Semantic Multimedia Analysis based on Region Types and Visual Context

Evaggelos Spyrou, Phivos Mylonas and Yannis Avrithis

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

Abstract. In this paper 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 the regions they consist of. More specifically, we construct a mid-level ontology, define its relations and integrate it in our knowledge modelling approach. In the past we have developed algorithms to address computationally efficient handling of visual context and extraction of mid-level characteristics and now we explain how these diverse algorithms and methodologies can be combined in order to approach a greater goal, that of semantic multimedia analysis. Early experimental results are presented using data derived from the *beach* domain.

1 Introduction

Although the well-known "semantic gap" [16] has been acknowledged for a long time, multimedia analysis approaches are still divided into two rather discrete categories; low-level multimedia analysis methods and tools, on the one hand (e.g. [13]) and high-level semantic annotation methods and tools, on the other (e.g. [20], [3]). It was only recently, that state-of-the-art multimedia analysis systems have started using semantic knowledge technologies, as the latter are defined by notions like ontologies [19] and whose advantages, when using them for the creation, manipulation and post-processing of multimedia metadata, are depicted in numerous research activities.

The main idea introduced herein relies on the integrated handling of concepts evident within multimedia content. 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 and previous multimedia research efforts. For instance, a region-based approach using MPEG-7 visual features and ontological knowledge is presented in [21] and a lexicon-driven approach is introduced in [4]. Among others, a region-based approach in content retrieval that uses Latent Semantic Analysis is presented in [17], whereas a mean-shift algorithm is used in [14], in order to extract low-level concepts, after the image is clustered.

In this work, our effort focuses on an integrated approach, offering unified and unsupervised manipulation of multimedia content. It acts complementary to the current state-of-the-art as it tackles both aforementioned challenges. Focusing on semantic analysis of multimedia, it contributes towards bridging the gap between the semantic and raw nature of multimedia content and tackles one of the most interesting problems in multimedia content analysis, i.e. detection of high-level concepts within multimedia documents, based on the semantics of each object, in terms of its visual context information. It proves that the use of mid-level information improves the results of traditional knowledge-assisted image analysis, based on both *visual* and *contextual* information. In the process, initial image analysis results are enhanced by the utilization of domain-independent, semantic knowledge in terms of region types and relations between them. In principle, mid-level information takes the form of an in-between description, which can be described semantically, but does not express high-level concepts and thus is included in a corresponding mid-level concept ontology.

The structure of this paper is as follows: In Section 2, we 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 subsection 3.2 describes a pre-processing contextualization step. Section 4 lists some preliminary experimental results and Section 5 briefly concludes our work.

2 Knowledge Representation

As can be found in the literature, the term *context* [9] may be interpreted and even defined in numerous ways, varying from the philosophical to the practical point of view [8]. However, since there is not a globally applicable aspect of context in the multimedia analysis chain, it is very important to establish a working representation for context, in order to benefit from and contribute to the proposed mid-level multimedia analysis. The problems to be addressed include how to represent and determine context, and how to use it to optimize the results of 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. In general, the lack of contextual information significantly hinders optimal analysis performance [12] and along with similarities in low-level features of various object types, results in a significant number of misinterpretations.

In this work we introduce a method for improving the results of low-level based multimedia analysis by using the notion of mid-level region types. The latter build an ontology, described by the set of region types and the relations between them. In general, we may decompose such an ontology O into two parts, the set T of all region types and the set R_{t_i,t_j} of all relations amongst any two given region types t_i, t_j. More formally:

$$O = \{T, R_{t_i,t_j}\}, \quad R_{t_i,t_j} : T \times T \rightarrow \{0,1\}, \quad i,j = 1 \ldots n \tag{1}$$

Any kind of 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 the concepts 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 region types, so as to exploit in an optimal manner 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 suitable for image analysis [2] and re-define them in a way to incorporate fuzziness, i.e. a degree of confidence is associated to each relation, and assist in discriminating between objects exhibiting similar visual characteristics.

Table 1. Contextual relations between region types.

