Enriching a context ontology with mid-level features for semantic multimedia analysis

Phivos Mylonas, Evaggelos Spyrou and Yannis Avrithis

Image, Video and Multimedia Laboratory,
National Technical University of Athens
Zographou Campus, PC 15773, Athens, Greece
{fmylonas, espyrou, iavr}@image.ntua.gr

Abstract. In this paper we focus on a contextual domain ontology representation aiding in the process of knowledge-assisted multimedia analysis. Previous work on the detection of high-level concepts within multimedia documents is extended by introducing a "mid-level" ontology as a means of exploiting the visual context of images, in terms of high-level concepts and mid-level region types they consist of. More specifically, we introduce a context ontology, define its components, its relations and integrate it in our knowledge modelling approach. In previous works we have developed algorithms to address computationally efficient handling of visual context and extraction of mid-level characteristics and now we expect these diverse algorithms and methodologies to be combined in order to exploit the proposed knowledge model. The ultimate goal remains that of efficient semantic multimedia analysis. Finally, a use case scenario derived from the *beach* domain is also presented, in order to demonstrate a possible application of the proposed knowledge representation.

1 Introduction

Recent advances in the research field of knowledge-assisted multimedia analysis along with the emerge of new content and metadata representations, have driven more and more researchers looking beyond solely low-level features (e.g. color, texture, and shape) in pursuit of more effective high-level multimedia representation and analysis methods. Current and previous multimedia research efforts have focused in combining both low-level descriptors computed automatically from raw multimedia content and semantics focusing in extracting high-level features.

The idea of combining formalized knowledge and a set of features to describe the visual content of an image has been presented for instance in [23], where a region-based approach using MPEG-7 visual features and ontological knowledge is presented. Moreover, in [17] a lexicon-driven approach is introduced. Some 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. Among others, a region-based approach in content

retrieval using Latent Semantic Analysis is presented in [18], whereas a mean-shift algorithm is used in [15], in order to extract low-level features, after the image is clustered. In [6] 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. In [5] visual categorization is achieved using a bag-of-keypoints approach. Finally, in [13] the authors train separate shape detectors using a shape alphabet, which is actually a dictionary of curve fragments.

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 [22]. 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 and/or scene classification [21] are unfortunately not producing reliable results and restricting the problem to a specific domain does not provide a global and satisfactory solution.

The notion of (visual) context is introduced in [22] and [10], 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. Visual context is a difficult notion to grasp and capture and thus we restrict it herein to the notion of ontological context, defined as part of the "fuzzified" context ontology presented in Section 3. Our choice is aligned with the clear research trend that exists in the literature [14] towards "fuzzification" of ontology description languages, like fuzzy DL and fuzzy OWL [20], as the representation and reasoning capabilities of fuzziness go clearly beyond classical.

In the following, we propose our initial research progress in implementing an RDF-based context representation approach, able to use within any knowledge-assisted image analysis methodology. The ultimate goal is to be able to apply our approach on top of any given image domain. We introduce a methodology to improve the results of high-level feature extraction, based on the introduced contextual information. In comparison to some of our previous research efforts [1], a novel multiple domain ontological representation for context is introduced, combining fuzzy theory and fuzzy algebra [8] with recent knowledge representation approaches, such as RDF [25] and reification [26]. In this process, the membership degrees among the ontology concepts are re-estimated appropriately, according to a context-based membership degree readjustment algorithm.

The structure of this paper is as follows: In Section 2, the proposed mid-level conceptualization is introduced, whereas in Section 3 the overall fuzzy context knowledge formalization is described, including some basic notation used throughout the paper. Section 4 describes the utilized contextualization step. Section 5 illustrates a preliminary use case scenario and Section 6 briefly concludes our work.

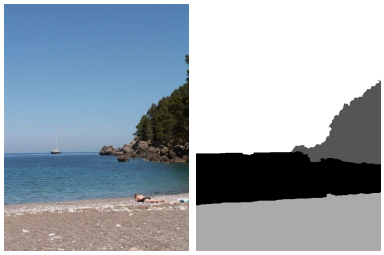


Fig. 1. An input image and its coarse segmentation.

