Using Visual Context and Region Semantics for High-Level Concept Detection

Phivos Mylonas, *Member, IEEE*, Evaggelos Spyrou, *Student Member, IEEE*, Yannis Avrithis, *Member, IEEE*, and Stefanos Kollias, *Member, IEEE*

***Abstract*—In this paper we investigate detection of high-level concepts in multimedia content through an integrated approach of visual thesaurus analysis and visual context. In the former, detection is based on model vectors that represent image composition in terms of region types, obtained through clustering over a large data set. The latter deals with two aspects, namely high-level concepts and region types of the thesaurus, employing a model of a priori specified semantic relations among concepts and automatically extracted topological relations among region types; thus it combines both conceptual and topological context. A set of algorithms is presented, which modify either the confidence values of detected concepts, or the model vectors based on which detection is performed. Visual context exploitation is evaluated on TRECVID and Corel data sets and compared to a number of related visual thesaurus approaches.**

***Index Terms*—Concept detection, contextualization, region thesaurus, region types, visual context.**

I. Introduction

As far as the current content-based image analysis and retrieval systems are concerned, most of them are limited by the existing state-of-the-art in image understanding, in the sense that they usually take a relatively low-level approach and fall short of higher-level interpretation and knowledge. Knowledge-based techniques, human perception and scene content understanding techniques have started to gain focus on the task of bridging the semantic and conceptual gap that exists between humans and computers. In this paper, we shall provide our research view on modelling and exploiting contextual information towards efficient high-level content understanding. This kind of information acts as a simulation of the human visual perception scheme, by taking into account all information relative to the visual content of a scene [4]. As a result, context may be used to improve the performance of automated systems for knowledge-assisted analysis, semantic indexing and retrieval of multimedia content.

It is true that combining context with semi-automated high-level concept detection or scene classification techniques, in order to achieve better semantic results during the multimedia content analysis phase, is a challenging and broad research area for any researcher. Although the well-known "semantic gap" [64] has been acknowledged for a long time, multimedia analysis approaches are still divided into two rather discrete categories; low-level multimedia analysis methods and tools, on the one hand (e.g., [59]) and high-level semantic annotation methods and tools, on the other (e.g., [78], [5]). It was only recently, that state-of-the-art multimedia analysis systems have started using semantic knowledge technologies, as the latter are defined by notions like ontologies [74], folksonomies [43] and the Semantic Web standards. Their advantages, when using them for creation, manipulation and post-processing of multimedia metadata, are depicted in numerous research activities. The core idea is to combine such formalized knowledge and a set of features to describe the visual content of an image or its regions, like, for instance, in [81], where a region-based approach using MPEG-7 visual features and ontological knowledge is presented.

The rest of this paper is structured as follows: In Section II, we describe briefly the problem this work attempts to address. In Section III, we place our work within related literature, whereas in Section IV, our motivation in utilizing the notion of visual context in concept detection and scene classification is described. Section V deals with the proposed enhanced visual conceptualization and the determination of an image's region types. In Section VI, the overall fuzzy context knowledge formalization is described, together with the proposed contextual adaptation in terms of the visual context algorithm and its optimization steps. Section VII lists our experimental results on well-known datasets and Section VIII concludes our work.

II. Problem Formulation

Conjunction of any form of contextual information with image content features in the means of either low-level features (e.g. color, texture, shape) or semantic concepts (e.g., *sky*, *sand*, *sea*, etc.) constitutes one of the most interesting research problems towards efficient concept detection and image classification. The use of enhanced visual information, such as information obtained from the application of supervised or unsupervised learning methodologies on low-level features is introduced in our work to improve the results of traditional knowledge-assisted image analysis, based both on low-level *visual* and high-level *contextual* information. In general, this information forms an intermediate description, which may be semantically described, but at the same time does not express the high-level concepts. Thus, a relative knowledge model is constructed, containing this kind of information, which may be described as follows:

Manuscript received April 17, 2008; revised October 06, 2008. Current version published January 16, 2009. This work was supported in part by the European Commission under contracts (FP6-027685 - MESH) and (FP7-215453 - WeKnowIt). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

The authors are with the Image, Video and Multimedia Laboratory, National Technical University of Athens, Zographou Campus, PC 157 80 Athens, Greece (e-mail: fmylonas@image.ntua.gr; espyrou@image.ntua.gr; iavr@image.ntua.gr; stefanos@image.ntua.gr).

Digital Object Identifier 10.1109/TMM.2008.2009681

- a *low-level* description, but then again one level *above* the one extracted from the multimedia content;
- a *high-level* conceptual description, but then again one level *below* the ultimate goal;
- an *in-between* description, which can be described semantically, but is not a high-level concept.

As a result, the main task of this work is the detection of high-level concepts in multimedia content through an integrated approach of both using visual thesaurus to represent the visual image properties and visual context in the detection process. The first part is based on the construction of model vectors that represent a given image in terms of the region types it contains and are obtained through clustering over a large, globally annotated, available image data set. The second part is modelled upon "fuzzified" contextual semantic relations among concepts (i.e., conceptual context), as well as "fuzzified" contextual topological relations among region types (i.e., topological context). We should emphasize at this point that our approach uses a globally annotated training dataset; thus, it detects whether a high-level concept exists within the image in question and not actual place of the concept with regard to a region.

III. Related Work

Being an appealing research problem, detection of high-level concepts within multimedia content is, in principle, a *bottom-up* approach. It focuses on local analysis to detect and recognize specific objects within an image, without utilizing explicit knowledge of the surrounding context (e.g., recognize a building or a tree). Acknowledging the need for providing such an analysis, many research efforts set focus on low-level feature extraction in a way to efficiently describe the 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- 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 (e.g., 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.

When focusing solely on the visual part of the analysis phase, it is true that high-level concept detection remains still an unsolved research task. Its two main and most interesting aspects appear the selection of the low-level features that will be extracted and the approach that will be used for assigning these low-level descriptions to high-level concepts. Plenty of works exist towards the solution of this problem. In [42], a multimedia analysis and retrieval system using multimodal machine learning techniques in order to model semantic concepts in video is presented. In [55], the structure of a scene image is estimated by the mean of global image features and is used to enhance object recognition in natural scenes. Some further research works fall in the category of the "bag-of-words" approach. There, an image is decomposed to a set of "visual words" derived after clustering or segmentation of the input image. Souvannavong *et al.* in [67] and [68] use a region-based approach in content retrieval that uses Latent Semantic Indexing (LSI) techniques. In [60], low-level features are extracted by segmented regions of an image, utilizing a mean-shift algorithm in the process. In [22], images are partitioned in regions, regions are clustered to obtain a codebook of region types, and a bag-of-regions approach is applied for scene representation. Moreover, in [27], the classification process starts with the extraction of Local Interest Points based on the extraction of SIFT features, quantized using a visual dictionary. Boujemaa *et al.* [6] present a hybrid thesaurus approach, whereas in [14], visual categorization is achieved using a bag-of-keypoints approach. In [56], the authors train separate shape detectors using a shape alphabet, which is actually a dictionary of curve fragments. A lexicon-driven approach, used within an interactive video retrieval system, is presented by Snoek *et al.* in [66] and [65]. Finally, in [34], an approach for texture and object recognition is presented. Therein, textures are represented using a visual dictionary found by quantizing appearance-based descriptors of local features.

In the field of segment-based image annotation, several research approaches adopt part-based visual features, either grids ([21], [29], [36]) or segmented regions ([3], [26], [28]), but they do not aim to establish the correspondences between individual parts and semantic labels. Fan *et al.* [18] and Yang *et al.* [85] have proposed different approaches for image-level annotation by firstly establishing correspondences between salient (or representative) regions and semantic labels. In [40], an exemplar-based algorithm for detection and segmentation is presented, while in [23], an approach to include contextual features for labeling images is discussed. There each pixel is assigned to one of a finite set of labels using multiscale conditional random fields. The advanced problem of image segmentation and region labeling is tackled in [83]. Therein, after an initial over-segmentation step, model-based recognition techniques are used to refine segmentation and to detect several concepts. Moreover, in [58], an object class recognition and segmentation system is presented. It is trained using weakly supervised training data, with the goal of examining the influence that different model choices can have on its performance.

On the other hand, both current and prior research activities focus either on low- or high-level interpretations in a totally discriminated manner. However, this kind of approach alone is not considered to be enough for efficient multimedia processing. Contextual information in terms of specific concepts, objects and events (e.g., typically present in an outdoor scenery) could be a considerable source of useful information [77]. A significant number of misclassifications usually occur because of the similarities in low-level characteristics of various object types and the lack of such high-level contextual information underlies as the major limitation of individual object detectors. Generic algorithms for automatic object recognition, region-level annotation, object recognition ([12], [15], [62]) and scene classification [75] are unfortunately not producing reliable results. Moreover, restricting the problem at hand to a specific domain of interest does not provide a global and thus satisfactory solution.