Name	Inverse	Symbol	Meaning
Similar	Similar	$Sim(a,b)$	region type similarity based on the i-th descriptor
Accompanier	AccompanierOf	$Acc(a,b)$	coexistence of two region types
Part	PartOf	$P(a,b)$	a region type is part of another region type
Component	ComponentOf	$Comp(a,b)$	combines two region types with each other
Combination	-	$Comb(a,b)$	combines more than two region types

As in [7], a fuzzy relation on T is a function $r_{t_i,t_j} : T \times T \rightarrow [0,1]$ and its inverse relation is defined as ${r_{t_i,t_j}}^{-1} = r_{t_j,t_i}$. Based on the above relations, a domain-specific, "fuzzified" version of a region type ontology may be described by O_F:

$$O_F = \left\{T, r_{t_i,t_j}\right\}, \quad i,j = 1 \ldots n, \quad i \neq j \tag{2}$$

where T represents again the set of all possible region types,

$$F(R_{t_i,t_j}) = r_{t_i,t_j} : T \times T \rightarrow [0,1] \tag{3}$$

denotes a fuzzy ontological relation amongst two region types t_i, t_j and

$$R_{t_i,t_j} = \{Sim, Acc, P, Comp, Comb, A, B, R, L\} \tag{4}$$

denotes any possible non-fuzzy relation amongst two region types. The final, meaningful combination of relations

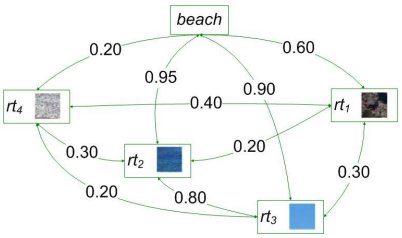
$$CR = (\bigcup_i {r_{t_i,t_j}}^{p_i}), \quad p_i \in \{-1,0,1\}, \quad i = 1 \ldots n \tag{5}$$

forms an RDF [22] graph and constitutes the abstract contextual knowledge model to be used during the analysis phase (Fig. 2). The value of p_i is determined by the semantics of each relation R_{t_i,t_j} used in the construction of CR. More specifically:

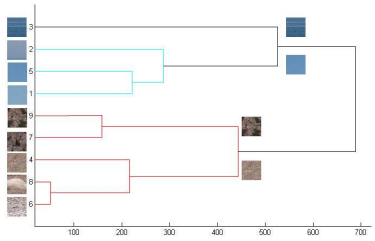
– $p_i = 1$, if the semantics of R_{t_i,t_j} imply it should be considered as is

– $p_i = -1$, if the semantics of R_{t_i,t_j} imply its inverse should be considered
– $p_i = 0$, if the semantics of R_{t_i,t_j} do not allow its participation in the construction of the combined relation CR.

The graph of the proposed model contains nodes (i.e. region types) and edges (i.e. contextual fuzzy relations between region types). 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). As each region type has a different probability to appear in the scene, a flat context model would not have been sufficient in this case.



(a) A fragment of the *beach* region type ontology.



(b) Region type selection using hierarchical clustering.

```
<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>
```

(c) RDF ontology fragment.

Fig. 1. From region thesaurus to mid-level ontology.

Describing the accompanying degree of confidence is carried out using RDF reification [23]. 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 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: "*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.

3 Semantic Multimedia Analysis

3.1 Region Type Analysis

The first step towards the construction of the mid-level ontology is the selection of the region types it will include. Thus, 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. The algorithm is a multiresolution implementation of the well-known RSST method [1], tuned to produce a coarse segmentation. We choose to tune the segmentation algorithm this way, since we want the produced segmentation to intuitively provide a qualitative description of the image. Then the segmentation results are used to define the candidate region types from each image.

These regions are then represented by a combination of their low-level visual features. Thereby, visual descriptors from the ISO/IEC MPEG-7 standard [5] are selected to capture a standardized description of their visual content. For representing the color features of the image regions, three MPEG-7 color descriptors are used: The *Color Layout Descriptor*, the *Scalable Color Descriptor* and the *Color Structure Descriptor* and for representing the texture features, the *Homogeneous Texture Descriptor* is selected. For the extraction of the aforementioned descriptors, the MPEG-7 eXperimentation Model (XM)[11] is used.

Given the entire set of regions, derived from the aforementioned segmentation process and their extracted low-level features, one can easily observe that those that belong to similar semantic concepts, also have similar low-level descriptions and also those images that contain the same high-level concepts are consisted of similar regions. As a natural sequence of this observation we apply a *hierarchical clustering* [6] algorithm on the regions of the given training set. 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*). A dendrogram describing the hierarchical clustering and the selection of the region types is depicted in figure 2(b). In this simplistic example an initial set of 9 candidate region types derived from 4 images is clustered and we choose to keep 4 region types to represent their visual content in terms of mid-level concepts.

