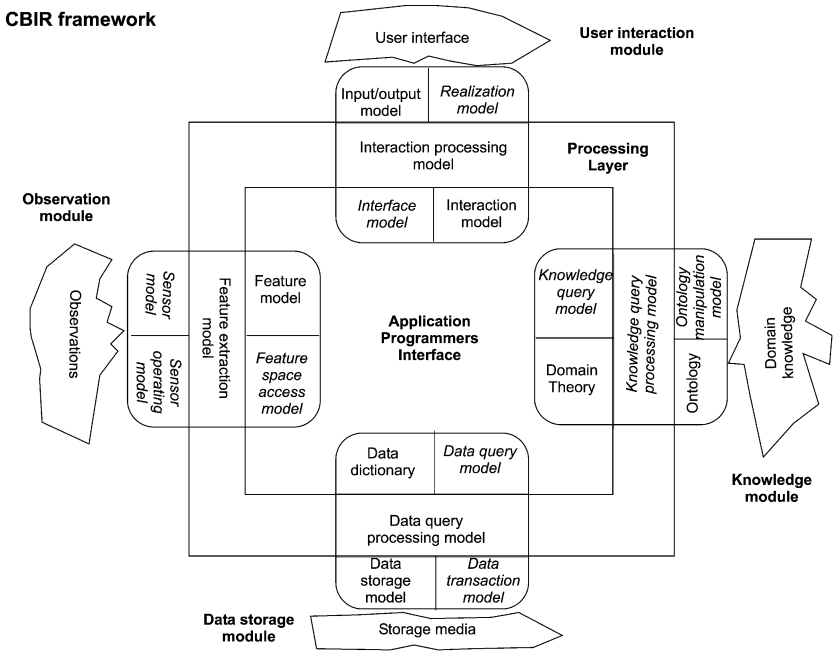


Fig. 13. The proposed framework for content-based image retrieval.

Although a query result is, in principle, an ordering of the *whole* database, the size of the answer set $A(q)$ is usually reduced to k most relevant images. When the number of relevant images is greater than k, recall is meaningless as a measure of the quality of the database.

In spite of these drawbacks, precision and recall (or other measures derived from them) are useful measurements in special circumstances. In particular, when the image database relies on the strong semantics provided by label or other textual description, precision and recall can be usefully employed [164].

In [119], the problem was considered of measuring the performance of a database without using the notion of relevant set. They assumed that an ideal database would take a query q and provide an ideal ordering of the database: $\mathcal{Z} = [z_1, \ldots, z_k]$. The ideal database will also provide a relevance measure for each image, $S(I_j) \in [0, 1]$. The database under test, on the other end, will order the image, given the same query q, as $\mathcal{Z}_\mathcal{Q} = [z_{\pi_1}, \ldots, z_{\pi_k}]$, where $[\pi_1, \ldots, \pi_k]$ is a permutation of $[1, \ldots, k]$. The displacement of the image I_j between the two orderings is given by $|j - \pi_j|$, and all the displacements can be added and weighted by the relevance of the respective images obtaining the weighed displacement

$$\omega = \sum_j S(I_j)|j - \pi_j|. \tag{27}$$

The weighed displacement gives a way of comparing the outputs of two databases, but this leaves open the problem of obtaining the ideal ordering $\mathcal{Z}$. In most cases, such ordering is obtained by performing experiments with human subjects.

With respect to experimental practices that use human subjects, a distinction can be made between the evaluation of a complete system and that of parts of a system [145]. In the first case, the system can be evaluated in the context of a well-defined activity by measuring the increased effectiveness resulting from the introduction of the database. Well-known techniques from social sciences can be used for the experimental design [19] and for the statistical analysis of the data [104]. In the second case, human subjects should be used to obtain the ground truth. Such an approach is followed, for instance, in [12], [124].

7.4 Discussion on System Aspects

As concerns system architecture, we maintain that a full-grown content-based retrieval system will result from the integration of a sensory and feature calculating part, a domain knowledge and interpretation module, an interaction and user interface module, and a storage and indexing module. For the system architectures discussed above, we conclude that most systems have an innovative emphasis understandably limited to one or two of these components. We feel there is a need for a framework for content-based image retrieval providing a more balanced view of the four constituent components. The framework would be based on explicit communication protocols to permit a discipline specific parlance within each of the modules, see Fig. 13. Such a framework follows the lines of object-oriented modular design, task differentiation, class abstractions, data hiding, and a communication protocol as CORBA.

8 Concluding Remarks

At the end of this review, we would like to present our view on a few trends:

1. *The driving force*. Content-based image retrieval came around quickly. Suddenly it was there, like the new economy. And it moves fast. In our review, most of the journal contributions are from the last five years. We are aware of the fact that much of what we have

said here will be outdated soon, and hopefully so, but we hope we laid down patterns in computation and interaction which may last a little longer.

The impetus behind content-based image retrieval is given by the wide availability of digital sensors, the Internet, and the falling price of storage devices. Given the magnitude of these driving forces, it is to us that content-based retrieval will continue to grow in every direction: new audiences, new purposes, new styles of use, new modes of interaction, larger data sets, and new methods to solve the problems.

What is needed most to bring the early years of content-based retrieval to an end is more precise foundations. For some of the reviewed papers, it was not clear what problem they were trying to solve or whether the proposed method would perform better than an alternative. A classification of usage-types, aims, and purposes would be very helpful here, including hard criteria for distinguishing among domain types. In spite of the difficulties intrinsic to the early years, it is now clear that content-based retrieval is not just old wine in new sacks. It will demand its own view of things as it is our belief that content-based retrieval in the end will not be part of the field of computer vision alone. The man-machine interface, domain knowledge, and database technology each will have their impact on the product.

2. *The heritage of computer vision*. An important obstacle to overcome before content-based image retrieval could take off was to realize that image retrieval does not entail solving the general image understanding problem. It may be sufficient that a retrieval system present similar images, similar in some user-defined sense. Strong segmentation of the scene and complete feature descriptions may not be necessary at all to achieve the similarity ranking. Of course, the deeper one goes into the semantics of the pictures, the deeper the understanding of the picture will have to be, but that could very well be based on categorizing pictures rather than on a precise understanding.

   We discussed applications of content-based retrieval in three broad types: target search, category search, and search by association. These user aims are rooted in the research tradition of the field.

   Target search builds on pattern matching and object-recognition. New challenges in content-based retrieval are the huge amount of objects among which to search, the incompleteness of the query specification and of the image descriptions, and the variability of sensing conditions and object states.

   Category search builds on object recognition and statistical pattern recognition problems. New challenges in content-based retrieval compared to the achievements of object recognition are the interactive manipulation of results, the usually very large number of classes, and the absence of an explicit training phase for feature selection and classification tuning.

   In the search by association, the goal is unspecified at the start of the session. Here, the heritage of computer vision is limited to feature sets and similarity functions. The association process is essentially iterative, interactive, and explicative. Therefore, association search is hampered most by the semantic gap. All display and relevance feedback has to be understood by the user, so the emphasis must be on developing features transparent to the user.

3. *The influence on computer vision*. In reverse, content-based image retrieval offers a different look at traditional computer vision problems.

