

Representation and Analysis of Multimedia Content: The BOEMIE Proposal

D.I. Kosmopoulos, V. Karkaletsis, S. Perantonis, G. Paliouras, C.D. Spyropoulos

Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
P.Grigoriou & Neapoleos str., 15310 Agia Paraskevi Attikis, Athens, Greece
{dkosmo,vangelis,sper,paliourg,costass}@iit.demokritos.gr

Abstract

We propose an approach to knowledge acquisition, which uses multimedia ontologies for fused extraction of semantics from multimedia content, and uses the extracted information to evolve the ontologies. We present the basic components of the proposed approach, describe an application scenario we currently examine, and discuss the open research issues focusing on knowledge representation and extraction techniques that will enable the development of scalable and precise knowledge acquisition technology.

1. Introduction

The objective of multimedia content analysis is the automated knowledge acquisition from various modalities, e.g., text, images, video etc. The high complexity that characterizes the multimedia content, along with the currently prevailing dearth of precise modeling for multimedia concepts, makes automatic semantics extraction a very challenging task.

Although latest advances in content analysis have improved capabilities for effective searching and filtering, a gap still remains between the low-level feature descriptions, and high-level semantic descriptions of concepts. A suitable approach to fill this gap is to use a semantic model in the extraction process. Moreover, the analysis of single modalities, in particular of visual content alone, is inadequate in all but a small number of restricted cases.

The proposed approach, which is envisaged in the framework of the IST project BOEMIE, is unique in that it links multimedia extraction with ontology evolution. This approach will be used to enrich digital maps with multimedia content related to city events. The content is collected from various proprietary or open sources and it becomes automatically semantically annotated. Driven by domain-specific multimedia ontologies, the information extraction systems implementing the proposed approach will be able to identify high-level semantic features in image, video, audio and text and fuse them for optimal extraction. The ontologies will be continuously populated and enriched using the extracted semantic content. This is a bootstrapping process since the enriched ontologies will in turn be used to drive the multimedia information extraction system.

This work provides the key ideas involved in the whole system and then focuses on the semantics extraction. Section 2 highlights the related research. Section 3 presents the main aspects of the proposed approach, the architecture and its basic components. Section 4 provides an application scenario we are currently examining for the evaluation of the proposed approach. Section 5 discusses some of the issues that arise under this bootstrapping framework. The paper concludes presenting our next steps for the implementation of the proposed approach.

2. State of the art

The involved technologies include the semantics extraction from multimedia content, the multimedia ontologies and techniques that exploit their synergy.

Semantics extraction from multimedia content is the process of assigning conceptual labels to either complete multimedia documents or entities identified therein. In general, extraction can be performed at the levels of *layout (structure)*, *content* and *semantics* (intended meaning of the author).

In the case where content is available in multiple related modalities, these can be combined for the extraction of semantics. The combination of modalities may serve as a verification method, a method compensating for inaccuracies, or as an additional information source (Snoek and Worring 2005). The processing cycle of combination methods may be iterated allowing for incremental use of context. The major open issues in the combination approaches concern the efficient utilization of prior knowledge, the specification of open architecture for the integration of information from multiple sources and the use of inference tools for efficient retrieval.

Most of the extraction approaches encountered in the literature are based on learning methods, e.g., naive Bayes classifiers, decision tree induction, k-Nearest neighbour, Hidden Markov model (Manning and Schutze 1999, Rabiner 1989). However, with the advent of promising methodologies in multimedia ontology engineering, knowledge-based approaches are expected to gain in popularity and be combined with the machine learning methods. This is also the case we will study in the proposed approach.

Ontologies can play a major role in multimedia content interpretation because they can provide high-level semantic information that helps disambiguating the labels assigned to multimedia objects. Indicative approaches for constructing multimedia ontologies are the ones presented in Hunter 2001, Mezaris et al 2004, and Troncy 2003. The major open issues here concern the automatic mapping between low level audio-visual features and high level domain concepts, the automated population from unconstrained content and when there are no metadata attached to the content. In cases of complex domains, multiple ontologies may be present and ontology

coordination techniques have to be employed (e.g., Castano 2004, Kotis and Vouros 2004, Gomez 2002).

