A Visual Context Ontology for Multimedia High-Level Concept Detection

Evaggelos Spyrou, Phivos Mylonas, and Yannis Avrithis

Image, Video and Multimedia Systems Laboratory,
National Technical University of Athens
9 Iroon Polytechniou Str., 157 80 Athens, Greece,
espyrou@image.ece.ntua.gr

Abstract. The notion of context plays a significant role in multimedia content search and retrieval systems. In this paper we focus our research efforts on a visual context knowledge representation, to be utilized for multimedia high-level concept detection. We propose and describe in detail types of contextual relations evident within the multimedia content, model them and provide a clear methodology on how to extract them. A visual context ontology is introduced, containing relations among different types of content entities, such as images, regions, region types and high-level concepts. In this manner, we facilitate traditional object detection approaches towards semantical interpretation. The application of the proposed knowledge structure provides encouraging initial results, improving the efficacy of related multimedia analysis techniques.

1 Introduction

Over the last decade, context is increasingly playing an important role in the multimedia analysis value chain. Different types of context are identified and exploited and their interrelations have been illustrated [3]. This applies also in the field of multimedia information retrieval and image and video analysis. The notion of visual context is introduced in [12] and [7], as an extra source of information for both object detection and scene classification. The truth is that the idea behind the use of such additional information refers to the fact that not all events are relevant in all situations and this holds also when dealing with image analysis problems. However, the context's usage and exploitation are the ones that define how satisfactory its modeling is.

In the research field of multimedia analysis, the problem of high-level object detection is still attracting a lot of attention. Due to the dynamic nature of multimedia content, efficient modeling and utilization of contextual information is considered to be among the best scientific tools towards tackling that scope. 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, these approaches are suffering from the well-known "semantic gap" effect, characterizing 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 multimedia 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. The idea of combining formalized knowledge and a set of features to describe the visual content of an image or its regions has been presented in [13], where a region-based approach using MPEG-7 visual features and ontological knowledge is presented. In [15], an attempt to exploit spatial context constraints for automated image region annotation is conducted. Finally, in [2] visual categorization is achieved using a bag-of-keypoints approach.

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, typically present in a beach, mountain or city scenery, could be a considerable source of useful information [12]. 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, which underlies as the major limitation of individual object detectors. Generic algorithms for automatic object recognition (e.g. [1]) and scene classification (e.g. [11]) are unfortunately not producing reliable results.

Consequently, it seems rather obvious that visual context is a difficult notion to grasp and capture. Thus, we restrict it herein to the notion of ontological context, defined as part of a "fuzzified" contextual ontology. Such an ontology can be viewed as a framework for knowledge representation in which the context determines the intended meaning of each entity; an entity used in different context may have different meanings[1]. A domain-independent, semantic knowledge in terms of content entities, such as images, regions, region types and high-level concepts, as well as fuzzy relations between them is introduced in the following sections. Describing the degree of each relation is carried out using the RDF reification technique [16], i.e. by making an additional statement about each statement, which contains the degree information. The proposed modeling of relations is based on fuzzy algebra principles and fuzzy sets and is aligned with the clear research trend that exists in the literature [9] towards "fuzzification" of ontology description languages (like fuzzy DL or fuzzy OWL), as the representation and reasoning capabilities of fuzziness go clearly beyond classical.

The structure of this paper is as follows: Section 2 refers to the multimedia analysis procedure that will be enhanced by the exploitation of the proposed infrastructure and explains the main motivation of our work. Section 3 deals with the basic notation to be used during the forthcoming knowledge formalization. Sections 5 and 6 present the corresponding inter- and intra-relations modeling,

[1] The formal definition of an ontology [5] supports also an inference layer, but this is outside the scope of this work.

whereas Section 7 presents some early experimental results. Finally, Section 8 briefly concludes this paper and discusses future aspects of our work.