2 Mid-level Conceptualization

As already mentioned in Section 1, among our main goals with this research work is to provide an ontological knowledge representation containing both high-level features (i.e. high-level concepts) and mid-level features. In this section we will describe the nature of the latter features and the way they are selected.

Generally, the visual features one can extract from an image or video document can be divided in two major categories. The first one contains the *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 the *high*-level features, which 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. Thus, this problem often referred to as the "Semantic Gap"[16] attracts a lot of interest within the research community.

In this sense, we try to enhance the notion of a visual context ontology 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 "mid-level" manner and attach it to the context ontology by defining certain relations, as described in Section 3.

To better understand the notion of these mid-level concepts, we present a visual example in Figure 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. a *"light blue" region*, a *"blue" region*, a *green* region and a *"grey" region*. We shall call these mid-level 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.

An arbitrary large number of candidate region types is initially needed. To gather it, a color segmentation algorithm is first applied on all images of the available training set, as a pre-processing step. This algorithm is a multi-resolution implementation of the well-known RSST [2], tuned to produce a coarse segmentation. This way, the resulting segmented regions facilitate a qualitative description of the image as the aforementioned of Figure 1. Then, from each region we extract certain low-level visual features. More specifically, color and texture descriptors from the MPEG-7 standard [4] are selected to capture a standardized description of their visual content as they have been effectively applied for such a use in various applications. For representing the color features, three MPEG-7 color descriptors are extracted: The *Color Layout Descriptor*, the *Scalable Color Descriptor* and the *Color Structure Descriptor*. Moreover, for representing the texture features, the *Homogeneous Texture Descriptor* is also extracted. For the extraction of these color and texture descriptors, the MPEG-7 eXperimentation Model (XM)[9] has been applied.

A simple observation of the training set of images and the set of the segmented regions reveals that images containing similar semantic concepts are consisted of similar regions. As a natural sequence of this observation we apply a *hierarchical clustering* algorithm on those regions. The general structure of this algorithm, adjusted for the problem at hand, is as follows:

1. Turn each input element into a singleton, i.e. into a cluster of a single element.
2. For each pair of clusters c_1, c_2 calculate a compatibility indicator $CI(c_1, c_2)$. The CI is also referred to as cluster similarity, or dissimilarity, measure.
3. Merge the pair of clusters that have the best CI. Depending on whether this is a similarity or a dissimilarity measure, the best indicator could be the *max* or *min* operator, respectively.
4. Continue at step 2, until the termination criterion is satisfied. The termination criterion most commonly used is the definition of a threshold for the value of the best compatibility indicator.

After this process, we should note that each cluster may or may not represent a high-level feature and each high-level feature may be represented by one or more clusters; i.e. the concept *sand* can have many instances differing e.g. in the color of the sand. 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 Figure 2. In this simplistic example an initial set of 14 candidate region types is clustered. Then, 6 region types are selected to represent the mid-level features.

3 Knowledge Formalization

In this section, we further advance the proposed conceptualization; we introduce a novel knowledge representation approach in the form of an extended context

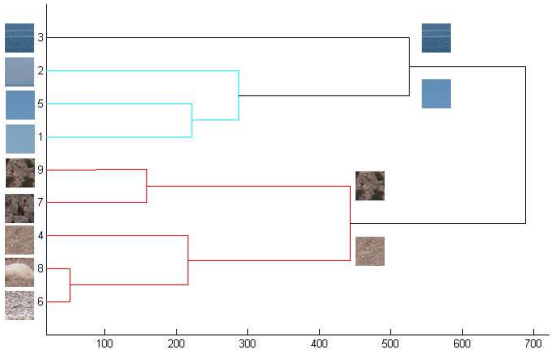


Fig. 2. Region type selection using hierarchical clustering. Selected region types are depicted within the red box.

ontology, initially presented in [12]. The proposed ontology is described by a set of high-level concepts, a set of region types and a set of relations among them. The set of concepts is defined by a domain expert. In general, this type of ontology O 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[1]. 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:

$$O = \{C, T, R_{x_i,y_j}\}, \quad R_{x_i,y_j} : X \times Y \rightarrow \{0,1\}, \quad i,j = 1 \ldots m, \quad i \neq j \qquad (1)$$

