Applying Media Semantics Mapping in a Non-linear, Interactive Movie Production Environment

Michael Hausenblas
(Institute of Information Systems & Information Management,
JOANNEUM RESEARCH Forschungsgesellschaft mbH
Steyrergasse 17, 8010 Graz, Austria
michael.hausenblas@joanneum.at)

Abstract: In this work we propose how to deal with the Semantic Gap in closed domains. That is, we propose to bridge the Semantic Gap by means of mapping well-known low-level feature patterns found in MPEG-7 descriptions to formal concepts. The key contributions of the proposed approach are (i) the utilisation of ontologies, and rules to enhance the retrieval capabilities (effectiveness), and (ii) the realisation of the feature matching process being carried out on the structural level through indexed MPEG-7 descriptions (efficiency). We discuss advantages and shortcomings of our approach, and illustrate its application in the realm of non-linear, interactive movie productions.
Key Words: Non-linear interactive media, Media Semantics, Semantic Gap
Category: H.5.1, H3.1

1 Introduction

One key problem of multimedia content understanding—bridging the **Semantic Gap**—still is not satisfactory solved. Following [Smeulders et al. 2000], the Semantic Gap is "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation". In the realm of non-linear, interactive movie productions, one major challenge is the dynamic matching of appropriate clips based on a formal expression describing the desired content. In our setup, a movie is—based on the users interaction—assembled on-the-fly, requiring the retrieval of the audio-visual content to be performed in near-real-time. For example, in one point of the narration, there could be a query for some material that *is about soccer*, *has an interview* in it and *starts with a PAN LEFT* camera motion.

We believe that our approach—driven by requirements that stem from a non-linear, interactive production environment—can offer a sound solution to the problem stated above.

In the next section we start with a discussion on related work and give a short overview on the production environment. We then discuss the theoretical underpinnings of our proposed solution, the *Media Semantics Mapping* in Section 3. In the following Section 4 we present the current (prototypical) application of the approach. Finally, we conclude on the work done so far and report on the next planned steps in Section 5.

2 Related Work and Environment

Thoughts on how to conceptually bridge the Semantic Gap are probably as old as the multimedia content itself [Grosky 1994]. Most of the work in the realm of multimedia content representation focuses on the integration of multimedia metadata—as MPEG-7 [MPEG-7 2001]—with logic-based ontology languages, typically OWL [OWL 2004].

The constitutive work of Hunter et.al. [Hunter 2001] has led to numerous related efforts [Troncy 2003, Garcia and Celma 2005] that all share the translational approach of mapping MPEG-7 to OWL. For the field of ontology-based video retrieval, for example [Tsinaraki et al. 2004] reports a methodology to support interoperability of OWL with MPEG-7[1].

Media Streams—developed by Davis [Davis 1995] in his PhD thesis— is a system for annotating, retrieving, repurposing, and automatically assembling digital video. It uses a stream-based, semantic representation of video content with an iconic visual language interface of hierarchically structured, composable, and searchable primitives. Nack and Putz presented the Authoring System for Syntactic, Semantic and Semiotic Modelling (A4SM) framework [Nack and Putz 2001] that includes the creation, and retrieval of media material. The project goal was to have a framework at hand that would allow for semi-automated annotation of audiovisual objects, and to demonstrate the applicability in a news production environment. Both the Media Streams system and the A4SM can be understood as precursor to our proposed architecture.

Motivated by the promising work reported in [Little and Hunter 2004] and [Hollink et al. 2005] the proposal presented inhere is based on our experiences with MPEG-7 annotation and retrieval [Bailer et al. 2005].

The environment. The New Media for a New Millennium (NM2) project [Rehatschek et al. 2006] targets at the creation of technologies for non-linear, interactive narrative-based movie production. NM2 is an Integrated Project of the EU's 6th Framework Programme running till summer 2007 with 13 partners from eight European countries.

The tools for personalised, reconfigurable media productions are elaborated in six audio-visual productions that range from news reporting and documentaries through a quality drama serial to an experimental television production. Targeted end-user devices are Windows Media Centre-PCs, game consoles, and mobile phones. For a detailed overview on the project objectives, system capabilities and the productions, the reader is referred to [Williams et al. 2006].