The introduction of contextual information in the above processes is not a straightforward task. As can be found in the literature, the term *context* [44] may be interpreted and even defined

in numerous ways, varying from the philosophical to the practical point of view [35]. It can take on many meanings and there is not one definition that is felt to be globally acceptable and that covers all the ways the term is used [16]. The notion of *visual context* is introduced in [49], [77] and [51], as an extra source of information for both object detection and scene classification. An ontology-based classification scheme is presented in [84]. There it is shown that ontology-based concept learning improves the accuracy of a concept by considering the possible influence relations between all concepts based on a predefined ontology hierarchy. Fan *et al.* [20] built a concept ontology using both semantic and visual similarity, in an effort to exploit the inter-concept correlations and to organize the image concepts hierarchically. In the process, they tried to effectively tackle the problem of intra-concept visual diversity by using multiple kernels. Moreover, in [19], they propose an algorithm called "product of mixture-experts" in order to model the contextual relationships between image concepts and several patterns of the relevant salient objects, with whom they co-appear.

Spatial relations among objects and regions appear also crucial in the concept detection process. In [25] an ontology of spatial relations is proposed that aims to facilitate image interpretations. In [87], an attempt to exploit spatial context constraints for automated image region annotation is conducted, utilizing topological relationships between regions in the same scene. Lipson *et al.* [37] presented earlier another spatial context modelling approach, called configuration-based scene modelling, suitable for content-based indexing and retrieval. Two further models of spatial context-aware object detection have been developed by Singhal *et al.* in [62] and by Boutell *et al.* in [10]. The former proposes a generic outdoor-scene model, whereas the latter one is specific to individual representative scene types (such as *beach*, *mountain* or *city* scenery). Moreover, in [86], multiple class object-based segmentation is achieved using the appearance and bag-of-keypoints models, integrated over mean-shift patches. The spatial relationships of keypoints are modeled and the Elliptical Fourier Descriptor is used to describe the global shapes.

An extension of Murphy's work presented in [38] was the additional use of temporal context, as given by the dependence between images captured within a short period of time [11]. In applications involving image collections, images are expected to be clustered sequentially, thus allowing surrounding images to be used as temporal context (e.g., family vacation photos). A general probabilistic temporal context model is also proposed in [9], where a first-order Markov property is utilized to integrate temporal context sequences. In principle, temporal context plays a significant role among different approaches varying from active object recognition [57] to temporally related events within natural language texts [47]. Another form of context, the so called imaging context, dealing with camera metadata tags about scene capture properties (e.g., EXIF [17] exposure time) is utilized in [7] to aid in a number of multimedia analysis problems, including indoor-outdoor classification and event detection. Finally, the approach proposed by Vailaya *et al.* [80] brings up the issue of utilizing context orientation information in object detection algorithms. However, this task is generally avoided due to the fact that such information is not always available and performance of classification algorithms is more than adequate despite this shortcoming.

According to the previous statements, visual context is strongly related to the core problem of image analysis, namely *high-level concept detection*. A significant number of misclassifications usually occur because of the similarities in low-level color and texture characteristics of various object types and also the lack of contextual information, which is a major limitation of individual object detectors. A lot of attention has been given among researchers on modeling relationships between objects and regions ([23], [32], [79]) or objects and other objects, e.g., in [33] or [76], where specifically an object detection and segmentation scheme is assisted by contextual information of other objects. Even a mechanism for estimating 3D scene geometry from a single image and use this information as additional improvement to object detection has been proposed by Hoiem *et al.* [24]. Towards the improvement of current object detectors, an interesting approach is the one presented in [39]. 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. Furthermore, in [61], Shotton *et al.* create a discriminative model of object classes by capturing simultaneous appearance, shape and context information. 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 still not considered globally sufficient, as it does not utilize the significant amount of available useful knowledge in the form of semantic or topological relations.

IV. Motivation and Overview

The idea behind the use of additional contextual information refers to the fact that not all events are relevant in all situations and this holds also when dealing with image analysis problems. Since there is not a 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. In the following, we shall refer 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*. 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 traditional content-based analysis, where the lack of contextual information significantly hinders its performance [50]. 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 modelling techniques, this work aims at a hybrid unification of them, in order to achieve optimized content analysis results and strengthen its high- and low-level correlation.

As the main drawback of previously discussed individual object detectors tends to be the fact that they only examine isolated

strips of pure object materials, taking into account the contextual information of the scene or individual objects themselves is of great value. However, this process is very important and also extremely challenging even for human observers. The herein proposed 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 (e.g., lighting, magnification); however, one important trait of humans is that they examine all the objects in the scene before making a final decision on the identity of individual objects. The use of visual context in the visual analysis process is the one that provides the necessary added value and forms the key for such a solid unambiguous detection process; thus it will be extensively presented and exploited in the following.

More specifically, our effort focuses on 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 towards 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 both semantic and topological relationships that are inherent within the visual part of the content. Our approach proves also that the use of such enhanced information improves the results obtained from traditional knowledge-assisted image analysis techniques, based on both *visual* and *contextual* information.

V. Image Representation and High-Level Concept Detection Using a Region Thesaurus

In the following subsections, we present our approach and implementation, in order to tackle the high-level concept detection problem from a different and at the same time innovative aspect, i.e., the one based on a region thesaurus and corresponding region types [69]. This research effort was expanded and further strengthened within [52], [53], [70], [72], and [73], by exploiting visual context in the process and by achieving promising research results. Our main focus remains to provide an ad-hoc "ontological" knowledge representation containing both high-level features (i.e., high-level concepts) and lower-level features and exploit it towards efficient multimedia analysis. Generally, the visual features one can extract from an image or video document may be divided in two major categories. The first one contains typical *low*-level visual features, which 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 those 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.

In this sense, we try to enhance the notion of a visual context knowledge model with *mid*-level concepts. These concepts may provide an in-between description, which can be described semantically, but does not express neither a high- nor a low-level concept. Thus, in this work we focus on a unified multimedia representation by combining low- 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, we present a visual example in Fig. 1. In this example, one could describe the visual content of the image either in a high-level manner (i.e., the image contains *sky*, *sea*, *sand* and *vegetation*) or in a lower level, but higher than a low-level description (i.e., an *"azure" region*, a *"blue" region*, a *"green" region* and a *"grey" region*). We shall call these features *region types* since in our belief each image can be intuitively and even efficiently described by a set of them. Thus, it is of crucial importance to define this set of region types in an effective manner, that can efficiently describe almost every image in the domain of interest. In the next subsections, we describe briefly the extraction of the low-level features and the region thesaurus construction. Fig. 2 presents the overall methodology.

A. Low-Level Feature Extraction

For the extraction of the color and texture features of a given image, an approach of extracting visual descriptors *locally*, i.e., from image regions, has been followed. Firstly, a color segmentation algorithm is applied. This algorithm is a multiresolution implementation of the well-known RSST method [2], tuned to produce a coarse segmentation. This way, one could intuitively provide a qualitative description of the image. The reason we choose such an under-segmentation is that it facilitates image description, without increasing the problem's complexity. After splitting the image in a small number of regions, low-level visual descriptors from the well-known MPEG-7 standard [13] are selected to capture a standardized description of their visual content. Since the high-level concepts we aim to detect fall to the categories of "materials" and "scenes", the only MPEG-7 descriptors that make sense to use are those that capture color and texture properties. For the extraction of these Descriptors we use the MPEG-7 eXperimentation Model (XM) [48]. More specifically, for the representation of color and texture features of the image regions, we select the following MPEG-7 descriptors: The *Dominant Color Descriptor*, the *Color Layout Descriptor*, the *Scalable Color Descriptor*, the *Color Structure Descriptor*, the *Homogeneous Texture Descriptor*, and the *Edge Histogram Descriptor* [41]. In order to obtain a single region description from all the extracted region descriptions, we follow an "early fusion" method and merge them after their extraction [71]. The vector formed for a given region r_i will be referred to as "feature vector" $f(r_i) \equiv f_i$.

B. Construction of a Region Thesaurus

After extracting the aforementioned MPEG-7 color and texture features we move to the next step that aims to bridge these low-level features to the high-level concepts. To achieve this,

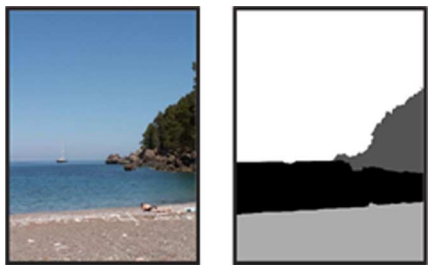


Fig. 1. Input image and its coarse segmentation.

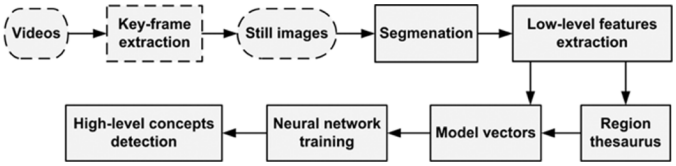


Fig. 2. High-level concept detection methodology.

we construct a *visual thesaurus* and with its aid we form an intermediate image description. This description will contain all necessary information to connect one image with every visual word of the dictionary. This way, we achieve to keep a fixed-size image description and tackle the problem that the number of segmented regions is not fixed. Moreover, this description will prove useful when contextual relations will be exploited. Given the entire training set of images and their extracted low-level features, one can easily observe that those regions that belong to similar semantic concepts, also have similar low-level descriptions. In addition, those images that contain the same high-level concepts are consisted of similar regions. This region similarity can be exploited as region co-existences often characterize the concepts that exist within an image [69].