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

As it is quite common in the literature, any kind of relation may be represented by an ontology; however, herein we restrict it to the notion of a "fuzzified" ad-hoc context ontology. In principle, given a universe U, a crisp set S of entities on U is described by a membership function $\mu_S : U \rightarrow \{0,1\}$. The crisp set S is defined as $S = \{s_i\}$, $i = 1, .., N$, $s_i \in S$, whereas a *fuzzy* set F on S is described by a membership function $\mu_F : S \rightarrow [0,1]$. The "fuzzified" ontology is introduced in order to express in an optimal way the real-world relationships that exist between the concepts and the region types of a scene. In order for this ontology type to be highly descriptive, it must contain a representative number of distinct and even diverse relations among high-level concepts, among region types, and among concepts and region types, so as to exploit in an optimal manner

[1] In the following, we shall use the term *entities* when referring to either concepts or region types.

the contextual information surrounding each one. Additionally, since modelling of real-life information is in most cases governed by uncertainty, it is our belief that these relations must incorporate fuzziness in their definition. Thus, we utilize a set of relations (Table 1), derived from the set of MPEG-7 relations, that are suitable for image analysis [3] and re-define them in a way to incorporate fuzziness. A degree of confidence is associated to each relation, and assists in discriminating between objects exhibiting similar visual characteristics. The set of utilized relations contains both topological and 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).

Table 1. Contextual relations between region types.

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

Each entity may be related to another using one or more of the aforementioned contextual fuzzy 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 any 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 pair of entities are depicted in the last three columns of Table 1.

Among the utilized relations, relation *Accompanier* denotes the coexistence of two entities within an image/video. *Component* denotes the combination of an entity with another, and is used when these entities combined form another one. Two or more high-level concepts may be combined to form another high-level concept, while two or more region types may be combined not only to form another region type, but also a high-level concept. For instance, *sky*, *sea* and *sand*, when combined, form the high-level concept *beach*. On the other side, it is obvious that a combination of two region types is another region type, however a *"brown"* and a *"green"* region type when combined form the high-level concept *tree*. The *PartOf* relation denotes that one entity is part of another. For example, *sky* may be a part of *outdoor*. *Specialization* allows to a high-level concept to specialize the meaning of another. For instance, *appletree* specializes *tree* which also specializes *vegetation*. *Greece* is an *ExampleOf country*. *sand* may be the *LocationOf umbrella*. Finally, a *"green"* region may be a *PropertyOf vegetation*.

All the above relations form a context model, that can be seen as a graph: every node of the graph represents a concept 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 concepts is produced based on the set of contextual fuzzy relations that are meaningful for the particular pair. For instance, the edge between concepts *rock* and *sand* is produced by the combination of relations *Location* and *Accompanier*, whereas the *water* and *sea* edge utilizes relations *Specialization*, *PartOf*, *Example* and *Location*, in order to be constructed. In the same sense, two region types, i.e. a *"green"* and a *"blue"* may utilize the relations *Similar*, *Accompanier* and *Component*.

As in [7], a fuzzy relation on T is a function $\mathcal{R}_{x_i,x_j} : X \times Y \rightarrow [0,1]$ and its inverse relation is defined as ${\mathcal{R}_{x_i,y_j}}^{-1} = \mathcal{R}_{y_j,x_i}$. Based on the above relations, a domain-specific, "fuzzified" version of the proposed ontology may be described by $\mathcal{O}$:

$$\mathcal{O} = \left\{C, T, \mathcal{R}_{x_i,x_j}\right\}, \quad i,j = 1,\dots m, \quad i \neq j \tag{2}$$

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

$$F(R_{x_i,y_j}) = \mathcal{R}_{x_i,x_j} : X \times Y \rightarrow [0,1] \tag{3}$$

denotes a fuzzy ontological relation amongst two entities x_i, y_j and

$$\mathcal{R}_{x_i,y_j} = \{Sim, Acc, P, Comp, Sp, Ex, Loc, Pr\} \tag{4}$$

A possible combination of relations

$$\mathcal{Z} = (\bigcup_{i,j} {\mathcal{R}_{x_i,y_j}}^{p_{ij}}), \quad p_{ij} \in \{-1, 0, 1\}, \quad i,j = 1 \dots m, \quad i \neq j \tag{5}$$

may then be used to form a single RDF graph [25], which constitutes the abstract contextual knowledge model formed herein and ready to be used during the analysis phase. The value of p_{ij} is determined by the semantics of each relation $\mathcal{R}_{x_i,y_j}$ used in the construction of $\mathcal{Z}$. More specifically:

- $p_{ij} = 1$, if the semantics of $\mathcal{R}_{x_i,y_j}$ imply it should be considered as is
- $p_{ij} = -1$, if the semantics of $\mathcal{R}_{x_i,y_j}$ imply its inverse should be considered
- $p_{ij} = 0$, if the semantics of $\mathcal{R}_{x_i,y_j}$ do not allow its participation in the construction of the combined relation $\mathcal{Z}$.

In Figure 5 we present a fragment of the aforementioned RDF graph. Only the relations among high-level concepts and region types are depicted, for the sake of presentation. In the same sense, we present the relations between high-level concepts in Figure 4(a) and those between region types in Figure 4(b). We should note here that we omit the relations among the entities of the ontology and the *beach* concept which actually is the root of the graph and all entities are connected with it.

```
<rdf:Description rdf:about="#Relation1">
    <rdf:subject rdf:resource="&dom;rt1"/>
    <rdf:predicate rdf:resource="&dom;Part"/>
    <rdf:object>rdf:resource="&dom;rt2"</rdf:object>
    <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Statement"/>
    <context:Part rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.85</context:Part>
</rdf:Description>
```

Fig. 3. RDF ontology fragment.

We should note here that the aforementioned graphs ,represent only a small fragments of the whole visual context ontology, for the sake of the presentation and the explanatory examples. That is because an ontology with 1 domain, 10 concepts and 25 region types (i.e. 36 entities) would require a maximum of 630 relations. Even though not all the semantic relations are applicable, as already explained, such a complicated graph is difficult to be presented in a figure.

The graph of the proposed model contains nodes (i.e. high-level concepts and region types) and edges (i.e. contextual fuzzy relations between high-level concepts and/or region types). The degree of confidence of each edge represents fuzziness in the model. Non-existing edges imply non-existing relations [2]. As each high-level concept and each region type have a

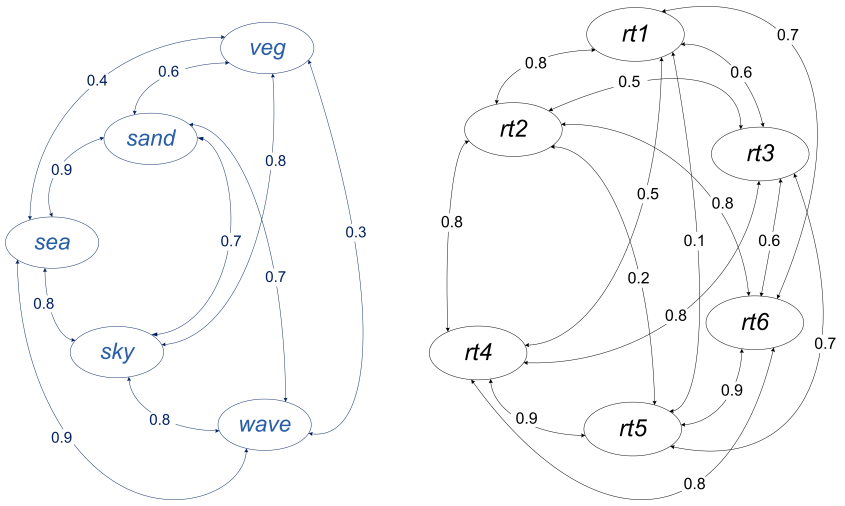
different probability to appear in the scene, a flat context model would not have been sufficient in this case; quite on the contrary, entities are related to each other, implying that the graph relations used are in fact transitive.

