

Semantic Processing

Phivos Mylonas

Department of Informatics, Ionian University, Corfu, Greece

Abstract

Semantic processing of multimedia content focuses in principle on the analysis of digital audiovisual content according to its high-level characteristics or entities to be derived in an (semi-)automated manner by suitable computational equipment. The notion of semantics is crucial in the process, since a “correct” machine-processable interpretation will allow for content providers to efficiently manipulate content and provide meaningful services to people in accordance with their individual standards, tastes, and preferences. The overall goal remains, of course, to contribute to the bridging of the gap between the semantic nature of user/people’s needs and raw multimedia content. The herein discussed approach analyzes visual content (such as still images or video sequences) and its associated textual annotation, in order to extract the underlying semantics and construct a meaningful semantic index, based on a unified knowledge model. Content of interest may then be retrieved from the semantics by carrying out semantic interpretation and expansion. All described processes are based on a semantic processing methodology, employing fuzzy algebra and principles of taxonomic knowledge representation, illustrating the semantic unification. As a result, the overall contribution of semantic processing to the improvement of multimedia content understanding and retrieval effectiveness is of great importance for the entire research community.

Keywords: Content analysis; Knowledge management; Multimedia semantics; Visual context.

INTRODUCTION

The task of multimedia content indexing and retrieval has been influenced during the last decade by the important progress in numerous fields, such as digital content production, archiving and standardization, multimedia database management, multimedia signal processing, analysis and coding, computer vision, artificial and computational intelligence, human–computer interaction, and information retrieval. One major obstacle, though, multimedia retrieval systems still need to overcome in order to gain widespread acceptance is the so-called *semantic gap*^[1,2] (Fig. 1).

This refers to the extraction of the semantic content of multimedia entities, the interpretation of user information needs and requests, as well as to the matching between the two. This obstacle becomes even harder when attempting to access vast amounts of multimedia information stored in different audiovisual (a/v) archives and represented in different formats. Among them, digital photographs and video sequences are the most demanding and complex data structures, due to their large amounts of spatiotemporal interrelations; video understanding is a key step toward more efficient manipulation of visual media, presuming semantic information extraction. Current and evolving international standardization activities, such as MPEG-4,^[3] MPEG-7,^[4] and MPEG-21^[5] for video, or JPEG-2000^[6] for still images, deal with aspects

related to a/v content and metadata coding and representation. Syntactic description seems to be well in hand in MPEG-7, but fleshing out the semantic description has not yet received the required attention. It becomes clear among the research community dealing with image processing and content-based retrieval, that the results to be obtained will be ineffective, unless major focus is given to the semantic information level, defining what most users desire to retrieve. Thus, in order to close the loop between the user and available content, the extraction of information at a semantic level is required.

In recent years, several research activities emerged in the direction of *knowledge acquisition and modeling*, capturing knowledge from raw information and multimedia content in distributed repositories to turn poorly structured information into machine-processable knowledge.^[7–11] A second direction is *knowledge sharing and use*, combining semantically enriched information with context, so as to provide inferencing for decision support and collaborative use of trusted knowledge between organizations.^[12,13] Finally, in the *intelligent content* vision, multimedia objects integrate content with metadata and intelligence and learn to interact with devices and networks.^[14] It is becoming apparent in all the above research fields that integration of diverse, heterogeneous, and distributed—preexisting—multimedia content will only be feasible through the design of *mediator systems*. In Biskup et al.,^[15] for

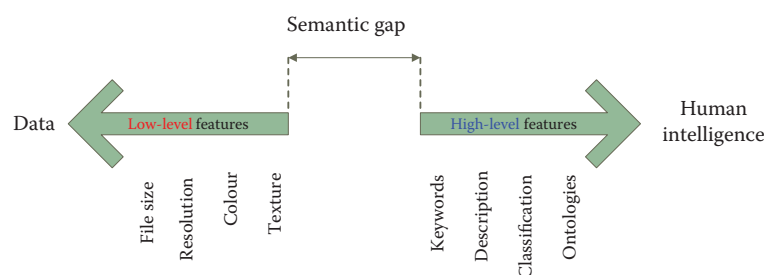


Fig. 1 The semantic gap

instance, a multimedia mediator is designed to provide a well-structured and controlled gateway to multimedia systems, focusing on schemas for semi-structured multimedia items and object-oriented concepts, while Altenschmidt et al.^[16] focuses on security requirements of such mediated information systems. On the other hand, Brink et al.^[17] deals with media abstraction and heterogeneous reasoning through the use of a unified query language for manually generated annotation, again without dealing with content or annotation semantics. A semantically rich retrieval model is suggested in the work of Glöckner and Knoll,^[18] based on fuzzy set theory with domain-specific methods for content analysis and allowing natural language queries. Finally, Cruz and James^[19] focus on the design of a single intuitive interface supporting visual query languages to access distributed multimedia databases.

In principle, the term *semantic processing* refers to a standalone, integrated approach, offering user-friendly and highly informative access to heterogeneous, distributed multimedia (audiovisual) pieces of information (archives). Focusing on a unified semantic analysis of digital multimedia information, as well as of their user-related needs, queries, and profiles, it contributes toward bridging of the gap between the semantic nature of initial user needs and raw multimedia content. As a result, it serves as a mediator between users and audiovisual archives, providing access to a/v content characterized by *semantic phrasing of the request*, *unified handling*, and *personalized response*. The core contribution of this task relies on the fact that it provides the missing link between low-level a/v features and high-level semantics that underlie in video and still images, on the one hand, and the purely semantic needs of users, on the other hand. To achieve this within the framework of semantic processing, we typically retrieve a/v content and associated textual annotation from participating a/v archives and perform visual and textual analysis to extract the underlying semantics and construct a semantic index, based on a unified knowledge model. We may then accept user queries, and, carrying out semantic interpretation and expansion, retrieve a/v content from the index, similarly to traditional text retrieval. Personalized ranking may be also supported, while user profiles are automatically generated and updated by monitoring and analyzing usage history. All above processes are typically based on

a common semantic processing methodology, employing fuzzy algebra and principles of novel taxonomic knowledge representations.

OVERVIEW

Architecture

The general system architecture of a semantic processing framework is briefly depicted in Fig. 2. Connections between its main subsystems are to be identified, that is, a single *user interface* provides a user-friendly access to all participating archives, whereas the *a/v archive interfaces* are responsible for the communication between the main system and each a/v archive. A suitably constructed database is typically used to store the knowledge of the system, the semantic index, and the user profiles. The main framework consists typically of three subsystems: (a) *semantic unification*, (b) *searching*, and (c) *personalization*. The semantic unification subsystem constructs and updates the semantic index, whereas the personalization subsystem updates potential user profiles. The searching subsystem typically analyzes user queries, carries out matching with the semantic index, and returns retrieved content to the end users.

Data Models

Since the above framework is aimed to operate as a mediator between the end user and diverse a/v archives, the mapping of archive content to a uniform data model is of crucial importance. The specification of the model itself is a challenging issue, as it needs to be descriptive enough, to adequately and meaningfully serve user queries, and at the same time, abstract and general enough, to accommodate the mapping of the content of any a/v archive at a semantic level. In the following, we provide the overview of such a data model, consisting mainly out of two components: (a) the *knowledge model* and (b) the *semantic index* (Fig. 3).