[1] For an overview on related MPEG-7 formalisations and multimedia ontologies, the reader is referred to [Hausenblas et al. 2007].

3 Media Semantics Mapping

In this section, we discuss the *Media Semantics Mapping* (MSM) foundations, the terminology used, and the possibilities gained from using this approach.

Modality in our understanding is a path of communication between the human and the computer; major modalities are vision and audition (others are tactition, olfaction, etc.). In this work, audio-visual data is referred to as *essence*, i.e., the actual piece of data that resides e.g. in the file system. A *media item* (MI) is a proxy for some essence and acts as a pivot for attaching low-level features as well as annotations stemming from the domains semantics. In NM2, MPEG-7 [MPEG-7 2001] is utilised for representing low-level features of the essence, as colour descriptors, etc. We head after extracting as much as possible automatically from the essence to produce comprehensive MPEG-7 descriptions based on technologies of our Multimedia Mining Toolbox [Bailer et al. 2005, Section 6]. In the visual domain we use the Dominant Color Descriptor and the Color Layout Descriptor to capture colour features. To describe textures, we make use of the Edge Histogram Descriptor. Shapes can be recognized via the Contour-Based Shape Descriptor. The Camera Motion Descriptor is utilised to describe camera movements (pan, tilt, zoom, etc.). Although a representation of low-level features on the ontological level would be possible, we do not lift MPEG-7 descriptions and description schemes onto the logical level, rather MPEG-7 fragments are referenced from within the ontology.

OWL-DL [OWL 2004] is used to formalise the domain semantics and functions as the interface to the Narrative Structure Language [Ursu and Cook 2005]. A *logical entity* (LE) is anything contained in a MI that can either directly be recognised w.r.t. a modality, or that is not directly observable. A more formal account of the terms is given below.

3.1 Media Semantics

What are media semantics? According to [Harel and Rumpe 2004], any language definition comprises syntax, semantic domain, and a semantic mapping from the syntactic elements to the semantic domain. When talking about media semantics inhere, we subscribe to this point of view. In our understanding the essence itself does not "have" semantics. A piece of essence may be consumed or manipulated, nevertheless, essence "carries" the semantics and it is up to the consumer of the essence to interpret what she understands from it. Hence we do not try to define what in the general case an object "looks like" or "sounds like". We therefore understand that the ontological constructs in combination with the rules are our syntactical framework, further the semantic domain is conceived as being the domain of the LE that can occur in the essence, and finally define the semantic mapping as described below.

3.2 Spaces of Abstraction

We allow for two orthogonal conceptual paradigms to model media semantics: spaces and the well-known class/instance pattern. A *space* represents a certain level of abstraction, ranging from low-level, as colour or shape to abstract entities such as human feelings. Classes and instances are used to define the actual LE. Therefore "the soccer ball" instance in the context of a soccer game is defined to be black, white, and round but this does not mean that "a ball" in general—referring to the class level—has these properties. Fig. 1[2] depicts the