2 Visual Dictionary Construction

In this section we present briefly previous work on image representation as a set of region types, using a visual dictionary. This is the first step of the analysis procedure. After the extraction of the low-level descriptions of all image regions, the approach of [10] is followed, in order to facilitate a semantically higher description, that will be used for the high-level concept detection. More specifically, a K-means clustering algorithm is used to cluster all regions derived from all images of the available training set. The number of clusters N_T is selected experimentally, after a trial and error process. This clustering is applied on the low-level descriptions of image regions, in the form of feature vectors, using the Euclidean distance as their similarity measure. Regions that lie closest to the centroids of the resulting clusters are selected to form the region thesaurus. These regions $w_i, \ i = 1, \ldots, N_T$ will be referred to as "region types". It should be made clear that each region type does not contain any high-level semantic information. However it provides a higher description in comparison to a low-level descriptor, i.e., one can describe a region type as "a green region with a coarse texture".

Using the above region thesaurus, the distances between each region of the image and all region types are calculated. This way, a "model vector" that semantically describes the visual content of an image, is formed, by keeping the smallest distance of all image regions r to each region type w_i. The j-th element of a model vector m_i describing image p_i is depicted in eq. 1:

$$m_i(j) = \min_{r \in R(k_i)} \Big\{ d\big(f(w_j), f(r)\big) \Big\} \tag{1}$$

where $i = 1 \ldots N_K$, $j = 1 \ldots N_T$, $d(\bullet)$ denotes the Euclidean distance function, $f(w_j)$ and $f(r)$ denote the feature vectors of a region type w_j and a region r, respectively.

We should make clear that the aforementioned methodology is used both to formalize a mid-level representation of a given image based on the region types of a visual dictionary and also to provide an initial estimation of the confidence with which the high-level concepts appear within it. This can be done by training appropriate detectors based on the model vectors for each concept.

3 Some Basic Notation

To begin, we define some notation that will be used throughout the description of the proposed context ontology. [2] Let e_1 and e_2 be two semantic entities and $R_1(e_1, e_2)$ be a relation between them. Then, we may define:

[2] We will refer to a high-level concept, a region or a region type as "semantic entity".

R_1^{-1} is the *Inverse Relation* of R_1: $R_1^{-1}(e_1, e_2) = R_1(e_2, e_1)$ (e.g. for the relation *sky co-occurs with sea*, the inverse relation is *sea co-occurs with sky.*

$\neg R_1$ is the *Opposite Relation* of R_1: $\neg R_1(e_1, e_2) = R_2(e_1, e_2)$ (e.g. "above" is the opposite relation of "below", "left" is the opposite relation of "right", etc).

Finally, for a given set S, its *cardinality* will be denoted by $\mid S \mid$.

4 Fundamental Sets

Before defining the contextual relations, we should begin by defining the sets of all entities that are encountered within the problems of interest. These fundamental sets, will be necessary for the definition of more specific sets and relations. More specifically, let:

$\mathcal{C} = \{c_k\},\ k = 1, \ldots, N_c$, be the set of all high-level concepts within the domain(s) of interest, as determined by a domain expert. An applicable subset of all these possible concepts may be selected from the ontology user/developer after the examination of the available training data.

$\mathcal{P} = \{p_k\},\ k = 1, \ldots, N_p$, be the set of all available images of the training set.

$\mathcal{S} = \{s_k\},\ k = 1, \ldots, N_s$ 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.

$\mathcal{Q} = \{q_k\},\ k = 1, \ldots, N_q$ be the set of all image pixels, of all regions, of all images.

$\mathcal{M} = \{m_k\},\ k = 1, \ldots, N_k$ be the set of all possible model vectors. For each image p_i, its model vector m_i is determined uniquely, using a specific visual dictionary.

$\mathcal{T} = \{t_k\},\ k = 1, \ldots, N_t$ be the set of all the region types of the given visual dictionary. $\mathcal{T}$ results by applying a clustering algorithm within all the elements of $\mathcal{S}$ and selecting the regions that lie closest to the centroids.

$\mathcal{D} = \{d_k\},\ k = 1, \ldots, N_d$ be the set of all possible visual descriptors of a given region. We should note that since in general the visual descriptors' values are quantized, this set may be significantly large, however it will always be finite. Visual descriptors are selected in order to be appropriate for the problem at hand.