Since more than one fuzzy relations are often applicable for a pair of entities, it appears difficult to visualize all of them simultaneously in a graph. Figures 5, 4(a) and 4(b) present fragments of *beach* context ontology, where each relation between two entities may be either unique or a combination of more than one relations, as depicted in eq. (5). To facilitate the visualization of the ontology, in Table 2, we present all possible relations between a high-level concept, depicted as C_1 and all other entities of the ontology. A zero value denotes the absence of the corresponding relation. Moreover, in Table 3 we present the corresponding fuzzy values for each pair of entities for the *Accompanier* fuzzy relation, which is applicable between any two given entities.

Table 2. Fuzzy relations between high-level concept C_1 and all other entities. The numbers indicate the fuzzy degree of confidence for each relation.

	C_1	C_2	...	C_N	T_1	T_2	...	T_M
Sim	0	0	...	0	0	0	...	0
Acc	1	0.5	...	0.9	0.7	0.8	...	0
P	1	0	...	0.3	0.7	0	...	0
$Comp$	1	0.2	...	0.9	0	0.5	...	0
Sp	0	0.8	...	0	0	0	...	0
Ex	0	0.7	...	0	0	0	...	0
Loc	0	0.9	...	0.8	0	0	...	0
Pr	0	0	...	0	0.5	0	...	0.7

[2] In other words relations with zero confidence values are omitted.



(a) High-level concepts and the relations among them.

(b) Region types and the relations among them.

Fig. 4. Two fragments of the visual context ontology.The numbers indicate the fuzzy degree of confidence for each relation.

Table 3. The degrees of the *Accompanier* relation for all pairs of entities.The numbers indicate the fuzzy degree of confidence for each relation.

	C_1	C_2	...	C_N	T_1	T_2	...	T_M
C_1	1	0.7	...	0	0.7	0.2	...	0.4
C_2	0.7	1	...	0.8	0.6	0.7	...	0.5
⋮	⋮	⋮		⋮	⋮	⋮		⋮
C_N	0	0.8	...	1	0.6	0.7	...	0.8
T_1	0.7	0.2	...	0.4	1	0.3	...	0.5
T_2	0.6	0.7	...	0.5	0.3	1	...	0.1
⋮	⋮	⋮		⋮	⋮	⋮		⋮
T_M	0.6	0.7	...	0.8	0.5	0.1	...	1

Describing each edge's accompanying degree of confidence may be carried out using a variety of methods; herein, we chose to use the methodology introduced by RDF reification [26]. Reification is used in knowledge representation to represent facts that must then be manipulated in some way; for instance, to compare logical assertions from different witnesses to determine their credibility. The message "Ben is the leader of the group" is an assertion of truth that commits the sender to the fact, whereas the reified statement, "Juliet reports that Ben is the leader of the group" defers this commitment to Juliet. In this way, statements may include fuzzy information (i.e. "Ben is the leader of the group with a degree of confidence equal to 0.85"), without creating contradictions in reasoning, since a statement is being made about the original statement, which