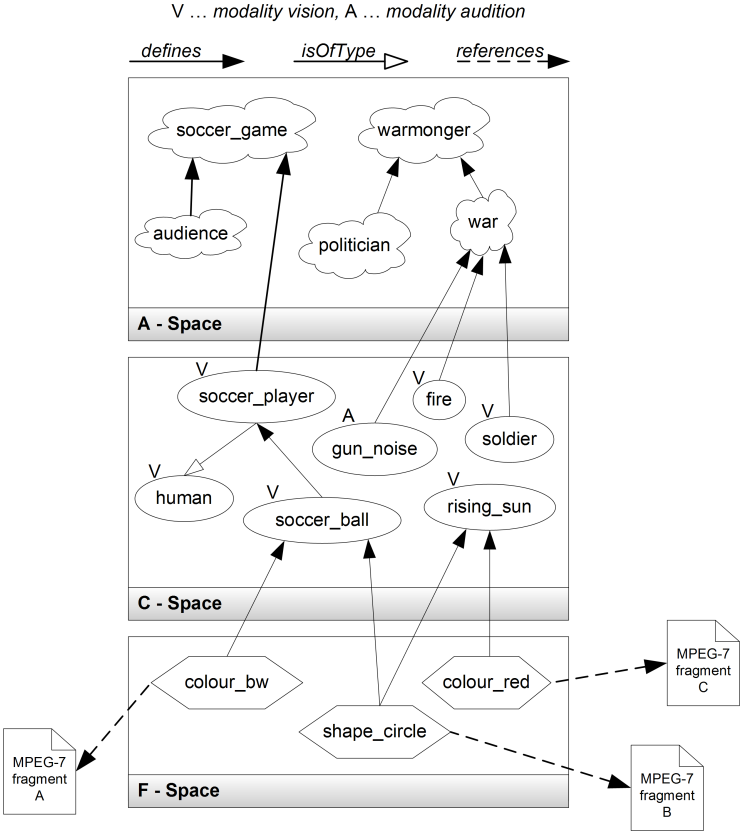

Figure 1: The Media Semantics Modelling Spaces.

spaces available in our approach, with V denoting the modality vision, and A the modality audition. Following [Grosky 1994] we introduce relationships that can

[2] The reader is invited to note that this figure has already been presented at a tutorial—What you Mean is What you Watch: Multimedia and the Semantic Web, cf. `http://gate.ac.uk/conferences/eswc2006/multimedia-tutorial/`—at the ESWC 2006.

hold between an essence and some logical entities. The spaces are defined—listed by increasing level of abstraction—as follows.

1. The Feature Space—**F-Space**. Contains LE that represent low-level features. A low-level feature (LLF) is a single aspect of a certain (spatio-temporal) part of a media item. For example the dominant colour of a spatial region (black and white) is represented as a LLF.

2. The Concrete Logical Entity Space—**C-Space**. Contains LE that can directly be recognized in the essence. A concrete logical entity (CLE) is a distinct object being defined by a combination of low-level features and their respective values (simple CLE) or using other concrete logical entities (composite CLE). For example, in the soccer domain, the CLE soccer ball may be defined by the LLF dominant colour black and white, and circular shape, i.e., a simple CLE. A table could be defined consisting of a CLE tabletop and four CLE table legs, resulting in a composite CLE.

3. The Abstract Logical Entity Space—**A-Space**. Contains LE that are not directly observable. An abstract logical entity (ALE) can be defined by a combination of CLE (simple ALE) or other ALEs (composite ALE). For example the ALE soccer game may be defined by the simultaneous presence of an ALE audience and a number of CLE soccer player.

Within each space, the class-instance modelling can be used to add further semantics, as taxonomies, object relations, etc. In each of the six NM2 productions, a domain-specific ontology is defined covering concepts and instances. This might be 'church', or 'painting' in the case of the documentary production about England's Golden Age in the 16^{th} century, or certain actors, moods, and keywords, as found in the drama Accidental Lovers [Hausenblas and Nack 2007].

Built-in rules. Due to the well-known limitations of DL-ontology languages [Horrocks et al. 2005] we utilise rules in addition to DL-ontologies (see also [Staab et al. 2003]) to define the semantics of a logical entity in the context of a production. However, using rules can lead to serious problems w.r.t. organisational issues. We therefore only provide a minimalistic set of so called built-in rules, and automatically generate the actual rules as described below.

Two properties, defined in the NM2 core ontology, enable the incorporation of rules, hence assisting to define the semantics of a logical entity. The `defines` property allows a combination of `ConcreteLogicalEntity` instances to define either another `ConcreteLogicalEntity` instance (composite pattern) or an `AbstractLogicalEntity` instance, hence an inter-space mapping. For each (partial) `defines`-property in the ABox of the ontology appropriate atoms are added to the corresponding rule.

A media item `contains` a number of `LogicalEntity` instances along with `LLFeature` instances representing an occurrence of a logical entity in a media item. Equally as above, for each occurrence atoms are added accordingly.

