12

Concept-Based Multimedia Processing Using Semantic and Contextual Knowledge

Evaggelos Spyrou, Phivos Mylonas, and Stefanos Kollias

12.1 Introduction

Most of today's content-based multimedia analysis and retrieval systems tend to follow a low-level approach when tackling both content analysis and retrieval tasks, thus falling short of benefits uprising from higher-level interpretation and knowledge. The role of additional information in the sense of semantics, context, and implicit or explicit knowledge is gaining focus on the task of bridging the semantic and conceptual gap that exists between humans and computers, in order to further facilitate human-computer interaction and scene content understanding. This chapter focuses on modeling and exploiting contextual knowledge toward efficient multimedia content understanding. As discussed below, this type of information acts as a simulation of the human visual perception scheme, by taking into account all contextual information relative to the visual content of a scene [1]. As a result, the notion of context, provided that it will be properly modeled and justified, may be used to improve the performance of knowledge-assisted analysis, semantic indexing, and retrieval of multimedia content.

When tackling the well-known problems of semi-automated high-level concept detection or scene classification, the researcher faces a challenging and broad research area. In order to achieve better semantic results during any multimedia content analysis phase, the influence of additional contextual information may be of great help, because although the well-known semantic gap [2] has been acknowledged for a long time, current multimedia analysis approaches are still divided into two rather discrete categories as low-level multimedia analysis methods and tools (for example, Reference [3]) and high-level semantic annotation methods and tools (for example, References [4] and [5]). Semantic knowledge technologies, like ontologies [6] and folksonomies [7], are only lately being successfully incorporated within multimedia analysis and retrieval frameworks, especially when using them for creation, manipulation, and postprocessing of multimedia metadata.

Still, one of the most interesting problems in multimedia content analysis is detection of high-level concepts within multimedia documents. Recognizing the need for such an analysis, many research works set focus on low-level feature extraction to efficiently describe various audiovisual characteristics of a multimedia document. However, the semantic gap often characterizes the differences between descriptions of a multimedia object by different representations and the linking from the low-level to the high-level features. Moreover, the semantics of each object depend on the context it is regarded within. For multimedia applications this means that any formal representation of real-world analysis and processing tasks requires the translation of high-level concepts and relations, for instance, in terms of valuable knowledge, into the elementary and extensively evaluated characteristics of low-level analysis, such as visual descriptions and low-level visual features.

An important step for narrowing this gap is to automate the process of semantic feature extraction and annotation of multimedia content objects, by enhancing image and video classification with semantic characteristics. The main idea introduced herein relies on the integrated handling of concepts evident within multimedia content. Recent advances in the research field of knowledge-assisted multimedia analysis, along with the emerge of new content and metadata representations, have driven more and more researchers looking

beyond solely low-level features (such as color, texture, and shape) in pursuit of more effective high-level multimedia representation and analysis methods. Current and previous multimedia research efforts are starting to focus on the combination of both low-level descriptors computed automatically from raw multimedia content and semantics focusing in extracting high-level features.

In the following, Section 12.2 describes the motivation in utilizing the notion of visual context in concept detection and scene classification. Section 12.3 presents the notions of bag-of-words image analysis techniques and visual context, and surveys the relevant state-of-the art methods. Section 12.4 deals with a novel proposition of an enhanced visual conceptualization of relative knowledge, as well as the instantiation of an image's region types. Section 12.5 presents three different types of context knowledge formalization, together with the proposed contextual adaptation in terms of the visual context algorithm and its optimization steps, according to the utilized knowledge. Some self-explanatory examples are presented in Section 12.6, whereas Section 12.7 lists experimental results derived from the *beach* domain. Finally, this chapter concludes with Section 12.8.

12.2 Motivation and Overview

Visual context forms a rather classical approach to context, tackling it from the scope of environmental or physical parameters that are evident in multimedia applications. The discussed context representation supports audiovisual information (for example, lighting conditions, environmental information) and is separately handled by visual context models. Research objectives in the field include 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 personalization aspects. The following text refers to the term *visual context*, by interpreting it as *all information related to the visual scene content of a still image or video sequence that may be useful during its analysis phase*.

Since there is no globally applicable aspect of context in the multimedia analysis chain, it is very important to establish a working representation for context, in order to benefit from and contribute to the proposed enhanced multimedia analysis. The problems to be addressed include how to represent and determine context, how to use it, and how to define and model corresponding analysis features to take advantage of it. Additionally, efficient ways to utilize the new content and context representations must be investigated, in order to optimize the results of content-based analysis. In general, the lack of contextual information significantly hinders optimal analysis performance [8] and, along with similarities in low-level features of various object types, results in a significant number of misinterpretations. Taken into account the current state-of-the-art, both in terms of works dealing with content classification and regional visual dictionaries, as well as context modeling techniques, this work aims at a hybrid unification of them, in order to achieve optimized content analysis results and strengthen its high-level and low-level correlation.

According to the previous statements, visual context is strongly related to two main problems of image analysis; that is, *scene classification* and *high-level concept detection/recognition*. Scene classification forms a *top-down* approach, where low-level visual features are typically employed to globally analyze the scene content and classify it in one of a number of predefined categories, such as indoor/outdoor, city/landscape, and so on. Concept 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). The above two major fields of image analysis actually comprise a chicken-and-egg problem. 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*.

However, a significant number of misclassifications usually occur because of the similarities in low-level color and texture characteristics of various object types and the lack of contextual 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 Reference [9]. 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. Unfortunately, such an approach alone is not considered sufficient, as it does not utilize the significant amount of available additional knowledge in the form of semantic relations.

So far, none of the existing methods and techniques utilizes the herein proposed contextual modeling 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. For instance, 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, such as lighting and 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 in the visual analysis process is the one that provides the necessary added value and forms the key for such a solid unambiguous recognition process and will be extensively presented and exploited in the following.

More specifically, this chapter presents an integrated approach, offering unified and unsupervised manipulation of multimedia content. It acts complementary to the current state-of-the-art, as it tackles both aforementioned challenges. Focusing on semantic analysis of multimedia, it contributes toward bridging the gap between the semantic and raw nature of multimedia content. It tackles one of the most interesting problems in multimedia content analysis, namely, detection of high-level concepts within multimedia documents, based on the semantics of each object, in terms of its visual context information. The latter is based on semantic relationships that are inherent within the visual part of the content. The pro-

posed approach proves also that the use of such information enhances the results obtained from traditional knowledge-assisted image analysis techniques, based on both *visual* and *contextual* information.

12.3 Image Analysis Based on Regions and Context

If focused solely on the visual part of analysis, it is rather true that high-level concept detection remains still a challenging and unsolved problem. Its most interesting aspects are first the low-level feature extraction, aiming to capture and describe the visual content of images or regions, and last the way that these features will be assigned to high-level concepts. This chapter deals with the latter part of the analysis process, and aims to create image descriptions from image regions, using standardized visual features, that is, the MPEG-7 (Moving Picture Experts Group) descriptors.

The most common approach in detection and recognition tasks begins with the extraction of a low-level description of the visual content of concepts. Then, for each concept, a detector is trained based on one or more examples. This is typically done using various machine learning techniques, such as neural networks, support vector machines (SVMs), and fuzzy systems.

In order to train the concept detectors, it is important to use or create a specific dataset, appropriately annotated either globally or locally. For a globally annotated image, one only gets the knowledge of the existence of certain concepts within it. For a locally annotated image, one also knows the exact location of concepts within it. However, despite the continuous growth of audiovisual content, the available locally annotated image collections remain few. This is not surprising, since such an annotation process is a difficult and tedious task. Two of the most important locally annotated collections are LabelMe [10], a collaboratively annotated collection for a very large number of concepts, and the PASCAL [11] collection. On the other hand, a global annotation of a given image is a much easier and less time-consuming task. There exist many such datasets, among which one should note the collaborative annotation of the LSCOM workshop [12], which focused on sequences from news videos of the TRECVID 2005 collection, and a similar attempt presented in Reference [13], focusing on cultural videos from the TRECVID 2007 collection. It should be noted here that in many popular social networks, such as Flickr,[1] many thousands of photos are globally annotated, while research efforts toward the annotation of such datasets are still increasing [14], [15], [16].

It is now clear that image analysis techniques that focus on the extraction of high-level concepts from globally annotated content are of greater importance and may have a broader field of applications. Thus, the algorithms that are presented in this chapter aim toward this exact problem, that is, how to extract and manipulate efficiently the visual content of globally annotated images in order to detect the presence of high-level concepts, without specifying their exact location.

[1] http://www.flickr.com

12.3.1 The Bag-of-Words Model

A very popular model that combines the aforementioned aspects of visual analysis is the *bag-of-words* model. Visual descriptions are extracted locally, from groups of image pixels. Using an appropriate visual dictionary, these descriptions are quantized to the corresponding visual words. An image is then described by a set of visual words, without considering any spatial or semantic relations. Finally, an appropriate detector is trained for each concept.

In order to develop a bag-of-words based technique, the first step to consider is how to select image parts, from whom visual descriptors should be extracted. Early approaches used a grid, dividing images to blocks. References [17] and [18] used square blocks of equal size. These techniques were very fast, but lacked in terms of the semantic interpretation of each block. To overcome this, Reference [19] used random sampling and a variable block size. Since this disintegration proved not to be very robust, later techniques were based on the extraction of points of interest. References [20] and [21] extracted features from the neighborhood of Harris affine points. Reference [18] selected points detected by the difference-of-Gaussian and extracted multi-resolution features. All these approaches aimed to selected invariant points under scale and some geometrical transforms and are very effective in the case of object detection. In parallel, Reference [22] applied a segmentation algorithm, in order to split an image to regions, based on their color and texture properties, with many advantages in material or scene detection.

The next step to consider is the extraction of the visual descriptions. In the case of grid-selected regions, color and texture descriptors are extracted. When regions are selected by a segmentation process, shape descriptors may also be extracted, if applicable to the targeted concepts. The MPEG-7 standard [23] contains many visual descriptors that are broadly used. Finally, in the point-of-interest-based regions, appropriate and popular features are those generated by the scale-invariant feature transform (SIFT) [24] and speeded-up robust feature (SURF) [25] methods and their various variations.

The success or failure of each technique is highly related to the creation of an appropriate codebook, based on which an image region is assigned to a visual word. Most of the techniques use typical clustering algorithms, such as the traditional K-means [26], or a hierarchical variation, such as the one proposed in Reference [27]. The selection of K is usually selected empirically or by a trial-and-error process on a validation set. However, certain techniques, such as the minimum description length (MDL) approach, are often applied in order to determine the appropriate dictionary size [28]. Typical sizes of dictionaries vary from a few tenths for grid approaches to many thousands of visual words for point-of-interest approaches.

The last step is to select an appropriate bag-of-words representation, with which the detectors will be trained. A number techniques have been proposed; many of them have been inspired by text categorization. A brief overview of the most important techniques is presented below.

Reference [29] created a texton library and trained texton histogram-based detectors. Similarly, Reference [30] constructed a codebook of prototype regions with specific geometric and photometric properties, namely, three-dimensional textons. Reference [31] used mean-shift clustering to split images to regions based on color features. Each image was

then represented by a binary vector indicating the presence of absence of visual words. In another work [32] pixel-based concept detectors were used in order to achieve semantic image clustering. Visual features were extracted from the resulting regions and scenes were detected using a codebook. Reference [20] replaced visual words with points of interest (bag-of-keypoints) and used K-means clustering to construct the visual dictionary. Both SVM and naive Bayesian techniques were then applied for concept detection. Reference [33] used semantic descriptions instead of low-level descriptors. In Reference [34], the authors used the bag-of-words model to a contour-based object recognition problem. They constructed the visual dictionary based on curve parts. Another work [35] divided images to subregions and calculated local histograms within these subregions. Then, the bag-of-words model was used. Reference [36] proposed the use of *keyblocks*, which is the equivalent of keywords in the field of images. A codebook, that contained those keyblocks in various resolutions, was built. Reference [18] investigated various methods of splitting images to parts and used a Bayesian hierarchical model.

