

11

Semantic Processing of Color Images

**Stamatia Dasiopoulou, Evangelos Spyrou, Yiannis Kompatsiaris,
Yannis Avrithis, and Michael G. Strintzis**

CONTENTS

11.1	Introduction.....	259
11.2	State of the Art	262
11.3	Knowledge-Assisted Analysis	264
11.4	Knowledge-Assisted Analysis Using MPEG-7 and Semantic Web Technologies	267
11.4.1	Overview of MPEG-7 Visual Descriptors	267
11.4.1.1	Color Descriptors	267
11.4.1.2	Texture Descriptors	268
11.4.1.3	Shape Descriptors	269
11.4.1.4	Descriptor Matching	269
11.4.2	Ontology Structure.....	271
11.4.2.1	Core Ontology	271
11.4.2.2	Visual Descriptor Ontology	272
11.4.2.3	Multimedia Structure Ontology.....	273
11.4.2.4	Domain Ontologies	273
11.4.3	Domain Ontologies Population.....	273
11.4.4	Semantic Multimedia Analysis	274
11.4.4.1	Image Representation	274
11.4.4.2	Image Segmentation	275
11.4.4.3	Low-Level Visual Descriptor Extraction.....	275
11.4.4.4	Spatial Relations Extraction.....	275
11.4.4.5	Descriptors Matching	275
11.4.4.6	Spatial Context Consistency Check.....	276
11.4.4.7	Knowledge-Base Retrieval.....	276
11.4.4.8	Semantic Metadata Creation	277
11.4.5	Results.....	277
11.5	Conclusions and Future Work	280
	Acknowledgments	281
	References.....	281

11.1 Introduction

Image understanding continues to be one of the most exciting and fastest-growing research areas in the field of computer vision. The recent advances in hardware and telecommunication technologies, in combination with the Web proliferation witnessed, have boosted

the wide-scale creation and dissemination of digital visual content. However, this rate of growth has not been matched by concurrent emergence of technologies to support efficient image analysis and retrieval. As a result, this ever-increasing flow of available visual content has resulted in the overwhelming of users with large volumes of information, thus hindering access to appropriate content. Moreover, the number of diverse, recently emerging application areas, which rely increasingly on image understanding systems, has further revealed the tremendous potential of the effective use of visual content through semantic analysis. Better access to image databases, enhanced surveillance and authentication support, content filtering, adaptation and transcoding services, summarization, and improved human and computer interaction, are among the several application fields that can benefit from semantic image analysis.

Acknowledging the need for providing image analysis at the semantic level, research efforts focus on the automatic extraction of image descriptions in a way that matches human perception. The ultimate goal characterizing such efforts is to bridge the so-called *semantic gap* between low-level visual features that can be automatically extracted from the visual content, and the high-level concepts capturing the meaning conveyed. The emerged approaches fall into two categories — data-driven and knowledge-driven — depending on the direction the creation processes of these high-level descriptions. The former adhere to the monolithic computational paradigm, in which the interpretation and retrieval to follow are based on some appropriately defined function computed directly from the data. No hierarchy of meaningful intermediate interpretations is created. By contrast, the latter follow the signals to symbol paradigm, in which intermediate levels of description are emphasized. They are based on the widely held belief that computational vision cannot proceed in a single step from signal-domain information to spatial and semantic understanding.

Data-driven approaches work on the basis of extracting low-level features and deriving the corresponding higher-level content representations without any prior knowledge apart from the developer's inherent one. Thus, these approaches concentrate on acquiring fully automated numeric descriptors from objective visual content properties, and the subsequent retrieval based on criteria that somehow replicate the human perception of visual similarity. The major weakness of such approaches is that they fail to interact meaningfully with the users' higher level of cognition, because the built-in associations between image semantics and its low-level quantitative descriptions are of no perceptual meaning to the users. Consequently, the underpinning linking mechanism remains a "black box" to the user, thus not allowing for efficient access or, more importantly, for the discovery of semantically related content. Systems based on the query-by-example paradigm, as well as traditional content-based image retrieval systems, are well-known application examples belonging in this category. Although they are efficient for restricted domains, such approaches lack the capability to adapt to different domains. Techniques like relevance feedback and incremental learning have been used to improve traditional content-based approaches by injecting some knowledge on user perception in the analysis and similarity matching process.

Knowledge-driven approaches, on the other hand, utilize high-level domain knowledge to extract appropriate content descriptions by guiding features extraction, analysis and elimination of the unimportant ones, descriptions derivation, and reasoning. These approaches form an interdisciplinary research area, trying to combine and benefit from the computer vision, signal processing, artificial intelligence, and knowledge management communities joined efforts for achieving automatic extraction of visual content semantics through the application of knowledge and intelligence. More specifically, the task of such image analysis approaches is to abstract users' visual content experience by means of computational models (i.e., reduce the volumes of multimodal data to concise representations that capture the essence of the data). Enabling intelligent processing of visual content

requires appropriate sensors, formal frameworks for knowledge representation, and inference support. The relevant literature considers two types of approaches, depending on the knowledge acquisition and representation process — explicit, realized by formal model definitions, or implicit, realized by machine learning methods.

The main characteristic of learning-based approaches is the capability to adjust their internal structure according to input and respective desired output data pairs in order to approximate the relations (rules) that are implicit in the provided training data, thus elegantly simulating a reasoning process. Consequently, the use of machine learning techniques to bridge the semantic gap between image features and high-level semantic annotations provides a relatively powerful method for discovering complex and hidden relationships or mappings, and a significant number of approaches have been proposed for a variety of applications, as presented in the next section. As illustrated in the following, neural networks, fuzzy systems, support vector machines, statistical models, and case-based reasoning are among the techniques that have been widely used in the area of object recognition and scene classification. However, the often employed “black box” method may be difficult to develop and maintain, as its effectiveness relies upon the design and configuration of multiple variables and options. In addition, extensive and detailed training data sets are required to ensure optimum tuning and performance. The main disadvantage of machine-learning-based image analysis systems is that they are built specifically for one particular domain and cannot be easily adapted to others or simply extended with further features for application on the same domain.

Following an alternative methodology, model-based image analysis approaches make use of prior knowledge in the form of explicitly defined models and rules/constraints. Such approaches attempt to bridge the gap between low-level descriptions and high-level interpretations by encompassing a hierarchical representation of objects, events, relations, attributes, and so forth, of the examined domain. Thus, because the terms of the employed language (ontologies, semantic nets, etc.) carry meaning that is directly related to the visual entities, they provide a coherent semantic domain model, required to support “visual” inference in the context specified by the current set of logical statements. However, the computational complexity of such systems increases exponentially with the number of objects of interest, restricting the applicability of such approaches to settings where only a small number of parts are to be identified within a scene. Although appropriate, parameterized or generic models can be employed for improving performance. Such systems are computationally infeasible for complex objects, because in such cases, the search space can become too large. As a result, in most model-based approaches, objects are first detected without prior usage of such models, and recognition takes place afterwards based on contextual knowledge and fusion of the extracted facts. Controlling the variability of the scene is still a necessary condition for keeping the problem tractable.

It is worth noticing that although there is no consensus on which of these two classes of approaches is superior to the other, studies have revealed that human perception organization includes some kind of preattentive stage of visual processing. During this stage, different image events are detected, which are joined into complex objects at a second stage. Treisman [1] hypothesized that the visual system starts with extracting a set of useful properties, and then a map of locations is formed in which the presence of discontinuities is registered. By focusing attention on this map, object hypotheses are created which are matched against stored object descriptions, for their recognition. In the latter stage, prior knowledge and expectations play an important role. Treisman further hypothesized that the preattentive visual system does not produce a single representation such as a single partitioned image. Rather, different image partitions are produced to support distinct channels in the human visual system which analyze the image along a number of different dimensions (such as color, depth, motion, etc.).

To conclude, semantic understanding of visual content is the final frontier in image retrieval. The difficulty lies in bridging the gap between low-level visual features and representations that can be automatically computed from visual content, and its associated high-level semantics as perceived by humans. In this chapter, semantic image analysis for the purpose of automatic image understanding and efficient visual content access and retrieval at the semantic level is discussed. The overview presented in Section 11.2 surveys current state-of-the-art analysis approaches aimed at bridging the “semantic gap” in image analysis and retrieval. It highlights the major achievements of the existing approaches and sheds light on the challenges still unsolved. Section 11.3 presents a generic framework for performing knowledge-assisted semantic analysis of images. Knowledge representation and modeling, content processing, and inferencing support aspects are detailed, providing further insight into requirement and specification issues for realizing automatic semantic descriptions generation from visual content. Section 11.4 begins with a brief overview of the MPEG-7 standardized descriptors used within the presented framework and a few methods used for matching followed by the ontology infrastructure developed. It also presents the way the knowledge-assisted analysis is performed, using Semantic Web technologies. Finally, conclusions are drawn in Section 11.5, and plans for future work are presented.

11.2 State of the Art

Enabling efficient access to still images based on the underlying semantics presents many challenges. Image understanding includes the extraction of global scene semantics and the recognition of the perceptual entities depicted. The former may refer to general annotations, such as indoors/outdoors and city/landscape classifications, while the latter considers finer-grained descriptions addressing the presence of particular semantic entities, objects, or events (e.g., sunset, beach, airplane, etc.). The visual content different information modalities in combination with the inherent uncertainty render the extraction of image semantics impossible without the use of considerable amounts of *a priori* knowledge. As illustrated in the following reviewed literature, numerous standardized and proprietary low-level feature descriptors have been applied to capture the information conveyed by the different modalities characterizing visual information, color, texture, and shape. Diverse approaches that consider knowledge representation, management, and inference realization have also been followed. Neural networks, expert systems, fuzzy logic, ontologies, decision trees, static and dynamic Bayesian networks, factor graphs, Markov random fields, and so forth, are among the popular mechanisms for storing and encoding high-level information.

