An Ontology Framework For Knowledge-Assisted Semantic Video Analysis and Annotation

S. Dasiopoulou[1,2], V. K. Papastathis[2], V. Mezaris[1,2], I. Kompatsiaris[2] and M. G. Strintzis[1,2] *

[1] Information Processing Laboratory, Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
[2] Informatics and Telematics Institute (ITI)/ Centre for Research and Technology Hellas (CERTH), 1st Km Thermi-Panorama Rd, Thessaloniki 57001, Greece
email: strintzi@iti.gr

Abstract. An approach for knowledge assisted semantic analysis and annotation of video content, based on an ontology infrastructure is presented. Semantic concepts in the context of the examined domain are defined in an ontology, enriched with qualitative attributes of the semantic objects (e.g. color homogeneity), multimedia processing methods (color clustering, respectively), and numerical data or low-level features generated via training (e.g. color models, also defined in the ontology). Semantic Web technologies are used for knowledge representation in RDF/RDFS language. Rules in F-logic are defined to describe how tools for multimedia analysis should be applied according to different object attributes and low-level features, aiming at the detection of video objects corresponding to the semantic concepts defined in the ontology. This supports flexible and managed execution of various application and domain independent multimedia analysis tasks. This ontology-based approach provides the means of generating semantic metadata and as a consequence Semantic Web services and applications have a greater chance of discovering and exploiting the information and knowledge in multimedia data. The proposed approach is demonstrated in the Formula One and Football domains and shows promising results.

1 Introduction

As a result of recent progress in hardware and telecommunication technologies, multimedia has become a major source of content on the World Wide Web, used in a wide range of applications in areas such as content production and distribution, telemedicine, digital libraries, distance learning, tourism, distributed CAD/CAM, GIS, etc. The usefulness of all these applications is largely determined by their accessibility and portability and as such, multimedia data sets

* This work was supported by the European Commission under contracts FP6-001765 aceMedia and FP6-507482 KnowledgeWeb.

present a great challenge in terms of storing, querying, indexing and retrieval. In addition, the rapid increase of the available amount of multimedia information has revealed an urgent need for developing intelligent methods for understanding and managing the conveyed information. To face such challenges developing faster hardware or more sophisticated algorithms has become insufficient. Rather, a deeper understanding of the information at the semantic level is required [1]. This results in a growing demand for efficient methods for extracting semantic information from such content, since this is the key enabling factor for the management and exploitation of multimedia content.

Although new multimedia standards, such as MPEG-4 and MPEG-7 [2], provide the needed functionalities in order to manipulate and transmit objects and metadata, their extraction, and that most importantly at a semantic level, is out of the scope of the standards and is left to the content developer. Extraction of features and object recognition are important phases in developing general purpose multimedia database management systems [3]. Significant results have been reported in the literature for the last two decades, with successful implementation of several prototypes [4]. However, the lack of precise models and formats for object and system representation and the high complexity of multimedia processing algorithms make the development of fully automatic semantic multimedia analysis and management systems a challenging task.

This is due to the difficulty, often mentioned as the *semantic gap*, in capturing concepts mapped into a set of image and/or spatiotemporal features that can be automatically extracted from video data without human intervention [5]. The use of domain knowledge is probably the only way by which higher level semantics can be incorporated into techniques that capture the semantics through automatic parsing. Such techniques are turning to knowledge management approaches, including Semantic Web technologies to solve this problem [6]. A priori knowledge representation models are used as a knowledge base that assists semantic-based classification and clustering [7, 8]. In [9] and [10] automatic associations between media content and formal conceptualizations are performed based on the similarity of visual features extracted from a set of pre-annotated media objects and the examined media objects. In [11], semantic entities, in the context of the MPEG-7 standard, are used for knowledge-assisted video analysis and object detection, thus allowing for semantic level indexing. In [12], the problem of bridging the gap between low-level representation and high-level semantics is formulated as a probabilistic pattern recognition problem. In [13], an object ontology, coupled with a relevance feedback mechanism, is introduced to facilitate the mapping of low-level to high-level features and allow the definition of relationships between pieces of multimedia information.