Knowledge Model

The knowledge model contains all semantic information used in the system. It supports structured storage of

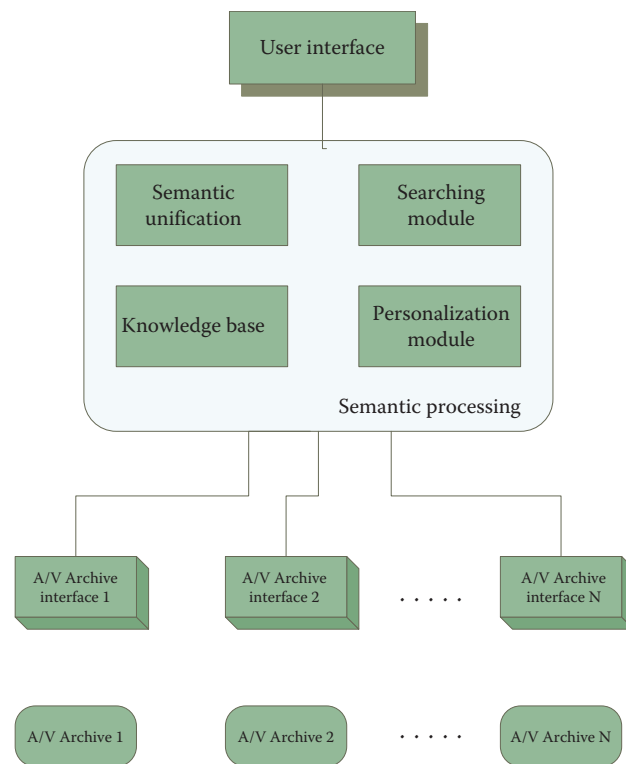


Fig. 2 General system architecture of a semantic processing framework

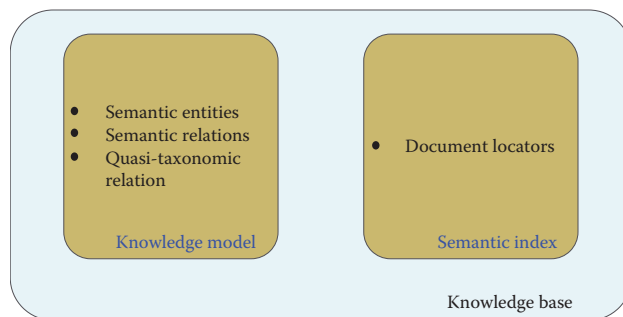


Fig. 3 Overview of the knowledge base data model

semantic entities and relations that experts have defined for indexing and retrieval purposes. Among other actions, it allows expanding a user query by looking for synonyms or related concepts. Three main types of information are introduced in the model:

1. *Semantic entities*: Entities such as thematic categories, objects, events, concepts, agents, and semantic places and times.
2. *Semantic relations*: Relations linking semantic entities, for example, “part of,” “specialization of,” and so on.
3. A *quasi-taxonomic relation*: A taxonomic knowledge representation to interpret the meaning of a multimedia item, composed of several elementary relations, also referred to as a taxonomy.

The knowledge model is manually constructed for a limited application set of specific multimedia content categories using the experts’ assessment.

Semantic Index

The semantic index is used to collect the results of multimedia content analysis in order to support unified access to archives. The index contains sets of information locators for each semantic entities, identifying which semantic entities have been associated to each available multimedia item. It is used for fast and uniform retrieval of content related to the semantic entities specified in, or implied by, the query and the user profile. Content locators associated to index entities may link to complete a/v content, objects, still images, or other video decomposition units that may be contained in the a/v databases.

KNOWLEDGE REPRESENTATION

The theoretical basis on top of which the framework under discussion is constructed derives from careful selection and definition of an appropriate knowledge model. A typical model contains a set of semantic entities and semantic relations between them, which form the basic elements toward semantic interpretation. Through this knowledge representation, a detailed content description of all potential multimedia archives, such as still images, is established in a unified manner. Due to the fact that relations among real-life entities are always a matter of degree, and are, therefore, best modeled using fuzzy relations instead of crisp ones, the best approach followed is based on a formal methodology founded on fuzzy relational algebra and the exploitation of contextual information.^[20]

The Notion of Context

It is common knowledge that the term *context* can take on many meanings and there is, in general, no solid definition that satisfies and covers the many aspects the term is used. The long history of the term appearance and usage varies from diverse areas of computer science to even philosophical and medical approaches.^[21] In the field of computer science, the interest in contextual information is of great importance in artificial intelligence, information retrieval, and image and video analysis.^[22] The nature of the applications in these fields is the one that signifies context, mostly dominated by rapid changes in the user’s context. An indicative example is formed by handheld and ubiquitous computing.^[24] Still, effective use of available contextual information within multimedia applications remains an open and challenging problem, although several researchers have tried to categorize context-aware applications in general according to subjective criteria, thus resulting in classes of applications.^[25]

A fundamental problem tackled via access to and processing of contextual information is the bridging of two fundamental gaps in the literature: the *semantic* and the *sensory gaps*.^[22] As already mentioned earlier, the *semantic gap* forms an issue inherent in most developments of multimedia systems and applications and may be described as the gap between the high-level semantic descriptions humans ascribe to images and the low-level features that machines can automatically parse. Given, for instance, the raw digital image of a wolf (Fig. 4), image analysis may extract feature or vectors (so-called descriptors) that focus on the segmented particles, salient regions, color histograms, etc.^[23] Still, semantic processing of multimedia content includes and may advance research toward steps including (Fig. 5) a prototypical combination of image descriptors, extraction of semantic concepts, and assignment of symbolic names to them, as well as the utilization and exploitation of metadata information and/or identification of higher level entities inter- and intra-relations.^[23]

As broader as the image domain gets, the wider the gap between the feature description and the semantic interpretation is. Narrowing down the domain to a specialized image one results into a smaller gap between features and their semantic, so domain-specific models may help in tackling the problem. However, the latter are not sufficient, but are considered as the first step toward an efficient

approach to the problem. To illustrate all of the above, consider, for example, a picture of a man tossing a blue ball to a dog on the beach (Fig. 6a), which would be “seen” by a vision system as a series of moving color regions or the analysis of a medical image, which would be tackled only as raw numerical medical data (Fig. 6b). Contextual information, such as the relationships between the man, the dog, the location where the ball is being thrown and the significance of this event to the person taking the picture, in the first case, or qualitative information, such as the perspective of the image, or even the age of the associated patient, in the second case, are all gone.

The *sensory gap* is described as the gap between an object and the computer’s ability to sense and describe that object. For example, for some computational systems, a *car* ceases to be a *car* if there is a tree in front of it, effectively dividing the car in two parts from the machine’s perspective. Characteristics different in nature and texture may determine the demands of the search and retrieval methods, for example, possible presence or absence of occlusion, illumination, and clutter. In other words, the sensory gap can be thought as the gap between the object in the world and the information in a computational description derived from a recording of the particular scene. The latter clearly yields uncertainty in what is known about the state of the object and is particularly poignant when a precise knowledge of the capturing conditions is missing. For instance, considering the famous Mona Lisa painting, there might be a sensory gap in the sense of the inability to record the scene due to, for example, too few colors or pixels, too low light conditions, too small memory, or too few frames per second imposed by the hardware (sensor) that attempts the video capture (Fig. 7) prior to the depiction of the captured content to its end user.^[26]

All in all, whenever there is a gap between an object and a computer’s ability to sense and describe the object, the sensory gap is present. And this is also the case when an infinite number of different “signals” can be produced by the same object and different objects can produce similar signals, like it is the case of an image of the same object taken from different viewpoints (Fig. 8). Human perceptual machinery excels at recognizing when different “signal patterns” are the same object or when similar patterns are different ones, but the problem is difficult for computers and clearly lies in their ability to “understand” this process.^[27]

Still, it is contextual knowledge that may enable computational systems to bridge these semantic and sensory gaps. With the advent of all kind of new multimedia-enabled devices and multimedia-based systems, new opportunities arise to infer the media semantics. Contextual metadata are capable of playing the important role of a “semantic mediator.” Toward that scope, two aspects of context seem to have special salience in most multimedia applications: *where* and *when*, that is, *spatial* and *temporal* context. By taking into account the spatial and to lesser extend the