Stochastic approaches include, among others, the work presented in Reference [2], where the problem of bridging the gap between low-level representation and high-level semantics is formulated as a probabilistic pattern recognition problem. A factor graph network of probabilistic multimedia objects, multijets, is defined in a probabilistic pattern recognition fashion using hidden Markov models and Gaussian mixture models. HMMs are combined with rules in the COBRA model described in Reference [3], where objects and events descriptions are formalized through appropriate grammars, and at the same time, the stochastic aspect provides the means to support visual structures that are too complex to be explicitly defined. A hierarchical model based on *Markov random fields* (MRF) was used in Reference [4] for implementing unsupervised image classification.

Histogram-based image classification is performed using a support vector machine (SVM) in Reference [5], while an object support vector machines' classifier that is trained once on a small set of labeled examples is presented in Reference [6]. An SVM is applied to represent conditioned feature vector distributions within each semantic class, and a Markov random

Au: Spell out
HMM.

Au: Spell out
COBRA.

field is used to model the spatial distributions of the semantic labels, for achieving semantic labelling of image regions in Reference [7]. To address cases in which more than one label fit the image data, Li, Wang, and Sung [8] propose a multilabel SVM active learning approach to address multilabel image classification problems.

In Reference [9], machine-learning techniques are used to semantically annotate images with semantic descriptions defined within ontologies, while in Reference [10], the use of the maximum entropy approach is proposed for the task of automatic image annotation. In Reference [11], a methodology for the detection of objects belonging to predefined semantic classes is presented. Semantic classes are defined in the form of a description graph, including perceptual and structural knowledge about the corresponding class objects, and are further semantically organized under a binary partition tree. Another nice example of a domain-driven semiautomated algorithm for semantic annotation is given in Reference [12], where a specific animal face tracker is formed from user-labeled examples utilizing an Ada-boost classifier and a Kanade–Lucas–Tomasi tracker. The semiautomatic image annotation system proposed in Reference [13] uses hints given in natural language to prune the search space of object detection algorithms. The user can give hints like “in the upper left corner there is a L-shaped building.” The system uses spatial constraints to reduce the area in which to search for an object, and other constraints to reduce the number of possible shapes or object types, supporting even complex queries describing several objects and their configurations.

Fuzziness is introduced in Reference [14], where an intelligent system using neuro-fuzzy networks is used to locate human faces within images. An object-oriented high-resolution image classification based on fuzzy rules is described in Reference [15]. Domain experts define domain-specific rules through a graphical interface, and the system using these rules can automatically generate semantic annotations for any image of the given domain. A rule-based fuzzy inference approach is also followed in Reference [16] for classifying building images. Knowledge representation is based on a fuzzy reasoning model in order to establish a bridge between visual primitives and their interpretations. A trainable system for locating clothed people in photographic images is presented in Reference [17]. Within this system, a tree is constructed, with nodes that represent potentially segmentable human parts, while the edges represent distributions over the configurations of those parts. This classifier adapts automatically to an arbitrary scene by learning to use context features. A context-aware framework for the task of image interpretation is also described in Reference [18], where constraints on the image are generated by a natural language processing module performing on the text accompanying the image.

A method for classifying images based on knowledge discovered from annotated images using WordNet is described in Reference [19]. Automatic class discovery and classifier combination are performed using the extracted knowledge (i.e., the network of concepts with the associated image and text examples). This approach of automatically extracting semantic image annotation by relating words to images was reported in a number of other research efforts, such as in Reference [20] using latent semantics analysis, [21], [22], and so forth.

Following the recent Semantic Web advances, several approaches have emerged that use ontologies as the means to represent the necessary-for-the-analysis tasks domain knowledge, and take advantage of the explicit semantics representation for performing high-level inference. In Reference [23], an ontology-based cognitive vision platform for the automatic recognition of natural complex objects is presented. Three distributed knowledge-based systems drive the image processing, the mapping of numerical data into symbolical data and the semantic interpretation process. A similar approach is taken in the FUSION project [24], where ontology-based semantic descriptions of images are generated based on appropriately defined RuleML rules that associate MPEG-7 low-level features to the concepts

included in the FUSION ontology. Also enhanced by rules is the user-assisted approach for automatic image annotation reported in Reference [25], while fuzzy algebra and fuzzy ontological information are exploited in Reference [26] for extracting semantic information in the form of thematic categorization. Ontology-based image classification systems are also presented in References [27] and [28]. In Reference [29], the problem of injecting semantics into visual data is addressed by introducing a data model based on description logics for describing both the form and the content of such documents, thus allowing queries on both structural and conceptual similarity.

Medical image understanding is another application field in which semantic image analysis has received particularly strong interest. Medical images interpretation is mainly required for diagnosis purposes in order to reduce repetitive work, and for providing assistance in difficult diagnoses or unfamiliar cases. Thus, the automatic acquisition of accurate interpretation is a strict requirement, and in addition, the efficient management of the huge volumes of information concentrated in medical image databases is vital. The approaches reported in the literature cover a wide variety of medical imaging cases such as tomography, mammography, ophthalmology, radiology, and so forth. Computer tomography images are analyzed in Reference [30] using two case-based reasoners, one for segment identification and another for a more holistic interpretation of the image. The system STARE, presented in Reference [31], is a management system for medical images that supports, among others, automated retinal diagnosis using Bayesian networks to realize an inference mechanism. KBIUS [32] is another knowledge-assisted rule-based image understanding system that supports x-ray bone images segmentation and interpretation.

Despite the sustained efforts in the last years, state of the art for semantic image understanding still cannot meet users' expectations for systems capable of performing analysis at the same level of complexity and semantics that a human would employ while analyzing the same content. Although a significant number of approaches with satisfactory results were reported, semantic image understanding remains an unsolved problem, because most state-of-the-art techniques make no attempt to investigate generic strategies for incorporating domain knowledge and contextual information, but rather rely on ad hoc, application-targeted, or hard-coded models, rules, and constraints [33]. Consequently, due to the unrestricted potential content and the lack of temporal context that would assist in the recognition of perceptual entities, the presented technical challenges render semantic image analysis a fascinating research area awaiting new advances.

Furthermore, recent studies revealed that apart from the need to provide semantic-enabled image access and management, the inherent dynamic interpretation of images under different circumstances should be taken into consideration in future efforts [34]. Perceptual similarity depends upon the application, the person, and the context of usage. Thus, machines not only need to learn the visual content and underlying meaning associations but also have to learn them online while interacting with users. Finally, in order for image understanding to mature, understanding how to evaluate and define appropriate frameworks for benchmark features, methods, and systems is of paramount importance.

11.3 Knowledge-Assisted Analysis

Building on the considerations resulting from the presented state of the art on semantic image analysis, this section presents a generic framework for performing semantics extraction from images based on explicitly defined *a priori* knowledge. The proposed semantic analysis framework does not consider global semantics extraction, but rather focuses on the

Au: Spell out
STARE.

recognition of salient perceptual entities at object level (e.g., sea and sand in a beach image or the presence of a mountain in an image depicting an outdoors scene). The extraction of higher-level semantic concepts is performed based on the available domain knowledge and appropriately defined rules that model the context on which such concepts occur (e.g., an image of a player scoring a goal presupposes a particular spatial arrangement of the ball, the goalpost, etc.).

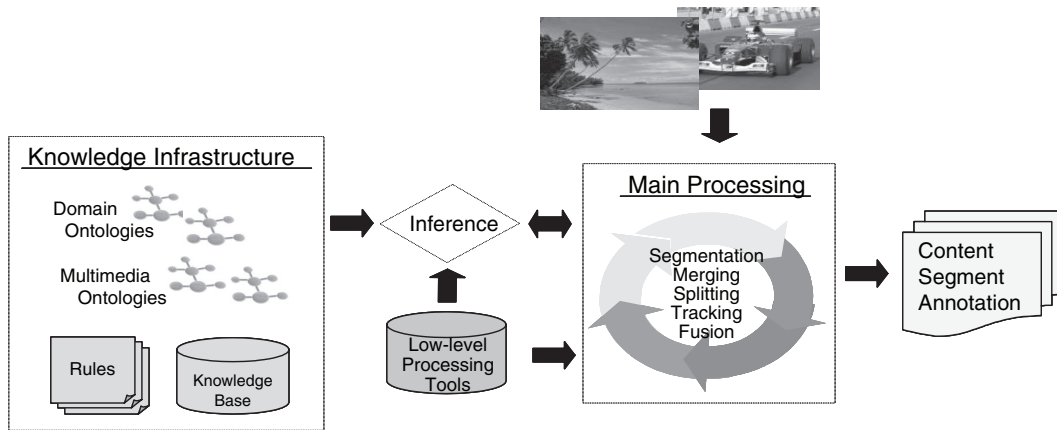
Before proceeding with the detailed description of the proposed analysis framework, the process of image analysis and understanding is briefly overviewed to better highlight the challenges and open issues involved and, thus, demonstrate how the proposed framework provides the means to address them.

The goal of knowledge-based semantic image analysis is to extract semantic descriptions from low-level image representations based on explicit prior knowledge about the domain examined. Such domain knowledge includes prototypical descriptions of the important domain concepts in terms of their visual properties and context of appearance, and thus allows for their identification. For this reason, visual descriptions of the image data need to be extracted and matched against the corresponding definitions included in the available domain knowledge. The resulted set of hypotheses (i.e., the set of semantic concepts possibly associated with each region) is further processed to determine the plausibility of each hypothesis and thereby to decide upon the final semantic labeling.