In this paper, an approach for knowledge assisted semantic content analysis and annotation, based on a multimedia ontology infrastructure, is presented. Content-based analysis of multimedia requires methods which will automatically segment video sequences and key frames into image areas corresponding to salient objects, track these objects in time, and provide a flexible framework for object recognition, indexing, retrieval and for further analysis of their relative

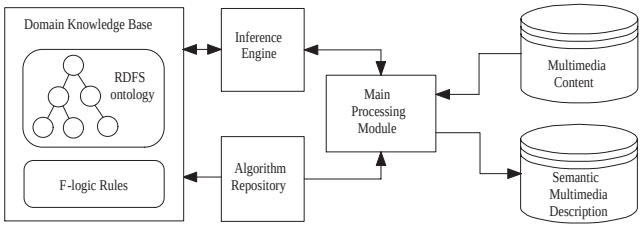

Fig. 1. Overall system architecture.

motion and interactions. This problem can be viewed as relating symbolic terms to visual information by utilizing syntactic and semantic structure in a manner related to approaches in speech and language processing [14]. In the proposed approach, semantic and low-level attributes of the objects to be detected in combination with appropriately defined rules determine the set of algorithms and parameters required for the objects detection. Semantic concepts within the context of the examined domain are defined in an ontology, enriched with qualitative attributes of the semantic objects, multimedia processing methods, and numerical data or low-level features generated via training. Semantic Web technologies are used for knowledge representation in RDF/RDFS language. Processing may then be performed by using the necessary processing tools and by relating high-level symbolic representations to extracted features in the signal (image and temporal feature) domain. F-logic rules are defined to describe how tools for multimedia analysis should be applied according to different object attributes and low-level features, aiming at the detection of video objects corresponding to the semantic concepts defined in the ontology. The proposed approach, by exploiting the domain knowledge modelled in the ontology, enables the recognition of the underlying semantics of the examined video, providing a first level semantic annotation. The general system architecture is shown in Fig. 1

Following this approach, the multimedia analysis and annotation process largely depends on the knowledge base of the system and as a result the method can easily be applied to different domains provided that the knowledge base is enriched with the respective domain ontology. Extending the knowledge base with spatial and temporal objects interrelations would be an important step towards the detection of semantically important events for the particular domain, achieving thus a finer, high-level semantic annotation. In addition, the ontology-based approach also ensures that semantic web services and applications have a greater chance of discovering and exploiting the information and knowledge in multimedia data.

The remainder of the paper is organized as follows: section 2 a detailed description of the ontology and rules developed is given, while in section 3, its application to the Formula One domain is described. Experimental results are presented in section 4. Finally, conclusions are drawn in section 5.

2 Multimedia Analysis Ontology Development and Rule Construction

In order to realize the knowledge-assisted multimedia content semantic analysis and annotation technique explained in the previous section, an analysis and a domain ontology are constructed. The *multimedia analysis ontology* is used to support the detection process of the corresponding domain specific objects. Knowledge about the domain under discourse is also represented in the form of an ontology, namely the *domain specific ontology*. The domain-independent, primitive classes comprising the analysis ontology serve as attachment points allowing the integration of the two ontologies. Practically, each domain ontology comprises a specific instantiation of the multimedia analysis ontology providing the corresponding color models, restrictions e.t.c as will be demonstrated in more detail in section 3.