The interaction between information extraction and ontology learning has also been modelled at a methodological level as a bootstrapping process that aims to improve both the conceptual model and the extraction system through iterative refinement. In Maedche and Staab 2000, the bootstrapping process starts with an information extraction system that uses a domain ontology. The system is used to extract information from text. This information is examined by an expert, who may decide to modify the ontology accordingly. The new ontology is used for further information extraction and ontology enrichment. Brewster et al. 2002 propose a slightly different approach to the bootstrapping process. Starting with a seed ontology, usually small, a number of concept instances are identified in the text. An expert separates these as examples and counter-examples which are then used to learn extraction patterns. These patterns are used to extract new concept instances and the expert is asked to re-assess these. When no new instances can be identified, the expert examines the extracted information and may decide to update the ontology and restart the process.

3. Methodology and architecture

We advocate an ontology-driven multimedia content analysis (semantics extraction from images, video, text, audio/speech) through a novel synergistic method that combines multimedia extraction and ontology evolution in a bootstrapping fashion (see Figure 1). In the following sub-sections we describe the proposed components.

3.1. Semantics Extraction from Multimedia Content

A suitable approach to bridge the semantic gap is to use a semantic model in the extraction process. Moreover, the analysis of single modalities, in particular of visual content alone, is inadequate in all but a small number of restricted cases. The effort required to provide problem-specific extraction tools makes single-media solutions non-scalable, while their precision is also rarely adequate. In the proposed approach, on the level of individual modalities, particular emphasis will be given to visual content, from images and video, due to the richness of this source and corresponding difficulty of extracting useful information. Non-visual content, audio/speech and text, will provide supportive evidence, in order to improve extraction precision. Since no single modality is powerful enough to encompass all aspects of the content and identify concepts precisely, fusing information from multiple media sources is needed.

3.2. Multimedia Ontologies

In our approach, we propose the development of a unifying representation for multimedia ontologies and related knowledge. This “multimedia semantic model” will serve as an integrated model for the different ontologies that are necessary to support the semantics extraction process:

- Multimedia content ontology: It represents the structure of the content of the multimedia documents. The top level hierarchy of a multimedia document is classified into: Image, Video, Audio, Audiovisual and Multimedia. Each of these types