An exemplary built-in rule defining the mapping from the F-Space to the C-Space is shown below. Given that a set of low-level features $\{llf_1 \ldots llf_i\}$ `defines` a certain logical entity *cle* (line 1), and it is known that a certain media item *mi* `contains` this set of low-level features (line 2), it can be inferred that *mi* `contains` *cle*.

$$1 \quad contains(mi, cle) \leftarrow defines(llf_1, cle) \wedge \ldots \wedge defines(llf_i, cle) \wedge$$
$$2 \quad contains(mi, llf_1) \wedge \ldots \wedge contains(mi, llf_i)$$

However, to ensure the correctness of the definition, some constraints must be put on the variables: $\forall llf_i \in LLFeature$, further $cle \in ConcreteLogicalEntity$, and $mi \in MediaItem$, which highlights the connection to the NM2 core ontology that defines each of the concepts. To enhance the domain-specific ontologies further, so called *user-defined rules* can be manually defined by the user.

The ontology and the rules together form the knowledge base $\mathcal{KB}_{MSM}$, which further is used to annotate the essence automatically. $\mathcal{KB}_{MSM}$ is defined as being a tuple $\langle \mathcal{O}_D, \mathcal{R} \rangle$, with $\mathcal{O}_D$ being an ontology that consists of an ABox and a TBox, and $\mathcal{R}$ a rule-base comprising *built-in rules* and *user-defined rules*.

4 Applying the Media Semantics Mapping

The **M**edia **S**emantics **M**apping Utility (MSM-Utility) is used to define instances based on the built-in rules, described above to generate $\mathcal{KB}_{MSM}$. For managing MPEG-7 documents we use our MPEG-7 Document Server [Bailer et al. 2005, Section 5.2], which provides access to MPEG-7 documents for a number of clients and allows the exchange of whole documents or fragments thereof utilising XPath. Access to parts of documents is crucial for the efficiency of the system, as MPEG-7 documents of larger media items tend to have considerable size. The MPEG-7 documents used in the system are compliant with the Detailed Audiovisual Profile (DAVP) [Bailer et al. 2005].

For processing the ontological information, we use a performant RDF-library, the Redland RDF library[3], wrapped up in an Object-Oriented-API (C++) that enables manipulation and query on the ontological level. Applying $\mathcal{R}$ onto $\mathcal{O}_D$ is done utilising Prolog. $\mathcal{O}_D$ represented in OWL-DL is converted into a number of SWI-Prolog[4] facts.

[3] `http://librdf.org/`
[4] `http://www.swi-prolog.org/`

Typically, users of the NM2 toolkit lay out their production-specific ontologies by means of creating concepts and instances. Through $\mathcal{KB}_{MSM}$ the system is then able to automatically tag the essence in two subsequent steps. Firstly, the low-level features are extracted automatically on the MPEG-7 level. Secondly, $\mathcal{KB}_{MSM}$ is used to match against the generated description of the essence, triggering an update of the ABox of $\mathcal{O}_D$.

5 Conclusion and Outlook

We have shown in this paper how to map low-level features extracted from multimedia essence to logical entities. This enables an effective and efficient retrieval of the essence. Another source for the entity definition process are scripts, shot-logs, etc., which are incorporated through the ingestion process. We also plan to include the support for *guided definitions*. This means to extract MPEG-7 features from a reference image or audio-clip, display the extracted values and let the user select a combination of the extracted values for definition purposes, quite similar to [Little and Hunter 2004].

To allow for queries as "find me all MI with an interview as establishing shot, **followed** by a ZOOM_IN onto a painting", we currently work on the integration of so called *temporal annotations* to be used within a media item, based on [Allen and Ferguson 1994].

Acknowledgements

The work reported herein was undertaken in the realm of the "New Media for a New Millennium" project (FP6-004124), partially funded under the 6th FP of the European Commission. Special thanks to Werner Bailer and Peter Schallauer for their decent support regarding MPEG-7, and to Rudi Schlatte for useful discussions, as well as to the NM2 production teams. The author would also like to gratefully acknowledge the comments from the reviewers.

References

[Allen and Ferguson 1994] Allen, J. F. and Ferguson, G. (1994). Actions and Events in Interval Temporal Logic. Technical Report TR521, University of Rochester.

[Bailer et al. 2005] Bailer, W., Schallauer, P., Hausenblas, M., and Thallinger, G. (2005). MPEG-7 Based Description Infrastructure for an Audiovisual Content Analysis and Retrieval System. In *Proceedings of SPIE - Storage and Retrieval Methods and Applications for Multimedia*, volume 5682, pages 284–295.