Descriptors <i>feature-vectors</i>	Segmented blobs, Salient regions, Pixel-level histograms, Fourier descriptors, etc...
Raw Media <i>images</i>	

Fig. 4 Raw media and primitive image analysis example

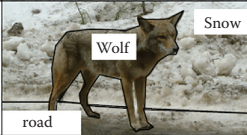
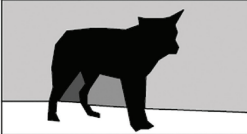
Semantics <i>object relationships and more</i>	Wolf on Road with Snow on Roadside in Yosemite National Park, California on 24/1/2004 at 23:19:11GMT
Object Labels <i>symbolic names of objects</i>	
Objects <i>prototypical combinations of descriptors</i>	

Fig. 5 Semantic processing of the previous example

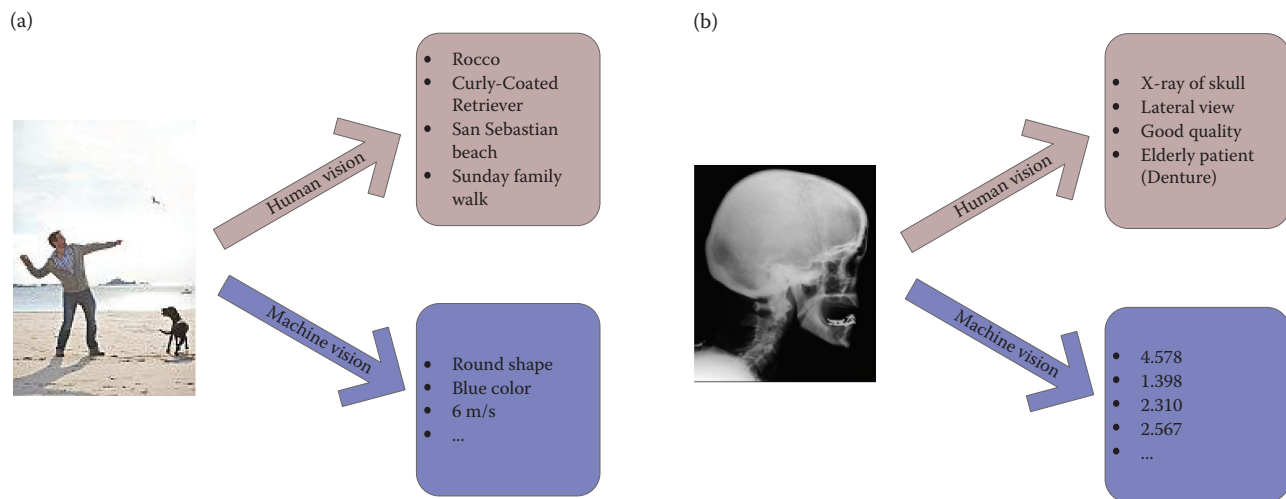


Fig. 6 Human versus machine vision interpretation: (a) a man tossing a ball to a dog; (b) a medical image of a human skull



Fig. 7 The sensory gap—limitations imposed by sensors



Fig. 8 The sensory gap—limitations imposed by different viewpoints

temporal context, we are able to gather and incorporate user interaction to adjust and add the extra information needed for interpretation. It is common knowledge though that context itself appears in various forms and modifications and researchers commonly emphasize distinctions between different types of context. Each context type illustrates different aspects of it and consequently has very little in common with the others.^[28] Depending on the specific objectives of the task at hand, different formalizations of

what is meant by context have been developed. For modern knowledge-based multimedia systems, context has a rather precise meaning and usage. In this framework, a constant enhancement of offered capabilities and functionalities is introduced, based on the always increasing contextual aspect of information provided by their end users. Context is exploited in pursue of a more efficient personalized approach, bringing the end user in the center of the application's interest.

Visual Context Identification

In general, the task of suitable visual context definition and identification is very important in the integration of a multimedia content-based semantic processing system. This is mainly due to the fact that all knowledge required for multimedia content analysis is thought to be context-sensitive, resulting in a specific need for formal definitions of context structures, prior to any static or dynamic context detection and analysis. The first objective formed within this task is the definition of the suitable aspect of context at hand, providing conceptual and audiovisual information. The latter will be used for context tailoring in several stages through the life cycle of the system's content, such as creation and analysis, consumption, and user interaction.

We may identify two related types of context with respect to its usage and applicability within multimedia systems, namely, the *context of content analysis* and the *context of use*. *Context of content analysis* refers to the context during the phase of analysis, including tasks such as knowledge-assisted analysis and reasoning. It is intended to be used to aid the extraction of semantic metadata both at the level of simple concepts and at the level of composite events and higher level concepts. In this manner, it forms the main employment of visual context in the semantic processing framework, for example, by performing scene classification to detect whether an image or video clip represents *city* or *landscape* content (Fig. 9), essentially aiding the analysis process to detect and recognize specific concepts or objects in the content. On the other hand, *context of use* is related to the use of content by a system's application modules, such as search/retrieval and personalization. In this case, given the multimedia content and metadata, contextual information from an external source are provided, consisting mainly of information about the particular user, network and client device, and so on. Ultimately, improved design and development of these applications is achieved, including retrieval, search, browsing, sharing, and management of content.

In multimedia computing applications, the aspects of context, that are thought to be the most suitable and appropriate for research and progress, are the ones of visual context described above. Therefore, from now on they may

be presented under a common approach, summarized in the notion of *visual context*. Visual context forms a rather classical approach to context, tackling it from the scope of environmental or physical parameters in multimedia applications. Different architectures, conceptual approaches, and models support dynamic and adaptive modeling of visual context. One of the main objectives in the field is the combination of context parameters extracted from low-level visual features with higher level concepts, like fuzzy set theory, to support reasoning. Specifically, the context description supports fuzziness in order to face the uncertainty introduced by content analysis or the lack of knowledge. This context representation also supports audiovisual information (e.g., lighting conditions and information about the environment) and is separately handled by visual context models. The second objective is visual context analysis, that is, to take into account the extracted/recognized concepts during content analysis in order to find the specific context, express it in a structural description form, and use it for improving or continuing the content analysis, indexing, and searching procedures, as well as for content personalization purposes.

In terms of *knowledge-assisted content analysis and processing*, a set of core functionalities of the multimedia application is defined, regarding the way such a system is expected to execute knowledge-assisted image analysis functions automatically or in a supervised mode, either to detect or to recognize parts of content. Additionally, it is thought to generate or assist end users classify their contents and metadata, through suggestions or sorting being performed in a sophisticated way, making quite naturally implicit use of context analysis functionalities. For example, in a face recognition scenario, visual clues help the system detect the right person. Issues relating more to the automatic creation of metadata even after analysis, for example, through inference, make use of context as different sources of information (different analysis modules, textual inputs) are also integrated.

As far as *retrieval* is concerned, a set of core functionalities of a multimedia search and retrieval system may also be defined; there are many distinct aspects suggested and commented by users, regarding the way of performing searches, the type of searches they expect to have and



Fig. 9 City/landscape classification

the constraints they imagine. Organizing multimedia data into meaningful categories marked by end users as being important could, for example, exploit contextual information. Additionally, several use case scenarios illustrate a system's capability of learning a user's interest in order to adapt its behavior to the assumed interests, taking into account *personalization* aspects. With respect to the learning functionality, users are in general concerned about privacy, while the adaptive system behavior can also be commented in terms of controllability. Retrieval is strongly related to context, when tackling textual query analysis, search by semantic, visual, or metadata similarity, semantic grouping, browsing and rendering of retrieved content, personalization, and relevance feedback. However, user browsing capabilities, together with retrieval capabilities, suppose detection of common metadata, which is not related to visual context. Another form of context, dealing mostly with the semantic part of the analysis would be more useful in this case.^[29] In any case, search by visual similarity may benefit by the use of visual context information, as in the cases of scene classification and object detection.