During the past few years, there have been a few notable attempts to enhance the bag-of-words model with spatial relations. Reference [37] used adaptive correlation histograms and presented a model robust to geometric transformations. Reference [38] suggested a hierarchical probabilistic model, based on a set of parts, each describing the expected position of objects. Reference [39] suggested a hierarchical bag-of-words model of spatial and spatiotemporal features of video sequences, whereas another work [40] suggested a segmentation algorithm as an extension of object detection techniques with the aid of a visual dictionary.

As the bag-of-words model is inspired by text processing techniques, it is not surprising that it has been enhanced by popular methods in this field. For example, the latent semantic analysis (LSA) approach [41] aims at exploiting the latent relations among the visual words. References [42] and [43] extended the bag-of-words model using LSA and probabilistic LSA (pLSA). Reference [21] modeled images as a mixture of concepts and also applied pLSA, in a fully unsupervised model.

12.3.2 The Role of Context

Scene context is probably the simplest way to model context in an image analysis problem. According to Reference [44], scene context may be defined as a combination of objects/concepts which are correlated under human perception and share the property to complement each other. Thus, in order to exploit scene context, one has first to detect present concepts and optionally their location within the image and then use contextual information to infer the scene the image depicts. For example, in a concept detection problem, if the concept *sea* is detected with high confidence, one could also expect the image to contain *sky* and depict a *beach* scene. Reference [45] used scene context to detect events. Reference [46] applied the expectation-maximization algorithm to the low-level features of image parts that depict concepts. Reference [47] investigated the use of Bayesian networks, Boltzmann machines, Markov random fields, and random fields to model the contextual relations between concepts. Reference [48] used scene context as an extra source in global descriptions and faced scene classification and object detection as a unified problem.

Spatial context aims at modeling the spatial relations among concepts. This step is used to enhance context models, after modeling scene context. It may be defined as a combination of objects/concepts, which, apart from their co-occurrences, occupy spatial positions within an image that are not independent. For example, to further extend the previous scene context example, if an image is detected to depict a *beach* scene and for a region it cannot be determined with high confidence whether it depicts *sky* or *sea*, two concepts that share similar visual features, its location should be considered. Thus, if it is located on the top of all other regions it should depict *sky*. Reference [49] used quantitative spatial and photometric relations among regions in a scene classification problem. Reference [50] used probabilistic models to model spatial relations among materials; this effort was further enhanced in Reference [9] by modeling spatial relations among concepts after a learning process. Reference [51] proposed an expectation-maximization model, which after a learning process low-level features to visual words that describe concepts. Reference [52] used a grid to split images to blocks and encoded both relations among low-level features and concepts and spatial relations among concepts. References [53] and [54] also used a grid and a two-dimensional hidden Markov model to form spatial relations among concepts. Another work [55] used graphs and statistics and showed that spatial relation-driven models increase precision. In Reference [56], the authors combined knowledge about scene layout and a generative model using material detectors and further improved their model in Reference [57].

Temporal context considers the relations of an image with other images from the same collection, taken with a small time difference, independently of their visual content. For example, if an image depicts a *beach* scene with a confidence, concepts present in images taken by the same user with a small time difference should also be in the context of *beach*. Reference [58] exploited the idea that the visual content of a photo is correlated to the one of photos taken with a small time interval. Reference [59] used many successive images from different viewpoints and exploited their temporal context to solve a three-dimensional photo retrieval problem by employing a Bayesian model. Reference [60] used temporal context in event detection in text documents. References [61] and [62] combined temporal context and photo location, whereas Reference [63] constructed hierarchical event models based on the visual content and temporal context of photo collections.

Metadata context involves the relations among the metadata that are available in digital images, such as the camera settings with which the image was taken. These metadata are embedded in image files according to the EXIF (Exchangeable Image File) standard [64]. For example, for an *indoor/outdoor* scene classification problem, the knowledge of the focus distance can let one assume the depicted scene; for example, a large distance usually indicates an *outdoor* scene. Reference [65] used metadata of camera settings combined with low-level features to automatically annotate photos. Reference [66] used a boosting algorithm and metadata in a scene classification problem. Reference [67] combined metadata with color features and *face* and *natural place* detectors. References [68], [69], and [70] showed that metadata information can significantly assist in the problems of *indoor/outdoor* classification and *sunset* detection, using Bayesian networks. Finally, Reference [71] proposed a system that combines metadata and visual features in order to automatically construct photo collections.

(a) (b)

FIGURE 12.1 (See color insert.)

An input image and its coarse segmentation.

12.4 Image Description and High-Level Concept Detection Using a Region Thesaurus

The following presents the proposed approach to tackle the problems of image description and high-level concept detection from a different and at the same time innovative aspect, that is, based on a region thesaurus containing a set of region types [72]. This research effort was expanded and further strengthened in References [73], [74], [75] and [76] by exploiting visual context in the process and achieving promising research results. The main focus remains to provide an ad hoc ontological knowledge representation containing both high-level features (that is, high-level concepts) and low-level features and exploit them toward efficient multimedia analysis.

Generally, the visual features extracted from an image or video can be divided into two major categories. The first contains typical *low-level* visual features that may provide a qualitative or quantitative description of the visual properties. Often these features are standardized in the form of a *visual descriptor*. The second category contains *high-level* features that describe the visual content of an image in terms of its semantics. One fundamental difference between these categories is that low-level features may be calculated directly from an image or video, while high-level features cannot be directly extracted, but are often determined by exploiting the low-level features. A human observer can easily recognize high-level features, even in situations when it could be rather difficult to provide their qualitative description and almost impossible to provide a quantitative one.

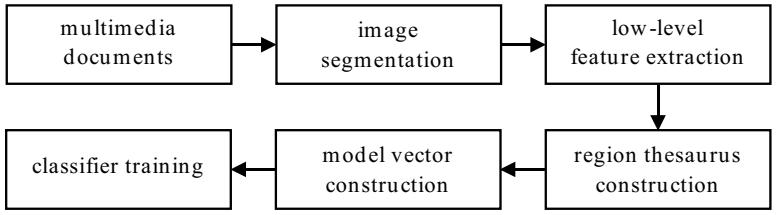


FIGURE 12.2

Offline part of the high-level concept detection algorithm.

FIGURE 12.3

Online part of the high-level concept detection algorithm.

In this sense, this chapter tries to enhance the notion of a visual context knowledge model with *mid*-level concepts. Those concepts are referred to as region types, for reasons clarified in Section 12.4.2. Such concepts may provide an in-between description, which can be described semantically, but does not express neither a high-level nor a low-level feature. Thus, this work will focus on a unified multimedia representation by combining low-level and high-level information in an efficient manner and attach it to the context model by defining certain relations. To better understand the notion of region types, Figure 12.1 presents a visual example. In this example, a human could easily describe the visual content of the image either in a high-level manner (that is, the image contains *sky*, *sea*, *sand*, and *vegetation*) or in a lower level, but higher than a low-level description (that is, an *azure* region, a *blue* region, a *green* region, and a *gray* region). Although a quantitative description cannot be provided, each image can be intuitively and even efficiently described by a set of such features, that is, the region types. Therefore, it is of crucial importance to encode the set of region types in an effective manner that can efficiently describe almost every image in a given domain. To achieve this, a *region thesaurus* needs to be constructed. The next sections briefly describe the extraction of the low-level features and the construction of the region thesaurus. Figure 12.2 presents the offline part of the overall methodology for the high-level concept detection process that leads to a trained set of classifiers, while Figure 12.3 presents the online part that leads to the extraction of high-level features.

12.4.1 Low-Level Feature Extraction

To represent the color and texture features of a given image, this chapter follows an approach of extracting visual descriptors *locally*, that is, from image regions. First, color segmentation is performed using a multi-resolution implementation of the well-known RSST

method [77], tuned to produce a coarse segmentation. Note that this under-segmentation both facilitates image description and prevents the problem from being too complex. After extracting image regions, low-level visual descriptors from the MPEG-7 standard [23] are extracted to capture a description of their visual content. Here, the extraction process focuses on color and texture features using appropriate MPEG-7 color and texture descriptors, since the employed high-level concepts belong to the categories of *materials* and *scenes*. To perform such feature extraction, a descriptor extraction tool [78], which is fully compatible with the MPEG-7 eXperimentation Model (XM) [79], is used.

More specifically, four color and two texture descriptors are selected: the *dominant color descriptor* (DCD), the *color layout descriptor* (CLD), the *scalable color descriptor* (SCD), the *color structure descriptor* (CSD), the *homogeneous texture descriptor* (HTD), and the *edge histogram descriptor* (EHD). To obtain a single region description from all the extracted region descriptions, features are merged after their extraction [80] into a feature vector. The feature vector f_i that corresponds to a region $r_i \in R$, for R denoting the set of all regions, is defined as follows:

$$f_i = f(r_i) = [DCD(r_i), CLD(r_i), SCD(r_i), CSD(r_i), HTD(r_i), EHD(r_i)]. \tag{12.1}$$

12.4.2 Construction of a Region Thesaurus

After extracting color and texture features, the next step aims to bridge these low-level features to the high-level concepts aimed at detection. To achieve this, first a region thesaurus will be constructed to assist with quantizing regions and forming an intermediate image description. This description will contain all the necessary information to connect one image with every region type of the dictionary. In this way, a fixed-size image description can be achieved, tackling the problem that the number of segmented regions is not fixed. Moreover, this description will prove again useful when contextual relations will be exploited, as described in Section 12.5.

Given the entire training set of images and their extracted regions, one can easily observe that regions belonging to similar semantic concepts also have similar low-level descriptions, and that images containing the same high-level concepts consist of similar regions. This gives a hint to exploit region similarity, as region co-existences often characterize the concepts that exist within an image [72].

The first step is the selection of region types that will form the region thesaurus. Based on the aforementioned observations, the proposed method starts from an arbitrary large number of segmented regions and applies a *hierarchical clustering* algorithm [81], adjusted for the problem at hand and with the clustering level empirically selected. After the clustering process, each cluster may or may not represent a high-level feature and each high-level feature may be contained in one or more clusters. This means that the concept *sand* can have many instances differing, for example, in color or texture. Moreover, in a cluster that may contain instances from a semantic entity (for example, *sea*), these instances could be mixed up with parts from another visually similar concept (for example, *sky*). Here, a single region is selected to represent each cluster, that is, the region type.

Finally, a region thesaurus T can be formally described as a set of N_T visual words t_i:

$$T = \left\{ t_i, \quad i = 1, 2, \ldots, N_T \right\}, \quad t_i \subset R, \tag{12.2}$$

$$\bigcup_i t = R, \quad i = 1,2,\ldots,N_T, \bigcap_{i,j} t = \emptyset, \quad i \neq j. \tag{12.3}$$

Generally, a thesaurus combines a list of every term in a given domain of knowledge and a set of related terms for each term in the list, which are the synonyms of the current term. In the proposed approach, the constructed region thesaurus contains all the region types that are encountered in the training set. Each region type is represented by its feature vector that contains all the extracted low-level information. As it is obvious, a low-level descriptor does not carry any semantic information. It only constitutes a formal representation of the extracted visual features of the region. On the other hand, a high-level concept carries only semantic information. It is now clear that a region type lies in-between those features. It contains the necessary information to formally describe the color and texture features, but can also be described with a *lower* description than the high-level concepts. Namely, one can describe a region type as *a green region with a coarse texture*.

12.4.3 Construction of Model Vectors

This section presents the algorithm, which is used here to describe each image with the aid of the region thesaurus. First, it must be noted that the MPEG-7 standard does not specify strict distance measures. It only suggests some, so as to allow for other measures to be used and test their efficiency. As depicted in the experiments presented in Reference [72], the use of the Euclidean distance provides a simple yet effective way to fuse all extracted low-level information, leading also to satisfactory results. Then, the distance $d(r_1, r_2)$ between two regions r_1 and r_2 defined in $\mathscr{R}$ is calculated by the Euclidean distance of their feature vectors f_1 and f_2 as follows:

$$d(r_1, r_2) = d(f_1, f_2) = \sqrt{\sum_{i=1}^{n} (f_1^i - f_2^i)^2}. \tag{12.4}$$

Having calculated the distance of each image region to all the words of the constructed thesaurus, the model vector that semantically describes the visual content of the image is formed by keeping the smaller distance for each intermediate concept (region type), thus containing all the necessary information to associate an image with the whole set of the region thesaurus. In particular, the model vector m_p describing image p is given by

$$m_p = [m_p(1), m_p(2), \ldots, m_p(j), \ldots, m_p(N_T)],\ i = 1,2,\ldots,N_K, \tag{12.5}$$

where

$$m_p(j) = \min_{r \in R(p)} \{d(f(t_j), f(r))\},\ i = 1,2,\ldots,N_T;\ j = 1,2,\ldots,N_T, \tag{12.6}$$

and $R(p)$ denotes the set of all regions of image p.