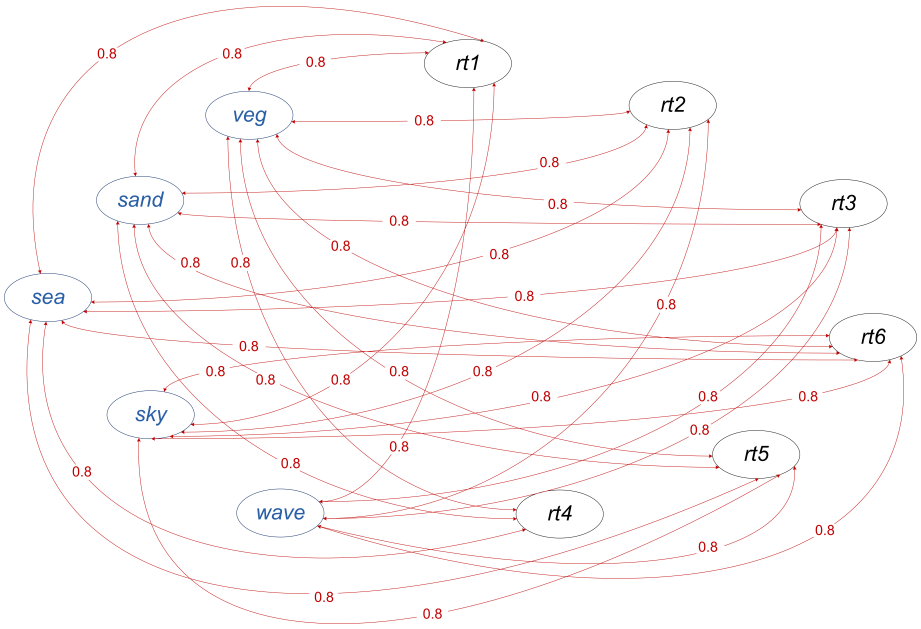


Fig. 5. A fragment of the visual context ontology. Only relations between high-level concepts and region types are depicted. The numbers indicate the fuzzy degree of confidence for each relation.

contains the degree information. Of course, the reified statement should not be asserted automatically, a fact that proves the use of the above technique to be acceptable. For instance, having a triple in RDF language such as: "*blue partOf green*" and a degree of confidence of "*0.85*" for this statement, does obviously not entail, that a *blue* region type will always be part of a *green* region type in the scene.

4 Visual Context Optimization

Based on the principles and mathematical foundations of fuzzy algebra [8] and the described knowledge conceptualization, we further present an ad hoc visual context optimization step and algorithm. Its core functionality 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 the dominant concept and/or 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. In this section, the remaining problems to be tackled 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 mid-level entity is derived from direct and indirect relationships of the latter with other entities in the constructed graph, using a meaningful compatibility indicator or distance metric. Depending on the nature of the domains provided in the domain ontology, the best indicator could be selected using the *max* or the *min* operator, respectively. Of course the ideal distance metric for two concepts or region types is again one that quantifies their semantic correlation. For the problem at hand, the *max* value is a meaningful measure of correlation for both of them. The general structure of the proposed degree of membership re-evaluation algorithm, using the standard t-conorm and the algebraic product as the t-norm, is as follows:

1. Identify a domain similarity (or dissimilarity) measure, imposed by the nature of the considered domain: $dnp \in [0, 1]$.
2. For each concept, $c \in C$ we describe the fuzzy set L_c, using the widely applied [8] sum notation: $L_c = \sum_{i=1}^{|C|} c_i/w_i = \{c_1/w_1, c_2/w_2, \ldots, c_n/w_n\}$, where w_i describes the membership function: $w_i = \mu_{L_c}(c_i)$
3. For each region type, $t \in T$ we describe the fuzzy set L_t: $L_t = \sum_{i=1}^{|T|} t_i/w_i = \{t_1/w_1, t_2/w_2, \ldots, t_n/w_n\}$, where w_i describes the membership function: $w_i = \mu_{L_t}(t_i)$
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 w_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 degree of membership w_i, taking into account each domain's similarity measure. 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 by considering the *context relevance* notion introduced in [11]: cr_{c_i} or cr_{t_i}, respectively.

We express the calculation of w_i for both cases with the recursive formula:

$$w_i^n = w_i^{n-1} - dnp(w_i^{n-1} - cr_{x_i}) \tag{6}$$

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:

$$w_i^n = (1 - dnp)^n \cdot w_i^0 + (1 - (1 - dnp)^n) \cdot cr_{x_i} \tag{7}$$

where w_i^0 represents the initial degree of membership for entity $x_i, x_i \in C \cup T$. Typical values for n reside between 3 and 5.

5 Experimental Results

In this section we will present initial results of the aforementioned knowledge base and the visual context optimization algorithm. We will try to show the usefulness of the visual context optimization algorithm when adopted in real multimedia problems and prove that the contextual knowledge in the form of the visual context ontology is able to complement the traditional approaches in high-level concept detection tasks, facilitate their process and refine their results.

A flowchart describing the interaction between the visual context optimization algorithm and the high-level feature detection process is depicted in Figure 6.