Object detection in general considers the exploitation of objects characteristic features in order to apply the most appropriate detection steps for the analysis process in the form of algorithms and numerical data generated off-line by training (e.g. color models). Consequently, the development of the proposed analysis ontology deals with the following concepts (RDFS classes) and their corresponding properties, as illustrated in Fig. 2:

- Class **Object**: the superclass of all video objects to be detected through the analysis process. Each object instance is related to appropriate feature instances by the **hasFeature** property and to one or more other objects through a set of appropriately defined spatial properties.
- Class **Feature**: the superclass of multimedia low-level features associated with each object.
- Class **Feature Parameter** which denotes the actual qualitative descriptions of each corresponding feature. It is subclassed according to the defined features, i.e. to **Connectivity Feature Parameter**, **Homogeneity Feature Parameter** e.t.c.
- Class **Limit**: it is subclassed to **Minimum** and **Maximum** and allows the definition of value restrictions to the various feature parameters.
- The **Color Model** and **Color Component** classes are used for the representation of the color information, encoded in the form of the Y, Cb, Cr components of the MPEG color space.
- Class **Distribution** and **Distribution Parameter** represent information regarding the defined **Feature Parameter** models.
- Class **Motion Norm**: used to represent information regarding the object motion.
- Class **Algorithm**: the superclass of the available processing algorithms (A_1, A_2,...,A_n) to be used during the analysis procedure. This class is linked to the **FeatureParameter** class through the *usesFeatureParameter* property in order to represent the potential argument list for each algorithm.
- Class **Detection**: used to model the detection process, which in our framework consists of two stages. The **CandidateRegionSelection** involves finding a set of regions which are potential matches for the object to be detected,

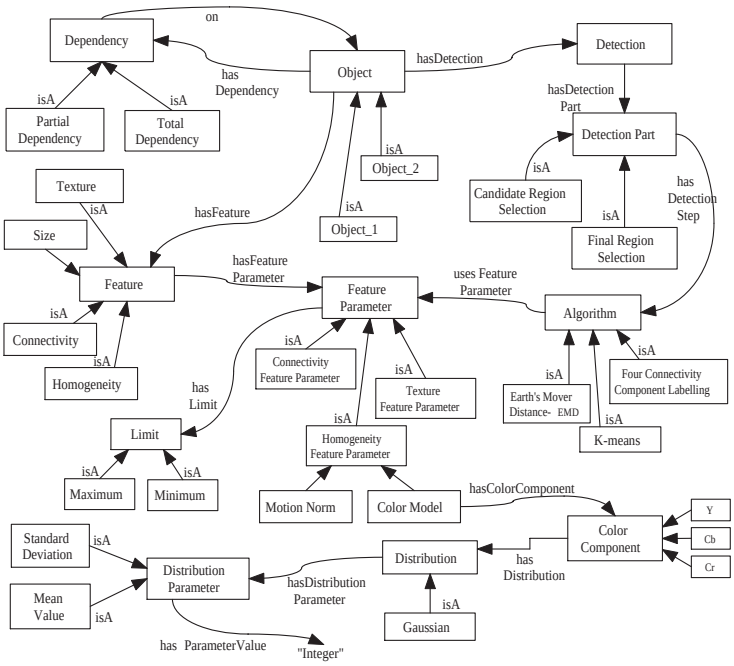

Fig. 2. Multimedia analysis ontology.

while **FinalRegionSelection** leads to the selection of only one region that best matches the criteria predefined for this object (e.g. size specifications).

– Class **Dependency**: this concept addresses the possibility that the detection of one object may depend on the detection of another, due to possible spatial or temporal interrelations between the two objects. For example in the Formula One domain, the detection of the car could be assisted and improved if the more dominant and characteristic region of road is detected first. In order to differentiate between the case where the detection of object O_1 requires the detection of the candidate regions of object O_2 and the case where the entire final region of object O_2 is required, **PartialDependency** and **TotalDependency** are introduced.

As mentioned before, the choice of algorithms employed for the detection of each object is directly dependent on its available characteristic features. This association is determined by a set of properly defined rules represented in F-logic. F-logic is a language that enables both ontology representation and reasoning about concepts, relations and instances [15, 16].