In order to better understand the above process, Figure 12.4 presents an indicative example, where an image is segmented in regions and a region thesaurus is formed by six region types. On the left, this figure presents the distances of each region type from the sky region; the distances of each image region from region type 5 are presented on the right. The model vector is constructed by the smallest distances for each region type. In this case and considering region type t_5, the minimum distance is equal to 0.1. The model vector for the specific image, given the region thesaurus, is defined as

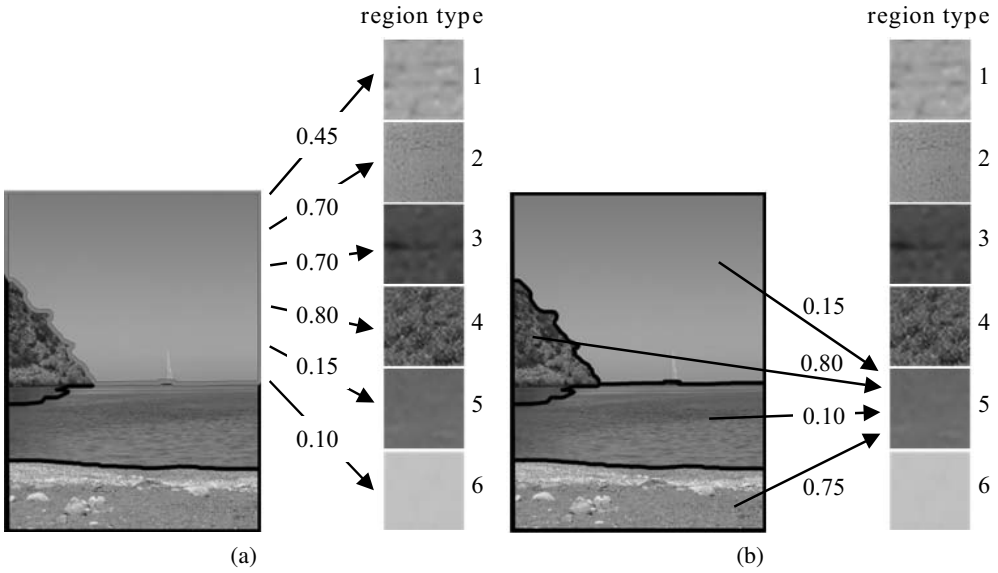


FIGURE 12.4 (See color insert.)

Distances between regions and region types: (a) distances between an image region and all region types; (b) distances between all regions and a specific region type.

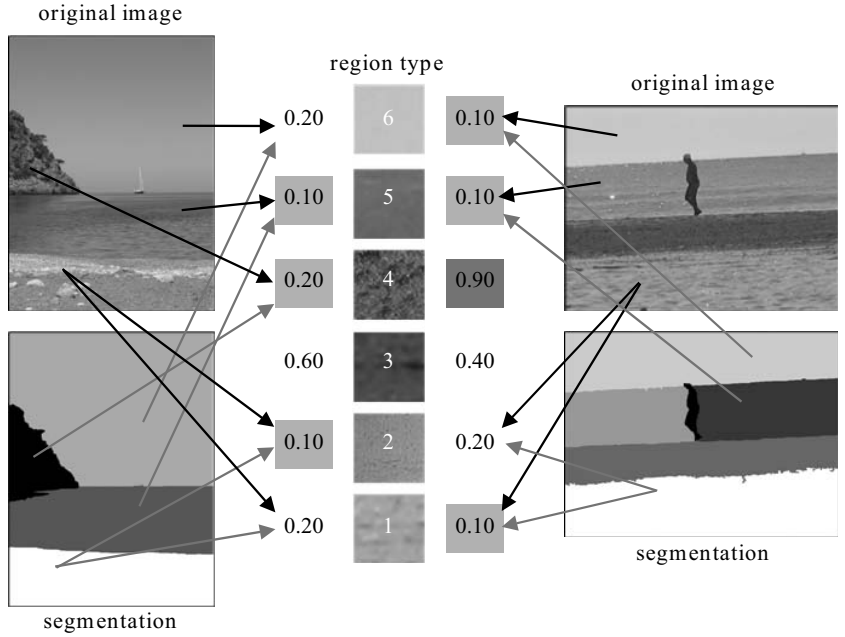


FIGURE 12.5 (See color insert.)

Construction of model vectors for two images and a visual thesaurus of six region types; lowest values of model vectors are highlighted (light gray) to note which region types of the thesaurus are most contained in each image, whereas a high value (dark gray) indicates a high distance between the corresponding region type and the image.

$$m = [m(1), m(2), \ldots, m(5), m(6)], \tag{12.7}$$

where $m(5)$ will be equal to 0.1. Taking into consideration all distances between all four image regions and all six region types (that is, a total of 24 distances), the corresponding model vector is formed. Figure 12.4 presents the model vectors for two images, using the region thesaurus shown in Figure 12.5.

12.4.4 High-Level Feature Detection

After extracting model vectors from all images in the (annotated) training set, an SVM-based detector is trained separately for each high-level concept. A model vector m_i, describing a keyframe in terms of the region thesaurus, is fed to the detectors. The output of the network is the confidence that the image in question contains the specific concept, this is done for all concepts. It is important to note that the detectors are trained based on annotation per image and not per region. The same stands for their output, thus providing the confidence that the specific concept exists *somewhere* within the keyframe in question.

12.5 Visual Context Optimization

In order to fully exploit the notion of visual context and combine it with the aforementioned bag-of-words technique, a threefold approach is introduced next. The proposed methodology could be divided into the following three sections, according to the effect of visual context regarding concepts and region types:

- a scene context approach that aims to refine initial high-level concept detection results by exploiting solely the contextual relations between high-level concepts;
- an approach that aims to refine the input of trained high-level concept detectors based on the contextual relations between region types of the given training set; and
- a unified approach that utilizes contextual relations among high-level concepts and region types.

It should be emphasized here that this research effort focuses on the integrated approach of the subject, which offers a unified and unsupervised management of multimedia content. It is proved that the use of enhanced intermediate information can improve the results of traditional, knowledge-assisted image analysis, based on both *visual* and *contextual* information.

12.5.1 Scene Context

The proposed approach differentiates itself from most of the related research works, because it deals with a global interpretation of the image and the concepts that are present in it. In other words, high-level concepts either exist or do not exist within the entire image under consideration and not within a specific region of interest (for example, the image might

contain concept *water*, but there is no information regarding its spatial location). Now, in order to further adapt the results of low-level and descriptor-based multimedia analysis, utilizing the notion of region types, a scene context method, founded on an enhanced high-level contextual ontology, is introduced. The proposed visual context application optimizes the high-level concept detection results (in terms of the classifiers' output) that were obtained based on the detailed methodology described in the previous sections.

12.5.1.1 A Scene Context Knowledge Model

The high-level concept ontology proposed herein is described as a set of concepts and semantic relations between concepts within a given universe. This set is introduced in order to efficiently express the real-world relations that exist between the concepts of the domain at hand. In general, one may decompose such an ontology O_c into two parts:

- Set $C = \{c_i\}$, for $i = 1,2,\ldots,n$, of all semantic concepts in the domain of interest.
- Set $R_c = \{R_{c_ij}\}, i,j = 1,2,\ldots,n$ of all semantic relations among concepts. Note that $R_{c,ij} = r_{c,ij}^{(k)}$ contains the K relations that can be defined among concepts c_i and c_j. Moreover, for a given relation $r_{c,ij}^{(k)}$, its inverse $\bar{r}_{c,ij}^{(k)}$ can be defined.

More formally:

$$O_c = \{C, R_c\}, \quad R_{c,ij} : C \times C \rightarrow \{0,1\}. \tag{12.8}$$

However, modeling of a domain using O_c is inappropriate, since it does not model relations among concepts as fuzzy as in real-world domains. Therefore, the aforementioned model is expanded to produce a fuzzified version of the scene context ontology, formally denoted as follows:

$$\mathscr{O}_c = \{C, \mathscr{R}_c\}, \tag{12.9}$$

where $\mathscr{R}_c = \{\mathscr{R}_{c,ij}\}$ is the set of fuzzy relations among concepts, with $\mathscr{R}_{c,ij} = r_{c,ij}^{(k)}$ and $\bar{r}_{c,ij}^{(k)}$ denoting again the corresponding inverse relation. Now, the following can be written:

$$r_{c_ij} : C \times C \rightarrow [0,1]. \tag{12.10}$$

Since for two concepts the existence of more than one relations is possible, a combination of all relations among c_i and c_j is defined as

$$\mathscr{U}_{c,ij} = \underset{k}{\cup}[\mathrm{r}_{c,ij}^{(k)}]^p, \; i,j = 1,2,\ldots,N, \; k = 1,2,\ldots,K. \tag{12.11}$$

The final combination of the MPEG-7 originating relations forms a resource description framework (RDF) graph and constitutes the abstract contextual knowledge model to be used (Figure 12.6). The value of p is determined by the semantics of each relation r_{c_ij} used in the construction of $\mathscr{U}_{c,ij}$. More specifically:

- $p = 1$, if the semantics of r_{c_ij} imply that it should be considered as is;
- $p = -1$, if the semantics of r_{c_ij} imply that its inverse $\bar{r}_{c_ij}$ should be considered; and
- $p = 0$, if the semantics of r_{c_ij} do not allow its participation in the construction of the combined relation $\mathscr{U}_{c,ij}$.

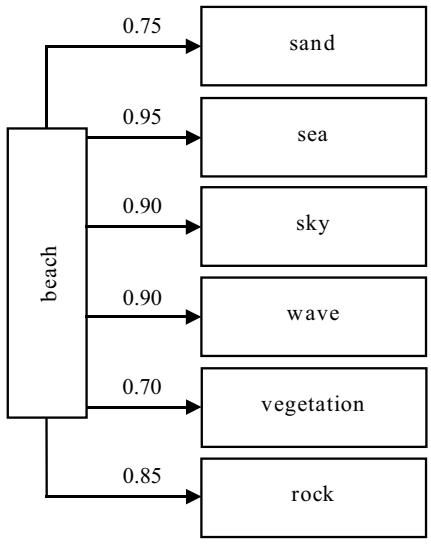


FIGURE 12.6

A fragment of the *Beach* domain ontology depicting the relations between concept *Beach* (the root element) and seven high-level concepts.

As indicated in Reference [82], any kind of semantic relation may be represented by such an ontology, however, herein it is restricted to a fuzzified ad hoc context ontology. The latter is introduced in order to optimally express the real-world relationships that exist between each domain's participating concepts. In order for this ontology to be highly descriptive, it must contain a representative number of distinct and even diverse relations among concepts, so as to scatter information among them and thus describe their context in a rather meaningful way. Moreover, the utilized relations need to be meaningfully combined, so as to provide a view of the knowledge that suffices for context definition and estimation. Since modeling of real-life information is usually governed by uncertainty and ambiguity, it is believed that these relations must incorporate fuzziness in their definition. Therefore, the proposed method extends a subset (Table 12.1) of the MPEG-7 semantic relations [83] that are suitable for image analysis and specified, in this case, by a domain expert. It should be noted at this point that since the proposed semantic relations are redefined in a way to represent fuzziness, a degree of confidence is associated to each of them. To further un-

TABLE 12.1

Fuzzy scene context semantic relations between concepts.