It should have been clear by now that visual context can be clearly exploited in *multimedia content processing*. During the phase of image metadata generation from one or more content analysis application modules, scene classification and/or object detection techniques can be of use, providing the necessary contextual information, for example, information about indoor/outdoor scenery at the metadata level. The same applies to knowledge-assisted image/video analysis and metadata generation; techniques and methodologies can be helpful in implementing scene classification and object detection. Moreover, when classifying and sorting content, degrees of confidence can be obtained by taking into consideration contextual information. The latter is useful for extraction of the semantics of content and detection of repeated content. It can help deal with identification of semantically similar content, as well as analyze relative metadata information stored or derived automatically from images. All aspects of the so far presented contextual information assist toward transparency and automation of knowledge extraction, providing the means to gather additional meaningful information in ontology mapping, image/video analysis, and metadata generation processes.

Visual Context in Image Analysis

By visual context in the sequel, we shall refer to all information related to the visual scene content of a still image (or video sequence) that may be useful for its analysis. Image analysis deals with a few well-known research problems shortly presented in the following, whereas visual context is mostly related to two of the problems in image analysis, namely, *scene classification* and *object detection* (Fig. 10). *Scene classification* forms a top-down approach,

where low-level visual features are employed to globally analyze the scene content and classify it in one of a number of predefined categories, for example, indoor/outdoor, city/landscape, and so on. On the other hand, *object detection/recognition* is a bottom-up approach that focuses on local analysis to detect and recognize specific objects in limited regions of an image, without explicit knowledge of the surrounding context, for example, recognize a building or a tree. These two major fields of image analysis actually comprise a chicken-and-egg problem as, for instance, detection of a building in the middle of an image might imply a picture of a city with a high probability, whereas pre-classification of the picture as "city" would favor the recognition of a building versus a tree. Solution to the above problem can be dealt through modeling of visual concept descriptors in one or more domain ontologies and ontology learning/visual concept detection techniques.

Another topic in the field of image processing is the automatic detection of important or interesting regions in an image; topic that has been tackled by a number of researchers over the past 20 years.^[30,31] An example methodology is illustrated in the work of Milanese,^[32] where a computational model of *visual attention* by combining knowledge about the human visual system with computer vision techniques is developed. The notion of region segmentation is aimed at generating regions of homogeneous properties such as color and texture, which are the basic units and also an efficient intermediate representation of the scene, for higher level reasoning and interpretation. This topic is considered in general as something that a human observer performs with relative ease, but at the same time it is difficult for an automated system to understand and make higher level analysis tractable. Moreover, *region-based segmentation* aids significantly in the process of reasoning on spatial relationships; a task that becomes meaningful and useful this way. Of course, the segmented regions need further processing in order to be able to provide semantically meaningful information. In this scope they need to be classified, if applicable, into semantic object classes that are encountered frequently in

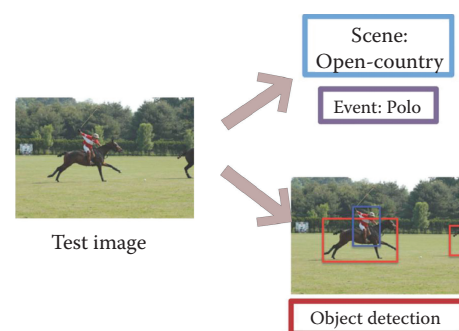


Fig. 10 Scene classification versus object detection tasks: A single test image is being used to either identify an open-country scene and a polo event or to detect interesting objects like the human rider and the horses

images, such as sky, cloud, grass, and tree. The purpose is to provide information about the existence of a few typical semantic object classes, which tend to belong either to the foreground or background, as well as an estimate of the scene category. The latter proves to be very useful for triggering top-down reasoning.

In *content-based image search and retrieval*, more and more researchers are looking beyond low-level color, texture, and shape features in pursuit of more effective searching methods. Natural object detection in indoor or outdoor scenes, that is, identifying key object types such as sky, grass, foliage, water, and snow, can facilitate content-based applications, ranging from image enhancement to coding or other multimedia applications. However, a significant number of misclassifications usually occur because of the similarities in color and texture characteristics of various object types and the lack of context information, which is a major limitation of individual object detectors. Toward the solution to the latter problem, an interesting approach is the one presented in the work of Luo et al.^[33] A spatial context-aware object-detection system is proposed, initially combining the output of individual object detectors in order to produce a composite belief vector for the objects potentially present in an image. Subsequently, spatial context constraints, in the form of probability density functions obtained by learning, are used to reduce misclassification by constraining the beliefs to conform to the spatial context models.

Other attempts in the area include the one proposed in the work of Naphade and Huang,^[34] where a list of semantic objects, including sky, snow, rock, water, and forest, is used in a framework for semantic indexing and retrieval of video. As already expected, color has been one of the central features of existing work on natural object detection. For example, in the work of Saber et al.,^[35] *color classification* is utilized in order to detect sky. In the context of content-based image retrieval, Smith and Li^[36] assumed that a blue extended patch at the top of an image is likely to represent clear sky. An exemplar-based approach is presented more recently that uses a combination of color and texture features to classify subblocks in an outdoor scene as sky or vegetation, assuming correct image orientation.^[38] The latter brings up the issue of utilizing context orientation information in object class detection algorithms, a task that is generally avoided due to the fact that such contextual information is not always available and the performance of the algorithms is more than adequate despite this shortcoming.

So far, none of the above methods and techniques utilize context in any form. This tends to be the main drawback of these individual object detectors, since they only examine isolated strips of pure object materials, without taking into consideration the context of the scene or individual objects themselves. This is very important and also extremely challenging even for human observers. The notion of visual context is able to aid in the direction of natural object

detection methodologies, simulating the human approach to similar problems. Many object materials can have the same appearance in terms of color and texture, while the same object may have different appearances under different imaging conditions (e.g., lighting, magnification). However, one important trait of humans is that they examine all the objects in the scene before making a final decision on the identity of individual objects. The use of visual context forms the key for this unambiguous recognition process, as it refers to the relationships among the location of different objects in the scene. In this manner, it is useful in many cases to reduce the ambiguity among conflicting detectors and eliminate improbable spatial configurations in object detection. As already discussed, visual context may be either spatial or temporal; *spatial context* is associated with spatial relationships between objects or regions in a still image or video sequence, whereas *temporal context* is associated with temporal relationships between objects, regions, or scenes in the case of a video sequence. In the sequel, discussion will be restricted to spatial context. One can identify two types of spatial contextual relationships that exist in natural images: (a) relationships that exist between co-occurrence of objects in natural images and (b) relationships that exist between spatial locations of certain objects in an image (Fig. 11).^[37]

The definition of spatial context is an important issue for the notion of visual context in general. In order to be able to use context in applications, a mechanism to sense the current context—when thought as location, identities of nearby people or objects and changes to those objects—and deliver it to the application is crucial and must be present. A significant distinction exists between methods trying to determine location in computing applications and research fields. Most of the existing approaches tend to restrict themselves, trying to infer the location where the image was taken (i.e., camera location); inferring the location of what the image was taken of (i.e., image content location) is a rather difficult and more complex task tackled by much less approaches.^[39] In the work of Davis,^[22] this challenge is addressed by leveraging regularities in a given user's and in a community of users' photo-taking behaviors. Suitable weights, based on past experience and intuition, are chosen in order to assist in the process of location-determining features and then adjusted through a process of trial and error. An example describing the notion behind the method considers the following: it seems rather intuitive that if two pictures are being taken in the same location within a certain time frame (e.g., a few minutes for pedestrian users), they are probably in or around the same location.