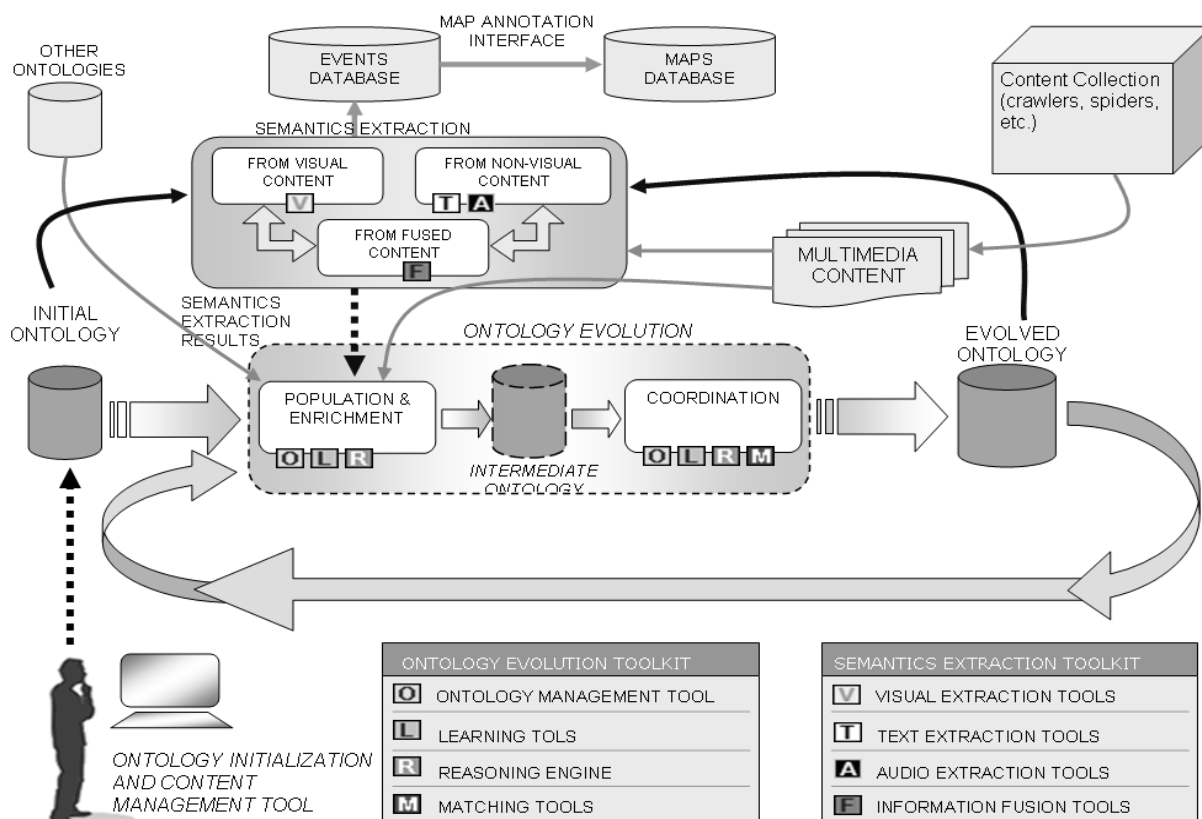


Figure 1: Architecture of the integrated system

has its own segment subclasses. These subclasses describe the specific types of multimedia segments, such as video segments, moving regions, still regions and mosaics.

- **Multimedia descriptor ontologies:** This ontology models concepts and properties that describe visual characteristics of objects in terms of low-level features and media structure descriptions. Sub-concepts will include MPEG-7 standard features like colour, shape, texture, motion, localization and basic descriptors. Separate descriptor ontologies apply to different modalities. Along with the multimedia descriptor ontologies is used the hierarchical “fusion model”, where the significance (weight) of each modality is defined for each concept.
- **Domain-specific ontologies:** These ontologies contain concepts and properties related to the knowledge of the domain of interest. In these concepts we assign instances, which are used to recognize semantic objects using the results of the content analysis process. These ontologies also contain detailed descriptions of objects using spatiotemporal and partonomic relations defined in the multimedia semantic model.

3.3. Evolution of Multimedia Ontologies

According to Stojanovic 2004 ontology evolution is “the timely adaptation of an ontology to the arisen changes and the consistent propagation of these changes to dependent artefacts”. Thus, ontology evolution is a complex process, involving the following sub-processes: ontology *population* and *enrichment*, i.e., addition and deletion of concepts, relations, properties and instances, *coordination* of homogeneous ontologies, e.g. when more than one ontologies for the same domain are available, and heterogeneous ontologies, e.g., updating the links between a modified domain ontology and a multimedia descriptor ontology, *maintenance* of semantic consistency, since any of the above changes may generate inconsistencies in other parts of the same ontology, in the linked ontologies or in the annotated content base.

Our approach for ontology population and enrichment will be based on machine learning techniques using the information from the semantics extraction process. More specifically, the extraction process will populate the ontologies with instances of the various concepts, together with their properties and will also provide unclassified entities extracted from the multimedia content which may lead to suggestions for the enrichment of the ontologies with new concepts and relations, through novelty detection. This novelty detection is based on information from all different types of media being processed.

Ontology coordination approaches will be devised to interlink ontologies with different levels of heterogeneity. Ontology coordination involves the use of matching techniques and tools for mapping, alignment and merging.

During ontology evolution, any of the changes may generate inconsistencies in other parts of the same ontology, in the linked ontologies or in the annotated resources. At the current state of the art, description logic

reasoning systems (e.g., RACER¹) are not tailored to these “incremental changes”. We will investigate how such changes can be much more efficiently supported. The aim is the development of models, techniques, and tools for semantic consistency checking of ontology content throughout the evolution process.

4. Application scenario

The application we are currently examining for the evaluation of the proposed bootstrapping approach concerns the enrichment of digital maps with semantic information. In other words, the results of the semantics extraction process will be displayed to the end-user, through an interactive digital map.

The user is interested to find “what” happens “where” in the city. The possible queries can be, events of a particular type in a specific time frame, events in a venue – location, persons related to event (e.g., actors, players etc), events at specific dates, events similar to a given one, events at nearby venues. It is assumed that we have discrete locations in the map, where all possible events are allowed to take place. As a concrete example of the application scenario, we propose the domain of sports where the user asks to know about a specific sport event (e.g., football game) in the city. He receives a list of games and is able to browse multimedia content related to game type, league, previous games, comments-gossip-interviews on the game to be played, concerning team history or the football ground.

