Chapter 1

Semantic Processing of color Images

S. DASIOPOULOU

Information Processing Laboratory
Electrical and Computer Engineering Department
Aristotle University of Thessaloniki
54124 Thessaloniki - Greece
Email: dasiop@iti.gr

E. SPYROU, Y. AVRITHIS

Image, Video and Multimedia Laboratory
School of Electrical and Computer Engineering
National Technical University of Athens
15773 Zografou - Greece
Email: espyrou@image.ece.ntua.gr
iavr@image.ece.ntua.gr

Y. KOMPATSIARIS, M.G. STRINTZIS

Informatics and Telematics Institute
Centre for Research and Technology Hellas
1st Km Thermi-Panorama Road
57001 Thessaloniki - Greece
Email: ikom@iti.gr
strintzi@eng.auth.gr

Invited Chapter to appear in "Color Image Processing: Methods and Applications," CRC Press, Rastislav Lukac and Konstantinos N. Plataniotis, Editors.

1.1 Introduction

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

Acknowledging the need for providing image analysis at semantic level, research efforts set focus on the automatic extraction of image descriptions matching 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 conveyed meaning. The emerged approaches fall into two categories, i.e. data-driven and knowledge-driven, depending on the direction of these high-level descriptions creation process. The former adhere to the monolithic computational paradigm, in which the interpretation corresponds to an extreme value of some global objective function computed directly from the data. No hierarchy of meaningful intermediate interpretations is created. The latter instead, 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 one 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 high-level content representations without any prior knowledge apart from the inherent developers one. Thus, these approaches concentrate on acquiring fully automated numeric descriptors from objective visual content properties and perform semantic annotation and the subsequent retrieval based on criteria that somehow replicate human perception of visual similarity. The major weakness of such approaches is that they fail to interact meaningfully with users high-level of cognition, since the built in associations between image semantics and its low-level features quantitative descriptions are of no perceptual meaning to the users. Consequently, the underpinning linking mechanism remains a “black box” to the user not allowing for efficient access and most importantly, for the discovery of semantically related content. Systems based on the query-by-example paradigm, as well as traditional keyword-based image retrieval systems, are well known application examples belonging in this category. Although efficient for restricted domains, such approaches lack capability to adapt to different domains. Techniques like relevance feedback and incremental learning have been used for improving 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. to 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, i.e. explicit, realized by formal model definitions or implicit, realized by machine learning methods.

The main characteristic of learning-based approaches is their ability to adjust their internal structure according to input and respective desired output data pairs in order to approximate the relations (rules) 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 of approaches for have been proposed for a variety of applications as presented in the next section. As will be 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 “black box” method often employed can 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 etc. of the examined domain. Thus, since the terms of the employed language (ontologies, semantic nets etc.) carry meaning 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 in settings where only a small number of parts are to be located within a scene. Although the strict assumption of fully specified geometric models can be relaxed by employing parameterized or generic models, 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 primary reliance on 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 pre-attentive stage of visual processing. During this stage different image events are detected, which are joined into complex objects at a second stage. Treisman [1] hypothesizes that the visual system start 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 hypothesizes that the pre-attentive 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.)

Concluding, 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. This chapter discusses semantic image analysis for the purpose of automatic image understanding and efficient visual content access and retrieval at semantic level. The overview presented in section 1.2 surveys current state of the art analysis approaches aiming at bridging the "semantic gap" in image analysis and retrieval. It highlights the major achievements of the existing approaches and sheds light to the challenges still unsolved. Section 1.3 presents a generic framework for performing knowledge-assisted semantic analysis of images. Knowledge representation and modelling, 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 1.4 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 on section 1.5 and plans for future work are presented.

1.2 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 impossible the extraction of image semantics 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 have also been followed considering knowledge representation, management and inference realization. Neural networks, expert systems, fuzzy logic, ontologies, decision trees, static and dynamic Bayesian networks, factor graphs, Markov random fields, etc., are among the popular mechanisms for storing and enforcing high-level information.

Stochastic approaches include among others, the work presented in [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, multijects, 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 [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) has been used in [4] for implementing unsupervised image classification.

Histogram-based image classification is performed using a Support Vector Machine (SVM) in [5], while an object support vector machines' classifier that is trained once on a small set of labelled examples is presented in [6]. An SVM is applied to represent conditioned feature vector distributions within each semantic class and a Markov random field is used to model the spatial distributions of the semantic labels, for achieving semantic labelling of image regions in [7]. To address cases where more than one labels fit the image data, [8] proposes a multilabel SVM active learning approach to address multilabel image classification problems.