   In the first place, content-based retrieval has brought large data sets. Where the number of test-images in a typical journal paper was well under a hundred until very recently, a state-of-the-art paper in content-based retrieval reports experiments on thousands of images. Of course, the purpose is different for computer vision and content-based retrieval. It is much easier to compose a general data set of arbitrary images rather than the specific ones needed in a computer vision application, but the stage has been set for more robustness. For one thing, to process a thousand images at least demands software and computational method be robust.

   In the second place, content-based retrieval has run into the absence of a general method for strong segmentation. Especially for broad domains and for sensory conditions where clutter and occlusion are to be expected, strong segmentation into objects is hard, if not impossible. Content-based retrieval systems have dealt with the segmentation bottleneck in a few creative ways. First, a weaker version of segmentation has been introduced in content-based retrieval. In weak segmentation, the result is a homogeneous region by some criterion, but not necessarily covering the complete object silhouette. Weak segmentation leads to the calculation of salient features capturing the essential information of the object in a nutshell. The extreme form of the weak segmentation is the selection of a salient point set as the ultimately efficient data reduction in the representation of an object, very much like the focus-of-attention algorithms for an earlier age. Weak segmentation and salient features are a typical innovation of content-based retrieval. It is expected that salience will receive much attention in the further expansion of the field, especially when computational considerations will gain in importance. The alternative to work around strong segmentation is to do no segmentation at all. Global features, such as wavelets and histograms, have been very effective. When the image is recorded with a photographic purpose, it is likely that the center of the image means something different than the surrounding parts of the image, so using that division of the picture could be of help too. Using no segmentation at all is likely to run dry on semantics

at the point where characteristics of the target are not specific enough in large databases to discriminate against the features of all other images.

In the third place, content-based retrieval has revitalized interest in color image processing. This is due to superior identification of trivalued intensities in identifying an object, as well as to the importance of color in the perception of images. As content-based is user-oriented, color cannot be left out. The purpose of most image color processing here is to reduce the influence of accidental conditions of the scene and the sensing (i.e., the sensory gap) by computing sensing and scene invariant representations. Progress has been made in tailored color space representation for well-described classes of variant conditions. Also, the application of local geometrical descriptions derived from scale space theory will reveal viewpoint and scene independent salient point sets, thus opening the way to similarity of images on a small number of most informative regions or points.

Finally, attention for invariance has been revitalized as well with many new features and similarity measures. For content-based retrieval, invariance is just one side of the coin, where discriminating power is the other. Little work has been reported so far to establish the remaining discriminating power of properties. This is essential as the balance between stability against variations and retained discriminating power determines the effectiveness of a property.

4. *Similarity and learning*. Similarity is an interpretation of the image based on the difference between two elements or groups of elements. For each of the feature types, a different similarity measure is needed. For similarity between feature sets, special attention has gone to establishing similarity between histograms due to their computational efficiency and retrieval effectiveness. Where most attention has gone to color histograms, it is expected that histograms of local geometric properties and texture will follow. Being such a unique computational concept, the histogram is receiving considerable attention from the database community for upgrading the performance on very large data sets. This is advantageous in the applicability of applying retrieval on very broad domains. To compensate for the complete loss of spatial information, new ways were recently explored as described above.

   Similarity of hierarchically ordered descriptions deserves considerable attention as it is one mechanism to circumvent the problems with segmentation while maintaining some of the semantically meaningful relationships in the image. Part of the difficulty here is to provide matching of partial disturbances in the hierarchical order and the influence of sensor-related variances in the description.

   We make a pledge for the importance of human-based similarity rather than general similarity. Also, the connection between image semantics, image data, and query context will have to be made clearer in the future. Similarity-induced semantics and the associated techniques for similarity adaptation (e.g., relevance feedback) are a first important step, but more sophisticated techniques, possibly drawing from machine learning, are necessary.

   Learning is quickly gaining attention as a means to build explicit models for each semantic term. Learning is made possible today by the availability of large data sets and powerful machines and allows one to form categories from captions, from partially labeled sets, or even from unlabeled sets. Learning is likely to be successful for large, labeled data sets on narrow domains first, which may be relaxed to broader domains and less standardized conditions as the available data sets will grow even more. Obviously, learning from labeled data sets is likely to be more successful than unsupervised learning first. New computational techniques, however, where only part of the data is labeled or the data is labeled by a caption rather than categories, open new possibilities. It is our view that, in order to bring semantics to the user, learning is inevitable.

5. *Interaction*. We consider the emphasis on interaction in image retrieval as one of the major departures from the computer vision tradition, as was already cited in the 1992 workshop [81]. Interaction was first picked up by frontrunners, such as NEC laboratories in Japan and the MIT Media Lab, to name a few. Now, interaction and feedback have moved into the focus of attention. Putting the user in control and visualization of the content has always been a leading principle in information retrieval research. It is expected that more and more techniques from traditional information retrieval will be employed or reinvented in content-based image retrieval. Text retrieval and image retrieval share the need for visualizing the information content in a meaningful way as well as the need to accept a semantic response of the user rather than just providing access to the raw data.

   User interaction in image retrieval has, however, some different characteristics from text retrieval. There is no sensory gap and the semantic gap from keywords to full text in text retrieval is of a different nature. No translation is needed from keywords to pictorial elements. In addition to the standard query types, six essentially different image based types have been identified in this paper. Each require their own user interface tools and interaction patterns. Due to the semantic gap, visualization of the query space in image retrieval is of great importance for the user to navigate the complex query space. While, currently, two- or three-dimensional display spaces are mostly employed in query by association, target search and category search are likely to follow. In all cases, an influx of computer graphics and virtual reality is foreseen in the near future.

   As there is no interactivity if the response time is frequently over a second, the interacting user poses high demands on the computational support. Indexing a data set for interactive use is a major challenge as the system cannot completely anticipate the user's actions. Still, in the course of the interaction, the

whole query space, i.e., the active image set, the features, the similarity, and the interpretations, can all change dynamically.

6. *The need for databases*. When data sets grow in size and when larger data sets define more interesting problems, both scientifically as well as for the public, the computational aspects can no longer be ignored.

   The connection between content-based image retrieval and database research is likely to increase in the future. Already, the most promising efforts are interdisciplinary, but, so far, problems like the definition of suitable query languages, efficient search in high dimensional feature space, search in the presence of changing similarity measures are largely unsolved.

   It is regrettable that little work cutting across the computer vision and database disciplines has been done so far, with a few notable exceptions. When truly large data sets come into view, with hundreds of thousands of images, databases can no longer be ignored as an essential component of a content-based retrieval system. In addition, when interactive performance is essential, storage and indexing must be organized in advance. Such large data sets will have an effect on the choice of features as the expressive power, computational cost, and hierarchical accessibility determine their effectiveness. For very large data sets, a view on content integrated with computation and indexing cannot be ignored. When speaking about "indexing" in computer vision, the emphasis is still on *what* to index, whereas the emphasis from the database side is on *how* to index. The difference has become smaller recently, but we believe most work is still to be done. Furthermore, in dealing with large feature vector sizes, the expansion of query definitions and query expansions in a useful manner for a variety of user aims is still mostly unanswered.

   For efficiency, more work on complete sets of feature calculations from compressed images is needed.

