High-Level Concept Detection based on Mid-level Semantic Information and Contextual Adaptation

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

Abstract

In this paper we propose the use of enhanced mid-level information, such as information obtained from the application of supervised or unsupervised learning methodologies on low-level characteristics, in order to improve semantic multimedia analysis. High-level, a priori *contextual knowledge about the semantic meaning of objects and their low-level visual descriptions are combined in an integrated approach that handles in a uniform way the gap between semantics and low-level features. Prior work on low-level feature extraction is extended and a region thesaurus containing all mid-level features is constructed using a hierarchical clustering method. A model vector that contains the distances from each mid-level element is formed and a neural network-based detector is trained for each semantic concept. Contextual adaptation improves the quality of the produced results, by utilizing fuzzy algebra, fuzzy sets and relations. The novelty of the presented work is the context-driven mid-level manipulation of region types, utilizing a domain-independent ontology infrastructure to handle the knowledge. Early experimental results are presented using data derived from the beach domain.*

1 Introduction

One of the most interesting problems in multimedia content analysis is detection of high-level concepts within multimedia documents. 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 well-known "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.

An important step for narrowing this gap is to automate the process of semantic feature extraction and annotation of multimedia content objects, by enhancing image and video classification with semantic characteristics. Combining both low-level descriptors computed automatically from raw multimedia content and semantics in the form of detection of semantic features in video sequences has been the ultimate task in current multimedia research efforts. Many approaches have been proposed, all sharing the common target and finally extracting high-level concepts from raw content. Among others, a region-based approach in content retrieval that uses Latent Semantic Analysis (LSA) is presented in [9], whereas in [4], a multi-modal machine learning technique is used in order to model semantic concepts within video sequences. A region-based approach using MPEG-7 visual features and ontological knowledge is presented in [11] and a lexicon-driven approach is introduced in [3]. Finally, a mean-shift algorithm is used in [8], in order to extract low-level concepts, after the image is clustered.

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. The use of mid-level information, such as information obtained from the application of supervised or unsupervised learning methodologies on low-level characteristics may be used to improve the results of traditional knowledge-assisted image analysis, based both on low-level *visual* and high-level *contextual* information. Initial image

analysis results are enhanced by the utilization of domain-independent, semantic knowledge in terms of concepts and relations between them. This mid-level information may often be described as: (i) a low-level description, but then again a level *above* the one extracted from the multimedia document, (ii) a high-level conceptual description, but then again a level *below* the ultimate goal and (iii) an in-between description, which can be described semantically, but does not express a high-level concept. This work focuses on this unified multimedia representation and processing, combining low- and high-level information in an efficient "mid-level" manner. We explore the interaction between intelligent local and global classification techniques, exploit both types of visual and contextual knowledge within the multimedia processing chain and investigate further potential content unification and adaptation approaches.

The structure of this paper is as follows: In Section 2, we briefly present the utilized fuzzy context knowledge representation, including some basic notation used throughout the paper. Section 3 is dedicated to the mid-level instantiation of an image's region types, whereas Section 4 describes the proposed contextual adaptation in terms of visual context algorithm optimization steps. Section 5 lists some preliminary experimental results derived solely from the *beach* domain and Section 6 concludes briefly our work.

2 Context Knowledge

As can be found in the literature, the term *context* may be interpreted and even defined in numerous ways, varying from the philosophical to the practical point of view, none of which is globally applicable or universal. Therefore, it is of great importance to establish a working representation for context, in order to benefit from and contribute to multimedia analysis and media adaptation. The problems to be addressed include how to represent and determine context, both in terms of low- and high-level visual characteristics, and how to use it to optimize the results of knowledge-assisted analysis. The latter are highly dependent on the domain an image belongs to and thus in many cases are not sufficient for the understanding of multimedia content. The lack of contextual information in the process [6] significantly hinders optimal analysis performance and together with similarities in low-level characteristics of various object types, results in a significant number of misinterpretations.

In this work we introduce a method for further adapting the results of low-level, descriptor-based multimedia analysis, utilizing the notion of mid-level region-types, based on a high-level contextual ontology. The latter is described as a set of concepts and semantic relations between concepts within a given universe. In general, we may decompose an ontology O into two parts, i.e.