After defining the aforementioned fundamental sets, we are now able to define more specific sets and relations for all the semantic entities. These sets will be in general subsets of the fundamental sets.

4.1 Sets within an Image *p*

To begin with a given each image p, we may define the following sets within it. More specifically, let:

$C_p = \{c_k^p\},\ k = 1, \ldots, N_c^p,\ p \in \mathcal{P}$, be the set of all high-level concepts *present* within image p and $C_p \subset \mathcal{C}$. C_p is determined by the provided annotation for the training set of images.

$S_p = \{s_k^p\},\ k = 1, \ldots, N_s^p,\ p \in \mathcal{P}$, be the set of all segmented regions of image p and $S_p \subset \mathcal{S}$.

$T_p = \{t_k^p\},\ k = 1, \ldots, N_t^p,\ p \in \mathcal{P}$, be the set of all region types (clusters) present in image p and $T_p \subset \mathcal{T}$.

$Q_p = \{q_k^p\},\ k = 1, \ldots, N_q^p,\ p \in \mathcal{P}$, be the set of all pixels of image p and $Q_p \subset \mathcal{Q}$.

$L_p = \{l_k^p\},\ k = 1, \ldots, N_l^p,\ p \in \mathcal{P}$, be the set of all labels of image p and $L_p \subset \mathcal{C}$. The labels of an image result from an application of appropriate high-level feature detectors, thus may not always be correct.

4.2 Sets and Relations for a Region s

For a given image region s, we may define the following sets and relations:

$p(s) : \mathcal{S} \rightarrow \mathcal{P}$, is a function that denotes the image that contains region s.

$t(s) : \mathcal{S} \rightarrow \mathcal{T}$, is a function that denotes the region type (cluster) t to which s "belongs" as in Eq. 1.

$C_s = C_{p(s)} = \{c_k^s\},\ s \in \mathcal{S},\ k = 1, \ldots, N_c^s$, is the set of high-level concepts contained within the image that contains region s. Herein, $N_c^p = N_c^{p(s)} = N_c^s$.

$L_s = \sum_k l_k^s / \mu_{L_s}(l_k^s)$[3], $k = 1, \ldots, N_c,\ l_k^s \in \mathcal{C},\ s \in \mathcal{S}$, denotes the fuzzy set of labels l_k^s for region s. The fuzzy membership function $\mu_{L_s}(l_k^s)$ denotes the confidence with which label l_k^s is assigned to region s.

$Q_s = \{q_k^s\},\ k = 1, \ldots, N_q^s,\ s \in \mathcal{S}$, is the entire set of pixels of region s.

$d(s) : \mathcal{S} \rightarrow \mathcal{D}$, is a function that extracts the visual descriptors from region s. This function may represent an appropriate tool used for the descriptor extraction.

4.3 Sets and Relations for a Region Type t

After defining sets and relations for a region, herein we define the following sets and relations for a region type t.

$L_t = \sum_k l_k^t / \mu_{L_t}(l_k^t),\ k = 1, \ldots, N_c,\quad l_k^t \in \mathcal{C},\ t \in \mathcal{T}$, denotes the fuzzy set of labels l_k^s for region type t, where the membership function $\mu_{L_s}(l_k^t)$ denotes the confidence that region type t corresponds to concept c_k^t.

$S_t = \{s_k^t\} = \{s \in \mathcal{S} : argmax_{t \in \mathcal{T}}(sim(d(s), d(t)))\},\ k = 1, \ldots, N_s^t,\ s \in \mathcal{S},\ t \in \mathcal{T}$, is the set of regions that are assigned to region type t. $d(t) : \mathcal{S} \rightarrow \mathcal{D}$, is the set of the visual descriptors extracted from region type t. $P_t = \{p \in \mathcal{P} : t \in T_p\},\quad t \in \mathcal{T}$ denotes the set of images that contain region type T

4.4 Sets and Relations for a Model Vector M_p

Let M_p be a model vector, as described in section 2. Before describing the necessary sets and relations concerning the model vectors, let $sim(d_1, d_2)$ denote a similarity function between two visual descriptors of the same type, d_1 and d_2. As necessary, $sim(d_1, d_2) \in [0, 1],\ d_1, d_2 \in \mathcal{D}$.