The first step towards the construction of the proposed unified knowledge model, which will exploit the notion of visual context, is the selection of its region types. Based on the aforementioned observations, we start from an arbitrary large number of segmented regions and we apply a *hierarchical clustering* algorithm [54] on them. After this process, we should note that each cluster may or may not contain (and thus represent) a high-level feature and each high-level feature may be contained in one or more clusters. Concept *sand*, for instance, may have many instances differing e.g., in the color or the texture. Moreover, in a cluster that may contain instances from a semantic entity (e.g., *sea*), these instances could be mixed up with parts from another visually similar concept (e.g., *sky*). We select the region type that represents each cluster as the closest region to its centroid. A dendrogram illustrating the described hierarchical clustering and the selection of the region types is depicted in Fig. 3. In this simplistic example an initial set of 14 candidate region types is clustered. Then, six region types are selected to represent the enhanced features.

Finally, we formally describe the constructed visual dictionary (thesaurus) T as a set of visual words w_i with the aid of (1) and (2) as follows:

$$T = \{w_i, \quad i = 1 \dots N_T\}, \quad w_i \subset R \tag{1}$$

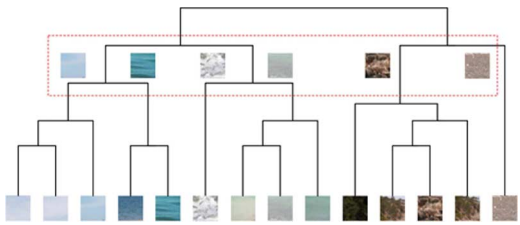


Fig. 3. Region-type selection using hierarchical clustering; selected region types are depicted within the red box.

$$\bigcup_{i=1,\dots,N} w_i = R \quad \text{and} \quad \bigcap_{i \neq j} w_i, w_j = \emptyset \tag{2}$$

where R is the set of all regions, N its cardinality, N_T is the number of region types of the thesaurus (and, obviously, the number of clusters) and w_i is the i-th cluster. Additionally, according to (2), the utilization of all clusters provides the entire R set, if and only if all regions are used for clustering and different clusters do not contain common regions (i.e., crisp clustering).[1]

We then choose to represent each region type by the feature vector that lies closer to the feature vector of each cluster. The calculation of each center is depicted in (3) as the centroid of all feature vectors belonging to the same cluster, where $|w_i|$ is the number of elements of cluster w_i. Thus, each feature vector selected to represent each region type is given by (4). The distance between two feature vectors is denoted by (5). These region types are the centroids of the clusters and all the other feature vectors of a cluster are their *synonyms*. By using a significantly large training set of images, the thesaurus is constructed. Its purpose is to formalize a conceptualization between the low and the high-level features and facilitate their association.

$$z(w_i) = \frac{1}{|w_i|} \sum_{r \in w_i} f(r) \tag{3}$$

$$f(w_i) = f\left(\arg\min_{r \in w_i} \{ d(f(r), z(w_i)) \} \right). \tag{4}$$

Each region type is represented by its feature vector that contains all the extracted low-level information for it. 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. 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, i.e., one may describe a region type as *a green region with a coarse texture*.

C. Model Vector Construction and High-Level Concept Detection

In the following, we present the algorithm used to describe each segmented image with the aid of the previously constructed region thesaurus. In particular, we apply the Euclidean distance on the vectors, in order to calculate the distance between two different regions, as far as their feature vectors are concerned.

[1]Obviously, in a further optimization stage, one could apply a fuzzy clustering approach, but this lies outside the scope of our current work.

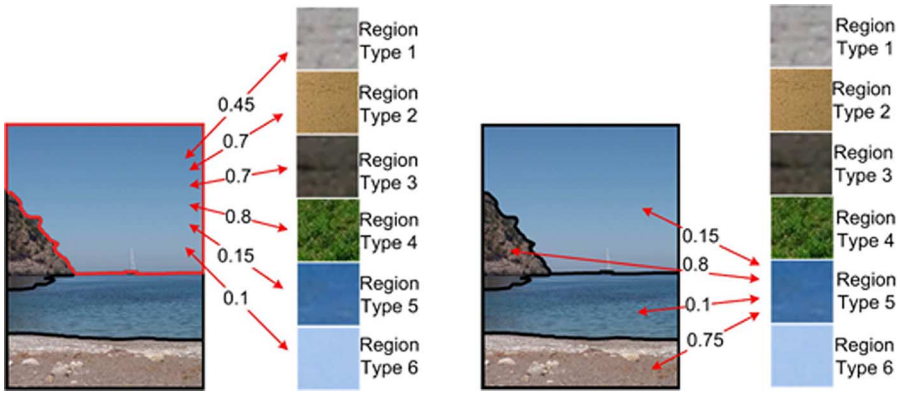


Fig. 4. Distances between regions and region types: on the left we present distances between an image region and all region types, whereas on the right, distances between all regions and a specific region type are depicted.

Let any two regions $r_1, r_2 \in \mathcal{R}, \quad R \subset \mathcal{R}$, with the feature vectors $f_1, f_2 \in \mathcal{F}, \quad F \subset \mathcal{F}$ and Euclidean distance $d(f_1, f_2)$, as the latter is denoted by (5). F is the set of feature vectors for the specific set of regions, whereas $\mathcal{F}$ is the entire feature vector space. The same applies for R and $\mathcal{R}$. At this point we must note that the MPEG-7 standard does not specify strict distance measures for the descriptors; it only suggests some, so as to allow for other measures to be used and test their efficiency. As depicted in experiments in our previous work [69], the use of the Euclidean distance provides a simple way to use all extracted low-level information, leading also to satisfactory results

$$d(f_1, f_2) = \sqrt{\sum_{i=1}^{n} (f_1^i - f_2^i)^2}. \tag{5}$$

Having calculated the distance of each region (cluster) of the image 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). After extracting model vectors from all images of the (annotated) training set, a neural network-based detector is trained separately for each high-level concept. The input of the detectors is a *model vector* m_i describing an image (or keyframe) in terms of the region thesaurus. The output of the network is the confidence that the image contains the specific concept. In particular, the model vector describing image p, will be:

$$m_p = [m_p(1), m_p(2), \ldots, m_p(j), \ldots, m_p(N_T)] \tag{6}$$

where

$$m_p(j) = \min_{r \in R(p)} \{d(f(w_j), f(r))\}, \quad j = 1 \ldots N_T \tag{7}$$

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

In order to better understand the above process, we present an indicative example in Fig. 4, where an image is segmented into regions and a region thesaurus is formed by 6 region types. On the left we present the distances of each region type from the *sky* region, whereas on the right we present the distances of each image region from *region type 5*. The model vector is constructed by the smallest distances for each region type. In this case and considering *region type 5*, the minimum distance is 0.1. The model vector for the specific image and the specific thesaurus is depicted in (8)

$$m = [m(1), m(2), m(3), m(4), m(5), m(6)] \tag{8}$$

where $m(5)$ will be 0.1. Taking into consideration all distances between all image regions (4) and all region types (6) (i.e., a total of $4 \times 6 = 24$ distances) we form the corresponding model vector.

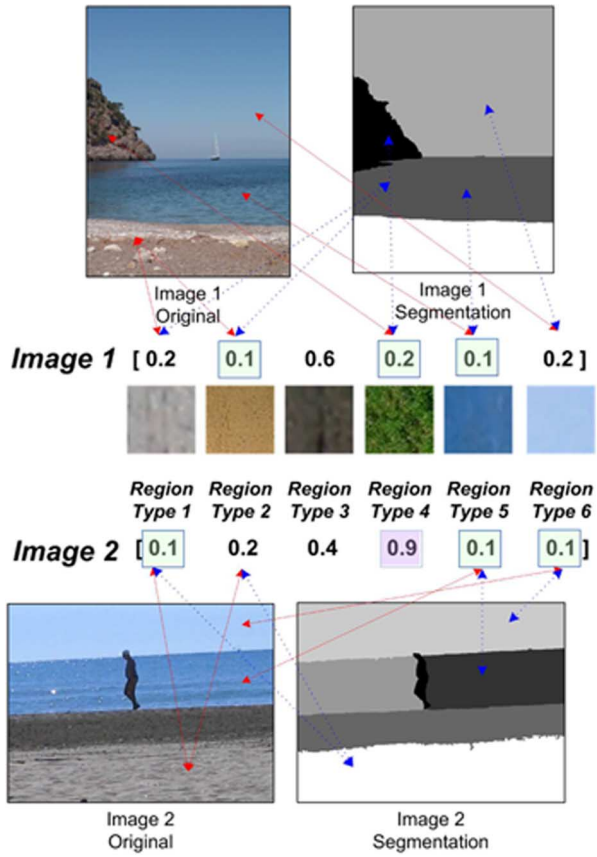


Fig. 5. Construction of model vectors for two similar images and a visual thesaurus of six region types; lowest values of model vectors are highlighted (light green) to note which region types of the thesaurus are most contained in each image, whereas a high value (light pink) indicates a high distance between the corresponding region type and the image.

Moreover, we present another indicative example in Fig. 5; this time, two semantically similar images are segmented into regions and their corresponding model vectors are formed using the same region thesaurus. In this example, one should observe that both images contain *sea*, *sky* and *sand*, while only the first one contains *vegetation*. Thus, the fourth element of their model vectors (i.e., the one corresponding to the fourth region type) is significantly different, while the rest have similar values.

VI. Visual Context

In order to fully exploit the notion of visual context and combine it with the aforementioned region types, we further introduce a twofold approach. Our methodology is divided into two subsections, according to the effect of visual context regarding concepts and region types. First, we present 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. We continue with a unified approach that utilizes contextual relations among both high-level concepts and region types.