1. the set C of all semantic concepts $c_i \in C, i = 1 \dots n$ and
2. the set R_{c_i,c_j} of all semantic relations amongst any two given concepts $c_i, c_j, j = 1 \dots n$

More formally:

$$O = \{C, R_{c_i,c_j}\}, \quad R_{c_i,c_j} : C \times C \rightarrow \{0, 1\} \qquad (1)$$

As indicated in our previous work [7], any kind of semantic relation may be represented by an ontology, however, herein we restrict it to a "fuzzified" ad-hoc context ontology. The latter is introduced in order to express in an optimal way the real-world relationships that exist between each domain's participating concepts. In order for this ontology type to be highly descriptive, it must contain a representative number of distinct and even diverse relations among concepts, so as to scatter information among them and thus describe their context in a rather meaningful way. The utilized relations need to be meaningfully combined, so as to provide a view of the knowledge that suffices for context definition and estimation. Additionally, since modelling of real-life information is usually governed by uncertainty and ambiguity, it is our belief that these relations must incorporate fuzziness in their definition. Thus, we extend a subset (Table 1) of the MPEG-7 semantic relations [2] suitable for image analysis and re-define them in a way to represent fuzziness, i.e. a degree of confidence is associated to each relation.

Consequently, a domain-specific, "fuzzified" version of a concept ontology may be described by O_F (eq. 2), where C represents again the set of all possible concepts, $F(R_{c_i,c_j}) = r_{c_i,c_j} : C \times C \rightarrow [0, 1]$ denotes a fuzzy ontological relation amongst two concepts c_i, c_j and R_{c_i,c_j} denotes the non-fuzzy semantic relation amongst the two concepts. The final combination of the MPEG-7 originating relations forms an RDF graph and constitutes the abstract contextual knowledge model to be used during the adaptation phase (Fig. 1).

$$O_F = \left\{C, r_{c_i,c_j}\right\}, \quad i, j = 1 \dots n \qquad (2)$$

The graph of the proposed model contains nodes (i.e. domain concepts) and edges (i.e. an appropriate combination[1] of contextual fuzzy relations between concepts). The degree of confidence of each edge represents fuzziness in the model. Non-existing edges imply non-existing relations (i.e. relations with zero confidence values are omitted). An existing edge between a given pair of concepts is produced

[1] In this knowledge representation the weakest part is the combination of different contextual fuzzy relations towards the generation of a practically exploitable knowledge view. At this point, combination of relations is performed by utilizing fuzzy algebra's operations, in general and the default *t*-norm, in particular.

Table 1. Fuzzy semantic relations utilized.

Name	Inverse	Symbol	Meaning	Example	
				a	b
Specialization	Generalization	$Sp(a,b)$	b is a specialization in the meaning of a	animal	boar
Part	PartOf	$P(a,b)$	b is a part of a	Australia	Sydney
Example	ExampleOf	$Ex(a,b)$	b is an example of a	leader	Ben
Instrument	InstrumentOf	$Ins(a,b)$	b is an instrument of or is employed by a	cut	knife
Location	LocationOf	$Loc(a,b)$	b is the location of a	tent	beach
Patient	PatientOf	$Pat(a,b)$	b is affected by or undergoes the action of a	give	peanut butter
Property	PropertyOf	$Pr(a,b)$	b is a property of a	banana	ripeness

based on the set of contextual fuzzy relations that are meaningful for the particular pair. For instance, the edge between concepts *rock* and *sand* is produced by the combination of relations *Location* and *Patient*, whereas the *water* and *sea* edge utilizes relations *Specialization*, *PartOf*, *Example*, *Instrument*, *Location* and *Patient*, in order to be constructed. Of course, each concept has a different probability to appear in the scene, thus a flat context model would not have been sufficient in this case; on the contrary, concepts are related to each other, implying that the graph relations used are in fact transitive.

Describing the accompanying degree of confidence is carried out using RDF reification [12]. 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 "Jack is six feet tall" is an assertion of truth that commits the sender to the fact, whereas the reified statement "Kate reports that Jack is six feet tall" defers this commitment to Kate. In this way, statements may include fuzzy information (i.e. "Jack is six feet tall with a degree of confidence equal to 0.90"), without creating contradictions in reasoning, since a statement is being made about the original statement, which 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 an RDF triple such as: "*sand partOf beach*" and a degree of confidence of "*0.85*" for this statement, does obviously not entail, that *sand* will always be part of a *beach* scene.

3 Region Type Analysis

For the extraction of the low-level features of an image, there are generally two categories of approaches: to extract the desired descriptors *globally* (from the entire image) or *locally* (from regions of interest). In our approach, a color segmentation algorithm is first applied on a given image as a pre-processing step. The algorithm is a multiresolution implementation of the well-known RSST method [1] tuned to produce a coarse segmentation. This way, the produced segmentation can intuitively provide a qualitative description of the image.