The rules required for the presented approach are: rules to define the mapping between algorithms and features (which implicitly define the object detection steps), rules to determine algorithms input parameters, if any, and rules to deal

with object interdependencies as explained above. The rules defined for each category have the following form:

- "IF an object O has features $F_1 \bigcap F_2 \bigcap \dots F_n$ as part of its qualitative description THEN algorithm A_1 is a step for the detection of O."
- "IF an object O has feature F AND O has algorithm A as detection step AND A uses feature F THEN A has as input the parameter values of F."
- "IF an object O_1 has partial dependency on object O_2 AND object O_2 has as **CandidateRegionSelection** part the set $S = \{A_1, A_2, \dots, A_m\}$ THEN execute the set of algorithms included in S before proceeding with the detection of O_1."
- IF an object O_1 is totally dependent on object O_2 THEN execute all detection steps for O_2 before proceeding with the execution of O_1 detection."

In order for the described multimedia analysis ontology to be applied, a domain specific ontology is needed. This ontology provides the vocabulary and background knowledge of the domain i.e. the semantically significant concepts and the properties among them. In the context of video understanding it maps to the important objects, their qualitative and quantitative attributes and their interrelations.

3 Domain Knowledge Ontology

As previously mentioned, for the demonstration of the proposed approach the Formula One and Football domains were used. The detection of semantically significant objects, such as the road area and the cars in racing video for example, is an important step towards understanding and extracting the semantics of a temporal segment of the video by efficiently modelling the events captured in it. The set of features associated with each object comprises their definitions in terms of low-level features as used in the context of video analysis. The selection of the attributes to be included is based on their ability to act as distinctive features for the analysis to follow, i.e. the differences in their definitions indicate the different processing methods that should be employed for their identification. As a consequence, the definitions used for the Formula One domain are:

- **Car**: a motion homogeneous (i.e. comprising elementary parts characterized by similar motion), fully connected region whose motion norm must be above a minimum value and whose size can not exceed a predefined maximum value.
- **Road**: a color homogeneous, fully connected region, whose size has to exceed a predefined minimum value and additionally to be the largest such region in the video.
- **Grass**: a color homogeneous, partly connected region with the requirement that each of its components has a minimum predefined size.
- **Sand**: a color homogeneous, partly connected region with the requirement that each of its components has a size exceeding a predefined minimum.

In a similar fashion, the corresponding definitions for the Football domain include the concepts **Player**, **Field** and **Spectators** and their respective visual descriptions. As can be seen, the developed domain ontologies focus mainly on the representation of the object attributes and positional relations and in the current version does not include event definitions. For the same object, multiple instances of the **Color Model** class are supported, since the use of more than one color models for a single object may be advantageous in some cases.

3.1 Compressed-domain Video Processing and Rules

The proposed knowledge-based approach is applied to MPEG-2 compressed streams. The information used by the proposed algorithms is extracted from MPEG sequences during the decoding process. Specifically, the extracted color information is restricted to the DC coefficients of the macroblocks of I-frames, corresponding to the Y, Cb and Cr components of the MPEG color space. Additionally, motion vectors are extracted for the P-frames and are used for generating motion information for the I-frames via interpolation. P-frame motion vectors are also necessary for the temporal tracking in P-frames, of the objects detected in the I-frames [17].

The procedure for detecting the desired objects starts by performing a set of initial clusterings, using up to eight dominant colors in each frame to initialize a K-means algorithm. ¿From the resulting mask, which contains a number of non-connected color-homogeneous regions, the non-connected semantic objects can be identified by color-model based selection. The application of a four connectivity component labelling algorithm results in a new mask featuring connected color-homogenous components. The color-model-based selection of an area corresponding to a color-homogeneous semantic object is performed using a suitable mask and the Earth Movers Distance (EMD). EMD computes the distance between two distributions represented as signatures and is defined as the minimum amount of work needed to change one signature into the other. Additional requirements as imposed by the models represented in the ontology, are checked to lead to the desired object detection. For motion-homogeneous objects a similar process is followed. At first, a mask containing motion-homogeneous regions is generated. Subsequently, the model- based selection depends on the information contained in the ontology (e.g. size restrictions, motion requirements).