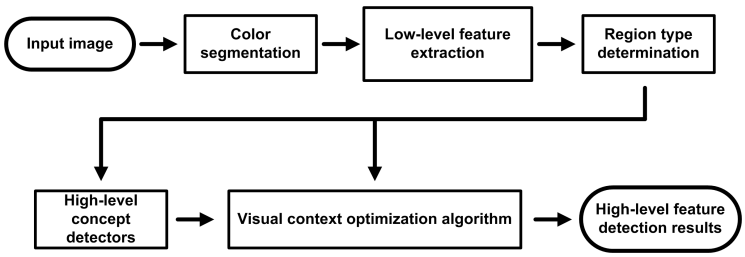


Fig. 6. Simple use case interaction flowchart.

We carried out experiments utilizing 287 images and 25 region types derived from the *beach* domain, acquired from personal collections and the World Wide Web. A ground truth was manually constructed, consisting of a number of region types associated to a unique concept. We utilized 57 images (merely 20% of the dataset) as our clustering training set and after an extensive try-and-error process selected $dnp = 0.12$ as the optimal normalization parameter for the given domain.

The problem we consider in this case is the detection of visual concepts. For certain high-level feature detection problems, only global annotation is available. In previous work [19] this problem has been tackled by an image representation based on a region thesaurus (i.e. a set of region types). This work is extended here by exploiting the visual context ontology, thus, aiming to improve the confidence values in an iterative way by taking into account the contextual relationships among the region types and the concepts that form the image.

An input image is depicted in Figure 7. The confidence values for all 6 region types t_i of this image are depicted in eq. (8).

$$\mathbf{T} = \{t_i\} = \left[0.89\ 0.62\ 0.21\ 0.68\ 0.67\ 0.31\right] \tag{8}$$

The confidences for the high-level concepts, *sea*, *sky*, *sand*, *wave* and *vegetation* as produced by specialized detectors are depicted in eq. (9) respectively.

$$\mathbf{C} = \{c_i\} = \left[0.32\ 0.91\ 0.12\ 0.87\ 0.35\right] \tag{9}$$

Fig. 7. The input image considered for the use case example.

As obvious to a human, the input image contains the high-level concepts *sea*, *sky* and *wave*. However, from the detectors' output as depicted in eq. (9), we may observe that the *sea* detector has failed to produce a significantly high confidence value. An explanation for this behavior is that this detector was trained using different *sea* positive examples than the the one of Figure 7. However, after the context optimization algorithm, the vector that contains the confidence values for the high-level concepts becomes the one of eq. (10).

$$\mathbf{C}' = \{c_i'\} = \begin{bmatrix} 0.62 & 0.95 & 0.18 & 0.90 & 0.29 \end{bmatrix} \tag{10}$$

In this simplistic example we are able to understand the power of the visual context optimization algorithm and its importance to multimedia analysis problems. Since the context algorithm had the information that

- This was a *beach* image
- *sky* had a high confidence value
- *wave* had a high confidence value
- There exists a *"blue"* region type
- There exists a *"white"* region type
- *sky* and *wave* have a high relation with *sea*
- "blue" region type and *"white"* region type have a high relation with *sea*

we may observe that it increased the confidence value of *sea*, while for instance the values of *vegetation* and *sand* which were small, remained small enough. Also, the confidence values of *sky* and *wave* which were already high, remained practically unchanged.

6 Conclusions and Future Work

The methodology presented in this paper can be exploited towards the development of a more efficient, context-based multimedia analysis environment. Among its core contributions, the notion of visual context interpretation utilizing a fuzzy, RDF-based, ontological representation of knowledge, as well as a visual context algorithm suitable for both high-level concepts and mid-level region types, forms an innovative approach, independent from the entities used. The proposed use case scenario indicates one of its possible utilizations. It is also the authors belief that further exploitation of the proposed approach may be considered in the fields of both multimedia knowledge representation and analysis.

7 Acknowledgements

This research was partially supported by the European Commission under contract FP6-001765 aceMedia. Evaggelos Spyrou is partially funded by PENED 2003 Project Ontomedia 03ED475.

