

# Contribution to Multimedia Ontology Framework Requirements

Yannis Avrithis

(with input from Stephan Bloehdorn & Carsten Saathoff)

10 December 2005

## 1. Introduction

This document contains an edited text from a part of a paper published in ESWC'05 [1]. It describes the requirements gathered during the design of a knowledge infrastructure in the aceMedia project, containing a core ontology, a visual descriptor ontology, a multimedia structure ontology, domain ontologies and a multimedia ontology annotation tool, M-OntoMat-Annotizer. The text comprises of two sections, namely (i) requirements for multimedia content representation, analysis and reasoning, and (ii) requirements for semantic multimedia content annotation.

## 2. Requirements for Multimedia Analysis & Annotation

The challenge in building an ontology framework for multimedia analysis and annotation arises from the fact that multimedia data comes in two separate though intertwined layers which need to be appropriately linked. On the one hand, *multimedia layer* deals with the semantics of properties and phenomena related to the presentation of content within the media-data itself, e.g. its spatio-temporal structure or visual features for analysis and is typically hard to understand for people who aren't trained in multimedia analysis. The *content layer*, on the other hand, deals with the semantics of the actual content contained in the media data as it is perceived by the human media consumer. The ontology framework should model the multimedia layer data so as to support extraction and inferencing of content layer data. This section analyzes a number of requirements for an integrated knowledge infrastructure and annotation environment for multimedia description, analysis and reasoning. To illustrate some of the requirements, we first present a simple scenario, with focus on direct exploitation:

*Multimedia content manager Samantha is working on a project on historic tennis matches. She has to prepare both the metadata infrastructure and the multimedia content. Samantha loads existing general sports ontologies into M-OntoMat-Annotizer and extends them by adding missing concepts of major interest. Next, she points M-OntoMat-Annotizer to images from the project, which are loaded and depicted in the user interface. One after another, Samantha then selects different objects in the images and drags them to the corresponding concepts in the domain ontology. The system extracts visual descriptors for these concepts and stores them in the application memory. Thus, Samantha has used M-OntoMat-Annotizer to describe the tennis domain and to describe the shape and the texture of tennis balls, rackets, nets, or courts.*

Note that this simple scenario has focused on simply providing conceptual information and the corresponding visual characteristics to the knowledge base which might be exploited directly. However, at the same time, the generated data would serve as a valuable a-priori source of information for multimedia analysis tools. These tools would use the descriptions in order to learn how to tag and relate segments of images and video keyframes with the domain ontology concepts.

## **2.1 Requirements for Multimedia Analysis**

In order to support linking between low level visual information and the higher level content domain, the above example scenario implicitly requires a suitable knowledge infrastructure tailored to multimedia descriptions:

*Low-Level Description Representation.* In order to represent the visual characteristics associated with a concept, one has to employ several different visual properties, depending on the concept at hand. For instance, in the tennis domain as was described in the scenario, the tennis ball might be described using its shape (“round”), color (“white”), or, in some cases of video sequences, motion. Similarly, a tennis racket has a distinctive and easily recognizable shape.

*Support for Multiple Visual Descriptions.* Visual characteristics of domain concepts can not be described using one single instance of the visual descriptors in question. For example, while the net of a tennis racket might be described in terms of its texture only once, its shape heavily depends on the viewing angle and occlusions (e.g. by the player in front of the net). The required conceptualization thus has to provide means for *multiple* prototypical descriptions of a domain concept.

*Spatiotemporal Relation Representation.* Simple visual properties may be used to model simple concepts. In some cases, however, decomposition of more complex concepts in terms of simpler object parts is desirable. A tennis player, for instance, is difficult to describe using a single shape, motion or texture description; it is more efficient to model and describe the characteristic parts (head, tennis shirt, racket) in terms of their visual properties first, and then define the human player as a spatial configuration of these parts. In other domains like beach holidays, it is more appropriate to describe the entire scene of a picture in terms of its color layout, depicting e.g. the sky at the top, the sand in the middle and the sea at the bottom. In such cases, modelling of spatiotemporal and partonomic relations is required apart from simple visual properties.