To capture the visual features of each region of the segmented image in a standardized way, we choose to extract color and texture features. Since a set of dominant colors in an image or a region of interest has the ability to efficiently capture its color properties, we choose to extract the MPEG-7 Dominant Color Descriptor (DCD) [5] from each region of the segmented image. The MPEG-7 Homogeneous Texture Descriptor [5] was used to capture texture properties of each region. The energy deviations of the descriptors were discarded, in order to simplify the description. All the low-level visual descriptions of an image are normalized and merged into a unique vector. This vector will be referred to as *feature vector*. We should note here that for the concepts we aim to detect and are depicted in Table 3 only color and texture descriptors make sense, thus shape descriptors have been omitted.

In general, it is expected that images with similar semantic features should have similar low-level descriptions. To exploit this, clustering is performed on the descriptions of a small portion of the entire dataset, used as a training set. We chose the well-known hierarchical clustering approach as we can easily modify the size of the thesaurus this way. 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. For example, the concept *sand* can have many instances differing e.g. in the color of the sand or in its texture since the bigger the sand grains are, their texture becomes more "complicated". 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 concept (e.g. *sky*), since both concepts often share similar low-level features.

A *thesaurus* combines a list of every term in a given domain of knowledge and a set of related terms for each term in the list. In our approach, the constructed *Region Thesaurus* contains all *region types* that are encountered in the training set. These region types are defined as the centroids of the clusters and all the other members of the cluster are

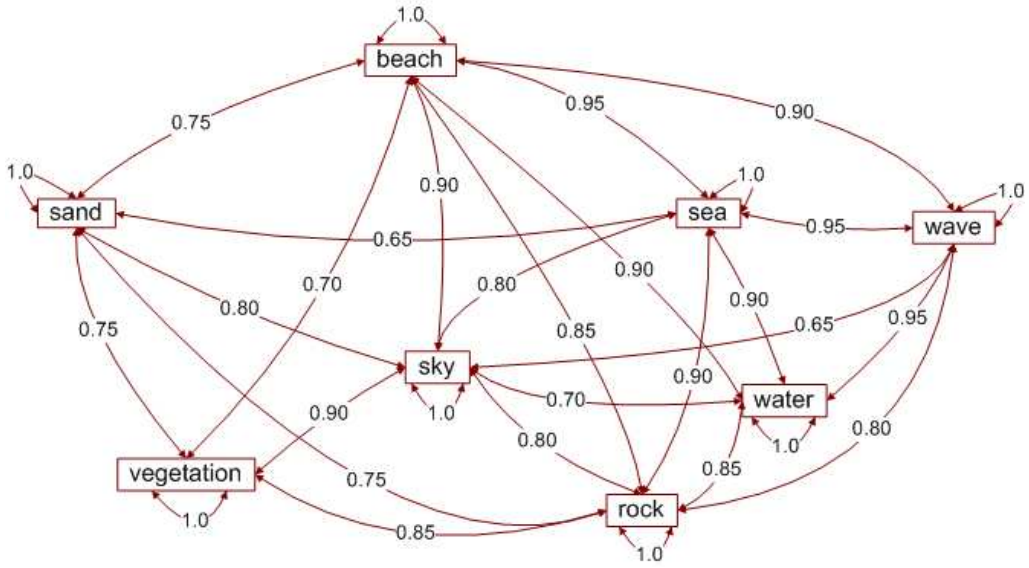


Figure 1. A fragment of the beach domain ontology. Concept *beach* is the "root" element.

their synonyms. The use of the thesaurus is to facilitate the association of the low-level features of the image with the corresponding high-level concepts. Thus, these region types are characterized as "mid-level" concepts, incorporating both low- and high-level information.

After the construction of the region thesaurus, a *model vector* is formed for each image. Its dimensionality is equal to the number of concepts constituting the thesaurus. The distance of a region to a region type is calculated as a linear combination of its average color and homogeneous texture distances, as in [10]. Having calculated the distance of each region of the image to all the region types 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 mid-level concept. More specifically, let: $d_i^1, d_i^2, ..., d_i^j, i = 1, 2, 3, 4$ and $j = N_C$, where N_C denotes the number of words of the thesaurus and d_i^j is the distance of the i-th region of the clustered image to the j-th region type. Then, the model vector D_m is the one depicted in equation 3.

$$D_m = [min\{d_i^1\}, min\{d_i^2\}, ..., min\{d_i^{N_C}\}], i = 1, 2, 3, 4 \quad (3)$$

For each semantic concept, a neural network-based classifier is then trained. Its input is a model vector and its output determines the confidence of the existence of the concept within the image.