7. *The problem of evaluation*. It is clear that the evaluation of system performance is essential to sort out the good and the not-so-good methods. Up to this point in time, a fair comparison of methods under similar circumstances has been virtually absent. This is due to the infancy of content-based retrieval, but also to objective difficulties. Where interaction is a necessary component in most systems, it is difficult to separate out the influence of the data set in the performance. Also, it may be the case that some queries may match the expressive power of the system, whereas others, similar at first glance, may be much harder. Searching for a sunset may boil down to searching for a large orange disc at about the center of the image. Searching for a lamp, which may seem similar to the general audience, is a much harder problem as there is a whole variety of designs behind a lamp. The success of the system heavily depends on the toolset of the system relative to the query. In addition, it is logical that a large data set is composed of several smaller data sets to get a sufficiently big size. Then, the difficulty is the internal coherence of the large data set with respect to the coherence of its constituents. When a data set is composed of smaller data sets holding interior decorations, prairie landscapes, ships, and pigeons, it is clear that the essential difficulty of retrieval is within each set rather than among them and the essential size of the data set is still one quarter. There is no easy answer here other than the composition of generally agreed upon data sets or the use of very very large data sets. In all cases, the vitality of the content-based approach calls for a significant growth of the attention to evaluation in the future.

   A reference standard against which new algorithms could be evaluated has helped the field of text recognition enormously, see http://trec.nist.gov. A comprehensive and publicly available collection of images, sorted by class and retrieval purposes, together with a protocol to standardize experimental practices, will be instrumental in the next phase of content-based retrieval. We hope that a program for such a repository will be initiated under the auspices of a funding agency.

   At any rate, evaluation will likely play an increasingly significant role. Image databases, with their strong interactional component, present very different problems from the present ones which will require borrowing concepts from the psychological and social sciences.

8. *The semantic gap and other sources*. A critical point in the advancement of content-based retrieval is the semantic gap, where the meaning of an image is rarely self-evident.

   Use of content-based retrieval for browsing will not be within the grasp of the general public as humans are accustomed to relying on the immediate semantic imprint the moment they see an image and they expect a computer to do the same. The aim of content-based retrieval systems must be to provide maximum support in bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics.

   One way to resolve the semantic gap comes from sources outside the image by integrating other sources of information about the image in the query. Information about an image can come from a number of different sources: the image content, labels attached to the image, images embedded in a text, and so on. We still have very primitive ways of integrating this information in order to optimize access to images. Among these, the integration of natural language processing and computer vision deserves attention.

ACKNOWLEDGMENTS

The authors are grateful for the generous suggestions by the three anonymous referees. In addition, they acknowledge the discussions with E. Pauwels and M. Kersten from the CWI, J. M. Geusebroek and T. Gevers from the University of Amsterdam, and A. Tam from Victoria University. The work was supported in part by the ICES MIA-project.

REFERENCES

[1] P. Aigrain, H. Zhang, and D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State of the Art Review," *Multimedia Tools and Applications,* vol. 3, pp. 179-202, 1996.

[2] S. Aksoy and R. Haralick, "Graph-Theoretic Clustering for Image Grouping and Retrieval," *Proc. Computer Vision and Pattern Recognition,* pp. 63-68, 1999.

[3] R. Alferez and Y-F. Wang, "Geometric and Illumination Invariants for Object Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 6, pp. 505-536, June 1999.

[4] D. Androutsos, K.N. Plataniotis, and A.N. Venetsanopoulos, "A Novel Vector-Based Approach to Color Image Retrieval Using a Vector Angular-Based Distance Measure," *Image Understanding,* vol. 75, nos. 1-2, pp. 46-58, 1999.

[5] E. Angelopoulou and L.B. Wolff, "Sign of Gaussian Curvature from Curve Orientation in Photometric Space," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 10, pp. 1,056-1,066, Oct.. 1998.

[6] L. Armitage and P. Enser, "Analysis of User Need in Image Archives," *J. Information Science,* vol. 23, no. 4, pp. 287-299, 1997.

[7] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu, "An Optimal Algorithm for Approximate Nearest Neighborhood Searching," *Proc. Symp. Discrete Algorithms,* pp. 573-582, 1994.

[8] F.G. Ashby and N.A. Perrin, "Toward a Unified Theory of Similarity and Recognition," *Psychological Rev.,* vol. 95, no. 1, pp. 124-150, 1988.

[9] D. Ashlock and J. Davidson, "Texture Synthesis with Tandem Genetic Algorithms Using Nonparametric Partially Ordered Markov Models," *Proc. Congress on Evolutionary Computation,* pp. 1,157-1,163, 1999.

[10] J. Assfalg, A. del Bimbo, and P. Pala, "Using Multiple Examples for Content Based Retrieval," *Proc. Int'l Conf. Multimedia and Expo,* 2000.

[11] R. Basri, L. Costa, D. Geiger, and D. Jacobs, "Determining the Similarity of Deformable Shapes," *Vision Research,* vol. 38, nos. 15-16, pp. 2,365-2,385, 1998.

[12] S. Berretti, A. del Bimbo, and E. Vicario, "Modeling Spatial Relationships between Color Sets," *Proc. Workshop Content-Based Access of Image and Video Libraries,* 1998.

[13] *Database Techniques for Pictorial Applications, Lecture Notes in Computer Science,* A. Blaser, ed., vol. 81, Springer Verlag GmbH, 1979.

[14] T. Bozkaya and M. Ozsoyoglu, "Distance-Based Indexing for High-Dimensional Metric Spaces," *Proc. SIGMOD Int'l Conf. Management of Data,* pp. 357-368, 1997.

[15] L. Brown and L. Gruenwald, "Tree-Based Indexes for Image Data," *J. Visual Comm. and Image Representation,* vol. 9, no. 4, pp. 300-313, 1998.

[16] R. Brunelli, O. Mich, and C.M. Modena, "A Survey on the Automatic Indexing of Video Data," *J. Visual Comm. and Image Representation,* vol. 10, pp. 78-112, 1999.

[17] G. Bucci, S. Cagnoni, and R. De Dominicis, " Integrating Content-Based Retrieval in a Medical Image Reference Database," *Computerized Medical Imaging and Graphics,* vol. 20, no. 4, pp. 231-241, 1996.

[18] H. Burkhardt and S. Siggelkow, "Invariant Features for Discriminating between Equivalence Classes," *Nonlinear Model-Based Image Video Processing and Analysis,* John Wiley and Sons, 2000.

[19] D. Campbell and J. Stanley, *Experimental and Quasi-Experimental Designs for Research.* Rand McNally College Publishing, 1963.

[20] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Region-Based Image Querying," *Proc. Int'l Workshop Content-Based Access of Image and Video libraries,* 1997.

[21] S.-F. Chang, J.R. Smith, M. Beigi, and A. Benitez, "Visual Information Retrieval from Large Distributed Online Repositories," *Comm. ACM,* vol. 40, no. 12, pp. 63-71, 1997.

[22] S.-K. Chang and A.D. Hsu, "Image-Information Systems—Where Do We Go from Here?" *IEEE Trans. Knowledge and Data Eng.,* vol. 4, no. 5, pp. 431-442, Oct. 1992.

[23] S.-K. Chang, Q.Y. Shi, and C.W. Yan, "Iconic Indexing by 2D Strings," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 9, pp. 413-428, 1987.

[24] H. Chen, B. Schatz, T. Ng, J. Martinez, A. Kirchhoff, and C. Lim, "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 771-782, Aug. 1996.