Consequently, partitioning the image into a set of meaningful regions is the prerequisite before any analysis can take place, because the analysis to follow is based on the visual features extracted from these regions. However, partitioning of an image into meaningful regions is a very challenging task [35]. The sensory data is inherently noisy and ambiguous, and the available segmentation approaches perform on a purely numerical basis, thus leading to segmentations that are unreliable and that vary in an uncontrollable way (i.e., regions may result as fragmented or falsely merged). In addition, the various domain objects can be characterized by diverse visual properties requiring more than one image partitioning scheme in order to capture them. For example, objects with indicative shape properties require shape-driven segmentation approaches, while texturized objects need segmentations based on, possibly different per object, texture descriptors.

From the above mentioned, it is discovered that semantic image analysis has to deal with multiple low-level representations based on the different modalities of visual information, overcome the syntactic nature of existing segmentation approaches, and exploit domain knowledge to control the complexity of the semantics extraction decision-making process. To assist in these extremely challenging tasks, the framework presented adopts a formal knowledge representation to ensure consistent inferencing, and exploits the knowledge available within each stage of the analysis process.

The main knowledge structures and functional modules of the proposed generic semantic analysis framework, as well as their interactions, are shown in Figure 11.1. As illustrated, ontologies are used for representing the required knowledge components. This choice is justified by the recent Semantic Web technologies advances and the consequent impacts on knowledge sharing and reuse. Several ontology languages that provide support for expressing rich semantics were developed, providing the formal definition framework required for making these semantics explicit [36]. Furthermore, ontology alignment, merging, and modularization are receiving intense research interest, leading to methodologies that further establish and justify the use of ontologies as knowledge representation formalism. In addition, tools for providing inference support have emerged that allow for reasoning about the existing facts and deriving knowledge that was previously implicit. If image content is to be fully exploited by search engines, services, and application agents within the Semantic Web context, semantic analysis should target the generation of annotations that meet the currently formulated semantics description standards. A detailed description of

**FIGURE 11.1**

Knowledge-assisted semantic image analysis framework.

each of the presented framework components and its respective role and contribution in the semantic analysis process are described in the sequel.

Due to the two-layer semantics of visual content (i.e., the semantics of the actual conveyed meaning and the semantics referring to the media, different kinds of ontologies are involved in the analysis process. More specifically, domain ontologies are used to model the conveyed content semantics with respect to specific real-world domains. They are defined in such a way as to provide a general model of the domain, with focus on the user-specific point of view. Consequently, the domain ontology includes those concepts that are of importance for the examined domain (i.e., the salient domain objects and events) and their interrelations. In addition, domain ontologies include qualitative and quantitative attributes of the defined concepts in order to support the various analysis tasks. Thus, the domain ontologies model the examined domain in a way that, on the one hand, makes the retrieval of images more efficient for end users and, on the other hand, the defined concepts can also be automatically extracted through image analysis. In other words, the concepts are recognizable by the automatic analysis methods, while they remain comprehensible to humans. Populating the domain ontologies results in enriching the knowledge base with the appropriate models (i.e., prototypical visual and spatial descriptions) of the domain concepts that need to be detected.

On the other hand, media analysis ontologies model the actual analysis process. They include knowledge specific to the media type, descriptive definitions for representing low-level visual features and attributes related to spatial topology, and in addition, the low-level processing algorithms definitions. By building this unifying model of all aspects of image analysis, all related parts can be treated as ontological concepts, thus supporting interoperability and reusability of the presented analysis framework. In addition, by associating the content-processing tools with visual properties, the analysis process gets decoupled from application-specific requirements and can be easily adapted to other domains.

To determine how the extraction of semantic concepts, the respective low-level features, and the processing algorithms execution order relate to each other, appropriate rules need to be defined. As a result, sufficiently expressive languages need to be employed for defining such rules and for allowing reasoning on top of the knowledge defined in the domain and the media analysis ontologies. Apart from the need for an inference engine, appropriate ontological knowledge management tools need to be investigated so that efficient and effective access and retrieval of the involved knowledge is ensured. This leads to ontological repositories and corresponding query languages issues.

To conclude, an ontology-based framework for knowledge-assisted domain-specific semantic image analysis was presented. The employed knowledge involves qualitative object attributes, quantitative low-level features generated by training, as well as low-level processing methods. Rules are used to describe how tools for image analysis should be applied, depending on object attributes and low-level features, for the detection of objects corresponding to the semantic concepts defined in the ontology. The added value comes from the coherent architecture achieved by using an ontology to describe both the analysis process and the domain of the examined visual content. Following this approach, the semantic image analysis process depends largely on the knowledge base of the system, and, as a result, the method can be easily applied to different domains provided that the knowledge base is enriched with the respective domain knowledge. In the following section, a specific implementation of a knowledge-assisted semantic image analysis system based on the MPEG-7 standard and the recently emerged Semantic Web technologies is presented.

11.4 Knowledge-Assisted Analysis Using MPEG-7 and Semantic Web Technologies

11.4.1 Overview of MPEG-7 Visual Descriptors

The goal of the ISO/IEC MPEG-7 standard [37] is to allow interoperable searching, indexing, filtering, and browsing of audiovisual (AV) content. Unlike its predecessors, the focus was on nontextual description of multimedia content, aiming to provide interoperability among applications that use audiovisual content descriptions.

In order to describe this AV content, the MPEG-7 standard specifies a set of various color, texture, shape, and motion standardized descriptors that extract visual, low-level, nonsemantic information from images and videos and use it to create structural and detailed descriptions of AV information. A descriptor defines the syntax and the semantics of an elementary AV feature, which may be low level (e.g., color) or high level (e.g., author). For tasks like image classification or object recognition, visual MPEG-7 descriptors [38] are considered. A brief overview of the most important MPEG-7 visual descriptors that are applicable to still color images follows.

11.4.1.1 Color Descriptors

Color is probably the most expressive of all the visual features. Thus, it has been extensively studied in the area of image retrieval during the last years. Apart from that, color features are robust to viewing angle, translation, and rotation of the regions of interest. The MPEG-7 color descriptors [39] comprise histogram descriptors, a dominant color descriptor, and a color layout descriptor (CLD). The presentation of the color descriptors begins with a description of the color spaces used in MPEG-7:

- *Color Space Descriptor* is introduced, as each color descriptor uses a certain color space, and therefore, a short description of the most widely used color spaces is essential. The color spaces supported are the monochrome, (red, green, blue) RGB, HSV, YCbCr, and the new HMMD [39]. These color space descriptors are also used outside of the visual descriptors (i.e., in specifying “media properties” in suitable description schemes).
- *Color layout descriptor* (CLD) is a compact MPEG-7 visual descriptor designed to represent the spatial distribution of color in the YCbCr color space. It can be used globally in an image or in an arbitrary-shaped region of interest. The given picture

or region of interest is divided into $8 \times 8 = 64$ blocks, and the average color of each block is calculated as its representative color. A discrete cosine transformation is performed in the series of the average colors, and a few low-frequency coefficients are selected using zigzag scanning. The CLD is formed after quantization of the remaining coefficients, as described in Reference [40].

- *Scalable color descriptor* (SCD) is a Haar-transform-based encoding scheme that measures color distribution over an entire image. The color space used is the HSV, quantized uniformly to 256 bins. To sufficiently reduce the large size of this representation, the histograms are encoded using a Haar transform, allowing also the desired scalability.
- *Color-Structure descriptor* (CSD) captures both the global color features of an image and the local spatial structure of the color. The latter feature of the CSD provides the descriptor the ability to discriminate between images that have the same global color features but different structure; thus, a single global color histogram would fail. An 8×8 structuring element scans the image, and the number of times a certain color is found within it is counted. This way, the local color structure of an image is expressed in the form of a “color structure histogram.” This histogram is identical in form to a color histogram but is semantically different. The color representation is given in the HMMD color space. The CSD is defined using four-color space quantization operating points (184, 120, 64, and 32 bins) to allow scalability, while the size of the structuring element is kept fixed.

11.4.1.2 Texture Descriptors

Texture refers to the visual patterns that may or may not have properties of homogeneity, which results from the presence of multiple colors or intensities in the image, is a property of virtually any surface, and contains important structural information of surfaces and their relationship to the surrounding environment. Describing textures in images by appropriate MPEG-7 texture descriptors [39] provides a powerful means for similarity matching and retrieval for both homogeneous and nonhomogeneous textures. The three texture descriptors, standardized by MPEG-7, are texture-browsing descriptor, homogeneous texture descriptor, and local edge histogram:

- *Texture-browsing descriptor* provides a qualitative characterization of a texture’s regularity, directionality, and coarseness. The regularity of a texture is described by an integer ranging from 0 to 3, where 0 stands for an irregular/random texture, and 3 stands for a periodic pattern. Up to two dominant directions may be defined, and their values range from 0° to 150° in steps of 30° . Finally, coarseness is related to image scale or resolution and is quantized to four levels, with the value 0 indicating a fine-grain texture and the value 3 indicating a coarse texture.
- *Homogeneous texture descriptor* (HTD) provides a quantitative characterization of texture and is an easy-to-compute and robust descriptor. The image is first filtered with orientation and scale-sensitive filters. The mean and standard deviations of the filtered outputs are computed in the frequency domain. The frequency space is divided in 30 channels, as described in Reference [40], and the energy and the energy deviation of each channel are computed and logarithmically scaled.
- *Local edge histogram* captures the spatial distribution of edges and represents local-edge distribution in the image. Specifically, dividing the image in 4×4 subimages, the local edge distribution for each subimage can be represented by a histogram. To generate the histogram, edges in the subimages are categorized

into five types: vertical, horizontal, 45° diagonal, 135° diagonal, and nondirectional edges. Because there are 16 subimages, a total of $5 \times 16 = 80$ histogram bins are required. This descriptor is useful for image-to-image matching, even when the underlying texture is not homogeneous.