Name	Inverse	Symbol	Meaning
Specialization	Generalization	$Sp(a,b)$	b is a specialization in the meaning of a
Part	PartOf	$P(a,b)$	b is a part of a
Example	ExampleOf	$Ex(a,b)$	b is an example of a
Instrument	InstrumentOf	$Ins(a,b)$	b is an instrument of or is employed by a
Location	LocationOf	$Loc(a,b)$	b is the location of a
Patient	PatientOf	$Pat(a,b)$	b is affected by or undergoes the action of a
Property	PropertyOf	$Pr(a,b)$	b is a property of a

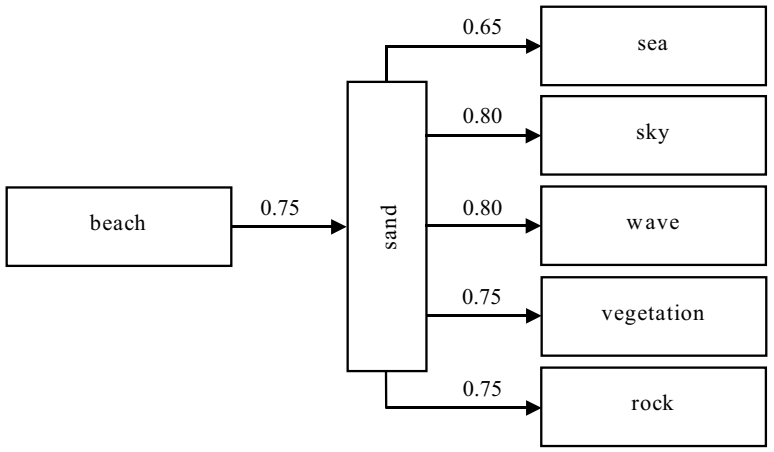


FIGURE 12.7

A fragment of the *Beach* domain ontology depicting the relations between concept *sand* (the root element) and six high-level concepts.

derstand the meaning of these semantic relations, some indicative examples are presented below. For example, a *bear* is a *Specialization* of an *animal*, whereas *tree* is a *part* of *forest*. Moreover, *clay* is an *Example* of a *material* and a *wheel* is the *Instrument* of a *car*. *Beach* might be the *Location* of an *umbrella*, *gun* is a *Patient* of the action of *soldier*, and *wavy* is a *Property* of a *sea*.

The graph of the proposed model contains nodes (that is, domain concepts) and edges (that is, an appropriate combination[2] of contextual fuzzy relations between concepts). The degree of confidence of each edge represents fuzziness in the model. Non-existing edges imply non-existing relations, meaning that relations with zero confidence values are omitted. An existing edge between a given pair of concepts is produced based on the set of contextual fuzzy relations that are meaningful for the particular pair. For instance, the edge between concepts *rock* and *sand* is produced by the combination of relations *Location* and *Patient*, whereas the edge between *water* and *sea* utilizes *Specialization*, *PartOf*, *Example*, *Instrument*, *Location*, and *Patient*, in order to be constructed. Since each concept has a different probability to appear in the scene, a flat context model would not have been sufficient in this case. On the contrary, concepts are related to each other, implying that the graph relations used are in fact transitive. The degree of confidence is implemented using the RDF reification technique [84].

12.5.1.2 Scene Context Optimization

Once the contextual knowledge structure is finalized and the corresponding representation is implemented, a variation of the context-based confidence value readjustment algorithm [8] is applied to the output of the neural network-based classifier. The proposed contextualization approach empowers a postprocessing step on top of the initial set of re-

[2]The combination of different contextual fuzzy relations toward the generation of a practically exploitable knowledge view is conducted by utilizing fuzzy algebraic operations in general and the default t-norm in particular.

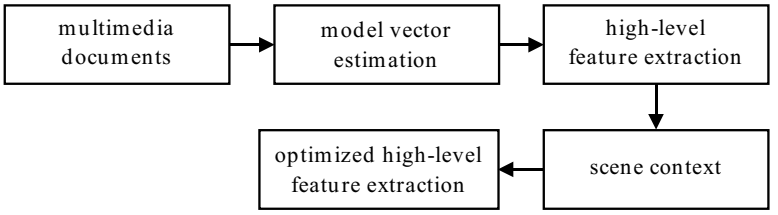


FIGURE 12.8

Contextual influence on high-level concepts.

gion types extracted. It provides an optimized re-estimation of the initial concepts' degrees of confidence for each region type and updates each model vector. In the process, it utilizes the high-level contextual knowledge from the constructed contextual ontology. The corresponding methodology is presented in Figure 12.8.

The degree of membership for each concept is estimated from direct and indirect relationships of the concept with other concepts using a meaningful compatibility indicator or distance metric. Again, depending on the nature of the domains provided in the domain ontology, the best indicator could be selected using the *max* or the *min* operator, respectively. Algorithm 12.1 depicts the general structure of the degree of membership re-evaluation procedure. The number of the iterations is defined empirically and usually three to five iterations are enough. The final output of the algorithm are the updated degrees of confidence for the presence of all concepts of the given domain within image p.

ALGORITHM 12.1 Procedure for evaluating the degree of membership.

1. The considered domain imposes the use of a domain similarity (or dissimilarity) measure $w_s \in [0,1]$.
2. For each image p, a fuzzy set L_p with the degrees of confidence $\mu_p(c_i)$ is defined for all concepts c_i of the domain.
3. For each c_i in the fuzzy set L_p with a degree of membership $\mu_p(c_i)$, the particular contextual information is obtained in the form of the set $\mathscr{U}_{c,i} = \{\mathscr{U}_{c,ij} : c_i, c_j \in C, \forall i \neq j\}$.
4. The new degree of membership $\mu'_p(c_i)$ is calculated by taking into account each domain's similarity measure. In the case of multiple concept relations in the ontology, when relating concept c_i to more concepts apart the *root* concept (Figure 12.6), an intermediate aggregation step should be applied for the estimation of $\mu'_p(c_i)$ by considering the *context relevance* notion, $cr_i = \max_j\{\mathscr{U}_{c,ij}\}$, $j = 1,2,\ldots,c_k$, defined in Reference [8]. The calculation of $\mu_p(c_i)$, which is the degree of confidence for c_i at the l-th iteration, is expressed with the recursive formula
$$\mu_p^l(c_i) = \mu_p^{l-1}(c_i) - w_s(\mu_p^{l-1}(c_i) - cr_i).$$
Equivalently, for an arbitrary iteration l:
$$\mu_p^l(c_i) = (1-w_s)^l \cdot \mu_p^0(c_i) + (1-(1-w_s)^l) \cdot cr_i,$$
where $\mu_p^0(c_i)$ represents the initial degree of membership for concept c_i.

12.5.2 Region Type Context

This section presents a different approach on scene context. The context optimization will have an effect on region types rather than concepts. It is relatively easy to prove that the utilization of this context information will improve the results of traditional image analysis. Indeed, initial analysis results are enhanced through the utilization of semantic knowledge, in terms of region-independent *region types* and semantic relations between them. In general, this information may be described by an intermediate description, which can be semantically described, but does not express the high-level concepts.

12.5.2.1 A Region Type Knowledge Model

The proposed methodology, that is presented in this section, follows precisely the steps of the scene context optimization from Section 12.5.1. Thus, an appropriate fuzzified ontology will be first defined in order to model in an appropriate way the real-world relations among the region types. In this case, the crisp ontology O_T may be described as set T of m region types and a set $R_{T,ij}$ of semantic relations among them. More specifically, let:

1. $T = \{t_i\}$, for $i = 1,2,\ldots,m$, be the set of all region types of the visual thesaurus used in the problem at hand, and
2. $R_T = \{R_{T,ij}\}$, for $i,j = 1,2,\ldots,m$, be the set of the semantic relations among region types. Set $R_{T,ij}^{(k)}$, for $k = 1,2,\ldots,K'$, includes K' relations that can be defined between region types t_i and t_j. Moreover, for a given relation $r_{T,ij}^{(k)}$, its inverse $\bar{r}_{T,ji}^{(k)}$ can be defined.

Thus, for a given ontology O_T and sets T and $R_{T,ij}$, the following can be written:

$$O_T = \{C, R_T\}, \tag{12.12}$$

$$R_{T,ij} : T \times T \rightarrow \{0,1\}. \tag{12.13}$$

Now, the ontology modeling should be redefined to include fuzziness. A fuzzified version $\mathscr{O}_T$ of O_T is defined as

$$\mathscr{O}_T = \{T, \mathscr{R}_T\}, \tag{12.14}$$

where $\mathscr{R}_T = \{\mathscr{R}_{T,ij}\}$ denotes the set of fuzzy semantic relations. Set $\mathscr{R}_{T,ij} = \mathrm{r}_{T,ij}^{(k)}$, for $k = 1,2,\ldots,K'$, includes all K' relations that can be defined between two region types t_i and t_j. Moreover, for each relation $\mathrm{r}_{T,ij}^{(k)}$, its inverse $\bar{\mathrm{r}}_{T,ji}^{(k)}$ can be defined. Finally, since relations are fuzzy, the following can be written:

$$r_{T,ij} : T \times T \rightarrow [0,1]. \tag{12.15}$$

Since it is possible that more than one relation may be valid simultaneously between two region types, a combination of relations can be defined as

$$\mathscr{U}_{T,ij} = \bigcup_k [\mathrm{r}_{T,ij}^{(k)}]^p,\ i,j = 1,2,\ldots,N;\ k = 1,2,\ldots,K'. \tag{12.16}$$

The value of p is once again determined by the semantics of each relation $\mathrm{r}_{,ij}$ used to construct $\mathscr{U}_{,ij}$. More specifically:

- $p = 1$, if the semantics of $r_{T,ij}$ imply that it should be considered as is;
- $p = -1$, if the semantics of $r_{T,ij}$ imply the use of its inverse $\bar{r}_{T,ji}$; and
- $p = 0$, if the semantics of $r_{T,ij}$ do not allow its participation in the construction of the combined relation $\mathscr{U}_{T,ij}$.

12.5.2.2 Relations between Region Types

Once again, semantic relations defined by the MPEG-7 standard [83] are chosen and redefined to include fuzziness. The relations that may be applicable between region types are summarized in Table 12.2.

In this case, these relations may be calculated after a statistical analysis in an appropriate training set, that is, the one used to form the region thesaurus. To make their semantics and the calculations clear, a few indicative examples are presented below.

- *Similar* denotes that a region type is similar to another region type, under a certain degree of confidence. To calculate this degree, their low-level features should be compared using an appropriate similarity function.
- *Accompanier* denotes the degree to which two region types co-occur in an image. It is calculated as the percentage of the images in the training set that contain both region types to the images that contain either of them.
- *PartOf* denotes that a region type is part of another. This relation is defined by an expert, when this knowledge derives from observations to the visual thesaurus construction.
- *Combination* denotes that two region types are combined to form another region type. This is a special case where the inverse relation cannot be defined.

It becomes obvious that modeling region type context with an ontology leads again to the construction of an RDF graph (Figure 12.9). Its nodes correspond to region types and its edges to their combined relations. RDF reification [84] is used again here to estimate the corresponding degrees of confidence. This way RDF triplets are formed, for instance, *blue partOf green*, with a degree of confidence equal to 0.85. This triplet does not imply that a *blue* region type will *always* be part of a *green* region type.

The region types of the ontology are those of the region thesaurus that have been constructed for the visual analysis. The final ontology relations are formed after calculations among these regions.

TABLE 12.2
Contextual relations between region types.

Relation	Inverse	Symbol	Meaning
Similar	Similar	$Sim(a,b)$	region type a is similar to region type b
Accompanier	Accompanier	$Acc(a,b)$	region type a is accompanier of region type b
Part	PartOf	$P(a,b)$	region type a is part of region type b
Combination	–	$Comb(a,b)$	combines two or more region types

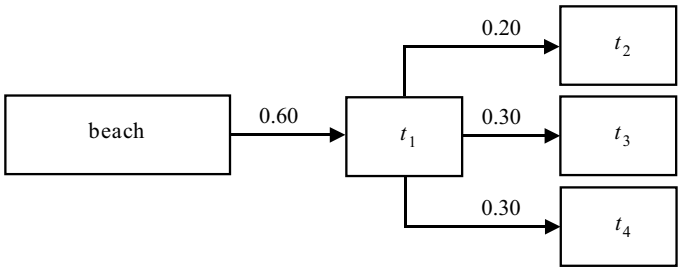


FIGURE 12.9

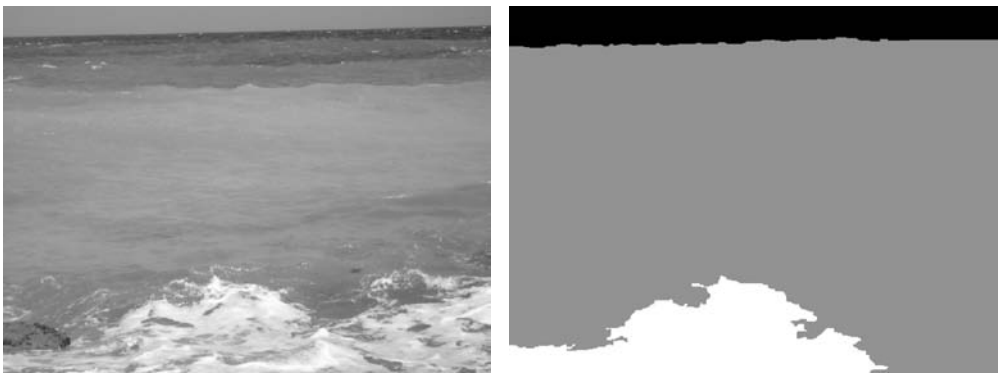
A fragment of the *Beach* region type knowledge model.