[3] The sum notation is used to describe a fuzzy set.

Then, for each model vector M_p, we are now able to define some basic sets and relations:

$T_s = \sum_k t_k^s / \mu_{T_s}(t_k^s),\ k = 1, \ldots, N_t^{(s)},\ t_k^s \in \mathcal{T}$, denotes the fuzzy set of region types t_k^s for region s, where the membership function $\mu_{T_s}(t_k^s) = sim(d(s), d(t))$, $s \in \mathcal{S},\ t \in \mathcal{T}$ represents the similarity between region s and region type t.

$t(s) = argmax_{t \in \mathcal{T}} \{\mu_{T_s}(t)\},\ s \in \mathcal{S},\ t \in \mathcal{T}$, is a function that determines the region type t where region s belongs.

$M_p = \sum_k m_k^p / \mu_{M_p}(m_k^p)\ k = 1, \ldots, N_t,\ p \in \mathcal{P},\ t \in \mathcal{T}$, denotes the fuzzy set that represents the model vector, where the membership function $\mu_{M_p}(t) = max_{s \in S_p} \{\mu_{T_s}(t)\},\ p \in \mathcal{P},\ t \in \mathcal{T}$, represents the maximum confidence among all those between the region s and the set of the region types $\mathcal{T}$.

5 Inter-relations among concepts, images, types and regions

Having defined the aforementioned sets and functions for all the entities of our framework, we continue in this section with a set of relations between different semantic entities (inter-relations).

5.1 Relations between concepts and region types

We begin with the relations among the high-level concepts and the region types that form the visual dictionary. These two semantic entities are linked, and as we have shown in previous work, the semantic content of an image may be described by the set of the region types from which the image is consisted of. Let:

$R_{ct} = \{r_{ct}\}$ be the set of the relations between a concept c and a region type t, where $r_{ct} = \mu_{L_t}(c),\ c \in \mathcal{C},\ t \in \mathcal{T}$. The membership function $\mu_{L_t}(c)$ denotes the confidence with which concept c is assigned to region type t. This confidence is calculated based on statistics of the training set. We should note here that the inverse relation cannot be defined. To calculate the membership function $\mu_{L_t}(c)$, Eq. 2 is used:

$$\mu_{L_t}(c) = \frac{\mid \{s \in S_t : c \in C_s\} \mid}{\mid S_t \mid} \tag{2}$$

5.2 Relations between concepts and regions

Between high-level concepts and image regions, there exists only one relation:

$R_{cs} = \{r_{cs}\} = \{\mu_{L_s}(c)\},\ c \in \mathcal{C},\ s \in \mathcal{S}$ denotes the set of the relations between a concept c and a region s. As it is obvious, r_{cs} is the confidence that concept c is assigned to region s. This confidence is calculated experimentally from the final output of a high-level concept detection scheme. We should note that the inverse relation cannot be defined.

5.3 Relations between region types and regions

To continue with the relations between the region types that form the visual dictionary and the image regions, we may define the following relations:

$t(s) : \mathcal{S} \rightarrow \mathcal{T}$: is a function that denotes the region type t to which region s is assigned. this function has already been defined.

$S_t,\ t \in \mathcal{T},\ s \in \mathcal{S}$: is the set of regions that are assigned to region type t. This function has already been defined.

5.4 Relations between images and concepts

In the same sense, we may define the following relations among images and high level concepts. These relations are defined in the sense that a certain number of concepts is assigned to an image based on a ground truth annotation and also by the output of a classification or detection scheme:

$C_p,\ p \in \mathcal{P}$, is the set of all high-level concepts present in image p, as defined in section 4.1.