*Multimedia Structure Representation.* The result of the annotation (or content analysis in a next step) should be able to express the structure of a multimedia document itself, depending on the type of document, e.g. image, video, audio, or multimedia presentation. For instance, an image is usually decomposed into a number of still regions corresponding to some semantic objects of interest, while a video clip may be decomposed into shots, each of which into associated

moving regions. A hierarchical structure of multimedia segments is thus needed in order to capture all possible types of spatiotemporal or media decompositions and relations.

*Alignment with MPEG-7 Standard.* The MPEG-7 multimedia content description standard already provides tools for representing fragments of the above information. For instance, the MPEG-7 Visual Part [2] supports color (e.g. dominant colors, color layout), texture, shape (e.g. region/contour-based), and motion (local or global) descriptors. Similarly, the MPEG-7 Multimedia Description Schemes (MDS) [3] supports spatial (directional or topological) and temporal multimedia segment relations, as well as hierarchical structures for multimedia segment decomposition. Given the importance of MPEG-7 in multimedia community, it is evident that in the design of an associated ontology, a large part of its structures should be appropriately captured, aligned and used.

*Support for Basic Data Types.* Finally, based on the previous requirement, and on the fact that MPEG-7 is built on XML Schema and supported by English-text semantic description but no associated data models, the implementation of an MPEG-7 ontology using an appropriate formalism like RDF Schema or OWL would have to deal with the representation of basic data types like numeric types (integer, float etc.), dates, vectors, arrays and so on. This is even more important when feature matching algorithms are employed on such data as part of the reasoning process during knowledge-assisted analysis.

## **2.2 Requirements for Semantic Annotation**

The described infrastructure requires appropriate authoring of the domain ontologies with respect to the domain and visual descriptor ontologies. In the following, we sketch the most important requirements in the context of conceptualization and annotation, influencing the overall ontology framework and the annotation tool design.

*Associate Visual Features with Concept Descriptions.* Visual descriptions are made on the conceptual level, i.e. certain visual descriptors should describe how a certain domain concept is expected to look like. The ontology and annotation framework should model this link in a way that is consistent with current semantic web standards and should avoid 2<sup>nd</sup> order statements, while

- preserving the ability to use reasoning on the ontology and the knowledge base respectively, and
- providing a clear distinction between the visual descriptions of a concept and its instances.

*User Friendly Annotation.* Domain ontologies are typically edited by trained indexers with little experience in multimedia analysis using standard ontology editing tools. Additionally, maintaining metadata about extracted low level

features is cumbersome and error-prone. An annotation framework thus has to integrate:

- management of reference multimedia content (images and videos)
- extraction of suitable low level features for objects depicted in the reference content
- automatic generation of fact statements describing the correspondence between a selected concept and the low level features
- while at the same time hiding the details of these mechanisms to the user behind an easy-to-use user interface.

*Modularization.* The links between domain ontology concepts and low level feature descriptions should form separate modules of the overall knowledge infrastructure. Specifically, updates of these fact statements should be possible without touching the integrity of the domain ontologies.

*Linking into Multimedia.* Visual Descriptors contain no information about their location in the original content. This becomes a problem if existing visual descriptors need to be visualized, e.g. to check them for appropriateness or to identify redundant descriptors. Additionally, in order to be able to exploit spatial relationships between objects within multimedia content, the objects have to be linked to the respective regions, they are depicted in. This combines to the more general requirement to provide means to describe regions in terms of their location within the content, i.e. to describe their spatial features, and to link them with objects representing both concepts from the domains and visual descriptors.

### 3. References

- [1] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab and M. G. Strintzis, "Semantic Annotation of Images and Videos for Multimedia Analysis", in Proc. of 2nd European Semantic Web Conference, (ESWC '05), Heraklion, Greece, May 29 – June 1, 2005.
- [2] T. Sikora. The MPEG-7 Visual standard for content description - an overview. IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7, 11(6):696–702, June 2001.
- [3] ISO/IEC 15938-5 FCD Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Schemes, March 2001, Singapore.