[25] H. Choi and R. Baraniuk, "Multiscale Texture Segmentation Using Wavelet-Domain Hidden Markov Models," *Proc. 32nd Asilomar Conf. Signals, Systems, and Computers,* vol. 2, pp. 1,692-1,697, 1998.

[26] C.K. Chui, L. Montefusco, and L. Puccio, *Wavelets: Theory, Algorithms, and Applications.* Academic Press, 1994.

[27] P. Ciaccia, M. Patella, and P. Zezula, "M-Tree: An Efficient Access Method for Similarity Search in Metric Spaces," *Proc. Very Large Data Bases Conf.,* 1997.

[28] G. Ciocca and R. Schettini, "Using a Relevance Feedback Mechanism to Improve Content-Based Image Retrieval," *Proc. Visual '99: Information and Information Systems,* pp. 107-114, 1999.

[29] P. Correira and F. Pereira, "The Role of Analysis in Content-Based Video Coding and Indexing," *Signal Processing,* vol. 66, no. 2, pp. 125-142, 1998.

[30] J.M. Corridoni, A. del Bimbo, and P. Pala, "Image Retrieval by Color Semantics," *Multimedia Systems,* vol. 7, pp. 175-183, 1999.

[31] I.J. Cox, M.L. Miller, T.P. Minka, and T.V. Papathomas, "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Pychophysical Experiments," *IEEE Trans. Image Processing,* vol. 9, no. 1, pp. 20-37, 2000.

[32] G. Csurka and O. Faugeras, "Algebraic and Geometrical Tools to Compute Projective and Permutation Invariants," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 1, pp. 58-65, Jan. 1999.

[33] J.F. Cullen, J.J. Hull, and P.E. Hart, "Document Image Database Retrieval and Browsing Using Texture Analysis," *Proc. Fourth Int'l Conf. Document Analysis and Recognition,* pp. 718-721, 1997.

[34] I. Daubechies, *Ten Lectures on Wavelets.* Philadelphia: SIAM, 1992.

[35] J. de Bonet and P. Viola, "Texture Recognition Using a Non-Parametric Multi-Scale Statistical Model," *Proc. Computer Vision and Pattern Recognition,* 1998.

[36] A. del Bimbo and P. Pala, "Visual Image Retrieval by Elastic Matching of User Sketches," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp. 121-132, Feb. 1997.

[37] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc.,* vol. 39, no. 1, pp. 1-38, 1977.

[38] D. Dubois and H. Prade, "A Review of Fuzzy Set Aggregation Connectives," *Information Sciences,* vol. 36, pp. 85-121, 1985.

[39] J.P. Eakins, J.M. Boardman, and M.E. Graham, "Similarity Retrieval of Trademark Images," *IEEE Multimedia,* vol. 5, no. 2, pp. 53-63, Apr.-June 1998.

[40] B. Eberman, B. Fidler, R. Ianucci, C. Joerg, L. Kontothanassis, D.E. Kovalcin, P. Moreno, M.J. Swain, and J.-M. van Thong, "Indexing Multimedia for the Internet," *Proc. Visual '99: Information and Information Systems,* pp. 195-202, 1999.

[41] F. Ennesser and G. Medioni, "Finding Waldo, or Focus of Attention Using Local Color Information," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 8, pp. 805-809, Aug. 1995.

[42] P.G.B. Enser, "Pictorial Information Retrieval," *J. Documentation,* vol. 51, no. 2, pp. 126-170, 1995.

[43] C. Esperanca and H. Samet, "A Differential Code for Shape Representation in Image Database Applications," *Proc. Int'l Conf. Image Processing,* 1997.

[44] L.M. Kaplan et al., "Fast Texture Database Retrieval Using Extended Fractal Features," *Storage and Retrieval for Image and Video Databases, VI,* vol. 3,312, pp. 162-173, SPIE Press, 1998.

[45] R. Fagin, "Combining Fuzzy Information from Multiple Systems," *J. Computer Systems Science,* vol. 58, no. 1, pp. 83-99, 1999.

[46] C. Faloutsos and K.I. Lin, "Fastmap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets," *Proc. SIGMOD, Int'l Conf. Management of Data,* pp. 163-174, 1995.

[47] G.D. Finlayson, "Color in Perspective," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 10, pp. 1,034-1,038, Oct. 1996.

[48] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC System," *IEEE Computer,* 1995.

[49] D.A. Forsyth, "A Novel Algorithm for Color Constancy," *Int'l J. Computer Vision,* vol. 5, no. 1, pp. 5-36, 1990.

[50] D.A. Forsyth and M.M. Fleck, "Automatic Detection of Human Nudes," *Int'l J. Computer Vision,* vol. 32, no. 1, pp. 63-77, 1999.

[51] G. Frederix, G. Caenen, and E.J. Pauwels, "PARISS: Panoramic, Adaptive and Reconfigurable Interface for Similarity Search," *Proc. Int'l Conf. Image Processing,* 2000.