In [9] machine-learning techniques are used to semantically annotate images with semantic descriptions defined within ontologies, while in [10] the use of the maximum entropy approach is proposed for the task of automatic image annotation. In [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 furthered semantically organized under a binary partition tree. Another nice example of domain driven semi-automated algorithm for semantic annotation is given in [12], where a specific animal face tracker is formed from user labelled examples utilizing Ada-boost classifier and Kanade-Lucas-Tomasi tracker. The semi-automatic image annotation system proposed in [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 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 configuration.

Fuzziness is introduced in [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 [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 [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 [17]. Within this system a tree is constructed, whose nodes 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 [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 [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 has been reported in a number of other research efforts such as in [20] using latent semantics analysis, [21], [22], etc.

Following the recent Semantic Web advances, several approaches 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 [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. Enhanced by rules is also the user-assisted approach for automatic image annotation reported in [25], while fuzzy algebra and fuzzy ontological information are exploited in [26] for extracting semantic information in the form of thematic categorization. Ontology-based image classification systems are also presented in [27] and [28]. In [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 content of such documents, thus allowing queries both on structural and conceptual similarity.

Medical image understanding is another application field in which semantic image analysis has received particularly strong interest. Medical images are mainly used for diagnosis purposes in order to reduce repetitive work, provide assistance in difficult diagnoses or unfamiliar cases, and efficiently manage the huge volumes of information concentrated in medical image databases, thus the automatic acquisition of accurate interpretation is a non-negotiable requirement. The approaches reported in the literature cover a wide variety of medical imaging cases such as tomography, mammography, ophthalmology, radiology, etc. Computer tomography images are analyzed in [30] using two case-based reasoners, one for segment identification and the second for a more holistic interpretation of the image. The system STARE, presented in [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 lag behind 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 has been reported, semantic image understanding remains an unsolved problem since most state of the art techniques make no attempt to investigate generic strategies for incorporating domain knowledge and/or 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 the recognition of perceptual entities, the presented technical challenges render semantic image analysis a fascinating research area urging for new advances.

Furthermore, recent studies revealed that apart from the need to provide semantic-enabled images access and management, the inherent dynamic interpretation of images under different circumstances should be taken into consideration as well 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 have also to learn them online 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.

1.3 Knowledge-Assisted Analysis

Building on the considerations resulting by 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 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 potential 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 examined domain. 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 plausibility of each hypothesis and thereby decide upon the final semantic labelling.

Consequently, partitioning the image into a set of meaningful regions is the prerequisite before any analysis can take place, since 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 purely numerical basis, thus leading to segmentations that are unreliable and vary in uncontrollable ways, i.e. regions result 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 for shape-driven segmentation approaches, while

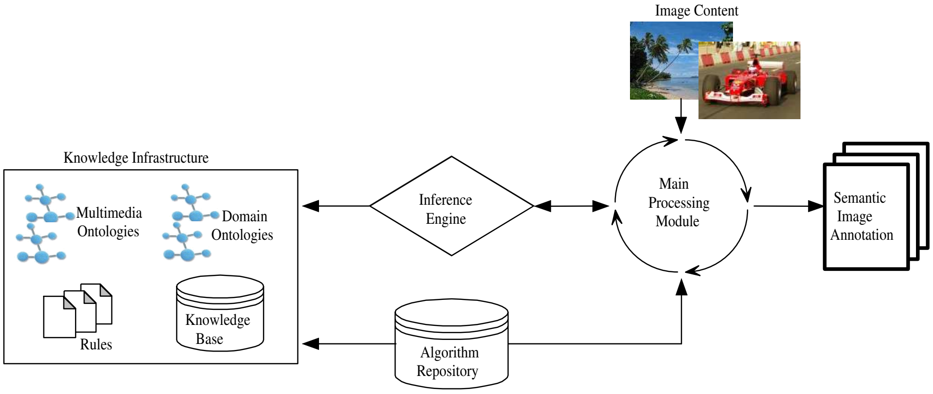


Figure 1.1: Knowledge-assisted semantic image analysis framework

texturized objects need segmentations based on, possibly different per object, texture descriptors.

From the above mentioned, it comes up 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 this extremely challenging tasks, the presented framework adopts formal knowledge representation to ensure consistent inferencing and exploits the available knowledge 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 Fig. 1.1. As illustrated, ontologies have been 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. Ontology languages have been developed that provide support for expressing rich semantics, while at the same time ensuring 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 to reason on the existing facts and derive new knowledge 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. The detailed description of 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 itself, different kind 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 a way 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, but still 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 itself, descriptors 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 de-coupled 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 needs to be investigated so that efficient and effective access and retrieval of the involved knowledge is ensured. This translates to ontological repositories and corresponding query languages issues.

Concluding, 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.