Another factor to be considered is the intersection of spatial (and temporal) metadata in determining the location of image content. For example, patterns of being in certain locations at certain times with certain people will help determine the probability of which building in an area a user might be in. Information on whether this particular

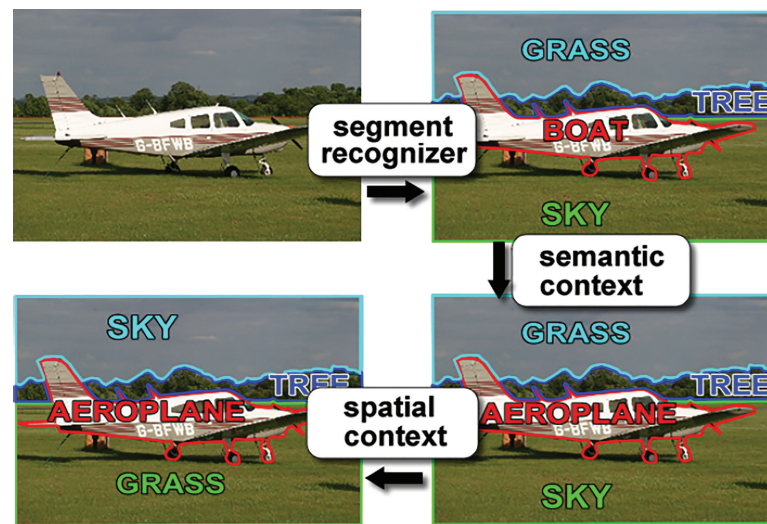


Fig. 11 Utilization of spatial context in semantic processing

building is the place he/she works in can also be derived in such a case. Rule-based constraint and inference engines can also be used to aid reasoning, as well as machine learning algorithms to learn from past performance to optimize and adjust the relative importance of the various location-determining features. Taking the process a step further into the field of context modeling transforms the problem into how to represent contextual information in a way that can help bridging the gap between applications using contextual information and the deployment of context-aware services. The development of such applications requires tools that are based on clearly defined context models. A simple approach is to use a plain model with context being maintained by a set of environment variables. Of course, visual context information can also come from an overall description of the whole scene. In that case, we are referring to the so-called *scene context*. In a number of studies, the context provided by a real-world scene has been claimed to have a mandatory, perceptual effect on the identification of individual objects in such a scene. This claim has provided a basis for challenging widely accepted data-driven models of visual perception. The so far discussed visual context, defined by the normal relationships among the locations of different materials in the scene without knowing exactly what the scene type is, is referred to as *spatial context*, and it is the one that is going to be used mostly in a multimedia system application. In the sequel, visual context analysis is discussed in relation to the problems of scene classification and object detection. With the increase in the number and size of digital archives and libraries, there is a need for automated, flexible, and reliable image search and retrieval algorithms, as well as for image and video database indexing. *Scene classification* provides solutions in the means of suitable applications for the problem. The ultimate goal is to classify scenes based on their content. However, scene classification remains a major open challenge. Most solutions

proposed so far, such as those based on color histograms and local texture statistics,^[40,41] lack the ability to capture a scene's global configuration, which is critical in perceptual judgments of scene similarity.

Initiatives have been taken in the field, whose main features can be summarized in the use of qualitative spatial and photometric relationships within and across regions. The emphasis on such qualitative measures leads to enhanced generalization abilities that are critical in achieving better coherence and efficiency for the final output/application. However, these similarity measures are rather inadequate for the problem at hand, if not combined together with additional information from other sources. And that is so, mainly because the previously defined similarity measures often produce results incongruent with human expectations, if the goal is to find images from a given object/scene class, such as snowy mountains or waterfalls. For example (Fig. 12), using color histograms to find the most similar images to a water scene at sunset could possibly return pictures of money, molten liquids, or even a watermelon! Obviously, all these images have the same overall gold color, although they differ in great degree in their semantic content.

Common standard approaches to object detection usually look at local pieces of the image in isolation when deciding if the object is present or not at a particular location. Of course, this is suboptimal and can be easily illustrated in the following example: consider the problem of finding a table in an office. A table is typically covered with other objects; indeed almost none of the table itself may be visible, and the parts that may be visible, such as its edge, are fairly generic features that may occur in many images. However, the table can be identified using contextual cues of various kinds. Of course, this problem is not restricted to tables or occluded objects: almost any object, when seen at a large enough distance, becomes impossible to recognize without using

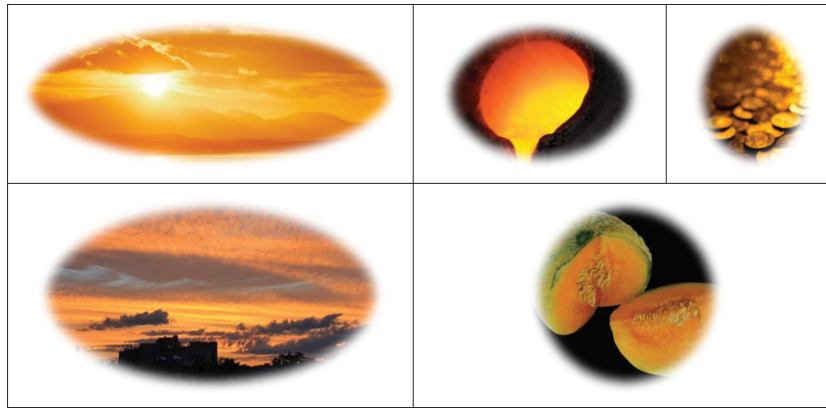


Fig. 12 All five images share a similar color histogram

context. Without doubt, there is a large amount of work in the field of *object detection and recognition* as well. However, techniques utilized in the field have usually positive results only in case of objects that have well-defined boundaries. Consequently, such strategies are not well suited for complex scenes, especially those that consist mostly of natural objects. The main difference between scene classification and object recognition techniques relies in the latter statement. Given these difficulties inherent in individual object recognition, scene classification approaches usually classify scenes without first attempting to recognize their components. This kind of strategy is also supported by psychophysical evidence showing that humans may holistically classify visual stimuli before recognizing the individual parts.^[42] Also, efforts have been made in using scene classification to facilitate object detection and vice versa.^[43] In general, scene classification methodology is characterized by the following three principles:

- The significance of the global scene configuration
- The use of qualitative measurements
- The sufficiency of low spatial frequency information

Of course, careful work has to be done to define what images might be indoor, what images might be outdoor,

and what kind of images might not be classified, notably close-up scenes, portraits, faces, animals, flowers, etc. The most addressed semantic image classification problem in the literature is the one of *indoor/outdoor classification*. One domain where automatic image classification is viable is personal photo collections, where specific categories of indoor or outdoor scenes can be identified (e.g., see Fig. 13).

The main features which are important in scene classification are *color* and *texture* and to a lesser extent, *shape*. Taking color into consideration, local color descriptors are in general detected in a straightforward manner. Clues such as blue sky, green plants, or red-tinted indoor scenes are the most commonly used.^[44] Furthermore, faster and more efficient global color histograms are also utilized toward that goal.^[45] In addition to color, texture and shape features are also of great importance. These features are capable of detecting regions with typical high frequencies in outdoor images, such as grass, leaves, sky, and water, and with typical vertical structures in indoor scenes, like wall corners or furniture, respectively. *City/landscape classification* (Fig. 9) is stated to be easier than indoor/outdoor classification.^[44] And that is so because city images are mainly characterized by continuous vertical and horizontal edges, whereas outdoor images are typically dominated by short edges in all directions. Usually,



Fig. 13 Indoor/outdoor classification

either polygonized edges of minimal length or direction coherence measures are used in order to take into account the presence of line-shaped, straight, continuous edges in city images.