A. Topological Context

First, we present an approach that utilizes high-level concept detectors for the entire image and not for a particular region of interest. The proposed methodology is part of the analysis process (and not e.g., a post-processing step) and does not utilize semantic relations for the knowledge construction, but only *topological* ones. It is rather easy to prove that the utilization of such information optimizes the results of traditional knowledge-assisted image analysis, based both on *visual*, as well as *contextual information*. Thus, the initial analysis results are enhanced through the utilization of semantic knowledge, in terms of region-independent *region types* and the relations between them.

1) Topological Context Model: In the following, we construct a contextual knowledge model, which encapsulates the above information type. We further introduce a method for improving the results of low-level based multimedia analysis by using the notion of region types. The latter form a knowledge model, described solely by the set of region types and the relations that characterize them. In general, we may decompose such a model $\mathcal{O}_T$ into two parts, the set T of all region types and the set R_{t_i,t_j} of all relations amongst any two given region types t_i, t_j. More formally

$$\mathcal{O}_T = \{T, R_{t_i,t_j}\}, \quad R_{t_i,t_j} : T \times T \to \{0,1\}, \quad i,j = 1 \ldots n. \tag{9}$$

We incorporate a rather classical subset of crisp topological relations between region types: {*adjacent*, *above*, *below*, *left*, *right*}. Their interpretation is rather simple and self-explanatory (see also Table I). We introduce a degree of confidence to each one of the above relations, in order to include the real-world vagueness in our approach. Before actually defining the contextual relations, we should begin by defining the sets of all entities that are encountered during their calculation process. More specifically and within this subsection let $\mathcal{T}$ be the set of all region types, $\mathcal{P}$ be the set of all images of the training set and $\mathcal{S}$ be the set of all regions of all images. For each image p we define the following.

- $\mathcal{T} = \{t_k\},\ k = 1, 2, \ldots, N_t$ be the set of all the region types of the given visual dictionary. As it has been described before, $\mathcal{T}$ may result by applying a clustering algorithm within all the elements of $\mathcal{S}$ and selecting the regions that lie closest to the cluster centroids.
- $\mathcal{S} = \{s_k\},\ k = 1, 2, \ldots, N_s$ be the set of all regions (segments), of all images. These regions may occur either after applying an image segmentation tool or splitting the image in orthogonal regions using a grid-based approach etc.
- $T_p = \{t_k^p\},\ k = 1, 2, \ldots, N_t^p,\ p \in \mathcal{P}$, be the set of all region types (clusters) present in image p. As obvious, $T_p \subset \mathcal{T}$.
- $D_p = \{d_k^p\},\ k = 1, 2, \ldots, N_d^p,\ p \in \mathcal{P}$, be the set of all initial detector values in image p. the initial detector values in an image are derived from the application of suitable high-level concept detectors.

TABLE I
Contextual Relations Between **Region Types**

Name	Inverse	Symbol	Meaning
Adjacent	-	$adj(a,b)$	a region type is adjacent to another region type
Co-occurrence	-	$co(a,b)$	a region type co-occurs with another region type
Left	Right	$left(a,b)$	a region type is left of another region type
Right	Left	$rgt(a,b)$	a region type is right of another region type
Above	Below	$ab(a,b)$	a region type is above of another region type
Below	Above	$bel(a,b)$	a region type is below of another region type

Letting $R_1(t_1, t_2)$ be a binary relation between any two given region types t_1 and t_2 and $R_2(t_1, t_2)$ be its opposite relation (e.g., "above" is the opposite relation of "below"), we finally define the *inverse* relation as: $\mathbf{R^{-1}}$: $R_1^{-1}(t_1, t_2) = R_1(t_2, t_1)$ and the *opposite* relation as: $\neg\mathbf{R}$: $\neg R_1(t_1, t_2) = R_2(t_1, t_2)$. The cardinality of a set is defined as $|\cdot|$.

In order to obtain a meaningful set of relations suitable to be used within the multimedia analysis value chain, when discussed from the global point of view, we extend the above relations and define a set of *fuzzy topological relations*. In this manner, let top denote any topological relation between any given region types t_1 and t_2

$$R_{tt}^{\text{top}} = \left\{r_{t_1,t_2}^{\text{top}}\right\} = \{top(t_1, t_2)\},\ t_1, t_2 \in \mathcal{T} \tag{10}$$

where $\text{top} \in \{adj, ab, \text{bel}, \text{left}, \text{rgt}\}$.

In order to define the *adjacency* relation, first we need to define the following sets:

$B_{t_1,t_2} = \{(s_1, s_2) \in \mathcal{S}^2 : s_1 \in S_{t_1}, s_2 \in S_{t_2}\}$ is the set of all pairs of regions that are assigned the first to region type t_1 and the latter to t_2.

$B_{t_1,t_2}^{co} = \{(s_1, s_2) \in B_{t_1,t_2} : co(s_1, s_2)\}$ is the subset of B_{t_1,t_2} that includes those pairs that co-exist within the same image.

$B^{\text{adj}}_{t_1,t_2} = \{(s_1, s_2) \in B_{t_1,t_2} : \text{adj}(s_1, s_2)\}$ is the subset of B_{t_1,t_2} that includes those pairs that are adjacent, within the same image.

Then, the *adjacency* relation is defined and calculated as: $R^{\text{adj}}_{tt} = \{r^{\text{adj}}_{t_1,t_2}\},\ t_1, t_2 \in \mathcal{T}$ is the *adjacency* relation between two region types t_1 and t_2. The degree to which this relation holds is calculated by (11) as

$$r^{\text{adj}}_{t_1,t_2} = \frac{|B^{\text{adj}}_{t_1,t_2}|}{|B^{co}_{t_1,t_2}|}. \tag{11}$$

The relations $r^{ab}_{t_1,t_2}$, $r^{\text{bel}}_{t_1,t_2}$, $r^{\text{left}}_{t_1,t_2}$ and $r^{rgt}_{t_1,t_2}$ are calculated based on the algorithm presented in [45], based on the assumption that the topological relations between two points may be determined by the angle made by the line passing through the two points and the x-axis.

Let $\{d_X\}$, $X \in \{\text{left}, \text{rgt}, ab, \text{bel}\}$ denote the calculated degree of confidence for each topological relation. Then, to calculate the appropriate relation for the given region types t_1 and t_2, we need to define the following subsets that include those pairs for which each topological relation holds

$$B^{\text{top}}_{t_1,t_2} = \{(s_1, s_2) \in B_{t_1,t_2} : \max\{d_X(t_1, t_2)\} = d_{\text{top}}(t_1, t_2)\}. \tag{12}$$

In other words we define the subsets of B_{t_1,t_2} containing the pairs, where t_1 is "above", "below", "left" or "right" in comparison to t_2. Now, the corresponding relations may be calculated by

$$r^{\text{top}}_{t_1,t_2} = \frac{|B^{\text{top}}_{t_1,t_2}|}{|B^{co}_{t_1,t_2}|},\ t_1, t_2 \in \mathcal{T}. \tag{13}$$

Finally, we introduce another important relation, i.e., the *co-occurrence* relation, which is defined statistically from the training dataset. We define

$$R^{co}_{tt} = \{r^{co}_{t_1,t_2}\} = \{co(t_1, t_2)\},\ t_1, t_2 \in \mathcal{T} \tag{14}$$

where

$$co(t_1, t_2) = \frac{|\{p \in \mathcal{P} : t_1 \in T_p \wedge t_2 \in T_p\}|}{|\{p \in \mathcal{P} : t_1 \in T_p \vee t_2 \in T_p\}|}. \tag{15}$$

As a result, we obtain a set of six fuzzy relations, all of them summarized in Table I.

As in [31], a fuzzy relation on T is a function $r_{t_i,t_j} : T \times T \to [0, 1]$ and its inverse relation is defined as $r_{t_i,t_j}{}^{-1} = r_{t_j,t_i}$. Based on the above relations, a domain-specific, "fuzzified" version of a region type knowledge model may be described by $\mathcal{O}^f_T$ as

$$\mathcal{O}^f_T = \{T, r_{t_i,t_j}\}, i, j = 1 \ldots n, \quad i \neq j, \quad r_{t_i,t_j} : T \times T \to [0, 1] \tag{16}$$

where T represents the set of all possible region types and $r \in \mathcal{R}$, denotes any possible non-fuzzy relation amongst two region types, where

$$R_{t_i,t_j} = \{R^{\text{top}}_{tt}, R^{co}_{tt}\} = \{adj, ab, \text{bel}, \text{rgt}, \text{left}, co\} \tag{17}$$

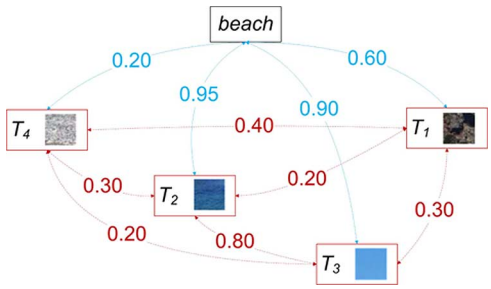


Fig. 6. Sample fragment of the *beach* region-type knowledge model.

is the set of all available relations. The final meaningful combination of relations

$$\mathcal{C}_{ij} = \left(\cup_{r \in \mathcal{R}} r_{t_i,t_j}{}^{p^r_{ij}}\right), \quad p^r_{ij} \in \{-1, 0, 1\}, \quad i, j = 1 \ldots n \tag{18}$$

leads to the final model

$$\mathcal{O}^{\mathcal{C}}_T = \{T, \mathcal{C}_{ij}\}, \quad i, j = 1 \ldots n, \quad i \neq j \tag{19}$$

which forms an RDF [88] graph (Fig. 6) and constitutes the abstract contextual knowledge model to be used during the analysis phase. The value of p^r_{ij} is determined by the semantics of each relation. To explain the need of p^r_{ij}, let the value of p for the *co-occurrence* relation be 0 for two region types t_1 and t_2 that never co-occur, i.e., $p^{co}_{12} = 0$. Nevertheless, in case, for example, t_1 is always *left* to t_2, $p^{rgt}_{12} = -1$, etc.