[Davis 1995] Davis, M. (1995). *Media streams: representing video for retrieval and repurposing*. PhD thesis, Massachusetts Institute of Technology.

[Garcia and Celma 2005] Garcia, R. and Celma, O. (2005). Semantic Integration and Retrieval of Multimedia Metadata. In *5^{th} International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot'05)*, Galway, Ireland.

[Grosky 1994] Grosky, W. I. (1994). Multimedia Information Systems. *IEEE MultiMedia*, 1(1):12–24.

[Harel and Rumpe 2004] Harel, D. and Rumpe, B. (2004). Meaningful Modeling: What's the Semantics of "Semantics"? *Computer*, 37(10):64–72.

[Hausenblas and Nack 2007] Hausenblas, M. and Nack, F. (2007). Interactivity = Reflective Expressiveness. *IEEE MultiMedia*, 14(2):1–7.

[Hausenblas et al. 2007] Hausenblas, M., Boll, S., Bürger, T., Celma, O., Halaschek-Wiener, C., Mannens, E., and R. Troncy (2007). Multimedia Semantics on the Web: Vocabularies *W3C Incubator Group Report*.

[Hollink et al. 2005] Hollink, L., Little, S., and Hunter, J. (2005). Evaluating the application of semantic inferencing rules to image annotation. In *3rd International Conference on Knowledge Capture (K-CAP 2005)*, pages 91–98, Banff, Alberta, Canada. ACM.

[Horrocks et al. 2005] Horrocks, I., Patel-Schneider, P. F., Bechhofer, S., and Tsarkov, D. (2005). OWL Rules: A Proposal and Prototype Implementation. *Journal of Web Semantics*, 3(1):23–40.

[Hunter 2001] Hunter, J. (2001). Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *First International Semantic Web Working Symposium (SWWS'01)*, Stanford, California, USA.

[Little and Hunter 2004] Little, S. and Hunter, J. (2004). Rules-By-Example - A Novel Approach to Semantic Indexing and Querying of Images. In *3rd International Semantic Web Conference (ISWC'04)*, volume 3298 of *Lecture Notes in Computer Science*, pages 534–548, Hiroshima, Japan.

[MPEG-7 2001] MPEG-7 (2001). Multimedia Content Description Interface. Standard No. ISO/IEC n15938.

[Nack and Putz 2001] Nack, F. and Putz, W. (2001). Designing annotation before it's needed. In *ACM Multimedia*, pages 251–260.

[OWL 2004] OWL (2004). Web Ontology Language Reference. W3C Recommendation.

[Rehatschek et al. 2006] Rehatschek, H., Hausenblas, M., Thallinger, G., and Haas, W. (2006). Cross media aspects in the areas of media monitoring and content production. In *5th International Conference on Language Resources and Evaluation (LREC) 2006, cross-media indexing workshop*, pages 25–31, Genoa, Italy.

[Smeulders et al. 2000] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

[Staab et al. 2003] Staab, S., Angele, J., Decker, S., Grosof, B., Horrocks, I., Kifer, M., and Wagner, G. (2003). Where are the rules? *IEEE Intelligent Systems, Trends & Controversies*, 18(5):76–83.

[Troncy 2003] Troncy, R. (2003). Integrating Structure and Semantics into Audio-visual Documents. In *2nd International Semantic Web Conference (ISWC'03)*, volume 2870 of *Lecture Notes in Computer Science*, pages 566–581, Sanibel Island, Florida, USA.

[Tsinaraki et al. 2004] Tsinaraki, C., Polydoros, P., and Christodoulakis, S. (2004). Interoperability support for Ontology-based Video Retrieval Applications. In *3rd International Conference on Image and Video Retrieval (CIVR'04)*, Dublin, Ireland.

[Ursu and Cook 2005] Ursu, M. F. and Cook, J. (2005). D5.3: Languages for the representation of visual narratives. Deliverable to EC (permission required), NM2 consortium.

[Williams et al. 2006] Williams, D., Ursu, M., Cook, J., Zsombori, V., Engler, M., and Kegel, I. (2006). ShapeShifted TV – Enabling Multi-Sequential Narrative Productions for Delivery over Broadband. In *The 2nd IET Multimedia Conference, 29-30 November 2006*. ACM Press.