Now, as already discussed in this section, several approaches of analyzing the content of images exist in the literature and many aspects of context are identified aiding in the process of image analysis. The main task of related research work and one of the main goals in the field is the effective combination of local and global information, toward implementing robust methods to use in typical image processing problems and techniques. It should be clear by now that visual context can play a key role in the procedure of combining this information; context should actually stand in the middle, being able to handle both types of information and providing the means to achieve better coherence and reliable research results. In order to achieve the latter, appropriate visual context models should be selected and designed in a straightforward and productive manner, utilizing the variations of the particular aspects of visual context; in the following sections we shall focus on such a model.

Fuzzy Taxonomic Relations

It is common knowledge that retrieval systems based on terms suffer from the problematic mapping of terms to semantic entities.^[46] Specifically, as more than one term may be associated to the same entity and more than one entity may be associated to the same term, the processing of query and index information is not trivial. In order to overcome such problems, one should work directly with *semantic entities*, rather than terms. In the sequel, we will denote by $S = \{s_1, s_2, \dots, s_n\}$, the set of semantic entities that are known. A knowledge representation model may consist of the definitions of these semantic entities, together with their textual descriptions, that is, their corresponding terms, as well as a set of *relations* among the semantic entities. The objective is to construct a model in which the context determines the intended meaning of each word, and a word used in different context may have different meanings. An initial formal definition of such a model may be given as follows:

$$M = \{S, \{R_i\}\}, \quad i = 1 \dots n$$

$$R_i : S \times S \rightarrow \{0,1\}, \quad i = 1 \dots n$$

where M is the knowledge model and R_i is the i th relation among the semantic entities. Although any type of relation may be included, the two main categories are *taxonomic* (i.e., *ordering*) and *compatibility* (i.e., *symmetric*) relations. Compatibility relations fail to assist in the determination of the context of a query; the use of ordering relations

is necessary for such tasks.^[1] Thus, a main challenge is the meaningful exploitation of information contained in taxonomic relations.

In addition, a knowledge model, in order to be highly descriptive, needs to contain a large number of distinct and diverse in relations among semantic entities. As a result, available information will be divided among them, making each one of them inadequate to fully describe a context. Thus, more than one such relation may need to be combined to provide a view of the knowledge that suffices for context definition and estimation. In order to overcome such problems, fuzzy semantic relations have been proposed for the modeling of real life information.^[1] In particular, several commonly encountered relations that can be modeled as *fuzzy ordering relations* can be combined for the generation of a meaningful, fuzzy, quasi-taxonomic relation. More formally, a new knowledge model M_F is thus constructed, denoting the fuzziness in the approach in comparison to the knowledge model presented above and summarized in the following:

$$M_F = \{S, \{R_i\}\}, \quad i = 1 \dots n$$

$$r_i = F(R_i) : S \times S \rightarrow [0,1], \quad i = 1 \dots n$$

where F denotes the fuzzification of the relations R_i . The existence of many relations has led to the need for utilization of more relations for the generation of an adequate *taxonomic relation* T . Based on the relations r_i , we construct the following relation:

$$T = Tr^t(\bigcup_i r_i^{p_i}), \quad p_i \in \{-1,1\}, \quad i = 1 \dots n$$

where $Tr^t(A)$ is the sup- t transitive closure of some relation A , and the role of p_i is depicted by the specific definition of each relation used in the construction of T . Depending on the semantics of the relation definition (e.g., order of arguments a, b in Table 1), some relations may need to be inverted before being used in the construction of T . The transitivity of relation T , a required property in order for it to be taxonomic, was not implied by the above definition as the union of transitive relations is not necessarily transitive. For the purpose of analyzing multimedia content descriptions, relation T may be generated with the use of the following fuzzy taxonomic relations, whose semantics are defined in MPEG-7 and summarized in Table 1.

Based on the above fuzzy relations, T is a new semantic relation that is calculated as follows:^[47]

$$T = Tr^t(Sp \cup P^{-1} \cup Ins \cup Pr^{-1} \cup Pat \cup L \cup Ex)$$

Based on the semantics of the participating relations, it is easy to see that T is ideal for the determination of the

Table 1 Fuzzy taxonomic relations used for generation of T

Name	Symbol	Meaning	Example	
			a	b
Part	$P(a, b)$	b is a part of a	Human body	Hand
Specialization	$Sp(a, b)$	a is a generalization of b	Vehicle	Car
Example	$Ex(a, b)$	b is an example of a	Player	Jordan
Instrument	$Ins(a, b)$	b is an instrument of a	Music	Drums
Location	$L(a, b)$	b is the location of a	Concert	Stage
Patient	$Pat(a, b)$	b is a patient of a	Course	Student
Property	$Pr(a, b)$	b is a property of a	Jordan	Star

topics that an entity may be related to, as well as for the estimation of the common meaning, that is, the context, of a set of entities. All relations used for the generation of T are partial ordering relations. Still, there is no evidence that their union is also antisymmetric, a property which is required for it to be taxonomic. Quite the contrary, T may vary from being a partial ordering to being an equivalence relation. This is an important observation, as true semantic relations also fit in this range (total symmetry, as well as total antisymmetry often have to be abandoned when modeling real life). Still, the semantics of the used relations indicate that T is very close to anti-symmetric. Therefore, we categorize it as quasi-ordering or *quasi-taxonomic*.

Taxonomic Context Model

When using a taxonomic knowledge representation to interpret the meaning of a multimedia entity, it is the context of a term that provides its truly intended meaning. In other words, the true source of information is the co-occurrence of certain entities and not each one independently. Thus, the common meaning of terms should be used in order to best determine the entities to which they should be mapped. We will refer to this as their *taxonomic context*; in general, term *context* refers to whatever is common among a set of entities. Relation T will be used for the detection of the context of a set of entities, as explained in the remaining of this subsection.

The fact that relation T is (almost) an ordering relation allows us to use it in order to define, extract, and use the context of a set of entities in general. Relying on the semantics of relation T , we define the context $K(s)$ of a single entity $s \in S$ as the set of its antecedents in relation T , where S is the set of all entities. More formally, $K(s) = T(s)$, following the standard superset/subset notation from fuzzy relational algebra. Assuming that a set of entities $A \subseteq S$ is crisp, that is, that all considered entities belong to the set

with degree one, the context of the set, which is again a set of entities, can be defined simply as the set of their common antecedents:

$$K(A) = \bigcap_i K(s_i), \quad s_i \in A$$

Obviously, as more entities are considered, the context becomes narrower, that is, it contains less entities and to smaller degrees, as illustrated in Fig. 14: $A \supset B \rightarrow K(A) \subseteq K(B)$.

When the definition of context is extended to the case of fuzzy sets of entities, this property must still hold. Moreover, we demand that the following are satisfied as well, basically because of the nature of fuzzy sets:

- $A(s) = 0 \Rightarrow K(A) = K(A - \{s\})$, that is, no narrowing of context.
- $A(s) = 1 \Rightarrow K(A) \subseteq K(s)$, that is, full narrowing of context.
- $K(A)$ decreases monotonically with respect to $A(s)$.

Taking this into consideration, we demand that, when A is a normal fuzzy set, the “considered” context $K(s)$ of s , that is, the entity’s context when taking its degree of participation to the set into account, is low when the degree of participation $A(s)$ is high or when the context of the crisp entity $K(s)$ is low.