1.4 Knowledge-Assisted Analysis using MPEG-7 and Semantic Web Technologies

1.4.1 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 audio-visual (AV) content and unlike its predecessors, focuses on non-textual description of multimedia content aiming to provide interoperability among applications that use audio-visual 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, non-semantic 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* of *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.

Color Descriptors

Color is probably the most expressive of all the visual features. Thus, it has extensively been 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 therefore, a short description of the most widely used color spaces is essential. The color spaces supported are the monochrome, 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 into 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 [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 images 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.

Texture Descriptors

Texture refers to the visual patterns that have properties of homogeneity or not, that result from 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 powerful means for similarity matching and retrieval for both homogeneous and nonhomogeneous textures. The three texture descriptors, standardized by MPEG-7 are: the *Texture Browsing Descriptor*, the *Homogeneous Texture Descriptor* and the *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 deviation of the filtered outputs are computed in the frequency domain. The frequency space is divided in 30 channels, as described in [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. Since 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.

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 3 shape descriptors [41], which are: *Region-Based Shape Descriptor*,*Contour-Based Shape Descriptor* and 2-*D/*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 object 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 of describing 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 whose contour is characteristic and therefore the Region-Based Shape descriptor is redundant. Apart from that, it can discriminate objects whose regions 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 of a multiview description of it.

Descriptor Matching

As described in section 1.4, 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 itself in terms of its structure and the modelled domain concepts in the form of visual descriptors and spatial characteristics definitions. Among the possible representations, the information considering low-level visual features information 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 similarity, 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 *Dominant Color* descriptor can be performed in the following way: Let

$$F_1 = \big\{\{c_{1i}, p_{1i}, v_{1i}\}, s_1\big\}, i = 1, \ldots, N_1$$
$$F_2 = \big\{\{c_{2i}, p_{2i}, v_{2i}\}, s_2\big\}, i = 1, \ldots, N_2$$

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

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

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 $d_{max} = aT_d$. More details about the determination of T_d and a for and also for a few modifications that can be made to take into account the variances and the spatial coherencies, can be found in [42].

MPEG-7 does not strictly standardize the distance functions to be used and sometimes does not propose a dissimilarity function leaving the developer 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 descriptors seems essential. This problem still 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 e.g. 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 [45] simply consists of merging all the descriptions into a unique vector. If $D_1, D_2, \ldots, D_n$ are the n descriptors to combine, then the merged descriptor is equal to:

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

This fusing method requires all features to have more or less the same numerical values to avoid scale effects. An alternative is to re-scale the data using principal component analysis for instance. Re-scaling is not necessary in the case of the MPEG-7 descriptors since they are already scaled to integer values of equivalent magnitude. Assigning *fixed weights* as in [46] can be a very efficient method especially when the number of the visual features is very small. The assignment of the weights can be done either manually, by simply observing the results and giving more weight on the descriptors that seem to have more discriminative power, or *statistically* as in [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.

1.4.2 Ontology Structure

As noted in section 1.4, 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 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 RDF have been used as the means for representing the various knowledge components involved.

Resource Description Framework Schema (RDFS) is a simple modelling language on top of the Resource Description Framework (RDF) formalism[1], both being developed by the W3C. *Web Ontology Lan-*

[1] RDF itself is not a Knowledge Representation system but tries to improve data interoperability on the Web. This is achieved

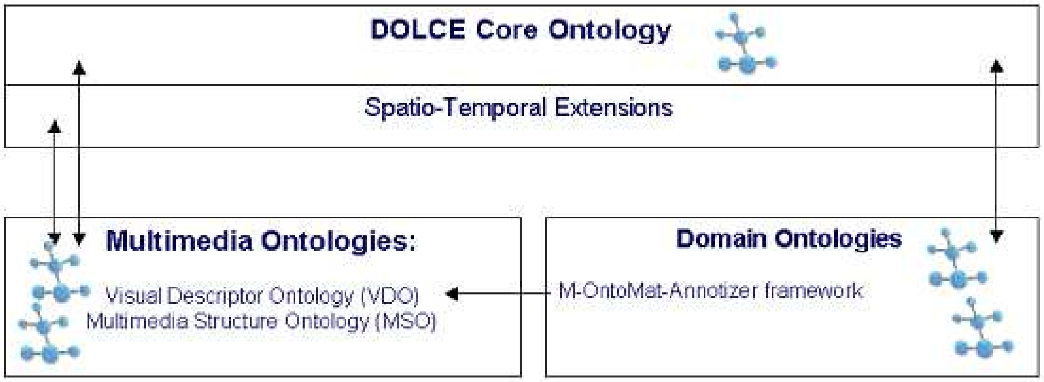

DOLCE Core Ontology
Spatio-Temporal Extensions
Multimedia Ontologies:
Visual Descriptor Ontology (VDO)
Multimedia Structure Ontology (MSO)
Domain Ontologies
M-OntoMat-Annotizer framework

Figure 1.2: Ontology Structure Overview

guage (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 modelling language due to the fact that a full usage of the increased expressiveness of OWL requires specialized and more advanced inference engines that are not yet widely available.