12.5.2.3 Region Type Context Optimization

After the construction of the model vector of an image p, an appropriately modified version of Algorithm 12.1 is applied. The algorithm now aims at refining the model vector by modifying its constituents, each corresponding to the degree of confidence for a region type. This is a preprocessing step that outputs an improved version of the model vector after considering the context of its region types. This leads to an increase in the detectors' precision, as the refined model vectors are closer to those used for their training.

To make this clear, a simplistic example is considered below. The *sea* detector in the *Beach* domain may have correlated this concept with the existence of a *blue*, a *light blue*, and a *brown* region, which corresponds to a typical *Beach* image that depicts *sea*, *sky*, and *sand*, respectively. If an image is presented where the region type that corresponds to *sea* is *green* (as in Figure 12.10) while the others remain as described before, the *sea* detector is certain to produce either a wrong result or a correct result with a small confidence, which will decrease overall precision.

Algorithm 12.2, which is a modified version of the previous algorithm, aims at this exact problem. In the given example, the model vector will be altered in a way that the confidence of the existence of a *blue* region type is increased, while the one of a *green* region type is decreased.



(a) (b)

FIGURE 12.10 (See color insert.)

An example from the *Beach* domain, where the region types of an image are different than a typical *Beach* image.

ALGORITHM 12.2 Improved procedure for evaluating the degree of membership.

1. The model vector for the image in question is calculated using the procedure presented in Section 12.4.3.
2. The considered domain imposes the use of a domain similarity (or dissimilarity) measure $w_T \in [0,1]$.
3. For each image p, the fuzzy set L_T with the degrees of confidence $\mu_p(t_i)$ is defined for all region types t_i, with $i = 1,2,\ldots,k'$ of the visual thesaurus.
4. For each region type t_i in the fuzzy set L with a degree of confidence $\mu_p(T_i)$, the particular contextual information is obtained in the form of the set $\mathscr{U}_{T,i} = \{\mathscr{U}_{T,ij} : t_i, t_j \in T, \forall i \neq j\}$.
5. The new degrees of confidence $\mu'_p(T_i)$ are calculated by taking into account the current domain's similarity measure. In the case of multiple concept relations, when t_i is related with one or more types, apart from the *root* of the ontology, an intermediate aggregation step should be applied in order to calculate $\mu'_p(T_i)$ using the context relevance notion cr_i defined in Reference [8] as $cr_i = \max_j\{\mathscr{U}_{T,ij}\}$, for $j = 1,2,\ldots,c_k$. The calculation of $\mu^l_p(t_i)$, which is the degree of confidence for t_i at the l-th iteration, is expressed with the recursive formula
$$\mu^l_p(t_i) = \mu^{l-1}_p(t_i) - w_t(\mu^{l-1}_p(t_i) - cr_i).$$
Equivalently, for an arbitrary iteration l:
$$\mu^l_p(t_i) = (1-w_t)^l \cdot \mu^0_p(t_i) + (1-(1-w_t)^l) \cdot cr_i,$$
where $\mu^0_p(c_i)$ represents the initial degree of confidence for t_i.

Figure 12.11 depicts a flowchart that describes the region type context. The number of the iterations is defined empirically and usually three to six iterations are enough also in this case. The final output of the algorithm is a refined model vector that is fed to the concept detectors instead of the one that is calculated by the visual features.

12.5.3 Unified Context

This section further advances the proposed conceptualization; it introduces a novel knowledge representation approach in the form of an extended mixed context model [74]. The classical notion of a contextual ontology is enhanced with *mid*-level concepts, that is, the region types and relations among different types of entities. These provide an intermediate description, which may be semantically described, but they do not express a high-level

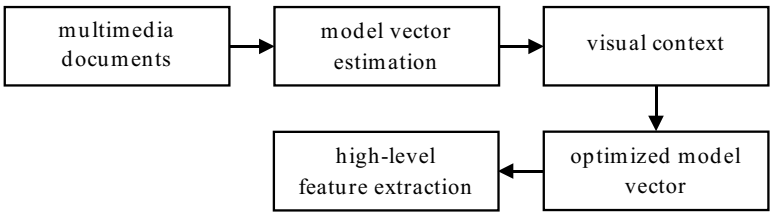


FIGURE 12.11

Contextual influence on region types.

nor a low-level concept. As a result, the focus here is on an integrated multimedia representation, combining efficiently low-level and high-level information and the description of a typical context model by defining new and expanding older relations. Within this section, both high-level concepts and region types will be simply referred to as *entities*.

12.5.3.1 A Unified Knowledge Model

This section describes a mixed fuzzified ontology that aims to model real-world relations among all entities present in images, that is, high-level concepts and region types. An ontology O that will model the unified context of a given domain contains:

- $C = \{c_i\}$, for $i = 1,2,\ldots,n$, which is the set of all high-level concepts of a given domain;
- $T = \{t_i\}$, for $i = 1,2,\ldots,m$, which is the set of all region types of the visual thesaurus used in the analysis process; and
- $R = R_{ij}$, for $i, j = 1,2,\ldots,n+m$, which is the set of all semantic relations among two entities x_i and x_j. The set $R_{ij} = r_{ij}^{(k)}$, for $k = 1,2,\ldots,K+K'$, includes at most $K+K'$ relations among x_i and x_j.

Thus, for a unified context ontology O and the aforementioned sets C, R, and T, the following can be written:

$$O = \{C, T, R_{ij}\}, \tag{12.17}$$

$$r_{ij}^{(k)} : (C \cup T) \times (C \cup T) \rightarrow \{0,1\}, \quad i, j = 1,2,\ldots m+n, \quad i \neq j. \tag{12.18}$$

As can be observed in Equation 12.18, unified context includes relations among concepts and region types. To model these relations as they exist in real-world problems, a fuzzified ontology $\mathscr{O}$ should be defined as follows:

$$\mathscr{O} = \{C, T, \mathscr{R}\}, \tag{12.19}$$

where $\mathscr{R}$ contains fuzzified relations among entities. As in the crisp ontology, $\mathscr{R} = \mathscr{R}_{ij}$. Set $\mathscr{R}_{ij} = r_{ij}^{(k)}$, for $k = 1,2,\ldots,K+K'$, includes $K+K'$ among two entities x_i and x_j. For a given relation $r_{ij}^{(k)}$, its inverse $\bar{r}_{ji}^{(k)}$ can be defined. Finally, r_{ij} can be formally expressed as

$$r_{ij} : (C \cup T) \times (C \cup T) \rightarrow [0,1]. \tag{12.20}$$

Since there often exist more than one relation among two entities, their combination is defined as

$$\mathscr{U}_{ij} = \bigcup_k [\mathrm{r}_{ij}^{(k)}]^p, \; i, j = 1,2,\ldots,N; \; k = 1,2,\ldots,K'. \tag{12.21}$$

This way allows constructing the model to be used in the analysis step. The value of p is defined by the semantics of each relation r_{ij}. More specifically:

- $p = 1$, if the semantics of r_{ij} imply that it should be considered as is;
- $p = -1$, if the semantics of r_{ij} imply the use of its inverse $\bar{\mathrm{r}}_{ji}$; and
- $p = 0$, if the semantics of r_{ij} do not allow its participation in the construction of the combined relation $\mathscr{U}_{ij}$.

TABLE 12.3
Semantic relations used in unified context.

Relation	Inverse	Symbol	Meaning	$C \times C$	$T \times T$	$C \times T$
Similar	Similar	$Sim(a,b)$	similarity between a and b	–	•	–
Accompanier	Accompanier	$Acc(a,b)$	co-occurrence between a and b	•	•	•
Part	PartOf	$P(a,b)$	a is part of b	•	•	•
Component	ComponentOf	$Comp(a,b)$	a is a component of b	•	•	•
Specialization	Generalization	$Sp(a,b)$	b specializes the meaning of a	•	–	–
Example	ExampleOf	$Ex(a,b)$	b is an example of a	•	–	–
Location	LocationOf	$Loc(a,b)$	b is a location of a	•	–	–
Property	PropertyOf	$Pr(a,b)$	b is a property of a	–	•	•

12.5.3.2 Relations between Entities

The relations between entities, as defined in Sections 12.5.1 and 12.5.2, are summarized in Table 12.1. Note that each entity may be related to another with more than one relation. However, it should be made clear that not all of the relations are appropriate for any two given entities. For instance, *Similar* cannot be defined between two concepts or a concept and a region type, whereas *sea* cannot be *Similar* to *sand* or to a *brown* region type. All applicable pairs of entities for each relation are summarized in Table 12.3.

The appropriate degrees of confidence of the semantic relations are either defined by an expert or calculated as described in Sections 12.5.1 and 12.5.2. For example:

- *Similar* may be defined only between two region types, with their visual similarity denoting the degree of confidence.
- *Accompanier* denotes the co-occurrence of any two entities in the same image. It should be noted here that a region type that co-occurs with a high degree of confidence with a concept does not necessarily depicts this concept. This degree is calculated statistically.
- *PartOf* is defined for any two given entities, when one is part of the other. For example, in case of two concepts, *sea* is *PartOf Beach*. In case of two region types, a *green and textured* region type is *PartOf* a *green* region type. Finally, in case of a concept and a region type, a *green* region type is *PartOf* a *tree*.
- *Component* is defined for any two given entities. For example, in case of two concepts, *tree* is a *Component* of *forest*, and in case of two region types, a *dark green* region type is a *Component* of a *green* region type. Finally, in case of a concept and a region type, an *orange* region type is a *Component* of a *sunset*. It should be noted here that there is also the case that a concept may be *Component* of a region type, due to undersegmentation. However, this case is not considered in this approach, since local annotation per image are unavailable.
- *Specialization* may be defined between two concepts, as defined in Section 12.5.1.
- *Example* may be defined between two concepts, as defined in Section 12.5.1.
- *Location* may be defined between two concepts, as defined in Section 12.5.1.

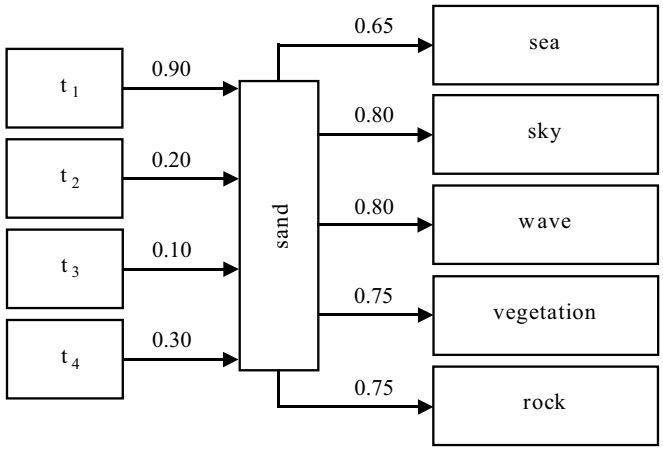


FIGURE 12.12

A fragment of unified context ontology that includes relations among *sand* and all other entities.