[52] G. Frederix and E.J. Pauwels, "Automatic Interpretation Based on Robust Segmentation and Shape Extraction," *Proc. Visual '99: Information and Information Systems,* pp. 769-776, 1999.
[53] C.-S. Fuh, S.-W. Cho, and K. Essig, "Hierarchical Color Image Region Segmentation for Content-Based Image Retrieval System," *IEEE Trans. Image Processing,* vol. 9, no. 1 pp. 156-163, 2000.
[54] B.V. Funt and G.D. Finlayson, "Color Constant Color Indexing," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 5, pp. 522-529, May 1995.
[55] J.M. Geusebroek, A.W.M. Smeulders, and R. van den Boomgaard, "Measurement of Color Invariants," *Proc. Computer Vision and Pattern Recognition,* 2000.
[56] T. Gevers and A.W.M. Smeulders, "Content-Based Image Retrieval by Viewpoint-Invariant Image Indexing," *Image and Vision Computing,* vol. 17, no. 7, pp. 475-488, 1999.
[57] T. Gevers and A.W.M. Smeulders, "Pictoseek: Combining Color and Shape Invariant Features for Image Retrieval," *IEEE Trans. Image Processing,* vol. 9, no. 1, pp. 102-119, 2000.
[58] G.L. Gimel'farb and A.K. Jain, "On Retrieving Textured Images from an Image Database," *Pattern Recognition,* vol. 29, no. 9, pp. 1,461-1,483, 1996.
[59] C.C. Gottlieb and H.E. Kreyszig, "Texture Descriptors Based on Co-Occurrences Matrices," *Computer Vision, Graphics, and Image Processing,* vol. 51, 1990.
[60] W.I. Grosky, "Multi-Media Information Systems." *IEEE Multimedia,* vol. 1, no. 1, Mar. 1994.
[61] A. Gupta and R. Jain, "Visual Information Retrieval," *Comm. ACM,* vol. 40, no. 5, pp. 71-79, 1997.
[62] J. Hafner, H.S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 7, pp. 729-736, July 1995.
[63] M. Hagendoorn and R.C. Veltkamp, "Reliable and Efficient Pattern Matching Using an Affine Invariant Metric," *Int'l J. Computer Vision,* vol. 35, no. 3, pp. 203-225, 1999.
[64] S. Hastings, "Query Categories in a Study of Intellectual Access to Digitized Art Images," *Proc. 58th Ann. Meeting Am. Soc. Information Science,* 1995.
[65] H. Hatano, "Image Processing and Database System in the National Museum of Western Art," *Int'l J. Special Libraries,* vol. 30, no. 3, pp. 259-267, 1996.
[66] G. Healey and D. Slater, "Computing Illumination-Invariant Descriptors of Spatially Filtered Color Image Regions," *IEEE Trans. Image Processing,* vol. 6, no. 7 pp. 1,002-1,013, 1997.
[67] K. Hirata and T. Kato, "Rough Sketch-Based Image Information Retrieval," *NEC Research and Development,* vol. 34, no. 2, pp. 263-273, 1992.
[68] A. Hiroike, Y. Musha, A. Sugimoto, and Y. Mori, "Visualization of Information Spaces to Retrieve and Browse Image Data," *Proc. Visual '99: Information and Information Systems,* pp. 155-162, 1999.
[69] N.R. Howe and D.P. Huttenlocher, "Integrating Color, Texture, and Geometry for Image Retrieval," *Proc. Computer Vision and Pattern Recognition,* pp. 239-247, 2000.
[70] C.C. Hsu, W.W. Chu, and R.K. Taira, "A Knowledge-Based Approach for Retrieving Images by Content," *IEEE Trans. Knowledge and Data Eng.,* vol. 8, no. 4, pp. 522-532, 1996.
[71] F.J. Hsu, S.Y. Lee, and B.S. Lin, "Similairty Retrieval by 2D C-Trees Matching in Image Databases," *J. Visual Comm. and Image Representation,* vol. 9, no. 1, pp. 87-100, 1998.
[72] M.K. Hu, "Pattern Recognition by Moment Invariants," *Proc. IRE Trans. Information Theory,* vol. 8, pp. 179-187, 1962.
[73] J. Huang, S.R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Spatial Color Indexing and Applications," *Int'l J. Computer Vision,* vol. 35, no. 3, pp. 245-268, 1999.
[74] B. Huet and E.R. Hancock, "Line Pattern Retrieval Using Relational Histograms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 12, pp. 1,363-1,371, Dec. 1999.
[75] F. Idris and S. Panchanathan, "Image Indexing Using Wavelet Vector Quantization," *Proc. Digital Image Storage and Archiving Systems,* vol. 2,606, pp. 269-275, 1995.
[76] L. Itti, C. Koch, and E. Niebur, "A Model for Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1,254-1,259, Nov. 1998.
[77] C.E. Jacobs and A. Finkelstein, S.H. Salesin, "Fast Multiresolution Image Querying," *Proc. SIGGRAPH,* 1995.
[78] A.K. Jain and A. Vailaya, "Image Retrieval Using Color and Shape," *Pattern Recognition,* vol. 29, no. 8, pp. 1,233-1,244, 1996.
[79] A.K. Jain and A. Vailaya, "Shape-Based Retrieval: A Case Study with Trademark Image Databases," *Pattern Recognition,* vol. 31, no. 9, pp. 1,369-1,390, 1998.
[80] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 4-37, Jan. 2000.
[81] *Proc. US NSF Workshop Visual Information Management Systems,* R. Jain, ed., 1992.
[82] R. Jain, "InfoScopes: Multimedia Information Systems," *Multimedia Systems and Techniques,* Kluwer Academic Publishers, 1996.
[83] L. Jia and L. Kitchen, "Object-Based Image Similarity Computation Using Inductive Learning of Contour-Segment Relations," *IEEE Trans. Iamge Processing,* vol. 9, no. 1, pp. 80-87, 2000.
[84] D.W. Joyce, P.H. Lewis, R.H. Tansley, M.R. Dobie, and W. Hall, "Semiotics and Agents for Integrating and Navigating through Multimedia Representations," *Proc. Storage and Retrieval for Media Databases,* vol. 3972, pp. 120-131, 2000.
[85] D. Judd, P. McKinley, and A.K. Jain, "Large-Scale Parallel Data Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 871-876, Aug. 1998.
[86] T. Kakimoto and Y. Kambayashi, "Browsing Functions in Three-Dimensional Space for Digital Libraries," *Int'l J. Digital Libraries,* vol. 2, pp. 68-78, 1999.
[87] N. Katayama and S. Satoh, "The SR-Tree: An Index Structure for High-Dimensional Nearest Neighbor Queries," *Proc. SIGMOD, Int'l Conf. Management of Data,* pp. 369-380, 1997.
[88] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A Sketch Retrieval Method for Full Color Image Database—Query by Visual Example," *Proc. ICPR, Computer Vision and Applications,* pp. 530-533, 1992.
[89] A. Kontanzad and Y.H. Hong, "Invariant Image Recognition by Zernike Moments," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 5, pp. 489-497, May 1990.
[90] F. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, "Efficient and Effective Nearest Neighbor Search in a Medical Image Database of Tumor Shapes," *Image Description and Retrieval,* pp. 17-54, 1998.
[91] S. Krishnamachari and R. Chellappa, "Multiresolution Gauss-Markov Random Field Models for Texture Segmentation," *IEEE Trans. Image Processing,* vol. 6, no. 2, 1997.
[92] A. Laine and J. Fan, "Texture Classification by Wavelet Packet Signature," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 11, pp. 1,186-1,191, Nov. 1993.
[93] L.J. Latecki and R. Lakämper, "Contour-Based Shape Similarity," *Proc. Visual '99: Information and Information Systems,* pp. 617-624, 1999.
[94] L.J. Latecki and R. Lakämper, "Convexity Rule for Shape Decomposition Based on Discrete Contour Evolution," *Image Understanding,* vol. 73, no. 3, pp. 441-454, 1999.
[95] T.K. Lau and I. King, "Montage: An Image Database for the Fashion, Textile, and Clothing Industry in Hong Kong," *Proc. Asian Conf. Computer Vision,* pp. 575-582, 1998.
[96] M. Leissler, M. Hemmje, and E.J. Neuhold, "Supporting Image Retrieval by Database Driven Interactive 3D Information-Visualization," *Visual Information and Information Systems,* pp. 1-14, Springer Verlag, 1999.
[97] M.S. Lew and N. Sebe, "Visual Websearching Using Iconic Queries," *Proc. Computer Vision and Pattern Recognition,* pp. 788-789, 2000.
[98] C.-S. Li and V. Castelli, "Deriving Texture Feature Set for Content-Based Retrieval of Satellite Image Database," *Proc. Int'l Conf. Image Processing,* 1997.
[99] H.C. Lin, L.L. Wang, and S.N. Yang, "Color Image Retrieval Based on Hidden Markov Models," *IEEE Trans. Image Processing,* vol. 6, no. 2, pp. 332-339, 1997.
[100] T. Lindeberg and J.O. Eklundh, "Scale Space Primal Sketch Construction and Experiments," *Image Vision Computing,* vol. 10, pp. 3-18, 1992.
[101] F. Liu and R.W. Picard, "Periodicity, Directionality, and Randomness: Wold Features for Image Modeling and Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 7, pp. 517-549, July 1996.
[102] W.Y. Ma and B.S. Manjunath, "Edge Flow: A Framework of Boundary Detection and Image Segmentation," *Proc. Computer Vision and Pattern Recognition,* pp. 744-749, 1997.

[103] M.K. Mandal, F. Idris, and S. Panchanathan, "Image and Video Indexing in the Compressed Domain: A Critical Review," *Image and Vision Computing,* 2000.