Then we form a region *thesaurus*, in order to combine and manipulate effectively a list of every region type in a given domain of knowledge (e.g. *beach*) and a set of related regions (synonyms) for each region type. These region types can be characterized as "mid-level" concepts, incorporating both low- and high-level information. Then, we use the thesaurus to facilitate the association of the low-level features of the image with the corresponding high-level concepts in the following way: 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 the average descriptor distances, as in [18]. Having calculated the distance of each region of the image to all the region types of the constructed thesaurus, the model vector D_m that semantically describes the visual content of the image is formed by keeping the bigger confidence value for each mid-level concept and is depicted in equation 6.

$$D_m = [min\{d_i^1\}, min\{d_i^2\}, ..., min\{d_i^{N_C}\}], i = 1, 2, ..., numOfRegions \quad (6)$$

Where d_i^j is the confidence value that the i-th region of the image corresponds to the j-th region type, $numOfRegions$ is the number of the segmented image regions and N_C the size of the region thesaurus.

3.2 Visual Context Optimization

Once a model vector for an image is calculated, a modified version of the context-based confidence value readjustment algorithm [12] is applied, so as to satisfy the needs of the problem at hand. The latter forms the last pre-processing step of the analysis process and provides an optimized re-estimation of the initial regions' degrees of confidence to the selected region types. Consequently, it updates the values of each model vector, allowing an optimized training process of the classifier, thus achieving significantly optimized evaluation results.

In a more formal manner, the problem that this work attempts to address is summarized in the following statement: the visual context analysis algorithm readjusts in a meaningful way the initial region type confidence values produced by the previous step of region type analysis. In this section, the remaining problems to be addressed include how to meaningfully readjust the initial membership degrees and how to use visual context to influence the overall results of knowledge-assisted image analysis towards higher performance.

An estimation of the degree of membership of each mid-level region type is derived from direct and indirect relationships of the latter with other region types in the 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 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 modified degree of membership re-evaluation algorithm is now 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 t we may describe the fuzzy set L_t using the widely applied [10] sum notation $L_t = \sum_{i=1}^{|T|} t_i/w_i = \{t_1/w_1, t_2/w_2, \ldots, t_n/w_n\}$, where w_i describes the membership function: $w_i = \mu_{L_t}(t_i)$.

3. for each region type t_i in the fuzzy set L_t with a degree of membership w_i, obtain the particular contextual information in the form of its relations to the set of any other region types: $\{r_{t_i,t_j} : t_i, t_j \in T, \quad i \neq j\}$.
4. Calculate the new degree of membership w_i, taking into account each domain's similarity measure. In the case of multiple mid-level region type relations, relating 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 cr_{t_i}, introduced in [12].

We express the calculation of w_i with the recursive formula:

$$w_i^n = w_i^{n-1} - dnp(w_i^{n-1} - cr_{t_i}) \tag{7}$$

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

$$w_i^n = (1 - dnp)^n \cdot w_i^0 + (1 - (1 - dnp)^n) \cdot cr_{t_i} \tag{8}$$

where w_i^0 represents the initial degree of membership for region type t_i. Typical values for n reside between 3 and 5.

4 Experimental Results

In this section we provide some early experimental results facilitating the proposed approach. We carried out experiments utilizing 287 images and 25 region types derived from the *beach* domain, 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 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. For the sake of space we present an indicative *beach* image use case example (Fig. 2): (a) the original input image and (b) the segmentation output of the image, where we consider a simpler region thesaurus, consisting of only 4 region types. The original model vector deriving from the comparison of the image regions to the region types of the region thesaurus is:

$$\mathbf{MV}_{before} = \begin{bmatrix} 0.723 & 0.220 & 0.753 & 0.364 \end{bmatrix} \tag{9}$$

Since we can observe that the given image consists of *sky* and *sea*, we should expect that region types that correspond to these semantic concepts should have larger values. The case here is that sea has a quite different color than the region type of the thesaurus that has occurred from a *sea* region. This color appears even similar to *rock* regions. We would like to increase this confidence to the region type and also decrease the confidence that corresponds to a *rock* region (2nd and 4th constituent of the model vector). After we apply the algorithm described in section 3.2, the model vector becomes:

$$\mathbf{MV}_{before} = \begin{bmatrix} 0.778 & 0.452 & 0.800 & 0.338 \end{bmatrix} \tag{10}$$

(a) Input image (b) Segmentation

Fig. 2. Indicative *beach* image example.