2) Topological Context Processing: Once the model vector for an image is calculated, a context-based confidence value readjustment algorithm is applied, so as to satisfy the needs of the problem at hand. The latter forms the last pre-processing step of the analysis process and provides an optimized re-estimation of the initial regions' degrees of confidence to the selected region types. Consequently, it updates the values of each model vector, allowing an optimized training process of the classifier, thus achieving significantly optimized evaluation results.

In a more formal manner, the problem that this work attempts to address is summarized in the following statement: the visual context analysis algorithm readjusts in a meaningful way the initial region type confidence values produced by the previous step of region type analysis. In this section, the remaining problems to be addressed include how to meaningfully readjust the initial membership degrees and how to use visual context to influence the overall results of knowledge-assisted image analysis towards higher performance. An estimation of the degree of membership of each region type is derived from direct and indirect relationships of the latter with other region types in the graph, using *max* as a meaningful compatibility indicator.

We may decompose the general structure of the proposed degree of membership re-evaluation algorithm into the following steps. It must be noted, that according to the characteristics of each considered domain, different domain similarity (or dissimilarity) measures are imposed, in terms of the normalization parameter $\mu \in [0, 1]$.

TABLE II
CONTEXTUAL RELATIONS BETWEEN BOTH TYPES OF **ENTITIES**

Name	Inverse	Symbol	Meaning	$C \times C$	$T \times T$	$C \times T$	Definition
Similar	Similar	$Sim(a,b)$	similarity between a and b		•		statistics
Accompanier	AccompanierOf	$Acc(a,b)$	coexistence of a and b	•		•	expert/statistics
Part	PartOf	$P(a,b)$	entity a is part of entity b	•	•	•	expert/statistics
Component	ComponentOf	$Comp(a,b)$	combines a with b	•	•	•	expert/statistics
Specialization	Generalization	$Sp(a,b)$	b specializes the meaning of a	•			expert
Property	PropertyOf	$Pr(a,b)$	b is a property of a		•	•	expert/statistics
Adjacent	-	$adj(a,b)$	a is adjacent to b		•		statistics
Co-occurrence	-	$co(a,b)$	a co-occurs with b		•		statistics
Left	Right	$left(a,b)$	a is left of b		•		statistics
Right	Left	$rgt(a,b)$	a is right of b		•		statistics
Above	Below	$ab(a,b)$	a is above b		•		statistics
Below	Above	$bel(a,b)$	a is below b		•		statistics

1) For each region type t describe the fuzzy set L_t using the widely applied [46] sum notation $L_t = \sum_{i=1}^{|T|} t_i/w_i = \{t_1/w_1, t_2/w_2, \ldots, t_n/w_n\}$, where w_i describes the membership function: $w_i = \mu_{L_t}(t_i)$.
2) For each region type t_i in the fuzzy set L_t with a degree of membership w_i, obtain the particular contextual information in the form of its relations to the set of any other region types: $\{r_{t_i,t_j} : t_i, t_j \in T, \quad i \neq j\}$.
3) In the case of multiple region type relations, relating region type t_i to more than the *root* concept, apply an intermediate aggregation step for the estimation of w_i, by considering the *context relevance* notion cr_{t_i} [50].
4) Calculate the new degree of membership w_i, taking into account each domain's similarity measure, based on the recursive formula

$$w_i^n = w_i^{n-1} - \mu \cdot (w_i^{n-1} - cr_{t_i}) \tag{20}$$

where n denotes the iteration used[2]. Equivalently, for an arbitrary iteration n

$$w_i^n = (1-\mu)^n \cdot w_i^0 + (1 - (1-\mu)^n) \cdot cr_{t_i} \tag{21}$$

where w_i^0 represents the initial degree of membership for region type t_i.

In this manner, we obtain the replacement of the initial model vector values with optimized ones, regarding their semantics and according to the guidelines of the contextual knowledge. The contribution of the proposed methodology is important, as it will be denoted within the Results Section VII.

B. Unified Context

We further advance the proposed conceptualization by introducing a hybrid knowledge representation approach in the form of an extended, unified context model [52]. We enhance the traditional notion of a contextual ontology [1] with *mid*-level entities and topological relations. These entities may provide an intermediate description, which may be semantically described, but they do not express a high-level nor a low-level entity. As a result, we focus on an integrated multimedia representation, combining low- and high-level information with an efficient manner and we combine it with the description of a typical context model by defining new (topological) and expanding previous (semantic) relations.

1) Unified Context Model: The proposed knowledge model is described by a set of high-level concepts, a set of region types and a set of relations among them. In general, this type of hybrid model $\mathcal{O}_M$ may be decomposed into three parts, the set C of all high-level concepts, the set T of all region types and the set R_{x_i,y_j} of all binary relations between any meaningful combination of concepts and region types[3]. More specifically there may exist one or several relations between two high-level concepts, two given region types or a high-level concept and a region type. More formally

$$\mathcal{O}_M = \{C, T, R_{x_i,y_j}\}, R_{x_i,y_j} : X \times Y \to \{0,1\}, \quad i,j = 1 \ldots m, \quad i \neq j \tag{22}$$

where $X, Y \subseteq C \cup T$. In other words, since the proposed model does not restrict the relations to be only amongst members of C or T, it is possible that $X \subseteq C$ and $Y \subseteq T$ or vice versa. Also, since for each applied semantic relation its inverse exists, $m = |X| = |Y|$, where $|\bullet|$ denotes the cardinality of a set.

Thus, we utilize a new set of relations (Table II) and redefine them by associating a degree of confidence to each one of them to incorporate fuzziness. This set of utilized relations contains both topological and MPEG-7 semantic relations, obtained by utilizing either a statistical approach on the training data set (used mainly for the definition of topological relations) or an expert's opinion (used mainly for the definition of the semantic relations).

However, it should be clear that not all relations are appropriate between any type of entity pairs. For example, the relation *Similar* does not make sense between two high-level concepts, or between a high-level concept and a region type, i.e. *sea* cannot be related to *sand* using this relation. However, similarity is a meaningful measure to relate two region types and may calculated by comparing their low-level features. The possible relations for each entities' pair are depicted on the right side of Table II.

All the above relations form a context model, that may be visualized as a graph (see Fig. 7); every node represents a concept

[2]According to our experimental results, typical values of n are: 3, 4, and 5.

[3]The term *entities* will be used in the following when referring to either *concepts* or *region types*.

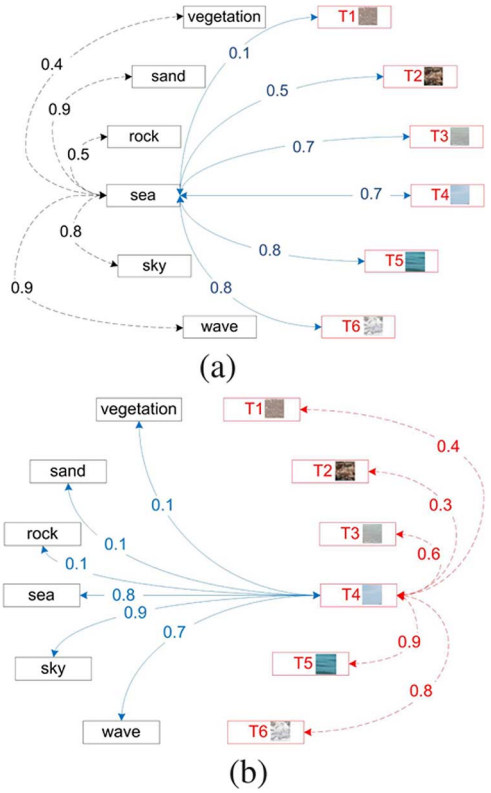


Fig. 7. Fragment of the unified context model. Numbers indicate the fuzzy degree of confidence for each relation. (a) Relations among *sea* and all other entities. (b) Relations among the fourth region type (T4) and all other entities.

or region type and each edge, between two nodes, a contextual relation between the respective entities. Additionally, a related degree of confidence is associated to each edge, expressing the desired fuzziness within the context model. An existing edge between a given pair of entities is produced based on the set of contextual fuzzy relations that are meaningful for the particular pair. For example, two region types, i.e., a "*green*" and a "*blue*" one, may utilize the relations *Similar*, *Accompanier*, and *Component*.