[104] J. Mandel, *The Statistical Analysis of Experimental Data.* Interscience Publishers, 1964. Republished in 1984, Mineola, N.Y.: Dover.

[105] B.S. Manjunath and W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 837-842, Aug. 1996

[106] J. Mao and A.K. Jain, "Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models," *Pattern Recognition,* vol. 25, no. 2, 1992.

[107] J. Matas, R. Marik, and J. Kittler, "On Representation and Matching of Multi-Colored Objects," *Proc. Fifth Int'l Conf. Computer Vision,* pp. 726-732, 1995.

[108] R. Mehrotra and J.E. Gary, "Similar-Shape Retrieval in Shape Data Management," *Computer,* vol. 28, no. 9, pp. 57-62, Sept. 1995.

[109] B.M. Mehtre, M.S. Kankanhalli, and W.F. Lee, "Shape Measures for Content Based Image Retrieval: A Comparison," *Information Procesing Management,* vol. 33, no. 3, pp. 319-337, 1997.

[110] C. Meilhac and C. Nastar, "Relevance Feedback and Category Search in Image Databases," *Proc. Int'l Conf. Multimedia Computing and Systems,* pp. 512-517, 1999.

[111] T.P. Minka and R.W. Picard, "Interactive Learning with a 'Society of Models,'" *Pattern Recognition,* vol. 30, no. 4, pp. 565-582, 1997.

[112] M. Mirmehdi and M. Petrou, "Segmentation of Color Texture," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 2, pp. 142-159, Feb. 2000.

[113] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.

[114] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond Eigenfaces: Probabilistic Matching for Face Recognition," *Proc. Third Int'l Conf. Automatic Face and Gesture Reognition,* 1998.

[115] A. Mojsilovic, J. Kovacevic, J. Hu, R.J. Safranek, and S.K. Ganapathy, "Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns," *IEEE Trans. Image Processing,* vol. 9, no. 1, pp. 38-54, 2000.

[116] F. Mokhtarian, "Silhouette-Based Isolated Object Recognition through Curvature Scale-Space," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 5, pp. 539-544, May 1995.

[117] *Applications of Invariance in Computer Vision, Lecture Notes in Computer Science,* J.L. Mundy, A. Zissermann, D. Forsyth, eds., vol. 825, Springer Verlag, 1994

[118] H. Murase and S.K. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance," *Int'l J. Computer Vision,* vol. 14, no. 1, pp. 5-24, 1995.

[119] D.A. Narasimhalu, M.S. Kankanhalli, and J. Wu, "Benchmarking Multimedia Databases," *Multimedia Tools and Applications,* vol. 4, no. 3, pp. 333-355, 1997.

[120] V.E. Ogle, "CHABOT—Retrieval from a Relational Database of Images," *Computer,* vol. 28, no. 9, pp. 40-48, Sept. 1995.

[121] S. Ornager, "Image Retrieval: Theoretical and Empirical User Studies on Accessing Information in Images," *Proc. 60th Am. Soc. Information Science Ann. Meeting,* vol. 34, pp. 202-211, 1997.

[122] P. Pala and S. Santini, "Image Retrieval by Shape and Texture," *Pattern Recognition,* vol. 32, no. 3, pp. 517-527, 1999.

[123] M.L. Pao and M. Lee, *Concepts of Information Retrieval.* Libraries Unlimited, 1989.

[124] T.V. Papathomas, E. Conway, I.J. Cox, J. Ghosn, M.L. Miller, T.P. Minka, and P.N. Yianilos, "Psychophysical Studies of the Performance of an Image Database Retrieval System," *Proc. Symp. Electronic Imaging: Conf. Human Vision and Electronic Imaging III,* 1998.

[125] G. Pass and R. Zabith, "Comparing Images Using Joint Histograms," *Multimedia Systems,* vol. 7, pp. 234-240, 1999.

[126] E.J. Pauwels and G. Frederix, "Nonparametric Clustering for Image Segmentation and Grouping," *Image Understanding,* vol. 75, no. 1, pp. 73-85, 2000.

[127] A. Pentland and T. Choudhury, "Face Recognition for Smart Environments," *Computer,* vol. 33, no. 2, pp. 50-, Feb. 2000.

[128] A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Content-Based Manipulation of Image Databases," *Int'l J. Computer Vision,* vol. 18, no. 3, pp. 233-254, 1996.

[129] E. Petrakis and C. Faloutsos, "Similarity Searching in Medical Image Databases," *IEEE Trans. Knowledge and Data Eng.,* vol. 9, no. 3, pp. 435-447, June 1997.

[130] R.W. Picard and T.P. Minka, "Vision Texture for Annotation," *Multimedia Systems,* vol. 3, pp. 3-14, 1995.

[131] J. Puzicha, T. Hoffman, and J.M. Buhmann, "Non-Parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval," *Proc. Computer Vision and Pattern Recognition,* 1997.

[132] W. Qian, M. Kallergi, L.P. Clarke, H.D. Li, D. Venugopal, D.S. Song, and L.P. Clark, "Tree-Structured Wavelet Transform Segmentation of Microcalcifications in Digital Mammography," *J. Medical Physiology,* vol. 22, no. 8, pp. 1,247-1,254, 1995.

[133] T. Randen and J.H. Husoy, "Filtering for Texture Classification: A Comparative Study," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 4, pp. 291-310, Apr. 1999.

[134] A. Rao, R.K. Srihari, and Z. Zhang, "Geometric Histogram: A Distribution of Geometric Configurations of Color Subsets," *Internet Imaging,* vol. 3,964, pp. 91-101, 2000.

[135] E. Riloff and L. Hollaar, "Text Databases and Information Retrieval," *ACM Computing Surveys,* vol. 28, no. 1, pp. 133-135, 1996.

[136] E. Rivlin and I. Weiss, "Local Invariants for Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 3, pp. 226-238, Mar. 1995.

[137] R. Rodriguez-Sanchez, J.A. Garcia, J. Fdez-Valdivia, and X.R. Fdez-Vidal, "The RGFF Representational Model: A System for the Automatically Learned Partitioning of 'Visual Pattern' in Digital Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 10, pp. 1,044-1,073, Oct. 1999.

[138] P.L. Rosin, "Edges: Saliency Measures and Automatic Thresholding," *Machine Vision Applications,* vol. 9, no. 7, pp.139-159, 1997.

[139] D. Roth, M.-H. Yang, and N. Ahuja, "Learning to Recognize Objects," *Proc. Computer Vision and Pattern Recognition,* pp. 724-731, 2000.

[140] I. Rothe, H. Suesse, and K. Voss, "The Method of Normalization of Determine Invariants," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 4, pp. 366-376, Apr. 1996.

[141] Y. Rui, T.S. Huang, and S.-F. Chang, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues," *J. Visual Comm. and Image Representation,* vol. 10, no. 1, pp. 39-62, 1999.

[142] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," *IEEE Trans. Circuits and Systems for Video Technology,* 1998.

[143] G. Salton, *Automatic Text Processing.* Addison-Wesley, 1988.

[144] H. Samet and A. Soffer, "MARCO: MAp Retrieval by Content," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 783-798, Aug. 1996.