11.4.1.3 Shape Descriptors

Humans can often recognize objects solely from their shapes, as long as they have a characteristic one. This is a unique feature of the shape descriptors, which discriminates them from color and texture. Thus, shape usually contains semantic information for an object. It is obvious that the shape of an object may be a very expressive feature when used for similarity search and retrieval. MPEG-7 proposes three shape descriptors [41]: region-based shape descriptor, contour-based shape descriptor, and two-dimensional (2-D)/three-dimensional (3-D) shape descriptor.

- *Region-based shape descriptor* expresses the 2-D pixel distribution within an object or a region of interest. It is based both on the contour pixel and the inner pixels of the object or region of interest; therefore, it is able to describe complex objects as well as simple objects with or without holes. The shape analysis technique used is based on moments, and a complex 2-D angular radial transformation (ART) is applied. Then the descriptor is constituted by the quantized magnitudes of the ART coefficients. In conclusion, the region-based shape descriptor gives a compact, efficient, and robust way to describe both complex and simple objects.
- *Contour-based shape descriptor* captures the characteristic features of the contours of the objects. It is based on an extension of the curvature scale-space (CSS) representation of the contour and can effectively describe objects with contours that are characteristic; therefore, the region-based shape descriptor is redundant. Apart from that, it can discriminate objects with regions that are similar but have different contours. This descriptor emulates the shape similarity perception of the human eye system and provides a compact and robust to nonrigid deformations and perspective transformations description of objects of region of interest. The descriptor's size adjusts to the contour complexity.
- *2-D/3-D shape descriptor* combines 2-D descriptors of a visual feature of an object or region of interest, seen from various different angles, thus forming an entire 3-D representation of it. Experiments have shown that a combination of contour-based shape descriptors of a 3-D object is an effective way to obtain a multiview description of it.

Au: Sentence
Ok?

11.4.1.4 Descriptor Matching

As described in Section 11.4, knowledge-assisted analysis approaches exploit *a priori* knowledge about the domain under consideration to perform semantic analysis tasks, such as object recognition and image classification. As detailed above, the provided knowledge includes information about the domain conceptualization, the image in terms of its structure, and the modeled domain concepts in the form of visual descriptors and spatial characteristics definitions. Among the possible representations, the information considering low-level visual features is often encoded using low-level descriptors similar to those that are proposed by the MPEG-7 standard. It is obvious that a key factor in such tasks is the selected measures used for the estimation of the distance between the descriptors. When the descriptors to be considered are MPEG-7 standardized, there are certain measures to

evaluate their similarities, which, in some cases, are explicit. This subsection presents a few similarity measures for some of the above described descriptors as they were defined by MPEG-7.

For example, matching with the *Dominant Color* descriptor can be performed in the following way: Let

$$\begin{aligned} F_1 &= \{ \{c_{1i}, p_{1i}, v_{1i}\}, s_1 \}, i = 1, \dots, N_1 \\ F_2 &= \{ \{c_{2i}, p_{2i}, v_{2i}\}, s_2 \}, i = 1, \dots, N_2 \end{aligned}$$

be two dominant color descriptors. Ignoring variances and spatial coherencies (which are optional), the dissimilarity between them may be defined as

$$D^2(F_1, F_2) = \sum_{i=1}^{N_1} P_1 i^2 + \sum_{j=1}^{N_2} P_2 j^2 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 2a_{1i,2j} p_{1i} p_{2j}$$

where a_{ij} is the similarity coefficient between two colors c_k and c_l , defined by

$$a_{k,l} = \begin{cases} 1 - d_{k,l}/d_{max} & \text{if } d_{k,l} \leq T_d \\ 0 & \text{if } d_{k,l} > T_d \end{cases}$$

and $d_{k,l}$ is the Euclidean distance between the two colors c_k and c_l , T_d is the maximum distance for two colors, and $d_{max} = a T_d$. More details about the determination of T_d and a , and also for modifications that can be made to take into account the variances and the spatial coherencies, can be found in Reference [42].

MPEG-7 does not strictly standardize the distance functions to be used and sometimes does not propose a dissimilarity function, leaving the developers with the flexibility to develop their own dissimilarity/distance functions. A few techniques can be found in the MPEG-7 eXperimentation Model (XM) [42]. Apart from that, there are many general-purpose distances that may be applied in order to simplify some complex distance function or even to improve the performance [43]. A large number of successful distance measures from different areas (statistics, psychology, medicine, social and economic sciences, etc.) can be applied on MPEG-7 data vectors.

However, in order to achieve better performance, combining more than one low-level descriptor seems essential. This problem remains open, and there are not any standardized methods to achieve it. Apart from that, fusion of the descriptors is necessary, as they would be otherwise incompatible and inappropriate to directly include, for example, in a Euclidean distance. A classic approach to combine the results of many descriptors is to normalize the distances between images according to the different descriptors, then add these distances to obtain a unique distance for each pair (additive fusion) [44]. A drawback of this additive fusion is that it computes the average of the distances (by summing them) and, therefore, risks neglecting the good performances of a given descriptor because of the poor performances of another. Merging fusion as in Reference [45] simply consists of merging all the descriptions into a unique vector. If D_1, D_2, \dots, D_n are the n descriptors to combine, then the merged descriptor is equal to

$$D_{merged} = [D_1 | D_2 | \dots | D_n]$$

This fusing method requires all features to have more or less the same numerical values to avoid scale effects. An alternative is to rescale the data using principal component analysis, for instance. Rescaling is not necessary in the case of the MPEG-7 descriptors because they are already scaled to integer values of equivalent magnitude. Assigning fixed weights as in Reference [46] can be an efficient method, especially when the number of the visual

features is small. The assignment of the weights can be done either experimentally, by simply observing the results and giving more weight to the descriptors that seem to have more discriminative power, or by using a statistical method as in Reference [47], where each feature is used separately, and the matching values assigned to the first two outputs of the system are added up. Then the average of this over the whole query set is found. The corresponding weight for each method is then inversely proportional to this average.

11.4.2 Ontology Structure

As noted in Section 11.4, among the possible knowledge representation formalisms, ontologies [36] present a number of advantages. They provide a formal framework for supporting explicit, machine-processable, semantics definitions, and they facilitate inference and the derivation of new knowledge based on rules and already existing knowledge. Thus, ontologies appear to be ideal for expressing multimedia content semantics in a formal machine-processable representation that will allow automatic analysis and further processing of the extracted semantic descriptions. Following these considerations, in the developed knowledge-assisted analysis framework, ontologies in a resource description framework (RDF) have been used as the means for representing the various knowledge components involved.

Resource Description Framework Schema (RDFS) is a simple modeling language on top of the RDF formalism¹, both being developed by the World Wide Web Consortium (W3C). *Web Ontology Language (OWL)*, a language inspired by description logics and also developed by the W3C, is designed to be used by applications that need increased expressive power compared to that supported by RDFS, by providing additional vocabulary along with formal semantics. In our framework, RDFS was chosen as the modeling language due to the fact that full usage of the increased expressiveness of OWL requires specialized and more advanced inference engines that are not yet widely available.

As shown in Figure 11.2, the developed knowledge infrastructure [48] consists of a *Core Ontology* whose role is to serve as a starting point for the construction of new ontologies, a *Visual Descriptor Ontology* that contains the representations of the MPEG-7 visual descriptors, a *Multimedia Structure Ontology* that models basic multimedia entities from the MPEG-7 Multimedia Description Scheme [49], and *Domain Ontologies* that model the content layer of multimedia content with respect to specific real-world domains.

11.4.2.1 Core Ontology

In general, core ontologies are typically conceptualizations that contain specifications of domain-independent concepts and relations based on formal principles derived from philosophy, mathematics, linguistics, and psychology. The role of the core ontology in this overall framework is to serve as a reference point for the construction of new ontologies, to provide a reference point for comparisons among different ontological approaches, and to serve as a bridge between existing ontologies. In the presented framework, the *DOLCE* [50] ontology is used for this purpose. DOLCE was explicitly designed as a core ontology, is minimal in the sense that it includes only the most reusable and widely applicable upper-level categories, rigorous in terms of axiomatization, and extensively researched and documented.

Although the DOLCE core ontology provides a means for representing spatiotemporal qualities, reasoning with such descriptions requires the coding of additional relations that

¹ RDF is not a knowledge representation system but tries to improve data interoperability on the Web. This is achieved by specializing the XML data model through a graph-based data model similar to the semantic networks formalism.