Based on the above relations, a domain-specific, "fuzzified" version of the proposed model may be described by $\mathcal{O}_M^f$

$$\mathcal{O}_M^f = \{C, T, r_{x_i,x_j}\}, \quad i,j = 1 \ldots m, \quad i \neq j, \quad r_{x_i,x_j} \in \mathcal{R} \tag{23}$$

where C represents again the set of all high-level concepts and T the set of all possible region types and

$$\mathcal{R} = \{Sim, Acc, P, Comp, Sp, Pr, adj, bel, ab, left, rgt, co\}. \tag{24}$$

A combination of relations

$$\mathcal{C}_{x_i,y_j} = \left(\cup_{i,j} r_{x_i,y_j}^{p_{r,ij}}\right), \quad p_{r,ij} \in \{-1, 0, 1\} \quad i,j = 1 \ldots m, i \neq j \tag{25}$$

leads to the final knowledge model

$$\mathcal{O}_M^{\mathcal{C}} = \{C, T, \mathcal{C}_{x_i,y_j}\}, \quad i,j = 1 \ldots m, \quad i \neq j, \quad r_{x_i,x_j} \in \mathcal{R}. \tag{26}$$

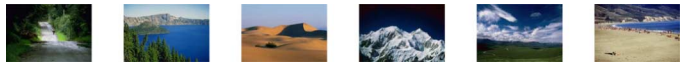


Fig. 8. Indicative Corel images.

Fig. 9. Indicative TRECVID images.

In Fig. 7(a) and (b), we present the visualization of two fragments of the aforementioned RDF graph. We should note here that we represent only two small fragments of the entire visual context knowledge model, for the sake of the presentation and the explanatory examples. That is because a model with 1 domain, 6 concepts, and 25 region types (i.e., 32 entities) would require a maximum of 496 relations. Even though not all relations are applicable, as already explained, such a complicated graph is difficult to be presented in a figure. The degree of confidence of each edge represents fuzziness in the model, implemented herein according to the RDF reification technique [89].

2) Unified Context Processing: Based on the principles and mathematical foundations of fuzzy algebra [46] and the described knowledge conceptualization, we further present the *ad hoc* unified context processing steps and algorithm. The core functionality of the above methodology is the meaningful readjustment of the membership degrees of each entity associated to a region or segment of an image, obtained from any kind of image analysis module. The novelty introduced herein deals with the context value, which is utilized in order to tackle cases where either the dominant concept or the region type is difficult to be identified. The problem that this step attempts to address is summarized in the following statement: it readjusts in a meaningful manner the initial concept and/or region type confidence values produced by an initial step of low-level multimedia analysis.

The composition of the proposed degree of membership re-evaluation algorithm is formed according to a meaningful adaptation and combination of the algorithmic principles of the visual case. Enhanced characteristics to the algorithm's structure are complementary imposed, stressing mainly on the exploitation of fuzzy algebra features for both high-level concepts and region types. The decomposition of the proposed degree of membership re-evaluation algorithm into steps for the unified case is as follows, using the standard t-conorm and the algebraic product as the t-norm.

1) Identify a domain similarity (or dissimilarity) measure, imposed by the nature of the considered domain: $\mu \in [0, 1]$.
2) For each concept, $c \in C$ describe the fuzzy set L_c using the sum notation $L_c = \sum_{i=1}^{|C|} c_i/w_i = \{c_1/w_1, c_2/w_2, \ldots, c_n/w_n\}$, w_i is the membership function: $w_i = \mu_{L_c}(c_i)$.
3) For each region type, $t \in T$ describe the fuzzy set L_t as $L_t = \sum_{i=1}^{|T|} t_i/z_i = \{t_1/z_1, t_2/z_2, \ldots, t_n/z_n\}$, z_i is the membership function: $z_i = \mu_{L_t}(t_i)$.

TABLE III
COMPARATIVE **PRECISION** P, **RECALL** R, AND F-**MEASURE** SCORES PER CONCEPT FOR SIX DIFFERENT CONCEPT DETECTION METHODOLOGIES APPLIED ON THE COREL DATASET

Concepts	RT			RT+LSA			RT+Top.Context			RT+Uni.Context			LIPs [27]			Rel. LSA [68]		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
rock	0.22	0.40	0.28	0.27	0.38	0.32	0.41	0.34	0.37	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	0.42	0.48	0.45	0.52	0.47	0.49	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	0.75	0.84	0.79	0.85	0.79	0.82	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	0.83	0.86	0.84	0.87	0.82	0.84	0.88	0.82	0.85	0.86	0.85	0.85	0.88	0.83	0.85
wave	0.48	0.68	0.56	0.53	0.64	0.58	0.65	0.58	0.61	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	0.70	0.78	0.74	0.72	0.78	0.75	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**	**0.58**	**0.67**	**0.62**	**0.67**	**0.63**	**0.65**	**0.71**	**0.62**	**0.66**	**0.63**	**0.65**	**0.64**	**0.67**	**0.63**	**0.65**

TABLE IV
COMPARATIVE **PRECISION** P, **RECALL** R, AND F-**MEASURE** SCORES PER CONCEPT FOR SIX DIFFERENT CONCEPT DETECTION METHODOLOGIES APPLIED ON THE TRECVID DATASET

Concepts	RT			RT+LSA			RT+Top.Context			RT+Uni.Context			LIPs [27]			Rel. LSA [68]		
	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	0.52	0.62	0.57	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	0.28	0.31	0.29	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	0.93	0.80	0.86	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	0.63	0.68	0.65	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	0.80	0.70	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	0.49	0.48	0.44	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	0.35	0.47	0.41	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**	**0.55**	**0.60**	**0.57**	**0.63**	**0.50**	**0.56**	**0.72**	**0.47**	**0.57**	**0.54**	**0.57**	**0.55**	**0.58**	**0.54**	**0.55**

4) For each concept c_i in the fuzzy set L_c with a degree of membership w_i, obtain the particular contextual information in the form of its relations to the set of any other entities: $\{R_{c_i,x_j} : c_i \in C, \quad x_j \in C \cup T, \quad i \neq j\}$.
5) For each region type t_i in the fuzzy set L_t with a degree of membership z_i, obtain the particular contextual information in the form of its relations to the set of any other entities: $\{R_{t_i,x_j} : t_i \in T, \quad x_j \in C \cup T, \quad i \neq j\}$.
6) Calculate the new degrees of membership w_i and z_i, taking into account each domain's similarity measure. Again, in the case of multiple relations, relating concept c_i or region type t_i to more than the *root* concept, an intermediate aggregation step should be applied for the estimation of w_i and z_i, by considering the appropriate *context relevance* notion cr_{c_i} or cr_{t_i}, respectively.
7) Express the calculation of w_i and z_i for both cases with the recursive formula

$$w_i^n = w_i^{n-1} - \mu \cdot (w_i^{n-1} - cr_{x_i}) \tag{27}$$

where n denotes the iteration used and x_i stands for either a concept c_i or region type t_i. Equivalently, for an arbitrary iteration n

$$w_i^n = (1-\mu)^n \cdot w_i^0 + (1-(1-\mu)^n) \cdot cr_{x_i} \tag{28}$$

where w_i^0 represents the initial degree of membership for entity $x_i \in C \cup T$.

VII. EXPERIMENTAL RESULTS

In the following we present an indicative selection of our experimental results. We include results from the application of the proposed visual context utilization methodology from the scope of the solely visual approach, as well as the unified scope. 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, we evaluate both approaches by utilizing parts of the well-known Corel and TRECVID datasets and compare their efficiency to related state-of-the-art techniques.

Initially, we present a set of experimental results from the application of our visual context approaches on a dataset containing 750 images, 40 region types, and 6 high-level concepts $\{$*sea, vegetation, sky, sand, rock, wave*$\}$. The number of the region types was selected based on the size of the region thesaurus and was verified using MDL [30]. The amount and type of concepts utilized is imposed by the problem/dataset at hand; the dataset utilized was a subset of the well-known Corel image collection [82], an indicative sample of which is presented in Fig. 8. We used a manually constructed ground truth and selected $\mu = 0.12$ as the best normalization parameter for the considered domain of interest. For the non-contextual detection of high-level concepts, we applied the already described methodology of Section V. We utilized 525 images to train 6 individual SVM concept classifiers and 225 images as the test set.

We also present some additional experiments from the application of the proposed unified contextualization approach on a second dataset, consisting of 4000 images from the TRECVID collection [63], 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 [30]. A characteristic sample of this dataset is presented in Fig. 9. We utilized 250 of those images to train seven individual SVM classifiers and the rest 997 images as our test set.

To evaluate the proposed approaches we compare them to similar techniques we have used in previous research work. The results of all approaches on both Corel and TRECVID datasets are summarized in Tables III and IV. Initially, by "*Region Types (RT)*" we present the results based only on the detection scheme presented in Section V, without exploiting any contextual knowledge. We further compare this approach with an extension that uses the Latent Semantic Analysis

technique [73]. This technique tries to take advantage of the latent (hidden) relations among the region types within the images of the training set, and modifies appropriately the formulated model vector. Results from this technique are denoted by "RT + LSA", whereas results from the application of both contextual approaches (i.e., "RT + Top. Context" and "RT + Uni. Context") follow.

To further evaluate our last approach, we also implemented two other techniques. The first one, denoted as "*Rel. LSA*" [68], adds directly structural constraints to the visual words of the thesaurus. The fundamental difference between the traditional LSA we have applied and this one is that every possible unordered pair of clusters is this time considered as a visual word. This way, a visual thesaurus with too many words (i.e., pairs of clusters) is created. Nevertheless, the low-level features extracted from each region are simpler than the MPEG-7 low-level features that we use. More specifically, a 64-bin color HSV histogram 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 approach [27] we have implemented starts with the extraction of *Local Interest Points* (LIPs) and is denoted as "*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-Gaussian (DoG)* 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 must be noted, that we chose to compare our proposed method to the above set of 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 latter 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 III and IV, the proposed contextual unified approach, denoted as "RT + Context", outperforms in principle all compared approaches, in terms of the achieved precision, whereas in some cases lacks, in terms of the recall.