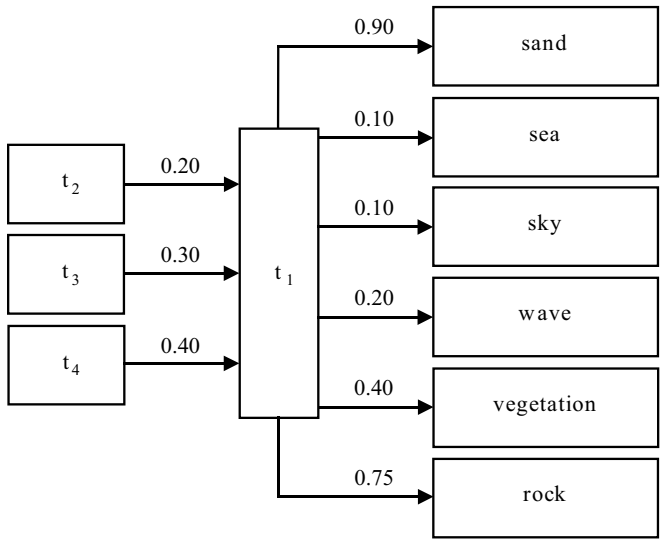


FIGURE 12.13

A fragment of unified context ontology that includes relations among region type t_1 and all other entities.

- *Property* may be defined between a concept and a region type, or between two concepts. In the first case, a *green* region type is a *Property* of *vegetation*. In the latter case, *wave* is a *Property* of *sea*.

The aforementioned relations model the unified context among all concepts and region types of a given domain. Between any two given entities x_i and x_j a single relation $\mathscr{U}_{ij}$ is formed and the occurring ontology $\mathscr{O}$ forms again an RDF graph, using RDF reification [85] to describe the degree of confidence for each edge. It is noted again that an edge between two entities is formed based on the set of valid relations for this pair. For example, the edge between *rock* and *sand* is formed by *Location* and *Accompanier*, while the

edge between *water* and *sea* is formed by *Specialization*, *PartOf*, *Example*, and *Location*. Similarly, a *green* and a *blue* region types are related by the combination of *Similar*, *Accompanier*, and *Component*, while a *blue* region type and *sea* are related by the combination of *Accompanier*, *PartOf*, and *Component*.

Figures 12.12 and 12.13 present two fragments of the graph built for the *Beach* domain. More specifically, Figure 12.12 depicts relations among *sea* and other entities, while Figure 12.13 depicts relations between region type T_4 and other entities.

12.5.3.3 Unified Context Optimization

Algorithm 12.3 is used for the optimization in the case of the unified context. This algorithm is a mixture of the ones presented in Sections 12.5.1 and 12.5.2. The target now is to refine model vectors and detector results in an iterative way.

ALGORITHM 12.3 Optimized iterative procedure for evaluating the degree of membership in the case of the unified context.

1. The considered domain imposes the use of a domain similarity (or dissimilarity) measure $w_m \in [0,1]$.
2. For each image p, the fuzzy set L_p with the degrees of confidence $\mu_p(c_i)$ is defined for all concepts c_i, with $i = 1,2,\ldots,k$ of the given domain.
3. For each image p, the fuzzy set L_T with the degrees of confidence $\mu_p(t_i)$ is defined for all region types t_i, with $i = 1,2,\ldots,k'$ of the visual thesaurus.
4. For each concept c_i in L_p with a degree of confidence $\mu_p(c_i)$, the particular contextual information is obtained in the form of the set $\mathscr{U}_{c,i} = \{\mathscr{U}_{c,ij} : c_i, c_j \in C, \forall i \neq j\}$.
5. For each region type t_i in L with a degree of confidence $\mu_p(T_i)$, the particular contextual information is obtained in the form of the set $\mathscr{U}_{T,i} = \{\mathscr{U}_{T,ij} : t_i, t_j \in T, \forall i \neq j\}$.
6. The new degrees of confidence $\mu'_p(c_i)$ and $\mu'_p(T_i)$ are calculated by taking into account the similarity measure of the given domain. In the case of multiple concept relations, when x_i is related with one or more types, apart from the root of the ontology, an intermediate aggregation step should be applied in order to calculate $\mu'_p(c_i)$ and $\mu'_p(T_i)$ by applying the context relevance notion [8], that is, $cr_i = \max_j\{\mathscr{U}_{ij}\}$, for $j = 1,2,\ldots,c_k$. The calculation of $\mu^l_p(c_i)$ and $\mu^l_p(t_i)$, which correspond to the degrees of confidence for the presence of c_i and t_i at the l-th iteration of the algorithm, is expressed with the recursive formula

$$\mu^l_p(x_i) = \mu^{l-1}_p(x_i) - w_m(\mu^{l-1}_p(x_i) - cr_i). \tag{12.22}$$

Equivalently, for the l-th iteration:

$$\mu^l_p(x_i) = (1 - w_m)^l \cdot \mu^0_p(x_i) + (1 - (1 - w_m)^l) \cdot cr_i,$$

where $\mu^0_p(x_i)$ denotes the initial degree of confidence for x_i.

The number of the iterations is defined empirically and usually three to six iterations are enough also in this case. A flowchart that describes the influence of the unified context is depicted in Figure 12.14.

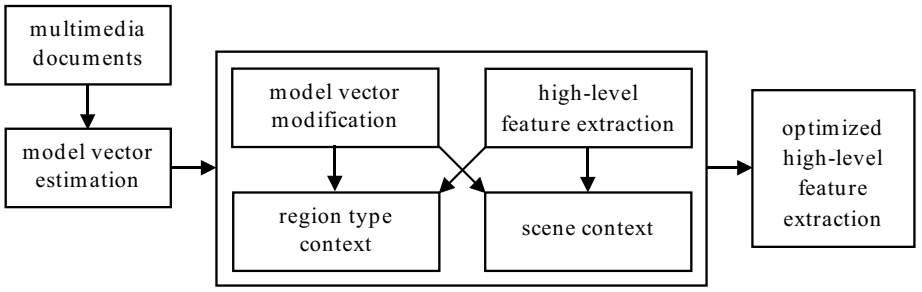


FIGURE 12.14
Unified contextual influence.

12.6 Context Optimization Examples

In order to clarify the influence of the context optimization on image analysis, this section presents simplistic examples for all three above cases.

12.6.1 Scene Context Example

Focused on scene context optimization, a simple example is presented to illustrate the way in which this optimization influences on the initial degrees of confidence. Based on the context ontology, whose fragments are depicted in Figures 12.6 and 12.7, and using appropriately trained detectors, Table 12.4 depicts degrees of confidence for all concepts before and after context optimization for the images shown in Figure 12.15. It can be observed that concepts detected with high confidence are considered to appear on images with a higher confidence. The opposite may be observed for concepts initially detected with low confidence.

12.6.2 Region Type Context Example

Next, a simple example is presented to illustrate region type context influence. In this case, context acts as a preprocessing step. For the image shown in Figure 12.10 and the

TABLE 12.4
Degrees of confidence before and after scene context optimization for the images shown in Figure 12.15.

	Figure 12.15a		Figure 12.15b		Figure 12.15c	
Concept	Before	After	Before	After	Before	After
sea	0.77	0.85	0.65	0.75	0.62	0.72
water	0.63	0.70	0.60	0.69	0.58	0.67
vegetation	0.35	0.43	0.35	0.40	0.62	0.72
sky	0.45	0.57	0.55	0.60	0.53	0.61
sand	0.69	0.75	0.45	0.56	0.52	0.60
rock	0.25	0.35	0.63	0.68	0.65	0.75
wave	0.00	0.00	0.25	0.34	0.20	0.27

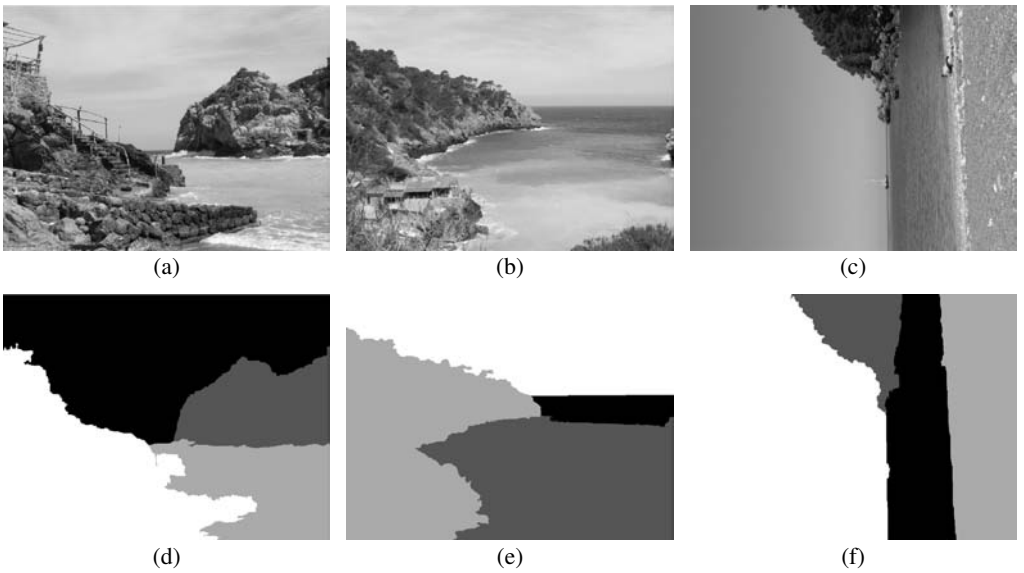


FIGURE 12.15 (See color insert.)

Three examples from the *Beach* domain. Initial images and their segmentation maps.

simplistic ontology, a fragment of which is depicted in Figure 12.9, the model vector calculated from the visual features is as follows:

$$\mathbf{MV}_{before} = \left[0.723\ 0.220\ 0.753\ 0.364\right]. \tag{12.23}$$

As it can be observed from the example image, it depicts *sky* and *sea* and intuitively one would expect that it should contain region types similar to those of the region thesaurus. However, in this case, *sea* is significantly different, perhaps more similar to a *rock*. After region type optimization, the model vector takes the following form:

$$\mathbf{MV}_{after} = \left[0.778\ 0.452\ 0.800\ 0.338\right]. \tag{12.24}$$

Thus, the degree of confidence for the region type that corresponds to *sea* (2nd constituent) is increased while the one that corresponds to *rock* (4th constituent) is decreased.

12.6.3 Unified Context Example

Finally, in the unified context case, the ontology with fragments depicted in Figures 12.12 and 12.13 is used. The proposed algorithm is applied to the image shown in Figure 12.10. Model vector T is initially set as

$$\mathbf{T} = \{T_i\} = \left[0.89\ 0.62\ 0.21\ 0.68\ 0.670.31\right], \tag{12.25}$$

while the degrees of confidence c_i are

$$\mathbf{C} = \{c_i\} = \left[0.32\ 0.91\ 0.12\ 0.87\ 0.35\right]. \tag{12.26}$$

As can be seen, the input image depicts *sea*, *sky*, and *wave*. However, the initial confidence for *sea* was low because no similar instances of *sea* were part of the training set. After unified context optimization, the improved value of the model vector T' is

$$\mathbf{T}' = \{T_i'\} = \left[0.89\ 0.62\ 0.21\ 0.68\ 0.670.31\right], \tag{12.27}$$

and the degrees of confidence for all concepts are

$$\mathbf{C}' = \{c_i'\} = \left[0.62\ 0.95\ 0.18\ 0.90\ 0.29\right]. \tag{12.28}$$

In brief, it should be emphasized that the unified context algorithm exploited the following information that was stored in the unified ontology:

- This was a *Beach* image, thus using the appropriate ontology.
- *Sky* was initially detected with a high confidence.
- *Wave* was initially detected with a high confidence.
- Image contains a *blue* region type.
- Image contains a *white* region type.
- *sky* and *wave* are related with a high degree with *sea*.
- *blue* and *white* region type are related with a high degree with *sea*.

Thus, the model vector and the degrees of confidence were modified in a way that:

- The confidence for the *blue* region type was increased.
- The confidence for the other region types remained invariable.
- The confidence for *sea* was increased.
- The confidence for the other concepts remained invariable.