The developed knowledge infrastructure 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 [48] and *Domain Ontologies* that model the content layer of multimedia content with respect to specific real-world domains.

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* [49] 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 means for representing spatio-temporal qualities, reasoning with such descriptions requires the coding of additional relations that describe the relationship between

by specializing the XML data model through a graph-based data model similar to the semantic networks formalism.

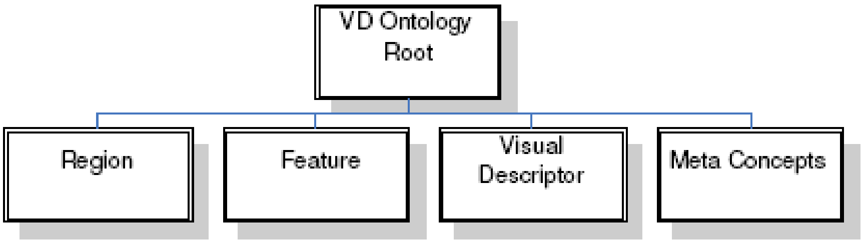


Figure 1.3: The Visual Descriptor Ontology (VDO).

space and/or time regions. Based on concepts taken from the ‘Region Connecting Calculus’ [50], Allen’s interval calculus [51] and directional models [52] [53], 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., touches and overlaps). 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.

Visual Descriptor Ontology

The Visual Descriptor Ontology (VDO) [54] represents the visual part of the MPEG-7, thus contains the representations of the set of visual descriptors used for knowledge-assisted analysis. Its modelled 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 [55]. Since 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 Visual Descriptor ontology consists of four main concepts, which are `VDO:Region`, `VDO:Feature`, `VDO:VisualDescriptor` and `VDO:Metaconcepts`, as illustrated in Fig. 1.3. 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. `VDO:VisualDescriptor` concept contains the visual descriptors as these are defined by MPEG-7. `VDO:Metaconcepts` concept on the other hand contains some additional concepts that were necessary for the Visual Descriptor Ontology 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.

For example, considering the `VDO:VisualDescriptor` concept, which consists of six subconcepts, 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 has been modified regarding the MPEG-7 standard and does not contain all the MPEG-7 descriptors.

Multimedia Structure Ontology

The Multimedia Structure Ontology (MSO) models basic multimedia entities from the MPEG-7 Multimedia Description Scheme [48] 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 [48] describes a spatial and/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 sub-segments through four types of decomposition: *spatial*, *temporal*, *spatiotemporal* and *media source*.

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. Since the *DOLCE* ontology has been selected as the core ontology of the ontology infrastructure, it is essential that all the domain ontologies are 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 users 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.

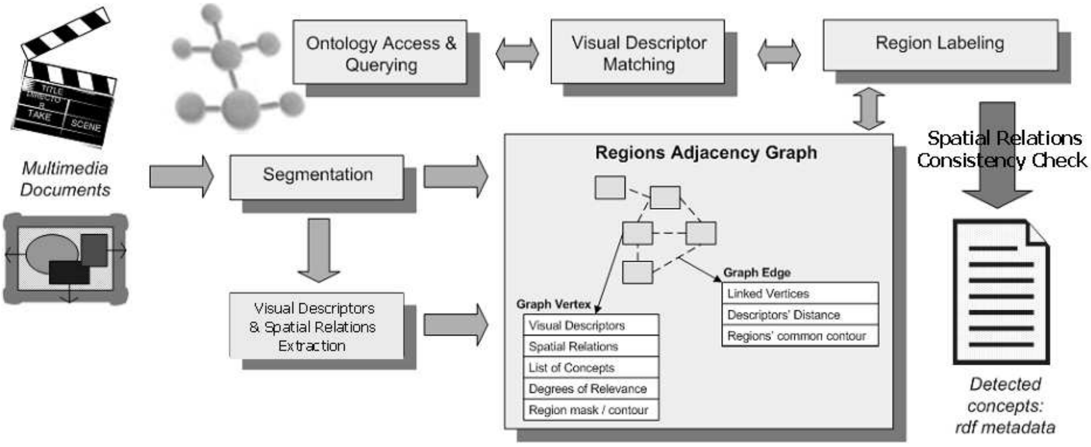


Figure 1.4: The developed knowledge-assisted semantic image analysis system architecture.