4 Visual Context Adaptation

Once the contextual knowledge structure is finalized and the corresponding representation is implemented, a variation of the context-based confidence value readjustment algorithm introduced in [6] is applied on the output of the neural network-based classifier. The proposed contextualization approach empowers a post-processing step on top of the initial set of mid-level region types extracted. It provides an optimized re-estimation of the initial concepts' degrees of confidence for each region type and updates each model vector. In the process, it utilizes the high-level contextual knowledge from the constructed contextual ontology.

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 concept confidence values produced by region type analysis. In designing such an algorithm, contextual information residing in the aforementioned domain ontology is utilized. As already depicted, the notion of context is strongly related to the notion of ontologies since an ontology can be seen as an attempt towards modeling real-world (fuzzy) entities, and context determines the intended meaning of each concept, i.e. a concept used in different contexts may have different meanings. In this section, the 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 each concept's degree of membership is derived from direct and indirect relationships of the concept with other concepts, 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

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 degree of membership re-evaluation algorithm is as follows:

1. The considered domain imposes the use of a domain similarity (or dissimilarity) measure: $dnp \in [0, 1]$.
2. For each region type r consider a fuzzy set L_r with a degree of membership $\mu_r(c)$, containing the possible concepts' degrees of confidence.
3. For each concept c_i in the fuzzy set L_r with a degree of membership $\mu_r(c_i)$, obtain the particular contextual information in the form of its relations to the set of any other concepts: $\{r_{c_i,c_j} : c_i, c_j \in C, \quad i \neq j\}$.
4. Calculate the new degree of membership $\mu_r(c)$, taking into account each domain's similarity measure. In the case of multiple concept relations in the ontology, when relating concept c to more than the *root* concept (Fig. 1), an intermediate aggregation step should be applied for the estimation of $\mu_r(c)$ by considering the *context relevance* notion, cr_c: $cr_c = \max\{r_{c,c_1}, \ldots, r_{c,c_k}\}$, $c_1 \ldots c_k \in C$. We express the calculation of $\mu_r(c)$ with the recursive formula:

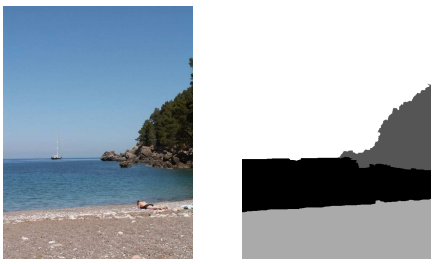
$$\mu_r^n(c) = \mu_r^{n-1}(c) - dnp(\mu_r^{n-1}(c) - cr_c) \quad (4)$$

where n denotes the iteration used. Equivalently, for an arbitrary iteration n:

$$\mu_r^n(c) = (1-dnp)^n \cdot \mu_r^0(c) + (1-(1-dnp)^n) \cdot cr_c \quad (5)$$

where $\mu_r^0(c)$ represents the initial degree of membership for concept c. The number of iteration steps is typically $n = 3 \ldots 5$.

5 Experimental Results



(a) Input image (b) Segmentation

Figure 2. 1^{st} **beach image example.**

In this section we provide early (i.e. derived solely from the *beach* domain) experimental results facilitating the proposed methodology. We carried out experiments utilizing 191 images and 7 beach-related concepts, acquired from personal collections and the Internet. A ground truth was manually constructed, consisting of a number of region types associated to a unique concept. We utilized 38 images (merely 20% of the dataset) as our clustering training set and after an extensive try-and-error process selected $dnp = 0.15$ as the optimal normalization parameter for the given domain. In the following we present indicative *beach* image examples (Fig. 2, 3, 4), where: (a) the original input image and (b) the segmentation output of the image are illustrated. In order to obtain the segmented output, we implemented segmentation using the RSST algorithm [1] and a distance threshold for termination in a meaningful number of regions.

(a) Input image (b) Segmentation

Figure 3. 2^{nd} **beach image example.**

(a) Input image (b) Segmentation

Figure 4. 3^{rd} **beach image example.**

The final concepts/degrees of membership pairs for the detected semantic concepts before and after visual context adaptation for the three *beach* image example are summarized in Table 2. Concepts that are not identified within the specific image sample have zero values.