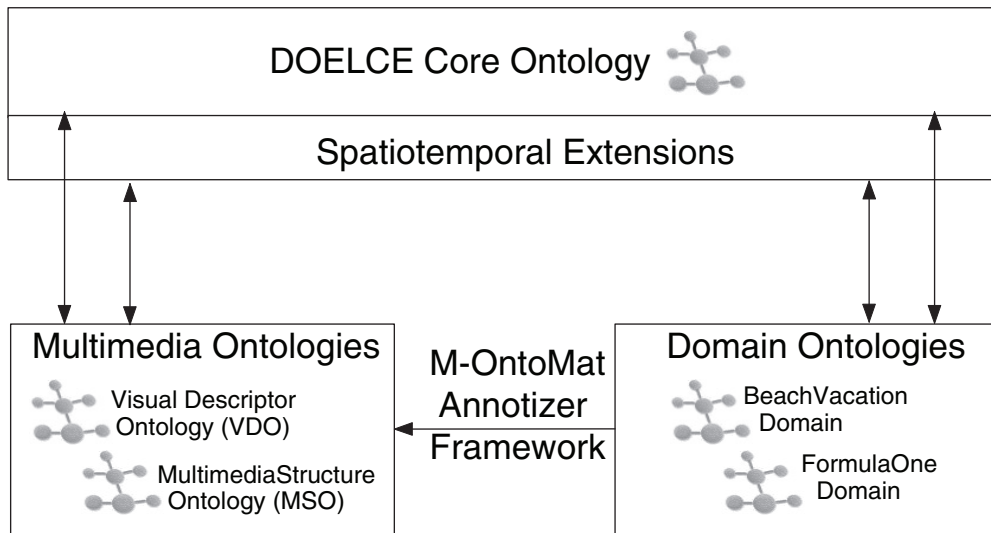


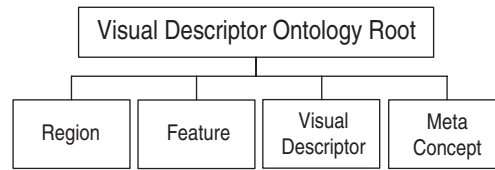
FIGURE 11.2
Ontology structure overview.

describe the relationship between space and time regions. Based on concepts taken from the “Region Connecting Calculus” [51], Allen’s interval calculus [52], and directional models [53], [54], the *Region* concept branch of DOLCE was extended to accommodate topological and directional relations between regions of different types, mainly *TimeRegion* and *2DRegion*. Directional spatial relations describe how visual segments are placed and relate to each other in 2-D or 3-D space (e.g., left and above). Topological spatial relations describe how the spatial boundaries of the segments relate (e.g., touch and overlap). In a similar way, temporal segment relations are used to represent temporal relationships among segments or events; the normative binary temporal relations correspond to Allen’s temporal interval relations.

11.4.2.2 Visual Descriptor Ontology

The visual descriptor ontology (VDO) [55] represents the visual part of the MPEG-7 and thus, contains the representations of the set of visual descriptors used for knowledge-assisted analysis. Its modeled concepts and properties describe the visual characteristics of the objects. The construction of the VDO attempted to follow the specifications of the MPEG-7 Visual Part [56]. Because strict attachment to the MPEG-7 Visual Part became impossible, several requisite modifications were made in order to adapt the XML schema provided by MPEG-7 to an ontology and the data-type representations available in RDFS.

The tree of the VDO consists of four main concepts, which are *VDO:Region*, *VDO:Feature*, *VDO:VisualDescriptor* and *VDO:Metaconcepts*, as illustrated in Figure 11.3. None of these concepts is included in the XML schema defined MPEG-7, but their need was vital in order to create a correctly defined ontology. The *VDO:VisualDescriptor* concept contains the visual descriptors, as these are defined by MPEG-7. The *VDO:Metaconcepts* concept, on the other hand, contains some additional concepts that were necessary for the VDO, but they are not clearly defined in the XML schema of MPEG-7. The remaining two concepts that were defined, *VDO:Region* and *VDO:Feature*, are also not included in the MPEG-7 specification, but their definition was necessary in order to enable the linking of visual descriptors to the actual image regions.

**FIGURE 11.3**

The visual descriptor ontology (VDO).

For example, consider the `VDO:VisualDescriptor` concept, which consists of six sub-concepts, one for each category of the MPEG-7-specified visual descriptors. These are *color*, *texture*, *shape*, *motion*, *localization*, and *basic descriptors*. Each of these subconcepts includes a number of relevant descriptors. These descriptors are defined as concepts in the VDO. Only the `VDO:BasicDescriptors` category was modified regarding the MPEG-7 standard and does not contain all the MPEG-7 descriptors.

11.4.2.3 Multimedia Structure Ontology

The multimedia structure ontology (MSO) models basic multimedia entities from the MPEG-7 Multimedia Description Scheme [49] and mutual relations like decomposition. Within MPEG-7, multimedia content is classified into five types: image, video, audio, audiovisual, and multimedia. Each of these types has its own segment subclasses. MPEG-7 provides a number of tools for describing the structure of multimedia content in time and space. The Segment DS [49] describes a spatial or temporal fragment of multimedia content. A number of specialized subclasses are derived from the generic Segment DS. These subclasses describe the specific types of multimedia segments, such as video segments, moving regions, still regions, and mosaics, which result from spatial, temporal, and spatiotemporal segmentation of the different multimedia content types. Multimedia resources can be segmented or decomposed into subsegments through four types of decomposition: spatial, temporal, spatiotemporal, and media source.

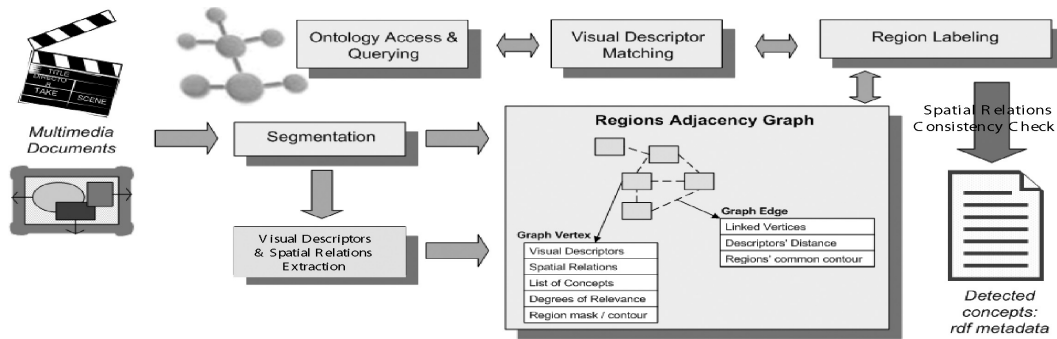
11.4.2.4 Domain Ontologies

In the presented framework, the domain ontologies model the content layer of multimedia content, with respect to specific real-world domains, such as Formula One or beach vacations. Because the DOLCE ontology was selected as the core ontology of the ontology infrastructure, it is essential that all the domain ontologies be explicitly based on or aligned to it, and thus connected by high-level concepts. This, in turn, assures interoperability between different domain ontologies.

In the context of this work, domain ontologies are defined in a way to provide a general model of the domain, with focus on the user's specific point of view. More specifically, ontology development was performed in a way that, on the one hand, the retrieval becomes more efficient for a user of a multimedia application, and on the other hand, the included concepts can also drive their automatic extraction from the multimedia layer. In other words, the defined semantic concepts are recognizable by automatic analysis methods while at the same time remaining comprehensible to users.

11.4.3 Domain Ontologies Population

In order to exploit the presented ontology infrastructure, the domain ontology should be populated with appropriate instances (i.e., visual descriptors and spatial relations of the defined domain objects), because, as described in Section 11.4, the produced semantic annotations are generated through matching against these objects' prototypes. To accomplish

**FIGURE 11.4**

The developed knowledge-assisted semantic image analysis system architecture.

this, the low-level descriptors that are included in the definition of each domain object need to be extracted for a sufficiently large number of corresponding object samples and be associated with the domain ontology. Within the described implementation, a user-oriented tool was developed. Users select regions that correspond to domain concepts and then choose the MPEG-7 Descriptor to be extracted. Triggered by the user's extraction command, the requested MPEG-7 Descriptors are extracted through calls to appropriate routines based on the MPEG-7 XM, a reference model utilizing the MPEG-7 visual descriptors [42].

11.4.4 Semantic Multimedia Analysis

The implemented semantic multimedia analysis architecture is presented in Figure 11.4. As illustrated, analysis starts by segmenting the input image content and extracting the low-level descriptors and the spatial relations in accordance with the domain ontology definitions. In the sequel, a first set of possible semantic concepts that might be depicted in each of the segmented regions is produced by querying the knowledge base and matching the previously extracted low-level descriptors with the ones of the objects prototype instances. To evaluate the plausibility of the produced hypotheses labels for each region and to reach the final semantic annotation, the object's spatial context information is used. Thereby, the image semantics are extracted, and respective content description metadata are generated. The implementation details of each of these processing steps are given in the following.

11.4.4.1 Image Representation

A region adjacency graph (RAG) has been selected as the means for the representation of the image: each vertex corresponds to a connected region of the image, while each edge represents the link between two regions. More specifically, each vertex of the graph holds the MPEG-7 visual descriptors of the image region it represents, the spatial relations between the region and its neighboring regions, and the degree of confidence to which this region matches a specific domain concept. Additionally, a list² of all the pixels that constitute the region and a list of all region's pixels that constitute its contour are also stored to improve performance. Finally, each edge of the graph stores the two linked regions, the distance of these regions estimated based on each visual descriptor, and a list of pixels that constitute the common contour of the two linked regions. Presently, the RAG is used only for efficient

² This list is more efficient than keeping the binary mask of the region, in terms of memory usage and time required for the analysis of an image.

representation and storage purposes and not in the actual process of the analysis (i.e., no graph matching takes place, but instead the subsequently described descriptor matching and spatial content consistency check are applied for generating semantic descriptions).

11.4.4.2 Image Segmentation

The first step should be to formulate a segmentation algorithm that will generate a few tens of connected regions and initialize the graph. The segmentation used is an extension of the well-known recursive shortest spanning tree (RSST) algorithm based on a new color model and so-called syntactic features [57].

11.4.4.3 Low-Level Visual Descriptor Extraction

The currently supported low-level descriptors, are the MPEG-7 Dominant Color and the Region Shape descriptors, and their extraction is based on the guidelines given by the MPEG-7 XM.