1.4.3 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, since, as described in section 1.4, the produced semantic annotations are generated through the matching against these objects prototypes. To accomplish 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 to the domain ontology. Within the described implementation, a user oriented tool was developed. Users select regions that correspond to domain concepts and then chose the MPEG-7 Descriptor to be extracted. Triggered by the users extraction command, the requested MPEG-7 Descriptors are extracted through calls to appropriate routines based on the MPEG-7 XM.

1.4.4 Semantic Multimedia Analysis

The implemented semantic multimedia analysis architecture is presented in Fig. 1.4. 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 objects 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.

Image Representation

A *region adjacency graph* 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, currently the Dominant Color and Region Shape, 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 list[2] 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.

Image Segmentation

The first step should be 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 [56].

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 eXperimentation Model (XM) [42].

Spatial Relations Extraction

As previously mentioned, apart from low-level descriptions it is necessary to include in the domain knowledge definitions information about objects spatial context as well, since it 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, other additional topological and directional information is needed in order to further assist the analysis and improve performance. The currently supported spatial relations are the *above of*, *below of*, *left of*, *right of* and *contained in*. In addition, two absolute relations were introduced,

[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

the *bottom-most* and *top-most* relations, since during experimentation they proved to be particular useful in the cases of particular semantic concepts, such as the *Sky*.

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 produce a unique fused distance. Since MPEG-7 does not provide a standardized method of combining these distances or to estimate a single distance based on more than one Visual Descriptors 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, since dominant color has been proven to have a better discriminative performance during the descriptor evaluation process

- A *back-propagation neural network* [45] which is trained to estimate the similarity between two regions. This network 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 on 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 worked fine. A typical normalization function is used and then the distance is inverted to *degree of confidence*, which is the similarity criterium 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 pre-defined 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, i.e. merging the graph’s vertices and updating all the necessary graph’s fields (the visual descriptors are again extracted, the contour of the regions is updated along with the edges of the graph etc).

Spatial Context Consistency Check

The descriptors matching step, by only examining low-level features information, often results in more than one possible semantic labels 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.

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 succeed 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 can be used to query OntoBroker's knowledge.

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 to

[3] see `http://www.ontoprise.de/products/ontobroker_en`

[4] see `http://www.ontoprise.de/documents/tutorial_flogic.pdf`

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.

1.4.5 Results

The presented knowledge-assisted semantic image analysis approach was tested in the Formula One and beach vacations 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 1.1. 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 the table 1.1. 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, since 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 Fig. 1.5 and 1.6 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.

Table 1.1: Formula One and Beach vacations domain definitions *((ADJ:adjacency relation), (ABV):above relation, (BLW):below relation, (INC):inclusion relation).*

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

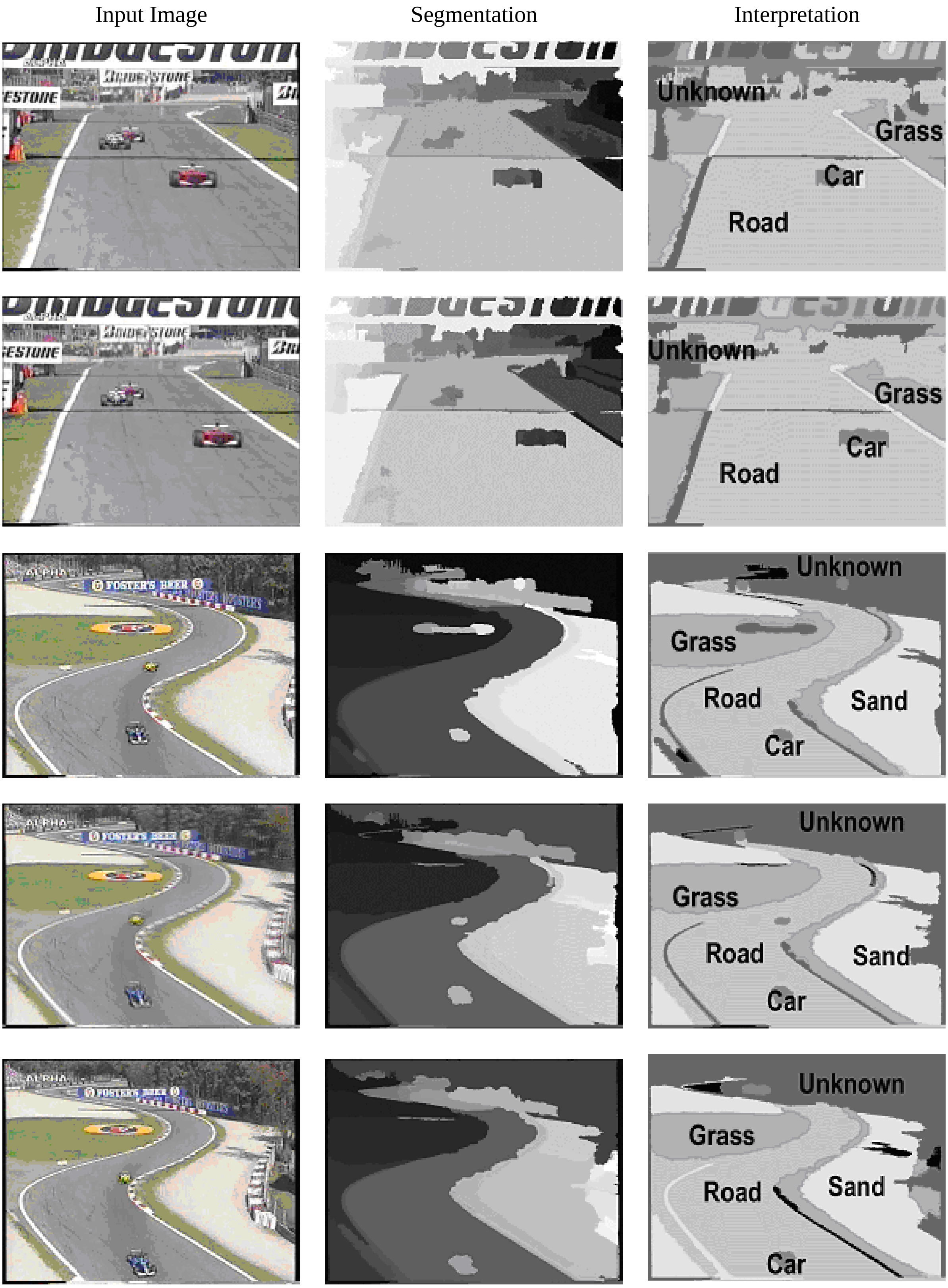
As previously mentioned, the use of spatial information captures part of the visual context, consequently resulting in the extraction of more meaningful descriptions provided that the initial color-based segmentation did not segment two objects as one atom-region. The benefits obtained by the use of spatial information are particularly evident in the beach 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 has been introduced to account for the cases where a region does not match any of the semantic objects definitions included in the domain ontology.