The construction of the domain specific rules derives directly from the aforementioned video processing methodology. For example, since color clustering is the first step for the detection of any of the three objects, a rule of the first category without any feature matching condition is used to add the k-means algorithm as the first detection step to all objects. A set of different algorithms could have been used as long as the respective instantiations are defined.

4 Experimental results

The proposed approach was tested in two different domains: the Formula One and the Football domain. In both cases, the exploitation of the knowledge con-

tained in the respective system ontology and the associated rules resulted to the application of the appropriate analysis algorithms using suitable parameter values, for the detection of the domain specific objects. For ontology creation the OntoEdit ontology engineering environment [18] was used, having F-logic as the output language. A variety of MPEG-2 videos of 720×576 pixels were used for testing and evaluation of the knowledge assisted semantic annotation system.

For the Formula One domain our approach was tested on a one-hour video. As was discussed in section 3, four objects were defined for this domain. For those objects whose homogeneity attribute is described in the ontology by the **Color Homogeneity** class, the corresponding color models were extracted from a training set of approximately 5 minutes of manually annotated Formula One video. Since we assume the model to be a Gaussian distribution for each one of the three components of the color space, the color models were calculated from the annotated regions of the training set accordingly. Results for the Formula One domain are presented both in terms of sample segmentation masks showing the different objects detected in the corresponding frames (Fig. 3) as well as numerical evaluation of the results over a ten-minute segment of the test set (Table. 1). For the Football domain, the proposed semantic analysis and annotation framework was tested on a half-hour video, following a procedure similar to the one illustrated for the Formula One domain. Segmentation masks for this domain are shown in Fig. 4, while numerical evaluation of the results over a ten-minute segment of the test set for this domain are given in Table. 1.

For the numerical evaluation, the semantic objects appearing on each I-frame were manually annotated and compared with the results produced by the proposed system. It is important to note that the regions depicted in the generated segmentation masks correspond to semantic concepts and this mapping is defined according to the domain specific knowledge (i.e. object models) provided in the ontology.

5 Conclusions

In this paper we have presented an ontology-based approach for knowledge assisted domain-specific semantic video analysis. Knowledge involves qualitative object attributes, quantitative low-level features generated by training as well as multimedia processing methods. The proposed approach aims at formulating a domain specific analysis model with the additional information provided by rules, appropriately defined to address the inherent algorithmic issues.

Future work includes the enhancement of the domain ontology with more complex model representations, including spatial and temporal relationships, and the definition of semantically important events in the domain of discourse. Further exploration of low-level multimedia features (e.g. use of the MPEG-7 standardized descriptors) is expected to lead to more accurate and thus efficient representations of semantic content. The above mentioned enhancements will allow more meaningful reasoning, thus improving the efficiency of multimedia content understanding. Another possibility under consideration is the use of a

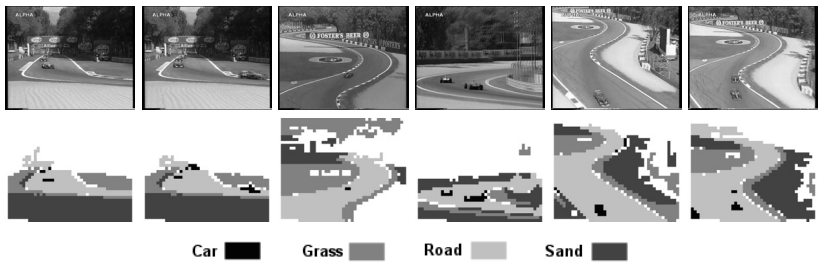

Fig. 3. Results of road, car, grass and sand detection for Formula One video. Macroblocks identified as belonging to no one of these four classes are shown in white.