11.4.4.4 Spatial Relations Extraction

As previously mentioned, apart from low-level descriptions, it is necessary to include in the domain knowledge definitions information about an object's spatial context as well, because this is the only way to discriminate between objects with similar visual appearance. Objects such as *Sky* and *Sea* are among the simplest and most common examples where spatial context is required to lead to correct interpretation. The information about neighboring (i.e., adjacent) regions can be found directly in the structure of the graph, as if there exists a link between two regions, these regions are connected, and thus neighboring. However, apart from the adjacency information provided by the graph, additional topological and directional information is needed in order to further assist the analysis and improve performance. The currently supported spatial relations are *above of*, *below of*, *left of*, *right of*, and *contained in*. In addition, two absolute relations were introduced, the *bottom-most* and *top-most* relations, because during experimentation, they proved to be particularly useful in the cases of particular semantic concepts, such as the *Sky*.

11.4.4.5 Descriptors Matching

After having extracted all information about the visual features of the regions of the image and their spatial relations, the next step is to calculate the degree of matching between the descriptors included in the domain knowledge and the ones extracted from the segmented regions, and thus generate possible labels for each region. To accomplish this, it is essential to estimate a distance between two regions based on these low-level features and a distance to each of the prototype instances stored in the VDO. A distance based on each descriptor may be estimated but remains useless without a method of combining all distances and producing a unique fused distance. Because MPEG-7 does not provide a standardized method of combining these distances or of estimating a single distance based on more than one Visual Descriptor, the following approaches are used:

- A *weighted sum* of the two distances, where the weight of the dominant color descriptor is greater than the one of the region shape descriptor, because dominant color has been proven to have a better discriminative performance during the descriptor evaluation process
- A *backpropagation neural network* [45], which is trained to estimate the similarity between two regions; it has as input a vector formed by the low-level descriptions of two regions or a region and a prototype instance, and responds with their "normalized" distance

It should be noted that the first method produces a single distance by combining the distances calculated for each descriptor with different weights, while from the latter, a distance is derived based solely on the low-level visual features that are extracted. In this simple scenario of only two descriptors, both approaches exhibited satisfactory performance. A typical normalization function is used, and then the distance is inverted to *degree of confidence*, which is the similarity criterion for all matching and merging processes. From this whole procedure, a list of possible concepts along with a degree of confidence for all regions is derived and stored appropriately in the graph.

In the case that two or more neighboring regions have been assigned to only one concept, or other possible concepts have a degree less than a predefined threshold, these regions are assumed to be part of a bigger region that was not segmented correctly due to the well-known segmentation limitations. This is then corrected by merging all those regions, that is, merging the graph's vertices and updating all the necessary graph's fields (the visual descriptors are again extracted, the contour of the region is updated along with the edges of the graph, etc.).

11.4.4.6 Spatial Context Consistency Check

The descriptors matching step, by only examining low-level features information, often results in more than one possible semantic label for each region of the image. To evaluate the plausibility of each of these hypotheses and to reach the final interpretation, spatial context is used. More specifically, for each region, the system checks whether the region's extracted spatial characteristics match the spatial context associated with the possible labels assigned to it.

11.4.4.7 Knowledge-Base Retrieval

Whenever new multimedia content is provided as input for analysis, the existing *a priori* knowledge base is used to compare, by means of matching the MPEG-7 visual descriptors and the spatial context information, each region of the graph to the prototype instances of the multimedia domain ontologies. For this reason, the system needs to have full access to the overall knowledge base consisting of all domain concept prototype instances. These instances are applied as references to the analysis algorithms, and with the help of appropriate rules related to the supported domains, the presented knowledge-assisted analysis system extracts semantic concepts that are linked to specific regions of the image or video shot.

For the actual retrieval of the prototypes and its descriptor instances, the *OntoBroker*³ engine is used to deal with the necessary queries to the knowledge base. *OntoBroker* supports the loading of RDFS ontologies, so all appropriate ontology files can be easily loaded. For the analysis purposes, *OntoBroker* needs to load the domain ontologies where high-level concepts are defined, the VDO that contains the low-level visual descriptor definitions, and the prototype instances files that include the knowledge base and provide the linking of domain concepts with descriptor instances. Appropriate queries are defined, which permit the retrieval of specific values from various descriptors and concepts. The *OntoBroker*'s query language is *F-Logic*.⁴ *F-Logic* is both a representation language that can be used to model ontologies and a query language, so it can be used to query *OntoBroker*'s knowledge.

³ See www.ontoprise.de/products/ontobroker_en.

⁴ See www.ontoprise.de/documents/tutorial_flogic.pdf.

TABLE 11.1

Formula One and Beach Vacations Domain Definitions

Concept	Visual Descriptors	Spatial Relations
Road	Dominant color	Road ADJ grass, sand
Car	Region shape, motion activity	Car INC road
Sand	Dominant color	Sand ADJ grass, road
Grass	Dominant color	Grass ADJ road, sand
Sky	Dominant color	Sky ABV Sea
Sea	Dominant color	Sea ABV, ADJ sand
Sand	Dominant color	Sand BEL, ADJ sea
Person	Region shape	Person INC sea, sand

Note: *ADJ*: adjacency relation; *ABV*: above relation; *BEL*: below relation; *INC*: inclusion relation.

11.4.4.8 Semantic Metadata Creation

Having identified the domain semantic concepts that correspond to the different image regions, the next step is to produce metadata in a form that can be easily communicated and shared among different applications. Taking into consideration the proliferation of the Semantic Web and the various emerging applications that use these technologies, the RDF Schema was chosen for representing the extracted annotation metadata. One could then read this RDF and use it directly as semantic annotation by associating the specific image with the number of detected concepts. One step further would be to produce new concepts through the process of fuzzy reasoning (or any other form of reasoning) utilizing both the degrees and the spatial relations. Although the second case seems a lot more interesting, it is more complicated, and so far, we have only used the first scenario to produce the semantic metadata.

11.4.5 Results

The presented knowledge-assisted semantic image analysis approach was tested in the Formula One and beach vacation domains. Analysis was performed by enriching the knowledge infrastructure with the appropriate domain ontology and by providing prototype instances for the corresponding defined domain objects. The defined semantic objects for each of the two examined domains, along with their visual descriptors and their spatial relations are given in Table 11.1. For example, the concept *Sea* in the beach vacations domain ontology is represented using the dominant color descriptor and is defined to be below the concept *Sky* and above or adjacent to the concept *Sand*. In a similar manner, the definitions of the other objects can be derived from Table 11.1. It must be noted that the results for the Formula One domain were obtained by analyzing image sequences and not still images. However, this does not discredit the proposed analysis framework, because each frame was processed separately following the above-described methodology, and the motion activity descriptor was employed only to further improve the attained performance for the *Car* concept. As illustrated in Figure 11.5 and Figure 11.6, respectively, the system output is a segmentation mask outlining the semantic description of the scene where different colors representing the object classes defined in the domain ontology are assigned to the segmented regions.

As previously mentioned, the use of spatial information captures part of the visual context, consequently resulting in the extraction of more meaningful descriptions, provided that the initial color-based segmentation has not segmented two objects as one region. The benefits obtained by the use of spatial information are particularly evident in the beach

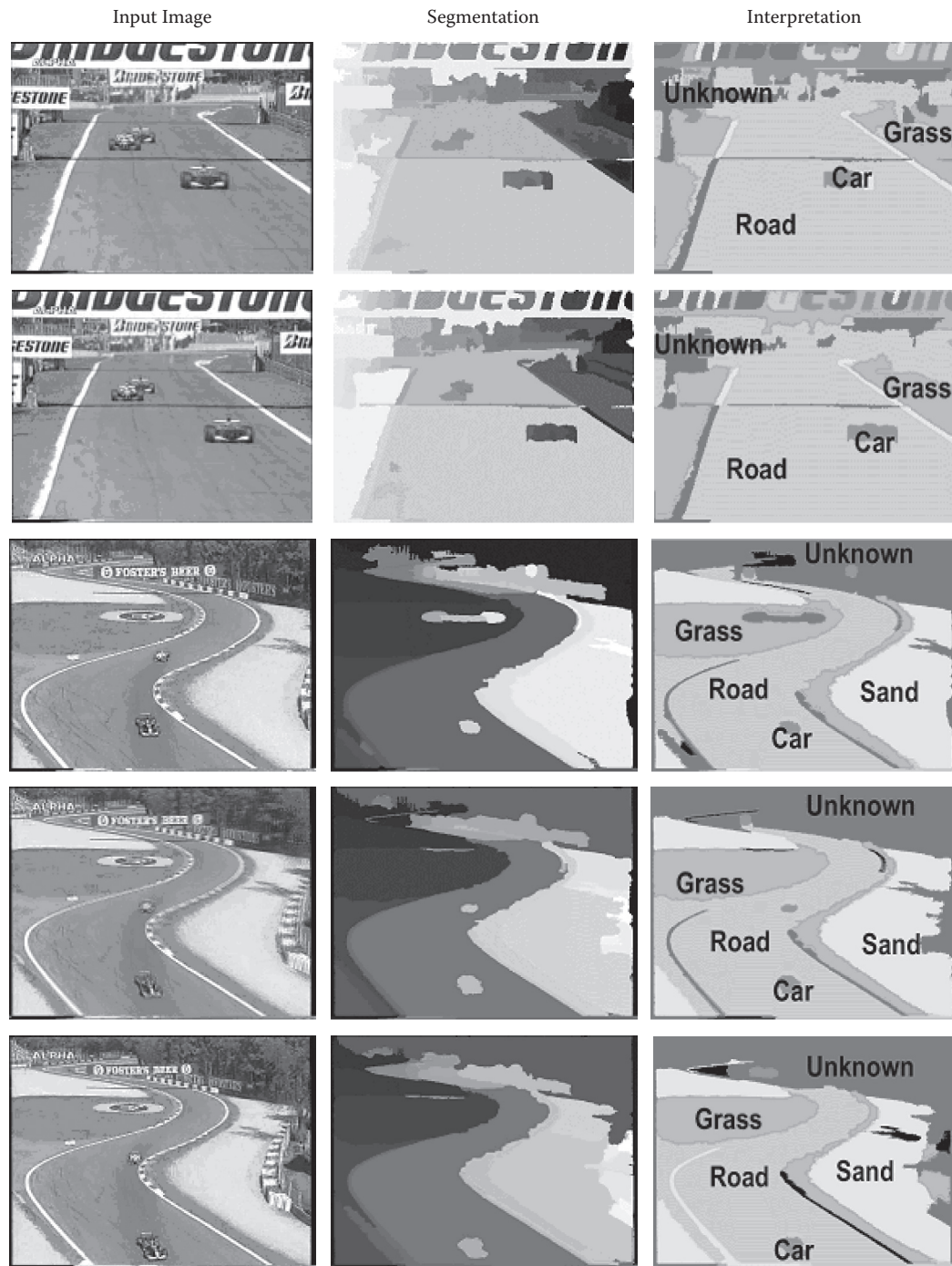


FIGURE 11.5
Semantic analysis results for the Formula One domain.