1.5 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 room for improvements. Since 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 the definition of more complex semantic concepts and the derivation of higher-level descriptions based on the already extracted ones such as the concept of an *Island*, which can ne detected as associated to the *Rock*, *Vegetation*, etc. 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 modelling the domain knowledge, i.e the definition of explicit models, there will be cases of semantic concepts whose description will be infeasible due to increased complexity or incomplete knowledge. To support for such cases the proposed approach can be appropriately extended to couple the domain ontology definitions with implicit representations using machine learning representations. Thereby, more accurate semantic descriptions will be available benefiting from the complementary functionalities provided by explicit and implicit knowledge modelling.



Input Image

Segmentation

Interpretation

Figure 1.5: Semantic analysis results for the Formula One domain.

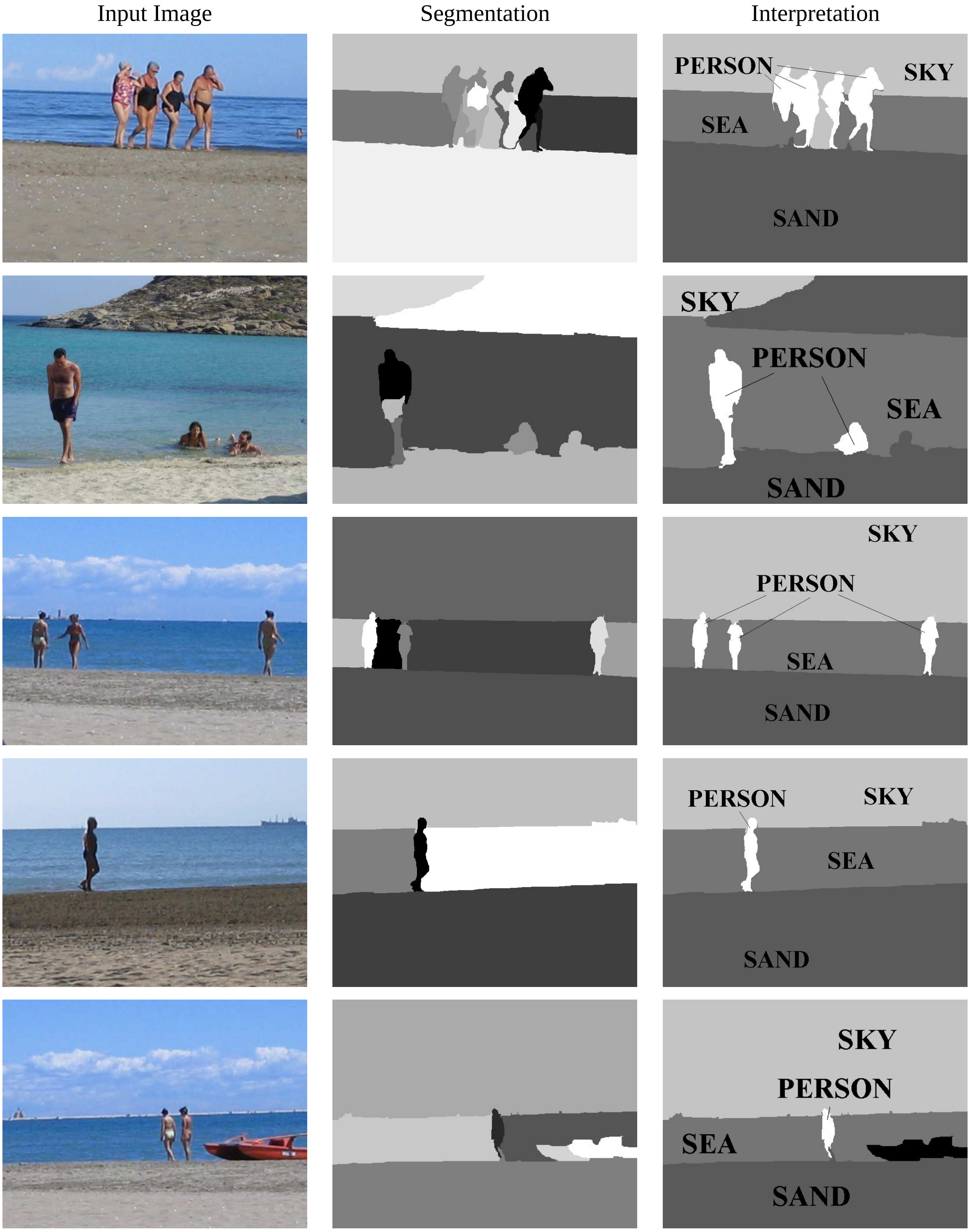


Figure 1.6: Semantic analysis results for the beach vacations domain.

List of Figures

List of Tables

Bibliography

[1] A. Treisman, “Features and objects in visual processing,” *Scientific American*, vol. 255, no. 5, pp. 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 Systems for Video Technology*, vol. 12, no. 1, pp. 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*, pp. 75–82, 2001.

[4] Z. KAto, J. Zerubia, and M. Berthod, “Unsupervised parallel image classification using a hierarchicalmarkovian model,” *Proc. Fifth International Conference on Computer Vision*, June 1995.

[5] O. Chapelle, P. Haffner, and V. Vapnik, “Support vector machines for histogram-based image classification,” *IEEE Trans. on Neural Networks*, vol. 10, pp. 1055–1064, September 1999.