$L_p = \sum_k l_k^p / \mu_{L_p}(l_k^p),\ k = 1, \ldots, N_c$, is the fuzzy set of the labels for image p. The membership function $\mu_{L_p}(l_k^p)$ is the confidence with which the high-level concept c_p (label) is assigned to image p. This function is calculated as the output of the high-level feature classifiers/detectors.

5.5 Relations between images and regions

Since we assume that a given image is decomposed to a number of regions, the only relation between image and regions is:

$S_p,\ p \in \mathcal{P}$, is the set of all regions of image p. This relation has already been defined in section 4.1.

5.6 Relations between images and region types

The final inter-relation is the one defined between images and region types, under the assumption that a given image may be described with the aid of a visual dictionary, as a set region types:

$T_p = \bigcup_{s \in S_p} t(s),\ p \in \mathcal{P}$, is the set of all region types present within image p. This relation has already been defined in section 4.1.

6 Intra-relations (within the same type of entities)

The final section deals with relations among the same kind of semantic entities. Therefore, we define relations among high-level concepts, among regions and among region types.

6.1 Relations among high-level concepts

To begin with the relations among semantic entities of the same kind, we define a set of relations R_{cc}, among high-level concepts C. First, the following semantic relations are defined by a domain expert:

$R_{cc}^{sim} = \{r_{c_1,c_2}^{sim}\} = \{sim(c_1, c_2)\}$, $c_1, c_2 \in \mathcal{C}$, is the semantic *Similarity* between high-level concepts c_1 and c_2. As obvious, $sim(c_1, c_2) = sim^{-1}(c_1, c_2)$. For example, the semantic *Similarity* would hold with a higher degree between concepts *sea* and *beach* than between *sea* and *outdoor*.

$R_{cc}^{part} = \{r_{c_1,c_2}^{part}\} = \{part(c_1, c_2)\}$, $c_1, c_2 \in \mathcal{C}$, is the *Part of* relation, i.e. concept c_1 is part of concept c_2. As obvious, $part(c_1, c_2) \neq part^{-1}(c_1, c_2)$. For example, *sky* may be a *Part of outdoor*, *wheel* may be a *Part of* a *car* and so on.

$R_{cc}^{spec} = \{r_{c_1,c_2}^{spec}\} = \{spec(c_1, c_2)\}$, $c_1, c_2 \in \mathcal{C}$, is the *Specialization* relation, i.e. concept c_1 is a *Specialization* of concept c_2. As obvious, $spec(c_1, c_2) \neq spec^{-1}(c_1, c_2)$. *Specialization* allows to a high-level concept to specialize the meaning of another. For instance, *appletree* specializes *tree* which also specializes *vegetation*.

To continue, we define a set of topological relations between high-level concepts. These relations are also defined by a domain expert, or calculated directly if an annotation per region is available, depending on the problem at hand. To avoid repetition in the definitions of those relations, since they are all rather similar, we summarize them in the following *Topological* relation:

$R_{cc}^{top} = \{r_{c_1,c_2}^{top}\} = \{top(c_1, c_2)\}$, $c_1, c_2 \in \mathcal{C}, top \in \{adj, ins, out, ab, bel, left, rgt\}$ is a *Topological* relation, i.e. concept c_1 is *Adjacent* to concept c_2. For the *Adjacency* relation, $adj(c_1, c_2) = adj^{-1}(c_1, c_2)$. As obvious in all other cases, $top(c_1, c_2) \neq top^{-1}(c_1, c_2)$.

The topological relations between concepts are summarized in Table 1.

Table 1. Contextual relations between entities.

Relation R	**Opposite** $\neg R$	**Symbol**	**Meaning**
Adjacent	-	$adj(a, b)$	adjacency between two entitiess
Inside	*Outside*	$ins(a, b)$	an entity is inside another entity
Outside	*Inside*	$out(a, b)$	an entity is outside another entity
Above	*Below*	$ab(a, b)$	an entity is above another entity
Below	*Above*	$bel(a, b)$	an entity is below another entity
Left	*Right*	$left(a, b)$	an entity is left to another entity
Right	*Left*	$rgt(a, b)$	an entity is right to another entity

The final relation between high-level concepts is defined statistically on the training set data.