12.7 Experimental Results

The following presents an indicative selection of experimental results. It includes results from the application of the proposed visual context utilization methodology, as presented in Sections 12.5.1.2 to 12.5.3.3. The utilized expert knowledge is rather ad hoc; however, this is not considered to be a liability, nor part of the discussed context model and is aligned to the current application datasets. More specifically, the evaluation focuses on both utilizing parts of the well-known Corel and TRECVID datasets and compares the efficiency of relevant state-of-the-art techniques.

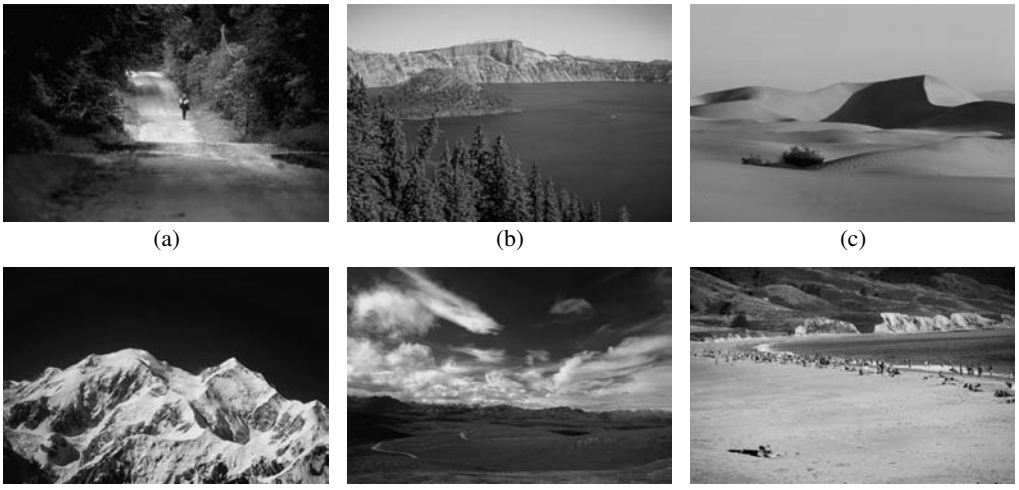


FIGURE 12.16 (See color insert.)

Indicative Corel images.

Initially, a set of experimental results are presented from the application of the proposed visual context approaches on a dataset containing 750 images, 40 region types, and 6 high-level concepts (*sea*, *vegetation*, *sky*, *sand*, *rock*, and *wave*). The number of the region types was selected based on the size of the region thesaurus and was verified using the minimum description length (MDL) method [28]. The amount and type of utilized concepts is imposed by the problem/dataset at hand; the employed dataset was a subset of the well-known Corel image collection [86], an indicative sample of which is presented in Figure 12.16. For the non-contextual detection of high-level concepts, the methodology described in Section 12.4 was applied. Overall, 525 images were used to train six individual SVM-based concept classifiers and 225 images were used as the test set.

Some additional results are also presented from the application of the proposed unified contextualization approach on a second dataset, consisting of 4000 images from the TRECVID collection [87], 100 region types and 7 high-level concepts (*vegetation*, *road*, *fire*, *sky*, *snow*, *sand*, *water*). The number of region types for this dataset was again decided based on experiments on the size of the region thesaurus and verified by using the minimum description length methodology introduced in Reference [28]. Figure 12.17 shows a characteristic sample of this dataset. In total, 250 of those images were used to train 7 individual SVM classifiers, whereas other 997 images were used for testing.

To evaluate the proposed approaches, they are compared to similar techniques used in previous research work. The results of all approaches on both Corel and TRECVID datasets are summarized in Tables 12.5 and 12.6. Note that *Region Types (RT)* refers to the results based only on the detection scheme presented in Section 12.4, without exploiting any contextual knowledge. Results from the application of contextual approaches correspond to *RT+Scene Context (SC)*, *RT+RT Context (RTC)*, and *RT+Unified Context* (UC).

To further evaluate the last proposed approach, two other techniques are implemented. The first technique, known as *relative LSA* (RLSA) [42], adds directly structural constraints

FIGURE 12.17 (See color insert.)

Indicative TRECVID images.

to the visual words of the thesaurus. The fundamental difference between the traditional LSA and RSLA is that every possible unordered pair of clusters is this time considered as a visual word. In this way, a visual thesaurus with too many words (that is, pairs of clusters) is created. Nevertheless, the low-level features extracted from each region are simpler than the MPEG-7 low-level features used here. More specifically, a 64-bin histogram expressed in the hue-saturation-value (HSV) color space is used to capture the color features of 24 Gabor filters whose energies capture the texture features. The number of the words that form the visual thesaurus is determined empirically. The second implemented technique [88] starts with the extraction of *local interest points* (LIPs). The local interest points, often denoted as *salient*, tend to have significantly different properties compared to all other pixels in their neighborhood. To extract these points, a method called *difference of Gaussians* is applied. From each LIP, a SIFT descriptor is extracted from an elliptic region. A visual thesaurus is generated by an offline quantization of LIPs. Then, using this thesaurus, each image is described as a vector of visual keywords. Finally, for each high-level concept, a classifier is trained. It should be noted that the proposed method is compared against the above techniques and methodologies mainly because they try to face the same problem with more or less the same motivation as the presented work. The first one tries to exploit the co-occurrence of region types and to incorporate structural knowledge when building a visual thesaurus, while the the other one defines the LIPs as the regions of interest and extracts therein appropriate low-level descriptors. Moreover, both works have been successfully applied to the TRECVID dataset. Finally, as it is obvious from the interpretation of Tables 12.5 and 12.6, the proposed contextual unified approach outperforms in principle all compared approaches in terms of the achieved precision, whereas in some cases it lacks in terms of the recall criterion.

TABLE 12.5
Comparative precision P, recall R, and F-measure scores per concept for six different concept detection methodologies applied on the Corel dataset.

	RT			SC			RTC			UC			LIPs [88]			RLSA [42]		
Concepts	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
road	0.22	0.40	0.28	0.25	0.37	0.30	0.39	0.32	0.35	0.43	0.35	0.39	0.34	0.37	0.35	0.42	0.35	0.38
sand	0.38	0.50	0.43	0.40	0.46	0.43	0.50	0.41	0.45	0.55	0.44	0.49	0.47	0.46	0.46	0.52	0.45	0.48
sea	0.72	0.85	0.78	0.71	0.81	0.76	0.81	0.78	0.79	0.89	0.80	0.84	0.77	0.83	0.80	0.80	0.82	0.81
sky	0.81	0.88	0.84	0.79	0.80	0.79	0.77	0.81	0.79	0.88	0.82	0.85	0.86	0.85	0.85	0.88	0.83	0.85
snow	0.48	0.68	0.56	0.51	0.62	0.56	0.62	0.57	0.59	0.72	0.57	0.64	0.58	0.61	0.59	0.64	0.57	0.60
vegetation	0.67	0.81	0.73	0.67	0.76	0.71	0.67	0.71	0.69	0.81	0.74	0.77	0.73	0.76	0.74	0.76	0.73	0.74
Total:	0.55	0.69	0.61	0.56	0.64	0.59	0.63	0.60	0.61	0.71	0.62	0.57	0.62	0.65	0.54	0.67	0.63	0.55

TABLE 12.6
Comparative precision P, recall R, and F-measure scores per concept for six different concept detection methodologies applied on the TRECVID dataset.

	RT			SC			RTC			UC			LIPs [88]			RLSA [42]		
Concepts	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
vegetation	0.50	0.64	0.56	0.50	0.60	0.55	0.68	0.49	0.57	0.78	0.45	0.57	0.50	0.59	0.54	0.52	0.55	0.54
road	0.22	0.31	0.26	0.25	0.30	0.27	0.41	0.27	0.33	0.43	0.24	0.31	0.30	0.30	0.30	0.37	0.27	0.31
sand	0.83	0.82	0.81	0.87	0.80	0.83	0.93	0.69	0.79	1.00	0.76	0.86	0.93	0.76	0.84	0.94	0.71	0.81
water	0.60	0.67	0.63	0.57	0.68	0.62	0.70	0.58	0.63	0.81	0.57	0.67	0.60	0.66	0.63	0.61	0.64	0.63
sky	0.60	0.79	0.68	0.62	0.77	0.69	0.74	0.68	0.71	0.90	0.57	0.70	0.59	0.79	0.67	0.60	0.76	0.67
snow	0.43	0.50	0.46	0.40	0.44	0.42	0.51	0.38	0.44	0.57	0.37	0.44	0.50	0.44	0.47	0.56	0.40	0.47
fire	0.30	0.47	0.37	0.22	0.44	0.29	0.46	0.37	0.41	0.55	0.36	0.43	0.38	0.45	0.41	0.45	0.43	0.44
Total:	0.50	0.60	0.55	0.49	0.58	0.54	0.63	0.50	0.56	0.72	0.47	0.57	0.54	0.57	0.55	0.58	0.54	0.55

Experimental results presented above show that existing relationships between concepts improve the precision of results, not only for the well trained, but also for weak SVM classifiers. The proposed unified context algorithm uses and exploits these relations and provides an expanded view of the research problem, which is based on a set of meaningful semantic relations. The interpretation of presented experimental results depicts that the proposed contextualization approach will favor rather certain degrees of confidence for the detection of a concept that exists within an image. On the contrary, it will also discourage rather uncertain or misleading degrees. It will strengthen the concepts' differences, but it will treat smoothly almost certain concepts' confidence values. Finally, based on the constructed knowledge, the algorithm is able to disambiguate cases of similar concepts or concepts being difficult to be detected from the simple low-level analysis steps.

12.8 Conclusion

Research effort summarized in this chapter clearly indicates that high-level concepts can be efficiently detected when an image is represented by a model vector with the aid of a visual thesaurus and visual context. The role of the latter is crucial and significantly aids the image analysis process. The core contributions of this work include, among others, the implementation of a novel threefold visual context interpretation utilizing a fuzzy, ontology-based representation of knowledge. Experimental results presented in this chapter indicate significant high-level concept detection optimization over the entire datasets. Although the improvement is not considered to be impressive, it is believed that the proposed approach successfully incorporates the underlying contextual knowledge and further exploits visual context in the multimedia analysis value chain. Moreover, minor enhancements of the implemented contextual model, for example, in terms of additional spatial, temporal, or semantic relationships exploitation, would further boost its performance.

Acknowledgment

The work presented in this book chapter was partially supported by the European Commission under contract FP7-215453 WeKnowIt.

References

[1] I. Biederman, R. Mezzanotte, and J. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive Psychology*, vol. 14, no. 2, pp. 143–177, April 1982.

[2] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, December 2000.

[3] K. Rapantzikos, Y. Avrithis, and S. Kollias, "On the use of spatiotemporal visual attention for video classification," in *Proceedings of the International Workshop on Very Low Bitrate Video Coding*, Sardinia, Italy, September 2005.

[4] G. Tsechpenakis, G. Akrivas, G. Andreou, G. Stamou, and S. Kollias, "Knowledge-assisted video analysis and object detection," in *Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems*, Algarve, Portugal, September 2002, pp. 497–504.

[5] A. Benitez and S. Chang, "Image classification using multimedia knowledge networks," in *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003, pp. 613–616.

[6] S. Staab and R. Studer, *Handbook on Ontologies*. New York, USA: Springer Verlag, 2004.

[7] A. Mathes, "Folksonomies-cooperative classification and communication through shared metadata," *Computer Mediated Communication*, vol. 47, pp. 1–28, December 2004.

[8] P. Mylonas, T. Athanasiadis, and Y. Avrithis, "Image analysis using domain knowledge and visual context," in *Proceedings of the 13th International Conference on Systems, Signals and Image Processing*, Budapest, Hungary, September 2006, pp. 483–486.

[9] J. Luo, A. Singhal, and W. Zhu, "Natural object detection in outdoor scenes based on probabilistic spatial context models," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Baltimore, USA, July 2003, pp. 457–460.

[10] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A database and Web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 157–173, May 2008.

[11] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge 2008 results." *International Journal of Computer Vision*, vol. 88, no. 2, June 2010, pp. 303–338.

[12] L. Kennedy, A. Hauptmann, M. Naphade, A. Smith, and S. Chang, "LSCOM lexicon definitions and annotations version 1.0," in *Proceedings of the DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia*, ADVENT Technical Report 217-2006-3, New York, USA, March 2006.