[6] B. Bose and E. Grimson, “Learning to use scene context for object classification in surveillance,” in *Proc. Joint IEEE Int’l Workshop on VS-PETS*, (Nice, France), pp. 94–101, October 2003.

[7] L. Wang and B. S. Manjunath, “A semantic representation for image retrieval.,” in *Proc. of International Conference on Image Processing (ICIP’02)*, pp. 523–526, 2003.

[8] X. Li, L. Wang, and E. Sung, “Multi-label svm active learning for image classification.,” in *Proc. of International Conference on Image Processing (ICIP’04)*, pp. 2207–2210, 2004.

[9] O. Marques and N. Barman, “Semi-automatic semantic annotation of images using machine learning techniques.,” in *International Semantic Web Conference*, pp. 550–565, 2003.

[10] S.Jeanin and A.Divakaran, “Mpeg-7 visual motion descriptors,” *IEEE trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 720–724, 2001.

[11] X. Giro and F. Marques, “Detection of semantic objects using description graphs,” in *Proc. 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 *Proc. European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT’04)*, 2004.

[13] R. K. Srihari and Z. Zhang, “Show&tell: A semi-automated image annotation system.,” *IEEE MultiMedia*, vol. 7, no. 3, pp. 61–71, 2000.

[14] A. Z. Kouzani, “Locating human faces within images.,” *Computer Vision and Image Understanding, Elsevier*, vol. 91, pp. 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*, vol. 9, pp. 233–242, 2003.

[16] A. Dorado and E. Izquierdo, “Exploiting problem domain knowledge for accurate building image classification.,” in *Proc. Conference on Image and Video Retrieval (CIVR’04)*, pp. 199–206, 2004.

[17] N. Sprague and J. Luo, “Clothed people detection in still images.,” in *Proc. International Conference on Pattern Recognition (ICPR’02)*, pp. 585–589, 2002.

[18] R. Chopra and R. K. Srihari, “Control structures for incorporating picture-specific context in image interpretation.,” in *IJCAI*, pp. 50–55, 1995.