In order to provide a first measure of overall evaluation for the proposed technique, we further present precision scores from its application to the entire dataset on a per high-level concept basis (i.e., after the final classification step is applied on the optimized model vector). Evaluation results for 6 high-level *beach* concepts are presented in Table 2. Each concept's row displays the precision value before and after the use of context.

Table 2. Overall precision scores per high-level *beach* concept

Concepts	**before**	**after**	**%**
sea	0.72	0.77	6.85%
water	0.36	0.38	5.56%
sky	0.85	0.97	11.69%
sand	0.70	0.74	6.06%
rock	0.68	0.73	6.15%
vegetation	0.43	0.48	10.87%
Overall	**0,62**	**0,68**	7.86%

5 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 high-level concept detection optimization (i.e. 5.56%-11.69% per concept - 7.86% overall) over the entire dataset utilized. Although the improvement is not impressive, we believe that minor enhancements on the implemented model should boost further its performance.

References

1. Avrithis, Y., Doulamis, A., Doulamis, N., Kollias, S.: A stochastic framework for optimal key frame extraction from mpeg video databases. (1999)
2. Benitez, A. B., Zhong, D., Chang, S.-F., Smith, J. R., *MPEG-7 MDS Content Description Tools and Applications*, Lecture Notes in Computer Science, 2001.
3. Benitez, A. B., and Chang, S.-F., *Image Classification Using Multimedia Knowledge Networks*, Proceedings of the IEEE Int. Conf. on Image Processing (ICIP'03), Barcelona, Spain, 2003.
4. Cees, D. C. K., Snoek, G.M., Worring, M., and Smeulders, A. W., *Learned lexicon-driven interactive video retrieval*, 2006.
5. Chang, S.F., Sikora, T., Puri, A.: Overview of the mpeg-7 standard. IEEE trans. on Circuits and Systems for Video Technology **11**(6) (2001) 688–695
6. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2 edn. Wiley Interscience (2000)
7. Klir, G., and Yuan, B., *Fuzzy Sets and Fuzzy Logic, Theory and Applications*, New Jersey, Prentice Hall, 1995.
8. Lewis, D., *Index, Context, and Content*, in Kanger, S. and Ohman, S. (Eds.), Philosophy and Grammar, Reidel Publishing, 1980.
9. McCarthy, J., *Notes on Formalizing Context*, in Proc. of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993), Chambéry, France, August-September 1993, pp. 81-98.
10. Miyamoto, S., *Fuzzy Sets in Information Retrieval and Cluster Analysis*, Kluwer Academic Publishers, Dordrecht / Boston / London, 1990.
11. MPEG-7: Visual experimentation model (xm) version 10.0. ISO/IEC/ JTC1/SC29/WG11, Doc. N4062 (2001)
12. Mylonas, P., Athanasiadis, T., & 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.
13. Rapantzikos, K., Avrithis, Y., Kollias, S., *On the use of spatiotemporal visual attention for video classification*, Proceedings of International Workshop on Very Low Bitrate Video Coding (VLBV '05), Sardinia, Italy, September 2005.
14. Saux, B., and Amato, G., *Image classifiers for scene analysis*, In Proc. of International Conference on Computer Vision and Graphics, 2004.
15. Skiadopoulos, S., Giannoukos, C., Sarkas, N., Vassiliadis, P., Sellis,T., Koubarakis, M., *2D topological and direction relations in the world of minimum bounding circles*, IEEE trans. on Knowledge and Data Engineering, Vol. 17(12), pp. 16101623, 2005.
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 Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 1349-1380, 2000.
17. 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.
18. Spyrou, E., LeBorgne, H., Mailis, T., Cooke, E., Avrithis, Y., O'Connor, N., *Fusing mpeg-7 visual descriptors for image classification*, In: International Conference on Artificial Neural Networks (ICANN), 2005.
19. Staab, S., and Studer, R., *Handbook on Ontologies*, International Handbooks on Information Systems, Springer-Verlag, Heidelberg, 2004.

20. Tsechpenakis, G., Akrivas, G., Andreou, G., Stamou, G., and Kollias, S., *Knowledge-Assisted Video Analysis and Object Detection*, Proceedings of European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems (Eunite02), Albufeira, Portugal, September 2002.
21. 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), April 13-15, 2005.
22. W3C, RDF, http://www.w3.org/RDF/
23. W3C, *RDF Reification*, `http://www.w3.org/TR/rdf-schema/#ch_reificationvocab`