Based on the above results it is rather obvious 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 hybrid combination of topological and semantic relations. The interpretation of all above experimental results depicts that the proposed contextualization approach will favor the rather confident degrees of confidence for the detection of a concept that exists within an image. Quite on the contrary, it will also discourage the 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.

VIII. Conclusion

Our research effort focus on an integrated approach offering a unified and unsupervised management of multimedia content. It is proved that the use of enhanced intermediate information is able to improve the results of traditional, knowledge-assisted image analysis, based on both *visual* and *contextual* information. It indicates clearly 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, as indicated from the evaluation of our approach in comparison to related state-of-the-art techniques on the well-known Corel and TRECVID datasets. Amongst the core contribution of this work has been the implementation of a novel, twofold visual context interpretation utilizing a fuzzy representation of knowledge. Experimental research results were presented, indicating a significant high-level concept detection optimization (i.e., precision improvement per concept varies from 12.04% to 95.45%) over the entire datasets utilized. Although the total improvement is not considered to be impressive, we believe that our approach successfully incorporates the underlying contextual knowledge and further exploits visual context in the multimedia analysis value chain. We also think that minor enhancements on the implemented contextual model, e.g., in terms of additional spatial, temporal or semantic relationships exploitation, would further boost its performance.

References

[1] Th. Athanasiadis, Ph. Mylonas, Y. Avrithis, and S. Kollias, "Semantic image segmentation and object labeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 298–312, Mar. 2007.

[2] Y. Avrithis, A. Doulamis, N. Doulamis, and S. Kollias, "A stochastic framework for optimal key frame extraction from MPEG video databases," *Comput. Vis. and Image Understand.*, vol. 75, no. 1/2, pp. 3–24, 1999.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[4] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cog. Psychol.*, vol. 14, no. 2, pp. 143–177, 1982.

[5] A.B. Benitez and S.-F. Chang, "Image classification using multimedia knowledge networks," in *Proc. IEEE Int. Conf. Image Processing (ICIP'03)*, Barcelona, Spain, 2003.

[6] N. Boujemaa, F. Fleuret, V. Gouet, and H. Sahbi, "Visual content extraction for automatic semantic annotation of video news," in *IS&T/SPIE Conf. Storage & Retrieval Methods & Applications for Multimedia*, 2004.

[7] M. Boutell and J. Luo, "Bayesian fusion of camera metadata cues in semantic scene classification," in *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR)*, Washington, DC, 2004, vol. 2, pp. 623–630.

[8] M. Boutell, J. Luo, X. Shena, and C. Brown, "Learning multi-label scene classification," *Patt. Recog.*, vol. 37, no. 9, pp. 1757–1771, Sept. 2004.

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

[10] M. Boutell, J. Luo, and C. Brown, "Improved semantic region labeling based on scene context," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Amsterdam, The Netherlands, 2005, pp. 980–993.

[11] M. Boutell, J. Luo, and C. Brown, "Learning spatial configuration models using modified Dirichlet priors," in *Proc. IEEE Workshop Statistical Relational Learning (in conjunction with ICML)*, Banff, AB, Canada, 2004.

[12] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *Proc. ECCV*, 2004, pp. 350–362.

[13] S.-F. Chang, T. Sikora, and A. Puri, "Overview of the MPEG-7 standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 688–695, Jun. 2001.

[14] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual categorization with bags of keypoints," in *Proc. ECCV Int. Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, 2004.

[15] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. ECCV*, 2002, pp. 97–112.

[16] B. Edmonds, "The pragmatic roots of context," in *Proc. 2nd Int. Interdisciplinary Conf. Modeling and Using Context*, 1999, vol. 1688, pp. 119–132, Springer Verlag LNAI.

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

[18] J. Fan, Y. Gao, and H. Luo, "Multi-level annotation of natural scenes using dominant image components and semantic concepts," in *Proc. ACM Multimedia*, 2004, pp. 540–547.

[19] J. Fan, Y. Gao, H. Luo, and R. Jain, "Mining multilevel image semantics via hierarchical classification," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 167–187, Feb. 2008.

[20] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation," *IEEE Trans. on Image Processing*, vol. 17, no. 3, pp. 407–426, Mar. 2008.

[21] A. Ghoshal, P. Ircing, and S. Khudanpur, "Hidden Markov models for automatic annotation and content-based retrieval of images and video," in *Proc. SIGIR 2003*, 2003, pp. 544–551.

[22] D. Gokalp and S. Aksoy, "Scene classification using bag-of-regions representations," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07)*, 2007.

[23] X. He, R. S. Zemel, and M. A. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2004.

[24] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *Proc. ICCV*, 2005.

[25] C. Hudelot, J. Atif, and I. Bloch, "Fuzzy spatial relation ontology for image interpretation," *Fuzzy Sets and Syst.*, vol. 159, no. 15, pp. 1929–1951, Aug. 2008.

[26] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. SIGIR 2003*, 2003, pp. 119–126.

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

[28] R. Jin, J. Y. Chai, and L. Si, "Effective automatic image annotation via a coherent language model and active learning," in *Proc. ACM Multimedia 2004*, 2004, pp. 892–899.

[29] J. Jiten, B. M'erialdo, and B. Huet, "Semantic feature extraction with multidimensional hidden Markov model," in *Proc. SPIE CMCAMR 2006*, 2006, vol. 6073, pp. 211–221.

[30] S. Kim and I. S. Kweon, "Simultaneous classification and visual word selection using entropy-based minimum description length," in *Proc. 18th Int. Conf. Pattern Recognition (ICPR)*, 2006.

[31] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[32] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *Proc. ICCV*, 2005.

[33] S. Kumar and M. Hebert, "Discriminative random fields: A discriminative framework for contextual interaction in classification," in *Proc. of ICCV*, 2003.

[34] S. Lazebnik, C. Schmid, and J. Ponce, "A discriminative framework for texture and object recognition using local image features," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2005.

[35] D. Lewis, , S. Kanger and S. Ohman, Eds., "Index, context, and content," in *Philosophy and Grammar*. Dordrecht, Holland: Reidel, 1980.

[36] J. Li, A. Najmi, and R. M. Gray, "Image classification by a two-dimensional hidden Markov model," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 517–533, Feb. 2000.

[37] P. Lipson, E. Grimson, and P. Sinha, "Configuration based scene classification and image indexing," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997.

[38] J. Luo and A. Savakis, "Indoor vs outdoor classification of consumer photographs using low-level and semantic features," in *Proc. IEEE Int. Conf. Image Processing (ICIP'01)*, 2001.

[39] J. Luo, A. Singhal, and W. Zhu, "Natural object detection in outdoor scenes based on probabilistic spatial context models," in *Proc. IEEE Int. Conf. Multimedia and Expo.*, 2002.

[40] T. Malisiewicz and A. Efros, "Recognition by association via learning per-exemplar distances," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[41] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.

[42] *Marvel, IBM Multimedia analysis and retrieval system*, [Online]. Available: http://mp7.watson.ibm.com/

[43] A. Mathes, "Folksonomies—Cooperative classification and communication through shared metadata," in *Computer Mediated Communication—LIS590CMC*. Urbana, IL: Grad. School of Library Info. Science, Univ. Illinois Urbana-Champaign, 2004.

[44] J. McCarthy, "Notes on formalizing context," in *Proc. of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993)*, Chambéry, France, Aug.-Sep. 1993, pp. 81–98.

[45] K. Miyajima and A. Ralescu, "Spatial organization in 2D images," in *Proc. 3rd IEEE Conf. Fuzzy Systems*, 1994.

[46] S. Miyamoto, *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Dordrecht/Boston/London: Kluwer, 1990.

[47] D. Moldovan, C. Clark, and S. Harabagiu, "Temporal context representation and reasoning," Proc. 19th Int. Joint Conf. Artificial Intelligence (IJCAI). Edinburgh, Scotland, U.K., 2005.

[48] *MPEG-7: Visual Experimentation Model (xm)*, , 2001, version 10.0. ISO/IEC/ JTC1/SC29/WG11, Doc. N4062.

[49] P. Murphy, A. Torralba, and W. Freeman, "Using the forest to see the trees: a graphical model relating features, objects and scenes," in *Adv. Neur. Inform. Process. Syst. 16 (NIPS)*. Cambridge, MA: MIT Press, 2003.

[50] Ph. Mylonas, Th. Athanasiadis, and Y. Avrithis, "Improving image analysis using a contextual approach," in *Proc. 7th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Seoul, Korea, 2006.

[51] Ph. Mylonas and Y. Avrithis, "Context modelling for multimedia analysis," in *Proc. 5th Int. Interdisciplinary Conf. Modeling and Using Context (CONTEXT'05)*, Paris, France, Jul. 2005.

[52] Ph. Mylonas, E. Spyrou, and Y. Avrithis, "Enriching a context ontology with mid-level features for semantic multimedia analysis," in *Proc. 1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies*, 2007, co-located with SAMT.