References

1. Athanasiadis, Th., Mylonas, Ph., Avrithis, Y. and Kollias, S.: Semantic Image Segmentation and Object Labeling, IEEE Trans. on Circuits and Systems for Video Technology, **17**(3), 298–312 (2007)
2. Avrithis, Y., Doulamis, A., Doulamis, N. and Kollias, S.: A stochastic framework for optimal key frame extraction from MPEG video databases, Computer Vision and Image Understanding, **75** (1/2), 3–24 (1999)
3. Benitez, A. B., Zhong, D., Chang, S.-F. and Smith, J. R.: MPEG-7 MDS Content Description Tools and Applications, Lecture Notes in Computer Science (2001)
4. Chang, S.F., Sikora, T. and Puri, A.: Overview of the MPEG-7 standard, IEEE Trans. on Circuits and Systems for Video Technology **11**(6) (2001) 688–695
5. Dance, C., Willamowski, J., Fan, L., Bray, C., and Csurka G.: Visual categorization with bags of keypoints, In Proc. of ECCV - International Workshop on Statistical Learning in Computer Vision, Prague, 2004
6. Gokalp, D., and Aksoy, S.: Scene Classification Using Bag-of-Regions Representations, In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07)
7. Klir, G., and Yuan, B.: Fuzzy Sets and Fuzzy Logic, Theory and Applications, Prentice Hall, New Jersey, 1995
8. Miyamoto, S.: Fuzzy Sets in Information Retrieval and Cluster Analysis, Kluwer Academic Publishers, Dordrecht / Boston / London (1990)
9. MPEG-7: Visual experimentation model (xm) version 10.0. ISO/IEC/ JTC1/SC29/WG11, Doc. N4062 (2001)
10. Mylonas, Ph. and Avrithis, Y.: Context modelling for multimedia analysis, In Proc. of 5th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT 05), Paris, France, July 2005.
11. Mylonas, Ph., Athanasiadis, Th. and Avrithis, Y.: Improving image analysis using a contextual approach, In Proc. of 7th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Seoul, Korea 2006
12. Mylonas, Ph., Athanasiadis, Th. and Avrithis, Y.: Image Analysis Using Domain Knowledge and Visual Context, In Proc. of 13th International Conference on Systems, Signals and Image Processing (IWSSIP 2006), Budapest, Hungary.
13. Opelt, A., Pinz, A., and Zisserman, A.: Incremental learning of object detectors using a visual shape alphabet, In Proc. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06), Washington, DC, USA
14. Sanchez, E.: Fuzzy Logic and the Semantic Web, Elsevier Science Inc., New York, NY, USA (2006)
15. Saux, B. and Amato, G.: Image classifiers for scene analysis, In Proc. of International Conference on Computer Vision and Graphics, 2004.
16. Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. and Jain, R.: Content-Based Image Retrieval at the End of the Early Years, IEEE Trans. on Pattern Analysis and Machine Intelligence, **22**: 1349–1380 (2000)

17. Snoek, C., Worring, M., Koelma, D. and Smeulders, A.: A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval, IEEE Trans. on Multimedia, 9(2): 280–292 (2007)
18. Souvannavong, F., Merialdo, B., and Huet, B.: Region-based video content indexing and retrieval, In Proc. of 4th International Workshop on Content-Based Multimedia Indexing, Riga, Latvia, 2005
19. Spyrou, E. and Avrithis, Y.,:A Region Thesaurus Approach for High-Level Concept Detection in the Natural Disaster Domain, In Proc of the 2nd International Conference on Semantics And digital Media Technologies (SAMT), 2007
20. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z. and Horrocks, I.: Fuzzy OWL: Uncertainty and the Semantic Web, International Workshop of OWL: Experiences and Directions, Galway, 2005
21. Torralba, A. *Contextual priming for object detection*, Int. J. Comp. Vis., vol. 53, pp. 169-191, 2003.
22. Torralba, A: Contextual influences on saliency, Neurobiology of attention, Academic Press Inc, London (2005)
23. Voisine, N., Dasiopoulou, S., Mezaris, V., Spyrou, E., Athanasiadis, T., Kompatsiaris, I., Avrithis, Y., and Strintzis, M. G.: Knowledge-assisted video analysis using a genetic algorithm, In Proc. of 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)
24. W3C, *OWL*, `http://www.w3.org/2004/OWL/`
25. W3C, *RDF*, `http://www.w3.org/RDF/`
26. W3C, *RDF Reification*, `http://www.w3.org/TR/rdf-schema/#ch_reificationvocab`