Table 2. Confidence values per *beach* **image.**

Concepts	1st		2nd		3rd	
	before	after	before	after	before	after
sea	0.77	0.85	0.65	0.75	0.62	0.72
water	0.63	0.70	0.60	0.69	0.58	0.67
vegetation	0.35	0.43	0.35	0.40	0.62	0.72
sky	0.45	0.57	0.55	0.60	0.53	0.61
sand	0.69	0.75	0.45	0.56	0.52	0.60
rock	0.25	0.35	0.63	0.68	0.65	0.75
wave	0.00	0.00	0.25	0.34	0.20	0.27

Overall precision scores from the application of the proposed methodology to the entire dataset on a per concept basis are presented in Table 3. Each concept's row displays the scores before and after the context adaptation step.

Table 3. Overall precision scores per concept

Concepts	before	after	improvement (%)
sea	0.68	0.73	18.04%
water	0.34	0.36	15.99%
vegetation	0.61	0.68	20.79%
sky	0.68	0.77	20.79%
sand	0.62	0.66	16.54%
rock	0.61	0.65	16.64%
wave	0.49	0.54	18.26%
Overall	**0.59**	**0.75**	18.33%

6 Conclusions

Our current research efforts indicate 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 context. Amongst the core contribution of this work has been the implementation of a novel, mid-level visual context interpretation utilizing a fuzzy, ontology-based representation of knowledge. Early research results were presented, indicating a significant aid (i.e. 15.99%-20.79% per concept - 18.33% overall) of visual context adaptation to the mid-level analysis chain. It is the authors' belief that such enhanced mid-level information forms a novel solution to the semantic multimedia analysis task.

Future work will consist of extensions supporting more color and texture descriptors and additional concepts. Also, shape descriptors will be used in order to detect certain concepts that cannot be described solely by their color and texture features. Further extension of the methodologies presented in this paper will be exploited towards the development of more intelligent and self-confident image processing frameworks, forming an interesting perspective to knowledge-assisted analysis.

7 Acknowledgments

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] Y. Avrithis, A. Doulamis, N. Doulamis and S. Kollias, *A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases*, Computer Vision and Image Understanding, vol. 75 (1/2), pp. 3-24, 1999.

[2] A. B. Benitez, D. Zhong, S.-F. Chang and J. R. Smith, *MPEG-7 MDS Content Description Tools and Applications*, Lecture Notes in Computer Science, vol. 2124, pp. 41-52, 2001.

[3] D. C. K. Cees, G. M. Snoek, M. Worring and A. W. Smeulders, *Learned lexicon-driven interactive video retrieval*, In Proc. of 5th International Conference on Image and Video Retrieval (CIVR), Tempe, Arizona, USA, July 13-15, 2006.

[4] IBM, *Marvel: Multimedia analysis and retrieval system*. `http://mp7.watson.ibm.com/`

[5] B. Manjunath, J. Ohm, V. Vasudevan and A. Yamada, *Color and texture descriptors*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 703-715, 2001.

[6] Ph. Mylonas, Th. Athanasiadis and Y. Avrithis, *Improving image analysis using a contextual approach*, In Proc. of 7th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Seoul, Korea, April 19-21, 2006.

[7] Ph. Mylonas and Y. Avrithis, *Using Multiple Domain Visual Context in Image Analysis*, In Proc. of 8th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Santorini, Greece, June 6-8, 2007.

[8] B. Saux and G.Amato, *Image classifiers for scene analysis*, In Proc. of International Conference on Computer Vision and Graphics (ICCVG), Warsaw, Poland, September 22-24, 2004.

[9] F. Souvannavong, B. Merialdo and B. Huet, *Region-based video content indexing and retrieval*, In Proc. of 4th International Workshop on Content-Based Multimedia Indexing (CBMI), Riga, Latvia, June 22-24, 2005.

[10] E. Spyrou, H. LeBorgne, T. Mailis, E. Cooke, Y. Avrithis and N. O'Connor, *Fusing MPEG-7 visual descriptors for image classification*, In Proc. of International Conference on Artificial Neural Networks (ICANN), Warsaw, Poland, September 11-15, 2005.

[11] 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*, In Proc. of 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux, Switzerland, April 13-15, 2005.

[12] W3C, *RDF Reification*, `http://www.w3.org/TR/rdf-schema/#ch_reificationvocab`