$R_{cc}^{co} = \{r_{c_1,c_2}^{co}\} = \{co(c_1, c_2)\}$, $c_1, c_2 \in \mathcal{C}$, is the *Co-occurrence* relation, i.e. concept c_1 *Co-occurs* with concept c_2. As obvious, $co(c_1, c_2) = co^{-1}(c_1, c_2)$. To calculate the degree of the *Co-occurrence* relation, Eq. 3 is applied:

$$co(c_1, c_2) = \frac{\mid \{p \in \mathcal{P} : c_1 \in C_p \wedge c_2 \in C_p\} \mid}{\mid \{p \in \mathcal{P} : c_1 \in C_p \vee c_2 \in C_p\} \mid} \tag{3}$$

6.2 Relations among image regions

We continue by presenting a set of relations R_{ss}, between image regions (segments) S. The first two relations are calculated directly:

$R^{sim}_{ss} = \{r^{sim}_{s_1,s_2}\} = \{sim(d(s_1), d(s_2))\},\ s_1, s_2 \in \mathcal{S}$, is the low-level (visual) *Similarity* between the descriptor(s) extracted from the two regions s_1 and s_2. The *Similarity* is calculated with the use of an appropriate similarity function such as the Euclidean, the L1 etc.

$R^{co}_{ss} = \{r^{co}_{s_1,s_2}\} = \{co(s_1, s_2)\},\ s_1, s_2 \in \mathcal{S}$, denotes the *Co-occurrence* of regions s_1 and s_2 within the same image p. The degree of the *Co-occurrence* relation may be determined by Eq. 4.

$$co(s_1, s_2) = \begin{cases} 1, & \exists p \in \mathcal{P} :\ s_1, s_2 \in S_p \\ 0, & otherwise \end{cases} \tag{4}$$

The following topological relations between two regions within the same image are calculated directly. We provide a single definition for all topological relations:

$R^{top}_{ss} = \{r^{top}_{s_1,s_2}\} = \{top(s_1, s_2)\},\ s_1, s_2 \in \mathcal{S}, top \in \{adj, ins, out, ab, bel, left, rgt\}$, denotes a *Topological* relation. As it is obvious also in this case, for the *Adjacency* relation $adj(s_1, s_2) = adj^{-1}(s_1, s_2)$ and for all other *Topological* relations, $top(s_1, s_2) \neq top^{-1}(s_1, s_2)$.

The *Topological* relations between two image regions are namely the same to those of high-level concepts, thus are also summarized in Table 1.

To determine whether two regions are *Adjacent* or not, Eq. 5 is applied.

$$adj(s_1, s_2) = \begin{cases} 1, & Q_{s_1} \cap Q_{s_2} = \varnothing \wedge \exists (q_1, q_2) \in Q_{s_1} \times Q_{s_2} : n_c(q_1, q_2) = 1 \\ 0, & otherwise \end{cases} \tag{5}$$

where, $n_c(q_1, q_2)$ denotes the c-connectivity between pixels q_1 and q_2, $c \in \{4, 8\}$.

To calculate the degree to which the remaining topological relations stand, we adopt the methodology of [6].

6.3 Relations among region types

This section presents the set of relations R_{tt} between two region types T. We begin with those relations that are calculated directly:

$R^{sim}_{tt} = \{r^{sim}_{t_1,t_2}\} = \{sim(d(t_1), d(t_2))\},\ t_1, t_2 \in \mathcal{T}$, is the *Similarity* between the extracted descriptors from the two region types t_1 and t_2. The Similarity is calculated with the use of an appropriate similarity function such as the Euclidean, the L1 etc.