Therefore, $cp(K(s)) = cp(K(s)) \cap (S \cdot A(s))$, where cp is an involutive fuzzy complement and $S \cdot A(s)$ is a fuzzy set defined as $[S \cdot A(s)]_{(x)} = A(s) \forall x \in S$. By applying De Morgan’s law, we obtain the following:

$$\kappa(s) = K(s) \cup cp(S \cdot A(s))$$

Then the set’s context is easily calculated as follows:

$$K(A) = \bigcap_i \kappa(s_i), \quad s_i \in A$$

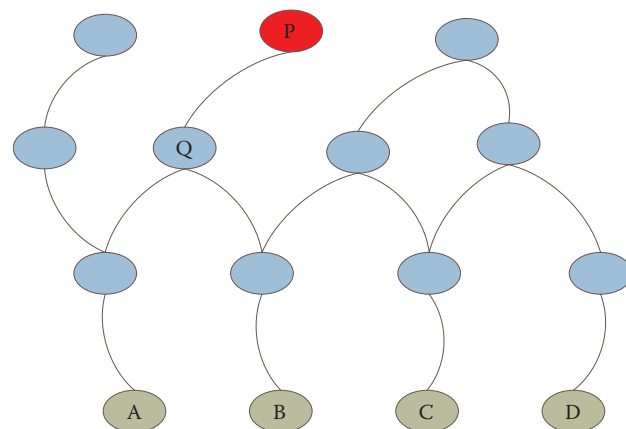


Fig. 14 As more entities are considered, the context it contains less entities and to smaller degrees: Considering only the first two leaves from the left (A, B), the context contains two entities (P, Q), whereas considering all the leaves (A, B, C) narrows the context to just one common ascendant (P)

Considering the semantics of the T relation and the process of context determination, it is easy to realize that when the entities in a set are highly related to a common meaning, the context will have high degrees of membership for the entities that represent this common meaning. Therefore, the height of the context $h(K(A))$ may be used as a measure of the semantic correlation of entities in set A . We refer to this measure as *intensity* of the context. The intensity of the context also represents the degree of relevance of the entities in the set.

SEMANTIC INDEXING

One of the main aspects of semantic processing tasks relies in the achievement of unified access to a/v content. The latter is accomplished through mapping of all a/v content and metadata to a semantically unified index, which is then used to serve user queries. In principle, the semantic index is constantly enriched and adapted to archive content changes. The main novelty introduced by this process is the fact that it allows an integrated detection of concepts both in multimedia content (i.e., images and videos) and text (i.e., in the form of metadata). Additionally, it sets the basis for efficient, intelligent clustering methods, for example, for the materialization of automatic thematic categorization of multimedia entities, using an appropriate knowledge model and the notions of semantic entities/relations and hierarchical contextual taxonomy. Thus, visual content analysis directly maps raw multimedia video to high-level semantic entities, integrating several algorithms and techniques toward an efficient image processing.

Detection of Semantic Entities

Very different strategies may be followed in order to detect the sought semantic entities. In general, techniques for semantic entities detection can be divided into two major fields: detection of semantic entities when dealing with multimedia content and textual semantic entities extraction when tackling text documents. In both cases, different content (e.g., digital images vs. text) is considered and as a result different analysis techniques are required, however, a uniform approach to their semantic handling is followed and same representations are obtained. The role of this semantic unification is to correlate the multimedia content descriptions provided by the a/v archives with the semantic entities stored in the knowledge base, so that user queries can be issued and handled at a semantic level. The result contains the correlations between multimedia content items and semantic entities. The built-in knowledge about entities permits robustness and uncertainty handling, whereas updating of knowledge ensures adaptation to environmental changes.

Visual Content Analysis

In a text document, the topic and semantics of the content are most often explicitly specified or at least contained within the document in a textual form. In a digital image, on the other hand, the entities to be indexed are not directly encountered in the image; recognizable features must be extracted and matched to the ones found in the knowledge base. Furthermore, abstract concepts, such as “sports” or “arts,” are not directly encountered and must be inferred from concrete objects and events, as well as features that are not attributed to a particular object or event, such as light.^[48] There are several options to implement this, even in the more complex case of video sequences, such as shot detection, key-frame extraction, and object localization and tracking techniques, so as to prepare a meaningful syntactic description of multimedia content.^[49] In this case, following such video preprocessing techniques, we may achieve a hierarchical, spatiotemporal partition of the video sequence into meaningful time entities or shots. Thus, detected objects are then linked together to form adjacency graphs, that are then matched to the so-called *description graphs*, that model, in principal, a complex semantic entity in a combination of simpler ones^[50] and are stored among objects, events, and other entities in the knowledge of the system. In this entry, we shall not examine more closely these problems encountered in visual content analysis, since it is not focusing on image processing and is more oriented toward the semantic interpretation of multimedia content.

Semantic Annotation Interpretation

The result of all (pre-)processing techniques mentioned in the previous section, including shot partitioning, key-frame extraction, and object detection and tracking, may be extended and integrated in a useful digital content annotation application; its annotation will have been semantically interpreted, that is, mapped to semantic entities and will be stored again within the semantic index. During this process, a query may be issued to each a/v archive for all content items that have not been indexed or whose description has been updated. The textual annotation contained in the MPEG-7 compliant description of each such item may be analyzed and semantic entities may be identified through matching with their definitions in the knowledge model. Links between detected semantic entities and the item in question may then be added to the index; weights may also be added depending on the location of each entity in the description and the degree of entity’s matching.

CONCLUSIONS

The core contribution of *semantic processing* in image processing tasks is the provision of uniform access to

heterogeneous a/v archives. This is accomplished by mapping all a/v content and metadata to a semantic index used to serve user queries, based on a common underlying knowledge model. A key aspect in these developments has been the exploitation of semantic metadata. In the following years, multimedia content management tasks are going to become even more complex. As it becomes obvious day by day, multimedia content itself will soon be a commodity, making the use of semantic metadata essential. Content providers will have to understand the benefits obtained from the systematic generation and exploitation of semantic information; service providers will have to accept them as the basis on which to build new services; and the producers of software tools for end users will redirect their imagination toward more appropriate integration of application software with content, taking advantage of semantic metadata. These developments clearly present some challenging prospects in technological, economic, standardization, and business terms and constitute semantic interpretation of multimedia content a key aspect of the conducted research activities.