[13] S. Ayache and G. Quenot, "TRECVID 2007 collaborative annotation using active learning," in *Proceedings of the TRECVID Workshop*, Gaithersburg, MD, USA, November 2007.

[14] C. Cusano, "Region-based annotation of digital photographs," in *Proceedings of the Computational Color Imaging Workshop*, Milan, Italy, April 2011, pp. 47–59.

[15] N. Aslam, J. Loo, and M. Loomes, "Adding semantics to the reliable object annotated image databases," *Procedia Computer Science*, vol. 3, pp. 414–419, February 2011.

[16] M. Wang and X. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 2, pp. 1–21, February 2011.

[17] J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," in *Proceedings of DAGM Pattern Recognition Symposium*, vol. 3175, September 2004, pp. 195–203.

[18] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, June 2005, pp. 524–531.

[19] S. Ullman and M. Vidal-Naquet, "Visual features of intermediate complexity and their use in classification," *Nature Neuroscience*, vol. 5, no. 7, pp. 682–687, June 2002.

[20] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of the International Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, May 2004.

[21] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Proceedings of the IEEE International Conference on Computer Vision*, Beijing, China, October 2005, pp. 370–377.

[22] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *The Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, February 2003.

[23] S. Chang, T. Sikora, and A. Purl, "Overview of the MPEG-7 standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688–695, June 2001.

[24] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999, pp. 1150–1157.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Lecture Notes in Computer Science*, vol. 3951, pp.404–417, February 2006.

[26] J. Hartigan and M. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[27] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, June 2006, pp. 2161–2168.

[28] S. Kim and I. Kweon, "Simultaneous classification and visual word selection using entropy-based minimum description," in *Proceedings of the International Conference on Pattern Recognition*, Hong Kong, August 2006, pp. 650–653.

[29] O. Cula and K. Dana, "Compact representation of bidirectional texture functions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, USA, December 2001, pp. 1041–1047.

[30] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29–44, June 2001.

[31] B. Le Saux and G. Amato, "Image recognition for digital libraries," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, New York, USA, October 2004, pp. 91–98.

[32] D. Gokalp and S. Aksoy, "Scene classification using bag-of-regions representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, June 2007, pp. 1–8.

[33] J. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *Proceedings of the International Conference on Multimedia and Expo*, Baltimore, MD, USA, July 2003, pp. 445–448.

[34] A. Opelt, A. Pinz, and A. Zisserman, "Incremental learning of object detectors using a visual shape alphabet," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, June 2006, pp. 3–10.

[35] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, June 2006, pp. 2169–2178.

[36] L. Zhu and A. Zhang, "Theory of keyblock-based image retrieval," *ACM Transactions on Information Systems*, vol. 20, no. 2, pp. 224–257, April 2002.

[37] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, June 2006, pp. 2033–2040.

[38] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky, "Learning hierarchical models of scenes, objects, and parts," in *Proceedings of the IEEE International Conference on Computer Vision*, Beijing, China, October 2005, pp. 1331–1338.

[39] J. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, June 2007, pp. 1–8.

[40] B. Leibe and B. Schiele, "Interleaved object categorization and segmentation," in *Proceedings of the British Machine Vision Conference*, Norwich, UK, September 2003, pp. 759–768.

[41] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, September 1990.

[42] F. Souvannavong, B. Merialdo, and B. Huet, "Region-based video content indexing and retrieval," in *International Workshop on Content-Based Multimedia Indexing*, Riga, Latvia, June 2005.

[43] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, April 2008.

[44] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychological Review*, vol. 113, no. 4, pp. 766–786, October 2006.

[45] M. Naphade and J. Smith, "A hybrid framework for detecting the semantics of concepts and context," *Lecture Notes in Computer Science*, vol. 2728, pp. 196–205, Springer, 2003.

[46] J. Fan, Y. Gao, and H. Luo, "Multi-level annotation of natural scenes using dominant image components and semantic concepts," in *Proceedings of the ACM International Conference on Multimedia*, New York, USA, ACM 2004, pp. 540–547.

[47] R. Yan, M. Chen, and A. Hauptmann, "Mining relationship between video concepts using probabilistic graphical model," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Toronto, ON, Canada, pp. 301–304.

[48] K. Murphy, A. Torralba, and W. Freeman, "Using the forest to see the trees: A graphical model relating features, objects and scenes," *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[49] P. Lipson, E. Grimson, and P. Sinha, "Configuration based scene classification and image indexing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, June 1997, pp. 1007–1013.

[50] A. Singhal, J. Luo, and W. Zhu, "Probabilistic spatial context models for scene content understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, June 2003, pp. 235–241.

[51] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition," *Lecture Notes in Computer Science*, vol. 3021, pp. 350–362, May 2004.

[52] W. Li and M. Sun, "Semi-supervised learning for image annotation based on conditional random fields," *Lecture Notes in Computer Science*, vol. 4071, pp. 463–472, July 2006.

[53] J. Li, A. Najmi, and R. Gray, "Image classification by a two-dimensional hidden Markov model," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 517–533, February 2000.

[54] J. Jiten, B. Merialdo, and B. Huet, "Semantic feature extraction with multidimensional hidden Markov model," *Proceedings of SPIE*, vol. 6073, pp. 211–221, January 2006.

[55] J. Yuan, J. Li, and B. Zhang, "Exploiting spatial context constraints for automatic image region annotation," in *Proceedings of the ACM International Conference on Multimedia*, Ausburg, Germany, September 2007, pp. 595–604.

[56] M. Boutell, C. Brown, and J. Luo, "Learning spatial configuration models using modified Dirichlet priors," in *Proceedings of the ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields*, Banff, AB, Canada, July 2004, pp. 29–34.

[57] M. Boutell, J. Luo, and C. Brown, "Improved semantic region labeling based on scene context," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Amsterdam, the Netherlands, July 2005, pp. 980–983.

[58] M. Boutell, J. Luo, and C. Brown, "A generalized temporal context model for classifying image collections," *Multimedia Systems*, vol. 11, no. 1, pp. 82–92, November 2005.

[59] L. Paletta, M. Prantl, and A. Pinz, "Learning temporal context in active object recognition using Bayesian analysis," in *Proceedings of the International Conference on Pattern Recognition*, Barcelona, Spain, September 2000, pp. 695–699.

[60] D. Moldovan, C. Clark, and S. Harabagiu, "Temporal context representation and reasoning," in *Proceedings of the International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, August 2005, pp. 1099–1104.

[61] N. O'Hare, C. Gurrin, H. Lee, N. Murphy, A. Smeaton, and G. Jones, "My digital photos: Where and when?," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, Singapore, November 2005, pp. 261–262.

[62] J. Pauty, P. Couderc, and M. Banâtre, "Using context to navigate through a photo collection," in *Proceedings of the ACM International Conference on Human Computer Interaction with Mobile Devices and Services*, Salzburg, Austria, September 2005, pp. 145–152.

[63] P. Mulhem and J. Lim, "Home photo retrieval: Time matters," *Lecture Notes in Computer Science*, vol. 2728, pp. 321–330, July 2003.

[64] Japan Electronics and Information Technology Industries Association, "Exchangeable image file format for digital still cameras: Exif Version 2.2," Technical report, JEITA CP-3451, April 2002.

[65] P. Sinha and R. Jain, "Classification and annotation of digital photos using optical context data," in *Proceedings of the International Conference on Content-Based Image and Video Retrieval*, Niagara Falls, ON, Canada, July 2008, pp. 309–318.

[66] X. Liu, L. Zhang, M. Li, H. Zhang, and D. Wang, "Boosting image classification with LDA-based feature combination for digital photograph management," *Pattern Recognition*, vol. 38, no. 6, pp. 887–901, June 2005.

[67] M. Tuffield, S. Harris, D. Dupplaw, A. Chakravarthy, C. Brewster, N. Gibbins, K. O Hara, F. Ciravegna, D. Sleeman, and N. Shadbolt, "Image annotation with photocopain," in *Proceedings of the Semantic Web Annotation of Multimedia Workshop at the World Wide Web Conference*, Edinburgh, Scotland, May 2006.

[68] M. Boutell and J. Luo, "Bayesian fusion of camera metadata cues in semantic scene classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, July 2004, vol. II, pp. 623–630.

[69] M. Boutell and J. Luo, "Beyond pixels: Exploiting camera metadata for photo classification," *Pattern Recognition*, vol. 38, no. 6, pp. 935–946, June 2005.

[70] M. Boutell and J. Luo, "Photo classification by integrating image content and camera metadata," in *Proceedings of the International Conference on Pattern Recognition*, Cambridge, UK, August 2004, vol. 4, pp. 901–904.

[71] S. Boll, P. Sandhaus, A. Scherp, and S. Thieme, "MetaXa – context- and content-driven metadata enhancement for personal photo books," in *Proceedings of the International Multi-Media Modeling Conference*, Singapore, January 2007, pp. 332–343.

[72] E. Spyrou and Y. Avrithis, "A region thesaurus approach for high-level concept detection in the natural disaster domain," in *Proceedings of the International Conference on Semantic and Digital Media Technologies*, Genova, Italy, December 2007, pp. 74–77.

[73] P. Mylonas, E. Spyrou, and Y. Avrithis, "Enriching a context ontology with mid-level features for semantic multimedia analysis," in *Proceedings of the 1st Workshop on Multimedia Annotation and Retrieval Enabled by Shared Ontologies*, Genova, Italy, December 2007, pp. 16–30.

[74] P. Mylonas, E. Spyrou, and Y. Avrithis, "High-level concept detection based on mid-level semantic information and contextual adaptation," in *Proceedings of the 2nd International Workshop on Semantic Media Adaptation and Personalization*, London, UK, December 2007, pp. 193–198.

[75] E. Spyrou, P. Mylonas, and Y. Avrithis, "Semantic multimedia analysis based on region types and visual context," in *Proceedings of the Artificial Intelligence and Innovations: From Theory to Applications*, Athens, Greece, September 2007, pp. 389–398.

[76] E. Spyrou, G. Tolias, P. Mylonas, and Y. Avrithis, "A semantic multimedia analysis approach utilizing a region thesaurus and LSA," in *Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, Klagenfurt, Austria, May 2008, pp. 8–11.

[77] Y. Avrithis, A. Doulamis, N. Doulamis, and S. Kollias, "A stochastic framework for optimal key frame extraction from MPEG video databases," *Computer Vision and Image Understanding*, vol. 75, no. 1, pp. 3–24, July 1999.

[78] G. Tolias, *VDE: Visual descriptor extraction*, 2008. Available online, http://image.ntua.gr/smag/tools/vde.

[79] A. Yamada, M. Pickering, S. Jeannin, L. Cieplinski, J. Ohm, and M. Kim, "MPEG-7 visual part of experimentation model version 9.0 ISO/IEC JTC1/SC29/WG11/N3914," International Organisation for Standardisation ISO, 2001, pp. 1–83.

[80] E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor, "Fusing MPEG-7 visual descriptors for image classification," *Proceedings of the Artificial Neural Networks: Formal Models and Their Applications-ICANN*, Warsaw, Poland, September 2005, pp. 847–852.

[81] P. Mylonas, M. Wallace, and S. Kollias, "Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering," *Lecture Notes in Artificial Intelligence*, vol. 3025, pp. 191–200, Springer 2004.

[82] P. Mylonas and Y. Avrithis, "Context modeling for multimedia analysis," in *Proceedings of the International and Interdisciplinary Conference on Modeling and Using Context*, Paris, France, July 2005.

[83] A. Benitez, D. Zhong, S. Chang, and J. Smith, "MPEG-7 MDS content description tools and applications," *Lecture Notes in Computer Science*, vol. 2124, pp. 41–52, September 2001.

[84] W3C, "RDF Reification," 2004. Available online, http://www.w3.org/TR/rdf-schema/#ch_reificationvocab.

[85] D. Beckett and B. McBride, "RDF/XML syntax specification (revised)," W3C Recommendation, vol. 10, 2004.

[86] J. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947–963, September 2001.

[87] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, CA, USA, October 2006, pp. 321–330.

[88] Y. Jiang, W. Zhao, and C. Ngo, "Exploring semantic concept using local invariant features," in *Proceedings of the Asia-Pacific Workshop on Visual Information Processing*, Beijing, China, November 2006.