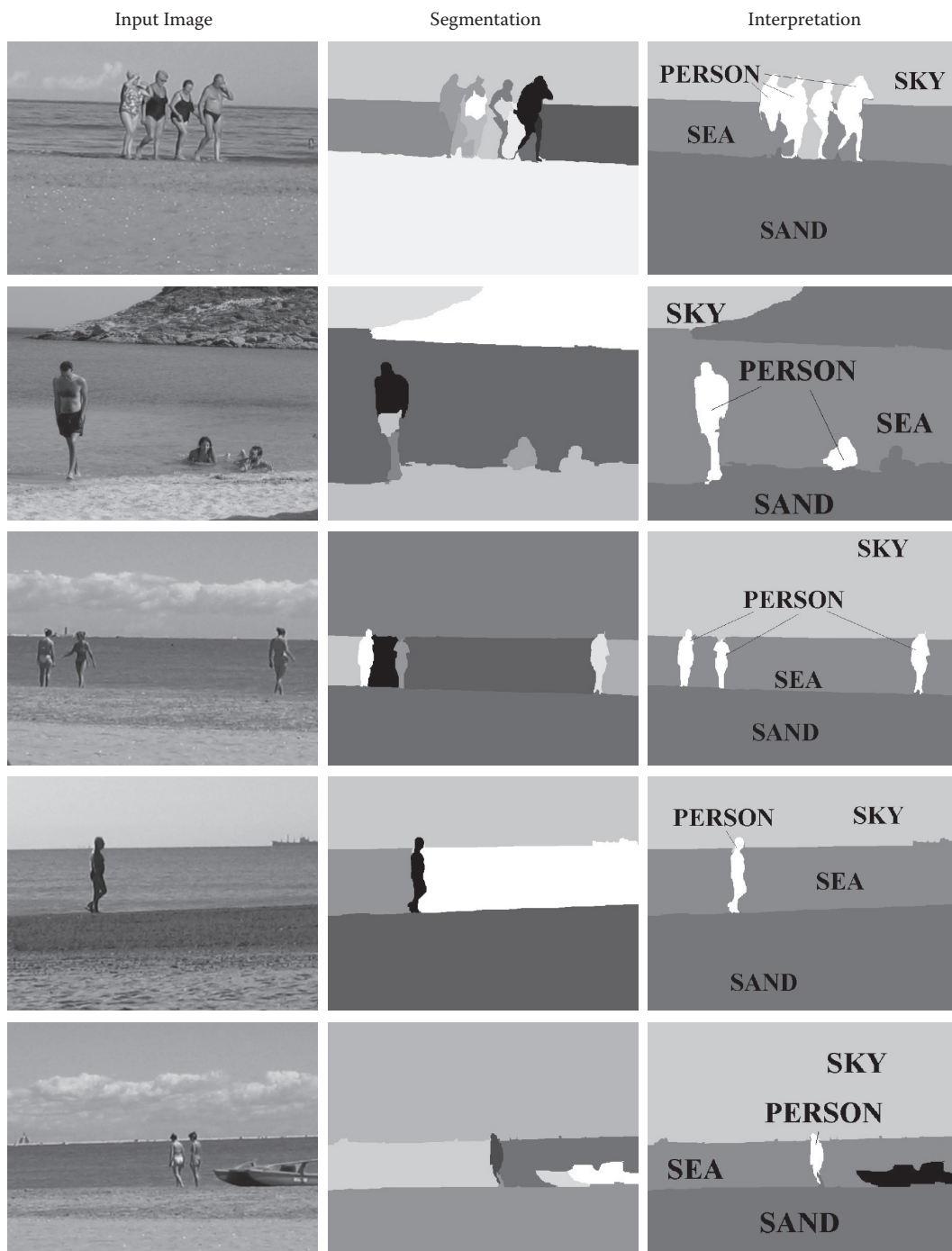


FIGURE 11.6
Semantic analysis results for the beach vacations domain.

vacations domain results, where the semantic concepts *Sea* and *Sky*, despite sharing similar visual features, are correctly identified due to their differing spatial characteristics. The unknown label shown in the produced semantic annotations was introduced to account for the cases where a region does not match any of the semantic objects definitions included in the domain ontology.

11.5 Conclusions and Future Work

This chapter reported on the challenges and current state of the art in semantic image analysis and presented an integrated framework for semantic multimedia content annotation and analysis. The employed knowledge infrastructure uses ontologies for the description of low-level visual features and for linking these descriptions to concepts in domain ontologies. Despite the early stage of experimentation, the first results obtained based on the presented ontological framework are promising and show that it is possible to apply the same analysis algorithms to process different kinds of images by simply employing different domain ontologies. In addition, the generation of the visual descriptors and the linking with the domain concepts is embedded in a user-friendly tool, which hides analysis-specific details from the user. Thus, the definition of appropriate visual descriptors can be accomplished by domain experts, without the need to have a deeper understanding of ontologies or low-level multimedia representations.

However, there is still plenty of space for improvements. Because the performance of the analysis depends on the availability of sufficiently descriptive and representative concepts definitions, among the first future priorities is the investigation of additional descriptors and methodologies for their effective fusion. Related to this is the development of methodologies to efficiently handle issues regarding the prototype instances management (i.e., how many are necessary, how can they be further processed to exploit the available knowledge, etc.). Furthermore, the use of additional spatial and partonomic relations will allow for the definition of more complex semantic concepts and for the derivation of higher-level descriptions based on the already extracted ones, such as the concept of an *Island*, which can be detected as being associated to the *Rock*, *Vegetation*, and so forth, concepts and being inside the *Sea* concept. Finally, apart from visual descriptions and relations, future focus will concentrate on the reasoning process and the creation of rules in order to detect more complex events. The examination of the interactive process between ontology evolution and use of ontologies for content analysis will also be the target of our future work, in the direction of handling the semantic gap in multimedia content interpretation.

To conclude, the proposed approach presents many appealing properties and produces satisfactory results even at this early stage of development. The implementation of the future directions described above will further enhance the achieved performance and contribute to semantic analysis. However, due to the approach followed in modeling the domain knowledge (i.e., the definition of explicit models), there will be cases of semantic concepts with description that will be infeasible due to increased complexity or incomplete knowledge. To support such cases, the approach proposed can be appropriately extended to couple the domain ontology definitions with implicit representations using machine learning representations. Thereby, more accurate semantic descriptions will become available, benefiting from the complementary functionalities provided by explicit and implicit knowledge modeling.

Acknowledgments

This research was partially supported by the European Commission under contracts FP6-001765 aceMedia and FP6-507482 KnowledgeWeb.

References

- [1] A. Treisman, Features and objects in visual processing, *Scientific American*, 255, 114–125, 1986.
- [2] M.R. Naphade, I.V. Kozintsev, and T.S. Huang, A factor graph framework for semantic video indexing, *IEEE Trans. on Circuits and Syst. for Video Technol.*, 12, 40–52, 2002.
- [3] M. Petkovic and W. Jonker, Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events, in *IEEE Workshop on Detection and Recognition of Events in Video*, 2001, pp. 75–82. **Au: Provide editors and publisher/location.**
- [4] Z. Kato, J. Zerubia, and M. Berthod, Unsupervised parallel image classification using a hierarchical Markovian model, in *Proceedings of the Fifth International Conference on Computer Vision*, June 1995. **Au: Provide editors, Page #s and publisher/location.**
- [5] O. Chapelle, P. Haffner, and V. Vapnik, Support vector machines for histogram-based image classification, *IEEE Trans. on Neural Networks*, 10, 1055–1064, September 1999.
- [6] B. Bose and E. Grimson, Learning to use scene context for object classification in surveillance, in *Proceedings of the Joint IEEE International Workshop on VS-PETS*, Nice, France, October 2003, pp. 94–101.
- [7] L. Wang and B.S. Manjunath, A semantic representation for image retrieval, in *Proceedings of International Conference on Image Processing (ICIP'02)*, 2003, pp. 523–526.
- [8] X. Li, L. Wang, and E. Sung, Multi-label svm active learning for image classification, in *Proceedings of International Conference on Image Processing (ICIP'04)*, 2004, pp. 2207–2210. **Au: Provide editors and publisher/location.**
- [9] O. Marques and N. Barman, Semi-automatic semantic annotation of images using machine learning techniques, in *International Semantic Web Conference*, 2003, pp. 550–565. **Au: Provide editors and publisher/location.**
- [10] S. Jeanin and A. Divakaran, MPEG-7 visual motion descriptors, *IEEE Trans. on Circuits and Syst. for Video Technol.*, 11, 6, 720–724, 2001.
- [11] X. Giro and F. Marques, Detection of semantic objects using description graphs, in *Proceedings of International Conference on Image Processing (ICIP'05)*, Genova, Italy, September 2005.
- [12] T. Burghardt, J. Calic, and B. Thomas, Tracking animals in wildlife videos using face detection, in *Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT'04)*, 2004. **Au: Provide editors and publisher/location.**
- [13] R.K. Srihari and Z. Zhang, Show&tell: A semi-automated image annotation system, *IEEE MultiMedia*, 7, 61–71, 2000.
- [14] A.Z. Kouzani, Locating human faces within images, *Comput. Vision and Image Understanding*, 91, 247–279, September 2003.
- [15] F.B.A.C. Lingnau and J.A.S. Centeno, Object oriented analysis and semantic network for high resolution image classification, *Boletim de Ciencias Geodesicas*, 9, 233–242, 2003.
- [16] A. Dorado and E. Izquierdo, Exploiting problem domain knowledge for accurate building image classification, in *Proceedings of the Conference on Image and Video Retrieval (CIVR'04)*, 2004, pp. 199–206. **Au: Provide editors and publisher/location.**
- [17] N. Sprague and J. Luo, Clothed people detection in still images, in *Proceedings of the International Conference on Pattern Recognition (ICPR'02)*, 2002, pp. 585–589. **Au: Provide editors and publisher/location.**
- [18] R. Chopra and R.K. Srihari, Control structures for incorporating picture-specific context in image interpretation, in *IJCAI*, 1995, pp. 50–55.
- [19] A.B. Benitez and S.F. Chang, Image classification using multimedia knowledge networks, in *Proceedings of International Conference on Image Processing (ICIP'03)*, 2003, pp. 613–616. **Au: Provide editors and publisher/location.**