[145] S. Santini, "Evaluation Vademecum for Visual Information Systems," *Storage and Retrieval for Image and Video Databases VIII,* vol. 3,972, 2000.

[146] S. Santini, A. Gupta, and R. Jain, "User Interfaces for Emergent Semantics in Image Databases," *Proc. Eighth IFIP Working Conf. Database Semantics (DS-8),* 1999.

[147] S. Santini and R. Jain, "Similarity Measures," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 9, pp. 871-883, Sept. 1999.

[148] C. Schmid and R. Mohr, "Local Grayvalue Invariants for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 5, pp. 530-535, May 1997.

[149] M. Schneier and M. Abdel-Mottaleb, "Exploiting the JPEG Compression Scheme for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, 849 -853, Aug. 1996.

[150] S. Sclaroff, "Deformable Prototypes for Encoding Shape Categories in Image Databases," *Pattern Recognition,* vol. 30, no. 4, pp. 627-641, 1997.

[151] S. Sclaroff, M. La Cascia, and S. Sethi, "Using Textual and Visual Cues for Content-Based Image Retrieval from the World Wide Web," *Image Understanding,* vol. 75, no. 2, pp. 86-98, 1999.

[152] S. Sclaroff, L. Taycher, and M. La Cascia, "Imagerover: A Content-Base Image Browser for the World Wide Web," *Proc. Workshop Content-Based Access to Image and Video Libraries,* pp. 1,000-1,006, 1997.

[153] D. Sharvit, J. Chan, H. Tek, and B.B. Kimia, "Symmetry-Based Indexing of Image Databases," *J. Visual Comm. and Image Representation,* vol. 9, no. 4, pp. 366-380, 1998.

[154] R.N. Shepard, "Toward a Universal Law of Generalization for Physical Science," *Science,* vol. 237, pp. 1,317-1,323, 1987.

[155] R.H. Shrihari, "Automatic Indexing and Content-Based Retrieval of Captioned Images," *Computer,* vol. 28, no. 9, Sept. 1995.

[156] C.-R. Shyu, C.E. Brodley, A.C. Kak, and A. Kosaka, "ASSERT: A Physician in the Loop Content-Based Retrieval System for HCRT Image Databases," *Image Understanding,* vol. 75, nos. 1/2, pp. 111-132, 1999.

[157] K. Siddiqi and B.B. Kimia, "Parts of Visual Form: Computational Aspects," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 3, pp. 239-251, Mar. 1995.

[158] D. Slater and G. Healey, "The Illumination-Invariant Recognition of 3D Objects Using Local Color Invariants," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 2, pp. 206-210, Feb. 1996.

[159] A.W.M. Smeulders, T. Gevers, J.M. Geusebroek, and M. Worring, "Invariance in Content-Based Retrieval," *Proc. Int'l Conf. Multimedia and Expo,* 2000.

[160] A.W.M. Smeulders, M.L. Kersten, and T. Gevers, "Crossing the Divide between Computer Vision and Data Bases in Search of Image Databases," *Proc. Fourth Working Conf. Visual Database Systems,* pp. 223-239, 1998.

[161] A.W.M. Smeulders, S.D. Olabariagga, R. van den Boomgaard, and M. Worring, "Interactive Segmentation," *Proc. Visual '97: Information Systems,* pp. 5-12, 1997.

[162] J.R. Smith and S.F. Chang, "Automated Binary Feature Sets for Image Retrieval," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* 1996.

[163] J.R. Smith and S.-F. Chang, "Integrated Sspatial and Feature Image Query," *Multimedia Systems,* vol. 7, no. 2, pp. 129-140, 1999.

[164] J.R. Smith and C.-S. Li, "Image Retrieval Evaluation," *Proc. Workshop Content-Based Access of Image and Video Libraries,* 1998.

[165] S.M. Smith and J.M. Brady, "SUSAN—A New Approach to Low Level Image Processing," *Int'l J. Computer Vision,* vol. 23, no. 1, pp. 45-78, 1997.

[166] M. Stricker and M. Orengo, "Similarity of Color Images," *Storage and Retrieval of Image and Video Databases III,* vol. 2,420, pp. 381-392, 1995.

[167] M. Stricker and M. Swain, "The Capacity of Color Histogram Indexing," *Proc. Computer Vision and Pattern Recognition,* pp. 704-708, 1994.

[168] M.J. Swain, "Searching for Multimedia on the World Wide Web," *Proc. Int'l Conf. Multimedia Computing and Systems,* pp. 33-37, 1999.

[169] M.J. Swain and B.H. Ballard, "Color Indexing," *Int'l J. Computer Vision,* vol. 7, no. 1, pp. 11-32, 1991.

[170] D.J. Swets and J. Weng, "Hierarchical Discriminant Analysis for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 5, pp. 386-401, May 1999.

[171] T.F. Syeda-Mahmood, "Location Hashing: An Efficient Method for Locating Object Queries in Image Databases," *Storage and Retrieval in Image and Video Databases,* vol. 3,656, pp. 366-378, 1999.

[172] H.D. Tagare, F.M. Vos, C.C. Jaffe, and J.S. Duncan, "Arrangement—A Spatial Relation between Parts for Evaluating Similarity of Tomographic Section," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 9, pp. 880-893, Sept. 1995.

[173] T. Tan, "Rotation Invariant Texture Features and Their Use in Automatic Script Identification," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 7, pp. 751-756, July 1998.

[174] P.M. Tardif and A. Zaccarin, "Multiscale Autoregressive Image Respresentation for Texture Segmentation," *Image Processing VIII,* vol. 3,026, pp. 327-337, 1997.

[175] J. Tatemura, "Browsing Images Based on Social and Content Similarity," *Proc. Int'l Conf. Multimedia and Expo,* 2000.

[176] K. Tieu and P. Viola, "Boosting Image Retrieval," *Proc. Computer Vision and Pattern Recognition,* pp. pp. 228-235, 2000.

[177] A. Treisman, P. Cavanagh, B. Fisher, V.S. Ramachandran, and R. von der Heydt, "Form Perception and Attention-Striate Cortex and Beyond," *Visual Perception: The Neurophysiological Foundation,* pp. 273-316, 1990.

[178] T. Tuytelaars and L. van Gool, "Content-Based Image Retrieval Based on Local Affinely Invariant Regions," *Proc. Visual '99: Information and Information Systems,* pp. 493-500, 1999.

[179] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Content-Based Hierarchical Classification of Vacation Images," *Proc. Int'l Conf. Multimedia Computing and Systems,* 1999.

[180] G.W.A.M. van der Heijden and M. Worring, "Domain Concept to Feature Mapping for a Plant Variety Image Database," *Image Databases and Multimedia Search,* vol. 8, pp. 301-308, 1997.

[181] V.N. Vapnik, *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[182] N. Vasconcelos and A. Lippman, "A Probabilistic Architecture for Content-Based Image Retrieval," *Proc. Computer Vision and Pattern Recognition,* pp. pp. 216-221, 2000.

[183] R.C. Veltkamp and M. Hagendoorn, "State-of-the-Art in Shape Matching," *Multimedia Search: State of the Art,* Springer-Verlag, 2000.

[184] J. Vendrig, M. Worring, and A.W.M. Smeulders, "Filter Image Browsing: Exploiting Interaction in Retrieval," *Proc. Visual '99: Information and Information Systems,* 1999.