[19] A. B. Benitez and S. F. Chang, “Image classification using multimedia knowledge networks.,” in *Proc. of International Conference on Image Processing (ICIP’03)*, pp. 613–616, 2003.

[20] R. Tansley, C. Bird, W. Hall, P. H. Lewis, and M. J. Weal, “Automating the linking of content and concept.,” in *ACM Multimedia*, pp. 445–447, 2000.

[21] K. Barnard, P. Duygulu, D. A. Forsyth, N. Freitas, D. M. Blei, and M. I. Jordan, “Matching words and pictures.,” *Journal of Machine Learning Research*, vol. 3, pp. 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*, pp. 398–405, 2003.

[24] J. Hunter, J. Drennan, and S. Little, “Realizing the hydrogen economy through semantic web technologies.,” *IEEE Intelligent Systems*, vol. 19, no. 1, pp. 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*, pp. 534–548, 2004.

[26] M. Wallace, G. Akrivas, and G. Stamou, “Automatic thematic categorization of multimedia documents using ontological information and fuzzy algebra.,” in *Proc. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE’03)*, (St. Liouis, MO, USA), 2003.

[27] L. Wang, L. Khan, and C. Breen, “Object boundary detection for ontology-based image classification.,” in *Proc. Multimedia Data Mining - Mining Integrated Media and Complex Data (MDM/KDD’02)*, pp. 51–61, 2002.

[28] C. Breen, L. Khan, and A. Ponnusamy, “Image classification using neural networks and ontologies.,” in *DEXA Workshops*, pp. 98–102, 2002.

[29] C. Meghini, F. Sebastiani, and U. Straccia, “Reasoning about the form and content of multimedia objects.,” in *Proc. 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*, pp. 164–178, 1996.

[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 *Proc. 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 *Proc. International Conference on Enterprise Information Systems (ICEIS’02)*, pp. 530–533, 2002.

[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. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[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 *Proc. Conference on Image and Video Retrieval (CIVR'03)*, pp. 1–8, 2003.

[35] P. Salembier and F. Marques, "Region-based representations of image and video: segmentation toolsfor multimedia services," *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, vol. 9, pp. 1147–1169, December 1999.

[36] S.Staab and R.Studer, *Handbook on Ontologies*. Springer Verlag, 2004.

[37] S.-F. Chang, T. Sikora, and A. Puri, "Overview of the mpeg-7 standard," *IEEE trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688–695, 2001.

[38] T. Sikora, "The mpeg-7 visual standard for content description - an overview," *IEEE trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 696–702, 2001.

[39] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703–715, 2001.

[40] MPEG-7, "Multimedia content description interface - part 3: Visual." ISO/IEC/ JTC1/SC29/WG11, Doc. N4062, 2001.

[41] M. Bober, "Mpeg-7 visual shape descriptors," *IEEE trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 716–719, 2001.

[42] MPEG-7, "Visual experimentation model (xm) version 10.0." ISO/IEC/ JTC1/SC29/WG11, Doc. 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, U.K.*, 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.

[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.

[47] F. Mokhtarian and S. Abbasi, "Robust automatic selection of optimal views in multi-view free-form object recognition," *Pattern Recognition*, no. 38, pp. 1021–1031, 2005.

[48] "ISO/IEC 15938-5 FCD Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Scemes, March 2001, Singapore,"

[49] 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.

[50] A. Cohn, B. Bennett, J. M. Gooday, and N. M. Gotts., *Representing and Reasoning with Qualitative Spatial Relations about Regions*, pp. 97–134. Kluwer Academic Publishers, 1997.

[51] J. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, no. 1, pp. 832–843, 1983.

[52] D. Papadias and Y. Theodoridis, "Spatial relations, minimum bounding rectangles, and spatial data structures," *International Journal of Geographical Information Science*, vol. 11, pp. 111–138, 1997.

[53] S. Skiadopoulos and M. Koubarakis, "Composing cardinal direction relations," *Artificial Intelligence*, vol. 152, pp. 143–171, 2004.

[54] N. Simou, V. Tzouvaras, Y. Avrithis, G. Stamou, and S. Kollias, “A Visual Descriptor Ontology for Multimedia Reasoning,” in *In Proc. of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS ’05), Montreux, Switzerland, April 13-15, 2005.*, (Montreux, Switzerland), April 13-15 2005.

[55] “ISO/IEC 15938-3 FCD Information Technology - Multimedia Content Description Interface - Part 3: Visual, March 2001, Singapore,”

[56] T. Adamek, N.O’Connor, and N.Murphy, “Region-based Segmentation of Images Using Syntactic Visual Features,” in *Proc. Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2005*, (Montreux, Switzerland), April 13-15 2005.