$R_{tt}^{co} = \{r_{t_1,t_2}^{co}\} = \{co(t_1,t_2)\}$ is the *Co-occurrence* within the same image p of region types t_1 and t_2. To calculate the degree of the *Co-occurrence* relation, Eq. 6 is applied.

$$co(t_1,t_2) = \frac{\mid P_{t_1} \cap\ P_{t_2} \mid}{\mid P_{t_1} \cup\ P_{t_2} \mid} \tag{6}$$

In order to define the *Adjacency* and the *Inside* relations, 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.

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

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

Then, *Adjacency* and *Inside* relations are defined and calculated as:

$R_{tt}^{adj} = \{r_{t_1,t_2}^{adj}\},\ t_1,t_2 \in \mathcal{T}$ is the *Adjacency* relation between two region types t_1 and t_2. The degree to which this relation holds is calculated by Eq. 7.

$$r_{t_1,t_2}^{adj} = \frac{\mid B_{t_1,t_2}^{adj} \mid}{\mid B_{t_1,t_2}^{co} \mid} \tag{7}$$

$R_{tt}^{ins} = \{r_{t_1,t_2}^{ins}\},\ t_1,t_2 \in \mathcal{T}$ is the *Inside* relation between two region types t_1 and t_2. The degree to which this relation holds is calculated by Eq. 8.

$$r_{t_1,t_2}^{ins} = \frac{\mid B_{t_1,t_2}^{ins} \mid}{\mid B_{t_1,t_2}^{co} \mid} \tag{8}$$

Finally, we provide the definition of the following topological relations between region types.

$R_{tt}^{top} = \{r_{t_1,t_2}^{top}\},\ t_1,t_2 \in \mathcal{T},\ top \in \{ab, bel, left, rgt\}$ denotes a *Topological* relation between region type t_1 and region type t_2. As obvious, $top(t_1,t_2) \neq top^{-1}(t_1,t_2)$.

The relations r_{t_1,t_2}^{ab}, r_{t_1,t_2}^{bel}, r_{t_1,t_2}^{left} and r_{t_1,t_2}^{rgt} are calculated based on the algorithm presented in [6]. This algorithm assumes that the spatial relations between two points are determined by the angle made by the line passing through the two points and the x-axis. Finally, for two given region types, t_1 and t_2 within the same image, we get one degree $d_X(t_1,t_2), X \in \{ab, bel, left, rgt\}$ for each relation. Then, we define the following subsets that include the pairs of region types for which each topological relation $top \in \{ab, bel, left, rgt\}$ holds:

$$B_{t_1,t_2}^{top} = \{(s_1,s_2) \in B_{t_1,t_2} : max\{d_X(t_1,t_2),\ X \in \{ab, bel, left, rgt\} = d_{top}(t_1,t_2)\} \tag{9}$$

Now, the corresponding relations may be calculated by:

$$r_{t_1,t_2}^{top} = \frac{\mid B_{t_1,t_2}^{top} \mid}{\mid B_{t_1,t_2}^{co} \mid},\ t_1,t_2 \in \mathcal{T} \tag{10}$$

7 Experiments

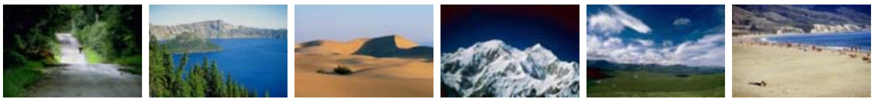

Fig. 1. Indicative Corel dataset images.

In the following we present our initial experimental results over a dataset of 750 images and 6 high-level concepts, derived from the well-known Corel [14] dataset and depicted in Figure 1. We utilized 525 images to train 6 different SVMs and 225 images as the test set. In Table 2, we summarize the concept detection results obtained from the utilization of the proposed knowledge formalization, based on the contextualization methodology presented in [8]. We observe a precision optimization for all 6 concepts.