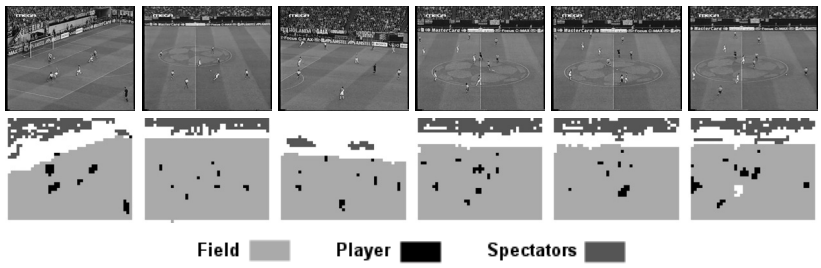

Fig. 4. Results of field, player, and spectators detection for Football video. Macroblocks identified as belonging to no one of these three classes are shown in white.

more expressive language, e.g. OWL, in order to capture a more realistic model of the specific domain semantics.

References

1. S.-F. Chang. The holy grail of content-based media analysis. *IEEE Multimedia*, 9(2):6–10, Apr.-Jun. 2002.
2. S.-F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):688–695, June 2001.
3. A. Yoshitaka and T. Ichikawa. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):81–93, Jan/Feb 1999.
4. P. Salembier and F. Marques. Region-Based Representations of Image and Video: Segmentation Tools for Multimedia Services. *IEEE Trans. Circuits and Systems for Video Technology*, 9(8):1147–1169, December 1999.
5. W. Al-Khatib, Y.F. Day, A. Ghafoor, and P.B. Berra. Semantic modeling and knowledge representation in multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):64–80, Jan/Feb 1999.
6. S. Little J. Hunter, J. Drennan. Realizing the hydrogen economy through semantic web technologies. *IEEE Intelligent Systems Journal - Special Issue on eScience*, 19:40–47, 2004.

Table 1. Semantic analysis results for the Formula One and Football domains

Object	correct detections	false detections	missed
Road	97%	2%	1%
Grass	87%	8%	5%
Sand	87%	9%	4%
Car	66%	27%	7%
Field	100%	0%	0%
Player	76%	5%	19%
Spectators	70%	2%	28%

7. A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa. Knowledge-assisted content-based retrieval for multimedia databases. *IEEE Multimedia*, 1(4):12–21, Winter 1994.
8. V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. An Ontology Approach to Object-based Image Retrieval. In *Proc. IEEE Int. Conf. on Image Processing (ICIP03)*, Barcelona, Spain, Sept. 2003.
9. A.B. Benitez and S.F. Chang. Image Classification Using Multimedia Knowledge Networks. In *Proc. IEEE Int. Conf. on Image Processing (ICIP03)*, Barcelona, Spain, Sept. 2003.
10. R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal. Automating the linking of content and concept. In *Proc. ACM Int. Multimedia Conf. and Exhibition (ACM MM-2000)*, Oct./Nov. 2000.
11. G. Tsechpenakis, G. Akrivas, G. Andreou, G. Stamou, and S.D. Kollias. Knowledge-Assisted Video Analysis and Object Detection. In *Proc. European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (Eunite02)*, Algarve, Portugal, September 2002.
12. M. Ramesh Naphade, I.V. Kozintsev, and T.S. Huang. A factor graph framework for semantic video indexing. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(1):40–52, Jan. 2002.
13. I. Kompatsiaris, V. Mezaris, and M. G. Strintzis. *Multimedia content indexing and retrieval using an object ontology.* Multimedia Content and Semantic Web - Methods, Standards and Tools, Editor G.Stamou, Wiley, New York, NY, 2004.
14. C. Town and D. Sinclair. A self-referential perceptual inference framework for video interpretation. In *Proceedings of the International Conference on Vision Systems*, volume 2626, pages 54–67, 2003.
15. J. Angele and G. Lausen. *Ontologies in F-logic.* International Handbooks on Information Systems. Springer, 2004.
16. M. Kifer, G. Lausen, and J. Wu. Logical foundations of object-oriented and frame-based languages. *J. ACM*, 42(4):741–843, 1995.
17. V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis. Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(5):606–621, May 2004.
18. Y. Sure, J. Angele, and S. Staab. *OntoEdit: Guiding Ontology Development by Methodology and Inferencing.* Springer-Verlag, 2002.