- Au: Provide editors and publisher/location.**
- [20] R. Tansley, C. Bird, W. Hall, P.H. Lewis, and M.J. Weal, Automating the linking of content and concept, in *ACM Multimedia*, 2000, pp. 445–447.
 - [21] K. Barnard, P. Duygulu, D.A. Forsyth, N. Freitas, D.M. Blei, and M.I. Jordan, Matching words and pictures, *J. Machine Learning Res.*, 3, 1107–1135, 2003.
 - [22] V. Lavrenko, R. Manmatha, and J. Jeon, A model for learning the semantics of pictures, in *Neural Information Processing Systems (NIPS'03)*, 2003.
 - [23] C. Hudelot and M. Thonnat, A cognitive vision platform for automatic recognition of natural complex objects, in *ICTAI*, 2003, pp. 398–405.
 - [24] J. Hunter, J. Drennan, and S. Little, Realizing the hydrogen economy through semantic web technologies, *IEEE Intelligent Syst.*, 19, 40–47, 2004.
 - [25] S. Little and J. Hunter, Rules-by-example—a novel approach to semantic indexing and querying of images, in *International Semantic Web Conference*, 2004, pp. 534–548.
 - [26] M. Wallace, G. Akrivas, and G. Stamou, Automatic thematic categorization of multimedia documents using ontological information and fuzzy algebra, in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'03)*, St. Louis, MO, 2003.
 - [27] L. Wang, L. Khan, and C. Breen, Object boundary detection for ontology-based image classification, in *Proceedings of Multimedia Data Mining—Mining Integrated Media and Complex Data (MDM/KDD'02)*, 2002, pp. 51–61.
 - [28] C. Breen, L. Khan, and A. Ponnusamy, Image classification using neural networks and ontologies, in *DEXA Workshops*, 2002, pp. 98–102.
 - [29] C. Meghini, F. Sebastiani, and U. Straccia, Reasoning about the form and content of multimedia objects, in *Proceedings of the AAAI Spring Symposium on the Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, 1997.
 - [30] M. Grimnes and A. Aamodt, A two layer case-based reasoning architecture for medical image understanding, in *EWCBR*, 1996, pp. 164–178.
 - [31] M. Goldbaum, S. Moezzi, A. Taylor, S. Chatterjee, E. Hunter, and R. Jain, Automated diagnosis and image understanding with object extraction, objects classification and inferencing in retinal images, in *Proceedings of International Conference on Image Processing (ICIP'96)*, 1996.
 - [32] S. Linying, B. Sharp, and C.C. Chibelushi, Knowledge-based image understanding: A rule-based production system for x-ray segmentation, in *Proceedings of the International Conference on Enterprise Information Systems (ICEIS'02)*, 2002, pp. 530–533.
 - [33] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content-based image retrieval at the end of the early years, *IEEE Trans. Patt. Anal. Machine Intelligence*, 22, 1349–1380, 2000.

Au: Provide editors and publisher/location.

 - [34] N. Sebe, M.S. Lew, X.S. Zhou, T.S. Huang, and E.M. Bakker, The state of the art in image and video retrieval, in *Proceedings of the Conference on Image and Video Retrieval (CIVR'03)*, 2003, pp. 1–8.
 - [35] P. Salembier and F. Marques, Region-based representations of image and video: Segmentation tools for multimedia services, *IEEE Trans. on Circuits and Syst. for Video Technol. (CSVT)*, 9, 1147–1169, December 1999.
 - [36] S. Staab and R. Studer, *Handbook on Ontologies*, Springer-Verlag, Heidelberg, 2004.
 - [37] S.-F. Chang, T. Sikora, and A. Puri, Overview of the MPEG-7 standard, *IEEE Trans. on Circuits and Syst. for Video Technol.*, 11, 688–695, 2001.

Au: Provide page#, editors and publisher/location For Ref. 43.

 - [38] T. Sikora, The MPEG-7 visual standard for content description—an overview, *IEEE Trans. on Circuits and Syst. for Video Technol.*, 11, 696–702, 2001.
 - [39] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, Color and texture descriptors, *IEEE Trans. on Circuits and Syst. for Video Technol.*, 11, 703–715, 2001.
 - [40] MPEG-7, Multimedia Content Description Interface—part 3: Visual, ISO/IEC/JTC1/SC29/WG11, Document N4062, 2001.
 - [41] M. Bober, MPEG-7 visual shape descriptors, *IEEE Trans. on Circuits and Syst. for Video Technol.*, 11, 716–719, 2001.

Au: Provide page#, editors and publisher/location.

 - [42] MPEG-7, Visual Experimentation Model (XM) Version 10.0, ISO/IEC/JTC1/SC29/WG11, Document N4062, 2001.
 - [43] H. Eidenberger, Distance measures for MPEG-7-based retrieval, in *ACM MIR03*, 2003.
 - [44] J. Stauder, J. Sirot, H. Borgne, E. Cooke, and N. O'Connor, Relating visual and semantic image descriptors, in *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, London, 2004.

- [45] E. Spyrou, H. LeBorgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor, Fusing MPEG-7 visual descriptors for image classification, in *International Conference on Artificial Neural Networks (ICANN)*, 2005. **Au: Provide page#s, editors and publisher/location.**
- [46] B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranic, Automatic selection and combination of descriptors for effective 3d similarity search, in *IEEE International Workshop on Multimedia Content-Based Analysis and Retrieval*, 2004. **Au: Provide page#s, editors and publisher/location.**
- [47] F. Mokhtarian and S. Abbasi, Robust automatic selection of optimal views in multi-view free-form object recognition, *Patt. Recognition*, 1021–1031, 2005. **Au: Provide volume#.**
- [48] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, I. Kompatsiaris, S. Staab, and M.G. Strintzis, Semantic annotation of images and videos for multimedia analysis, in *Proceedings of the Second European Semantic Web Conference (ESWC)*, Heraklion, Greece, May 2005. **Au: Provide page#s, editors and publisher/location.**
- [49] "ISO/IEC 15938-5 FCD Information Technology—Multimedia Content Description Interface—Part 5: Multimedia Description Schemes, March 2001, Singapore. **Au: Provide page#s, editors and publisher/location.**
- [50] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, Sweetening ontologies with DOLCE, in *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Proceedings of the 13th International Conference on Knowledge Acquisition, Modeling and Management, EKAW 2002*, Vol. 2473 of *Lecture Notes in Computer Science*, Siguenza, Spain, 2002. **Au: Provide page#s, editors and publisher/location.**
- [51] A. Cohn, B. Bennett, J.M. Gooday, and N.M. Gotts, *Representing and Reasoning with Qualitative Spatial Relations about Regions*, Kluwer, Dordrecht, 1997, pp. 97–134.
- [52] J. Allen, Maintaining knowledge about temporal intervals, *Commun. ACM*, 26, 832–843, 1983.
- [53] D. Papadias and Y. Theodoridis, Spatial relations, minimum bounding rectangles, and spatial data structures, *Int. J. Geographical Inf. Sci.*, 11, 111–138, 1997. **Au: Provide page#s, editors and publisher/location.**
- [54] S. Skiadopoulos and M. Koubarakis, Composing cardinal direction relations, *Artif. Intelligence*, 152, 143–171, 2004.
- [55] N. Simou, V. Tzouvaras, Y. Avrithis, G. Stamou, and S. Kollias, A visual descriptor ontology for multimedia reasoning, in *Proceedings of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '05)*, Montreux, Switzerland, April 13–15, 2005. **Au: Provide page#s, editors and publisher/location.**
- [56] ISO/IEC 15938-3 FCD Information Technology—Multimedia Content Description Interface—Part 3: Visual, March 2001, Singapore.
- [57] T. Adamek, N. O'Connor, and N. Murphy, Region-based segmentation of images using syntactic visual features, in *Proceedings of the Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2005*, Montreux, Switzerland, April 13–15, 2005. **Au: Provide page#s, editors and publisher/location.**

P1: Binaya Dash

May 10, 2006 21:20 9774 9774 Book