Table 2. Precision (P)/recall (R) per concept.

	before		**after**		**%**	
concepts	**P**	**R**	**P**	**R**	**P**	**R**
road	0.22	0.25	0.43	0.21	+95%	-16%
sand	0.38	0.33	0.55	0.28	+45%	-15%
sea	0.78	0.71	0.89	0.68	+14%	-4%
sky	0.81	0.72	0.91	0.67	+12%	-7%
snow	0.48	0.58	0.72	0.45	+50%	-22%
vegetation	0.74	0.62	0.87	0.54	+18%	-13%
total	**0.57**	**0.54**	**0.73**	**0.47**	**+28%**	**-13%**

8 Conclusions and future work

This work proposed an integrated novel type of contextual knowledge to be utilized within the multimedia analysis value chain. The introduced context model is suitable for use and significantly aids in knowledge extraction, when handling high-level concept detection problems. The herein presented effort forms a small piece of work at the beginning of our research on contextually-assisted mid-level image/video analysis. It places itself in the process, as it relates to object identification and image classification and will be exploited in the form of driving the analysis process of our work by selecting suitable algorithms, detectors and classifiers. Future work will include all above mentioned issues, along with large-scale experimental results on the Corel and TREC datasets, indicating its benefits and contributing to the overall usage of context in multimedia analysis.

9 Acknowledgements

This research was partially supported by the European Commission under contracts (FP6-027685 - MESH), (FP6-027026 - K-Space) and (FP7-215453 - WeKnowIt). Evaggelos Spyrou is partially funded by PENED 2003 Project Ontomedia 03ED475.

References

1. K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei and M. I. Jordan, *Matching words and pictures*, J. Mach. Learn. Res., 3:11071135, 2003.
2. C. Dance, J. Willamowski, L. Fan, C. Bray and G. Csurka, *Visual categorization with bags of keypoints*, In Proc. of ECCV - International Workshop on Statistical Learning in Computer Vision, Prague, 2004
3. B. Edmonds, *The Pragmatic Roots of Context*, 2nd International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-99), LNAI, vol. 1688, Berlin: Springer, 1999.
4. G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice Hall, 1995.
5. A. Maedche, B. Motik, N. Silva and R. Volz, *MAFRA - An Ontology MApping FRAmework in the Context of the SemanticWeb*, Proceedings of the Workshop on Ontology Transformation ECAI2002, Lyon, France, July 2002.
6. Miyajima, K. and Ralescu, A., *Spatial organization in 2D images*, Proc. of the Third IEEE Conference on Fuzzy Systems, 1994.
7. Ph. Mylonas and Y. Avrithis, *Context modeling for multimedia analysis*, 5th International and Interdisciplinary Conference on Modeling and Using Context, 2005.
8. Ph. Mylonas, E. Spyrou and Y. Avrithis, *Enriching a context ontology with mid-level features for semantic multimedia analysis*, 1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies, co-located with SAMT '07, Genova, Italy, December 2007.
9. E. Sanchez, *Fuzzy Logic and the Semantic Web*, Elsevier Science Inc., New York, NY, USA, 2006.
10. E. Spyrou and Y. Avrithis, *A region thesaurus approach for high-level concept detection in the natural disaster domain*, 2nd International Conference on Semantics And digital Media Technologies (SAMT), Genova, Italy, December 2007.
11. A. Torralba, *Contextual priming for object detection*, International Journal on Computer Vision, vol. 53, pp. 169-191, 2003.
12. A. Torralba, *Contextual influences on saliency*, Neurobiology of attention, Academic Press Inc., London, 2005.
13. N. Voisine, S. Dasiopoulou, V. Mezaris, E. Spyrou, T. Athanasiadis, I. Kompatsiaris, Y. Avrithis and M. G. Strintzis, *Knowledge-assisted video analysis using a genetic algorithm*, 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005).
14. J.Z. Wang, J. Li, G. Wiederhold, *SIMPLIcity: Semanticssensitive Integrated Matching for Picture LIbraries*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 23, No.9, 947-963, 2001.
15. J. Yuan, J. Li and B. Zhang, *Exploiting spatial context constraints for automatic image region annotation*, In Proceedings of the 15th international conference on Multimedia, pp. 595-604, Augsburg, Germany, 2007.
16. W3C, RDF reification, `http://www.w3.org/TR/rdf-schema/`, 2004.