[185] L.Z. Wang and G. Healey, "Using Zernike Moments for the Illumination and Geometry Invariant Classification of Multi-Spectral Texture," *IEEE Trans. Image Processing,* vol. 7, no. 2, pp. 196-203, 1991.

[186] M. Weber, M. Welling, and P. Perona, "Towards Automatic Discovery of Object Categories," *Proc. Computer Vision and Pattern Recognition,* pp. 101-108, 2000.

[187] J. Weickert, S. Ishikawa, and A. Imiya, "Linear Scale Space Has First Been Proposed in Japan," *J. Math., Imaging and Vision,* vol. 10, pp. 237-252, 1999.

[188] M. Werman and D. Weinshall, "Similarity and Affine Invariant Distances between 2D Point Sets," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 8, pp. 810-814, Aug. 1995.

[189] D. White and R. Jain, "Similarity Indexing with the SS-Tree," *Proc. 12th Int'l Conf. Data Eng.,* 1996.

[190] D.A. White and R. Jain, "Algorithms and Strategies for Similarity Retrieval," *Storage and Retrieval in Image, and Video Databases,* vol. 2,060, pp. 62-72, 1996.

[191] R.C. Wilson and E.R. Hancock, "Structural Matching by Discrete Relaxation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 6, pp. 634-648, June 1997.

[192] H.J. Wolfson and I. Rigoutsos, "Geometric Hashing: An Overview," *IEEE Trans. Computational Science Eng.,* vol. 4, no. 4, pp. 10-21, 1997.

[193] J.K. Wu, A.D. Narasimhalu, B.M. Mehtre, C.P. Lam, and Y.J. Gao, "CORE: A Content Based Retrieval System for Multimedia Information Systems," *Multimedia Systems,* vol. 3, pp. 25-41, 1995.

[194] Y. Wu, Q. Tian, and T.S. Huang, "Discriminant-EM Algorithm with Application to Image Retrieval," *Proc. Computer Vision and Pattern Recognition,* pp. 222-227, 2000.

[195] P.N. Yanilos, "Data Structures and Aalgorithms for Nearest Neighbor Search in General Metric Spaces," *Proc. Third Ann. Symp. Discrete Algorithms,* pp. 516-523, 1993.

[196] N. Yazdani, M. Ozsoyoglu, and G. Ozsoyoglu, "A Framework for Feature-Based Indexing for Spatial Databases," *Proc. Seventh Int'l Working Conf. Scientific and Statistical Database Management,* pp. 259-269, 1994.

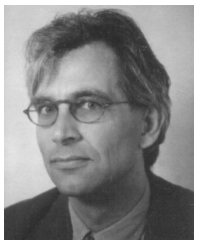
[197] P.C. Yuen, G.C. Feng, and D.Q. Tai, "Human Face Image Retrieval System for Large Database," *Proc. 14th Int'l Conf. Pattern Recognition,* vol. 2, pp. 1,585-1,588, 1998.

[198] Q.L. Zhang, S.K. Chang, and S.S.T. Yau, "A Unified Approach to Iconic Indexing, Retrieval and Maintenance of Spatial Relationships in Image Databases," *J. Visual Comm. and Image Representation,* vol. 7, no. 4, pp. 307-324, 1996.

[199] R. Zhao and W. Grosky, "From Features to Semantics: Some Preliminary Results," *Proc. Int'l Conf. Multimedia and Expo,* 2000.

[200] Y. Zhong, K. Karu, and A.K. Jain, "Locating Text in Complex Color Images," *Pattern Recognition,* vol. 28, no. 10, pp. 1,523-1,535, 1995.

[201] P. Zhu and P.M. Chirlian, "On Critical Point Detection of Digital Shapes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 8, pp. 737-748, Aug. 1995.



Arnold W.M. Smeulders has been in image processing since 1975 when he received the MSc degree in physics. He received the PhD degree in medical image analysis in 1982. He is a full professor of multimedia and the director of the Informatics Institute of the University of Amsterdam. He heads the 25 person Intelligent Sensory Information Systems group (ISIS) working on the theory of computer vision, image retrieval, and industrial vision. He has published extensively on vision and recognition. His current topics are: image retrieval, color, intelligent interaction, the role of engineering in vision, and the relation between language and vision. He is member of the board of the IAPR and cochair of TC12 on Multimedia. Previously, he served as an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and currently serves as an associate editor of *Cytometry* and *BioImaging*. He is a senior member of the IEEE.

Marcel Worring received a degree in computer science (honors) from the Free University of Amsterdam. His PhD was on digital image analysis and was obtained from the University of Amsterdam. He is an assistant professor in the Intelligent Sensory Information Systems group of the University of Amsterdam. His current interests are in multimedia information analysis, in particular, user interaction and browsing, video, and document analysis. He has been a visiting researcher in the Department of Diagnostic Imaging at Yale University (1992) and at the Visual Computing Lab at the University of California, San Diego (1998).

Simone Santini (M '98) received the Laurea degree from the University of Florence, Italy, in 1990, the MSc and the PhD degrees from the University of California, San Diego (UCSD) in 1996 and 1998, respectively. In 1990, he was a visiting scientist at the Artificial Intelligence Laboratory at the University of Michigan, Ann Arbor, and, in 1993, he was a visiting scientist at the IBM Almaden Research Center. Currently, he is a project scientist in the Department of Electrical and Computer Engineering, UCSD, and a researcher at Praja, Inc. His current research interests are interactive image and video databases, behavior identification and event detection in multisensor stream, query languages for event-based multimedia databases, and evaluation of interactive database systems. He is a member of the IEEE.

Amarnath Gupta received the BTech degree in mechanical engineering from the Indian Institute of Technology, Kharagpur, in 1984, the MS degree in biomedical engineering from the University of Texas at Arlington in 1987, and the PhD degree (engineering) in computer science from Jadavpur University, India, in 1994. He is an assistant research scientist at the Center for Computational Science and Engineering, University of California, San Diego (UCSD). Before joining UCSD, he was a scientist at Virage, Inc., where he worked on multimedia information systems. His current research interests are in multimedia and spatiotemporal information systems, heterogeneous information integration, and scientific databases. He is a member of the IEEE.

Ramesh Jain (F '92) received the BE degree from Nagpur University in 1969 and the PhD degree from the Indian Institute of Technology, Kharagpur in 1975. He was a professor of electrical and computer engineering, and computer science and engineering at the University of California, San Diego. Before joining UCSD, he was a professor of electrical engineering and computer science, and the founding Director of the Artificial Intelligence Laboratory at the University of Michigan, Ann Arbor. He also has been affiliated with Stanford University, IBM Almaden Research Labs, General Motors Research Labs, Wayne State University, the University of Texas at Austin, the University of Hamburg, West Germany, and the Indian Institute of Technology, Kharagpur, India. His current research interests are in multimedia information systems, image databases, machine vision, and intelligent systems. He is a fellow of the IEEE, AAAI, and Society of Photo-Optical Instrumentation Engineers, and a member of ACM, Pattern Recognition Society, Manufacturing Engineers. He has been involved in the organization of several professional conferences and workshops, and served on the editorial boards of many journals. He was the editor-in-chief of *IEEE Multimedia* and is on the editorial boards of *Machine Vision and Applications*, *Pattern Recognition*, and *Image and Vision Computing*.