REFERENCES

- Dorai, C.; Venkatesh, S. Computational media aesthetics: Finding meaning beautiful. *IEEE Multimedia* **2001**, 8 (4), 10–12.
- Smeulders, A.W.M.; Worring, M.; Santini, S.; Gupta, A.; Jain, R. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, 22, 1349–1380.
- Battista, S.; Casalino, F.; Lande, C. MPEG-4: A multimedia standard for the third millenium Part 2. *IEEE Multimedia* **2000**, 7 (1), 76–84.
- Sikora, T. The MPEG-7 visual standard for content description—An overview. *IEEE Trans. Circuits Syst. Video Technol. Special issue on MPEG-7*, **2001**, 11 (6), 696–702.
- MPEG-21 Overview v.3, ISO/IEC JTC1/SC29/WG11 N4511, December 2001.
- ISO/IEC JTC1/SC29/WG1 N1646R: JPEG 2000 Part I Final Committee Draft Version 1.0, 2000.
- Ramesh Naphade, M.; Kozintsev, I.V.; Huang, T.S. A factor graph framework for semantic video indexing. *IEEE Trans. Circuits Syst. Video Technol.* **2002**, 12 (1), 40–52.
- Mich, O.; Brunelli, R.; Modena, C.M. A survey on video indexing. *J. Visual Commun. Image Represent.* **1999**, 10, 78–112.
- Yang, Y.; Zha, Z.-J.; Gao, Y.; Zhu, X.; Chua, T.-S. Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Trans. Multimedia* **2014**, 16 (6), 1677–1689.
- Naphide, H.R.; Huang, T.S. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. Multimedia* **2001**, 3 (1), 141–151.
- Jun, W.; Worring, M. Efficient genre-specific semantic video indexing. *IEEE Trans. Multimedia* **2012**, 14 (2), 291–302.
- Euzenat, J.; Fensel, D.; Lara, R.; Gómez-Pérez, A. Knowledge web: Realising the semantic web, all the way to knowledge enhanced multimedia documents. In *Proceedings of European Workshop on the Integration of Knowledge; Semantics and Digital Media Technology (EWIMT)*: London, UK, 2004, 25–26.
- Nikolopoulos, S. Semantic Multimedia Analysis Using Knowledge and Context. PhD Thesis, Queen Mary University of London, 2012. Available at <http://qmro.qmul.ac.uk/jspui/handle/123456789/3148>.
- Kompatsiaris, I.; Avrithis, Y.; Hobson, P.; Strinzis, M.G. Integrating knowledge, semantics and content for user-centred intelligent media services: The aceMedia project. In *Proceedings of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*; Lisboa, Portugal, 2004, 21–23.
- Biskup, J.; Freitag, Y.; Karabulut, Sprick, B. A mediator for multimedia systems. In *Proceedings 3rd International Workshop on Multimedia Information Systems*; Como, Italy, 1997.
- Altenschmidt, C.; Biskup, J.; Flegel, U.; Karabulut, Y. Secure mediation: Requirements, design, and architecture. *J. Comput. Secur.* **2003**, 11 (3), 365–398.
- Brink, A.; Marcus, S.; Subrahmanian, V. Heterogeneous multimedia reasoning. *IEEE Comput.* **1995**, 28 (9), 33–39.
- Glöckner, I.; Knoll, A. Natural language navigation in multimedia archives: An integrated approach. In *Proceedings of the 7th ACM International Conference on Multimedia*; Orlando, FL, 1999, 313–322.
- Cruz, I.; James, K. A user-centered interface for querying distributed multimedia databases. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*; Philadelphia, PA, 1999, 590–592.
- Klir, G.; Yuan, B. *Fuzzy Sets and Fuzzy Logic, Theory and Applications*; Prentice Hall: Upper Saddle River, NJ, 1995.
- Winograd, T. Architectures for context. *Hum.-Comput. Interact.* **2001**, 16 (2), 401–419.
- Davis, M.; Good, N.; Sarvas, R. From context to content: Leveraging context for mobile media metadata. In *Workshop on Context Awareness at the Second International Conference on Mobile Systems, Applications, and Services (MobiSys 2004)*, Boston, MA, 2004.
- Raieli, R. *Multimedia Information Retrieval: Theory and Techniques*, Chandos Information Professional Series; Chandos Publishing: New Delhi, India, 2013.
- Weiser, M. Some computer science issues in ubiquitous computing. Special Issue, *Comput.-Augmented Environ. CACM* **1993**, 36 (7), 74–83.
- Schilit, B.; Adams, N.; Want, R. Context-aware computing applications. In *Proceedings of IEEE Workshop on Mobile Computing Systems and Applications*, Palo Alto Research Center: Santa Cruz, CA, 1994.
- Lux, M. Visual Information Retrieval. Klagenfurt University, Graz, Austria, May 2009.
- Glushko, R. Information Organization & Retrieval. School of Information, University of California, Berkeley, October 6, 2014.
- Edmonds, B. The pragmatic roots of context 119–132. In *Proceedings of the 2nd International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-99)*; Bouquet, P.; Serafini, L.; Brézillon, P.;

- Benerecetti, M.; Castellani, F.; Eds.; LNAI; Springer: Berlin, Germany, 1999, Vol. 1688.
29. Wallace, M.; Akrivas, G.; Mylonas, Ph.; Avrithis, Y.; Kollias, S. Using context and fuzzy relations to interpret multimedia content. In *Proceedings of 3rd International Workshop on Content-Based Multimedia Indexing (CBMI)*; IRISA, Rennes, France, 2003.
 30. Zhao, J.; Shimazu, Y.; Ohta, K.; Hayasaka, R.; Matsushita, Y. An outstandingness oriented image segmentation and its applications. In *Proceedings of 4th International Symposium on Signal Processing and Its Applications*; Gold Coast, QLD, 1996.
 31. Osberger, W.; Maeder, A.J. Automatic identification of perceptually important regions in an image. In *Proceedings of IEEE International Conference on Pattern Recognition*, Brisbane, QLD, 1998.
 32. Milanese, R. Detecting salient regions in an image: From biology to implementation. PhD Thesis, University of Geneva, Switzerland, 1993.
 33. Luo, J.; Singhal, A.; Zhu, W. Natural object detection in outdoor scenes based on probabilistic spatial context models. In *Proceedings of IEEE International Conference on Multimedia and Expo*; Baltimore, MD, 2002.
 34. Naphade, M.; Huang, T.S. A factor graph framework for semantic indexing and retrieval in video. In *CVPR Workshop on Content-based Image and Video Retrieval*, 2000.
 35. Saber, E.; Tekalp, A.M.; Eschbach, R.; Knox, K. Automatic image annotation using adaptive colour classification. *CVGIP: Graph. Models Image Process.* **1996**, *58*, 115–126.
 36. Smith, J.R.; Li, C.-S. Decoding image semantics using composite region templates. In *Proceedings IEEE International Workshop on Content-based Access of Image and Video Database*, Santa Barbara, CA, 1998.
 37. Galleguillos, C.; Rabinovich, A.; Belongie, S. Object Categorization using Co-Occurrence, Location and Appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, 23–28, 2008.
 38. Vailaya, A.; Jain, A. Detecting sky and vegetation in outdoor images. In *Proceedings of SPIE*, Vol. 3972, January 2000.
 39. Kalantidis, Y.; Tolias, G.; Avrithis, Y.; Phinikettos, M.; Spyrou, E.; Mylonas, Ph.; Kollias, S. VIRaL: Visual Image Retrieval and Localization. *Multimedia Tools Appl.* **2011**, *51* (2), 555–592.
 40. Ashley, J.; Flickner, M.; Lee, D.; Niblack, W.; Petkovic, D. Query by Image Content and Its Applications. IBM Research Report, RJ 9947 (87906) Computer Science/Mathematics, March, 1995.
 41. Smith, J.R.; Chang, S. Local color and texture extraction and spatial query. In *Proceedings of IEEE International Conference on Image Processing*, 1996.
 42. Tanaka, J.W.; Farah, M. Parts and wholes in face recognition. *Q. J. Exp. Psychol.* **1993**, *46A* (2), 225–245.
 43. Murphy, K.; Torralba, A.; Freeman, B. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *Advances in Neural Information Processing Systems 16 (NIPS 2003)*.
 44. Vailaya, A.; Figueiredo, M.; Jain, A.; Zhang, H.-J. Content-based hierarchical classification of vacation images. In *Proceedings IEEE International Conference on Multimedia Computing and Systems*; Florence, Italy, 1999, 7–11.
 45. Stauder, J.; Gouzien, G.; Chupeau, B.; Vigouroux, J.R.; Kijak, E. Semantic image browsing using hidden categories and confidence values. In *Storage and Retrieval for Media Databases 2003*, Santa Clara, CA, January 20–24, 2003.
 46. Salembier, P.; Smith, J.R. MPEG-7 multimedia description schemes. *IEEE Trans. Circuits Syst. Video Technol.* **2001**, *11* (6), 748–759.
 47. Wallace, M.; Akrivas, G.; Mylonas, Ph.; Avrithis, Y.; Kollias, S. Using context and fuzzy relations to interpret multimedia content. In *Proceedings of the Third International Workshop on Content-Based Multimedia Indexing (CBMI)*; IRISA: Rennes, France, 2003.
 48. Zhao, R.; Grosky, W.I. Narrowing the semantic gap-improved text-based web document retrieval using visual features. *IEEE Trans. Multimedia, Special Issue on Multimedia Database* **2002**, *4* (2), 189–200.
 49. Tsechpenakis, G.; Akrivas, G.; Andreou, G.; Stamou, G.; Kollias, S. Knowledge-assisted video analysis and object detection. In *Proceedings of European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems*, Albufeira, Portugal, 2002.
 50. Giro, X.; Marques, F. Semantic entity detection using description graphs. In *Workshop on Image Analysis for Multimedia Application Services (WIAMIS'03)*, London, UK, 2003.