[53] Ph. Mylonas, E. Spyrou, and Y. Avrithis, "High-level concept detection based on mid-level semantic information and contextual adaptation," in *Proc. 2nd Int. Workshop on Semantic Media Adaptation and Personalization (SMAP 2007)*, London, U.K., Dec. 17–18, 2007.

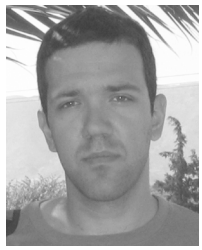
[54] Ph. Mylonas, M. Wallace, and S. Kollias, "Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering," in *Proc. 3rd Hellenic Conf. Artificial Intelligence*, Samos, Greece, May 2004.

[55] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progr. Brain Res.: Vis. Percept.*, no. 155, pp. 23–36, 2005.

[56] A. Opelt, A. Pinz, and A. Zisserman, "Incremental learning of object detectors using a visual shape alphabet," in *Proc. CVPR 2006*, Washington, DC, 2006.

[57] L. Paletta, M. Prantl, and A. Pinz, "Learning temporal context in active object recognition using bayesian analysis," in *Proc. 15th Int. Conf. Pattern Recognition (ICPR'00)*, 2000, vol. 1, p. 1695.

[58] C. Pantofaru and M. Hebert, "A framework for learning to recognize and segment object classes using weakly supervised training data," in *Proc. British Machine Vision Conf.*, Sep. 2007.
[59] K. Rapantzikos, Y. Avrithis, and S. Kollias, "On the use of spatiotemporal visual attention for video classification," in *Proc. Int. Workshop on Very Low Bitrate Video Coding (VLBV '05)*, Sardinia, Italy, Sep. 2005.
[60] B. Saux and G. Amato, "Image classifiers for scene analysis," in *Proc. Int. Conf. on Computer Vision and Graphics*, 2004.
[61] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," in *Proc. European Conf. on Computer Vision*, 2006.
[62] A. Singhal, J. Luo, and W. Zhu, "Probabilistic spatial context models for scene content understanding," in *Proc. CVPR 2003*, Madison, WI, 2003, pp. 235–241.
[63] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. 8th ACM Int. Workshop on Multimedia Information Retrieval*, Santa Barbara, CA, Oct. 26–27, 2006.
[64] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1349–1380, 2000.
[65] C. Snoek, M. Worring, D. Koelma, and A. Smeulders, "A learned lexicon-driven paradigm for interactive video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 280–292, 2007.
[66] C. Snoek, M. Worring, D. Koelma, and A. Smeulders, Learned Lexicon-Driven Interactive Video Retrieval 2006.
[67] F. Souvannavong, B. Mérialdo, and B. Huet, "Region-based video content indexing and retrieval," in *CBMI 2005 4th Int. Workshop on Content-Based Multimedia Indexing*, Riga, Latvia, 2005.
[68] F. Souvannavong, L. Hohl, B. Mérialdo, and B. Huet, "Structurally enhanced latent semantic analysis for video object retrieval," *IEE Proc. Vision, Image, and Signal Processing*, vol. 152, no. 6, pp. 859–867, 2005.
[69] E. Spyrou and Y. Avrithis, "A region thesaurus approach for high-level concept detection in the natural disaster domain," in *Proc. 2nd Int. Conf. Semantics And digital Media Technologies (SAMT)*, 2007.
[70] E. Spyrou *et al.*, "The COST292 experimental framework for TRECVID 2007," in *Proc. 5th TRECVID Workshop*, Gaithersburg, MD, Nov. 2007.
[71] E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor, "Fusing MPEG-7 visual descriptors for image classification," in *Proc. Int. Conf. Artificial Neural Networks ICANN 05*, 2005.
[72] E. Spyrou, Ph. Mylonas, and Y. Avrithis, "Semantic multimedia analysis based on region types and visual context," in *Proc. 4th IFIP Conf. Artificial Intelligence Applications and Innovations (AIAI)*, Athens, Greece, Sep. 19–21, 2007.
[73] E. Spyrou, G. Tolias, Ph. Mylonas, and Y. Avrithis, "A semantic multimedia analysis approach utilizing a region thesaurus and LSA," in *Proc. 9th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2008.
[74] S. Staab and R. Studer, "Handbook on ontologies," in *Int. Handbooks on Information Systems*. Heidelberg, Germany: Springer, 2004.
[75] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, pp. 169–191, 2003.
[76] A. Torralba, K. P. Murphy, and W. T. Freeman, *Contextual Models for Object Detection Using Boosted Random Fields*. Vancouver, BC, Canada: Neural Information Processing Systems (NIPS), Dec. 2004.
[77] A. Torralba, "Contextual influences on saliency," in *Neurobiology of Attention*. London, U.K.: Academic, 2005.
[78] G. Tsechpenakis, G. Akrivas, G. Andreou, G. Stamou, and S. Kollias, "Knowledge-assisted video analysis and object detection," in *Proc. European Symp. Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems (Eunite02)*, Albufeira, Portugal, Sep. 2002.
[79] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 113–140, 2005.
[80] A. Vailaya and A. Jain, "Detecting sky and vegetation in outdoor images," in *Proc. SPIE*, Jan. 2000, vol. 3972.
[81] N. Voisine, S. Dasiopoulou, V. Mezaris, E. Spyrou, Th. Athanasiadis, I. Kompatsiaris, Y. Avrithis, and M. G. Strintzis, "Knowledge-assisted video analysis using a genetic algorithm," in *Proc. 6th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)*, 2005.
[82] J.Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semanticssensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. . Mach. Intell.*, vol. 23, no. 9, pp. 947–963, 2001.
[83] L. Wang, J. Shi, G. Song, and I.-F. Shen, "Object detection combining recognition and segmentation," in *Proc. 8th Asian Conf. Computer Vision, (ACCV)*, 2007.
[84] Y. Wu, B. L. Tseng, and J. R. Smith, "Ontology-based multi-classification learning for video concept detection," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2004.
[85] C. Yang, M. Dong, and F. Fotouhi, "Region based image annotation through multiple-instance learning," in *Proc. ACM Multimedia 2005*, 2005, pp. 435–438.
[86] L. Yang, P. Meer, and D. J. Foran, "Multiple class segmentation using a unified framework over mean-shift patches," in *Proc. 2007 Computer Vision and Pattern Recognition Conf. (CVPR)*, Minneapolis, MN, 2007.
[87] J. Yuan, J. Li, and B. Zhang, "Exploiting spatial context constraints for automatic image region annotation," in *Proc. 15th Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 595–604.
[88] RDF [Online]. Available: http://www.w3.org/RDF/ W3C
[89] RDF Reification [Online]. Available: http://www.w3.org/TR/rdf-schema/#ch_reificationvocab W3C



Evaggelos Spyrou was born in Athens, Greece, in 1979. He received the diploma in electrical and computer engineering from the National Technical University of Athens (NTUA) in 2003. He is currently pursuing the Ph.D. degree at the Image, Video and Multimedia Laboratory of NTUA.

His research interests include semantic analysis of multimedia, high-level concept detection, visual context, and neural networks. He has published two book chapters and 20 papers in international conferences and workshops.

Phivos Mylonas (S'99–M'09) received the diploma in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 2001, the M.S. degree in advanced information systems from the National and Kapodestrian University of Athens (UoA) in 2003, and the Ph.D. degree from NTUA in 2008.

He is currently a Researcher at the Image, Video and Multimedia Laboratory of NTUA. His research interests lie in the areas of content-based information retrieval, visual context representation and analysis, knowledge-assisted multimedia analysis, issues related to multimedia personalization, user adaptation, user modelling, and profiling. He has published 19 articles in international journals and book chapters and has published three books.

Dr. Mylonas is a member of ACM. He is a Guest Editor for two international journals, a reviewer for seven international journals, an author of 34 papers in international conferences and workshops, and has been involved in the organization of 13 international conferences and workshops.

Yannis Avrithis (S'96–A'01–M'03) received the diploma in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1993, the M.Sc. degree (with distinction) in communications and signal processing from the Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, University of London, London, U.K., in 1994, and the Ph.D. degree in digital image processing from NTUA in 2001.

He is currently a Senior Researcher at the Image, Video, and Multimedia Systems Laboratory (IVML) of NTUA, coordinating the R&D activities for Greek and European Union projects, and lecturing at NTUA. His research interests include image/video segmentation and interpretation, knowledge-assisted multimedia analysis, annotation, content-based and semantic indexing and retrieval, video summarization, and personalization. He has published one book and 15 book chapters.

Dr. Avrithis is a member of ACM and EURASIP. He has published 13 articles in international journals and 62 in conferences and workshops.

Stefanos Kollias (S'81–M'85) received the diploma in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1979, the M.Sc. degree in communication engineering in 1980 from UMIST, Manchester, U.K., and the Ph.D. degree in signal processing from the Computer Science Division, NTUA.

He has been with the Electrical Engineering Department, NTUA, since 1986, where he currently serves as a Professor. Since 1990, he has been the Director of the Image, Video, and Multimedia Systems Laboratory, NTUA. He has published more than 120 papers, 50 of which are in international journals. Ten graduate students have completed their doctorate under his supervision, while another ten are currently performing their Ph.D. thesis. He and his team have been participating in 38 European and national projects.

Dr. Kollias has been a member of the Technical or Advisory Committee or invited speaker in 40 international conferences. He is a reviewer of ten IEEE Transactions and of ten other journals.