4.1. Initialization

We will start by collecting, extending and merging existing ontologies for sub-domains referring. These ontologies will also be linked to the appropriate multimedia descriptor ontologies. This process will be accomplished using the ontology initialization and content annotation tool and will result to the initial multimedia semantic model for the domain.

4.2. Training

The various semantics extraction and ontology evolution tools are trainable to the domain. Therefore, a training dataset needs to be collected and used to customise the system. This training set should contain representative and annotated multimedia content, as expected to be encountered by the system at run time.

4.3. Information gathering

Having customised the system, the first step of its run-time use is to collect content from various Web and proprietary sources. In the case of sports events, such sources may include TV and news programmes, on-line magazines, sports-related sites, specialized discussion fora and Weblogs, as well as generic content sources.

4.4. Semantics extraction

The trained semantics extraction tools will be applied at regular intervals to the incoming stream of multimedia content, performing extraction of the relevant information from each piece of content.

¹ <http://www.sts.tu-harburg.de/~r.f.moeller/racer/index.html>

We define some city-specific concepts in the multimedia ontologies. For each concept we should have already defined features (in the multimedia descriptor ontology) and some classification parameters that enable a decision about the content with a certain probability or certainty factor or fuzzy membership (hard rules are inappropriate due to uncertainties in processing). This applies to all available modalities (text, images, video, audio), which means that individual modality – specific features (and classifiers) are available and decoupled from other modalities.

There is a “fusion model” with a hierarchical structure, which receives a decision input from the individual modalities (see Figure 2). For each node it holds the information about the weight of the decision taken by each modality. Its role is to combine the decisions taken by each modality-specific processing by applying the respective weights.

For the sports scenario we assume that the multimedia ontology defines the following city-specific classes-concepts (from general to more specific): *Indoor-outdoor*. The outdoor may decompose to *concert, sports, theater*. Each of the above categories may decompose to relevant subcategories. For example, the sport decomposes to *football, swimming* in the initial ontology. More football-like sports may be identified progressively through the evolution of the initial ontology.

In the example scenario the system

- (i) receives the input, which happens to be a video scene of football.
- (ii) The multimodal information is separated and processed separately to the visual part, and the audio part.
- (iii) The information is processed hierarchically assuming no prior knowledge.
- (iv) Using the visual features we classify the content with respect to the highest concept in the hierarchy, i.e., the indoor – outdoor. The appropriate feature for this task is the color. We find that the percentage of “green” is high in the average image histogram so the classifier gives a probability of 0.7 for outdoor and 0.3 for indoor ($P(\text{outdoor})=0.7$, $P(\text{indoor})=0.3$). Other uncertainty representations instead of probabilities could be applicable too.
- (v) Using the audio ontology we find that there are no sounds that are typical for outdoor environment, e.g., sounds of birds, waves, wind etc, so the classifier gives $P(\text{outdoor})=0.45$ and $P(\text{indoor})=0.55$.
- (vi) The text processing (after OCR) does not provide any relevant info so the probability is shared between the two classes ($P(\text{outdoor})=0.50$ and $P(\text{indoor})=0.50$).
- (vii) The “fusion model” has predefined through training that the weights for the visual, audio and text modalities (W_v , W_a , W_t , given in Figure 2) and based on them it decides that $P(\text{outdoor})=0.565$. So we proceed to the next layer examining the “child” of the outdoor concept.

Similarly we examine the three modalities to classify to event categories.

- (i) The visual modality finds very high motion and gives $P(\text{sport})=0.7$, $P(\text{concert})=0.2$, $P(\text{theatre})=0.1$.
- (ii) The audio detects speech and crowd sounds and gives $P(\text{sport})=0.5$, $P(\text{theatre})=0.4$, $P(\text{concert})=0.1$.
- (iii) The text does not find relevant features so the probability is shared to the three concepts.
- (iv) The “fusion model” decides using the related weights that $P(\text{sport})=0.589$.

The next level has to do with classification into football and swimming.

- (i) The visual ontology defines human motion features, color histogram and we calculate $P(\text{football})=0.9$.
- (ii) The audio ontology identifies patterns related to football such as “goal”, “corner”, “foul” and therefore gives probability $P(\text{football})=0.85$.
- (iii) The text identified gives the score, the team names and thus $P(\text{football})=0.7$.
- (iv) The “fusion model” decides that $P(\text{football})=0.86$. So the shot is classified as “football”.

Additional information that could be used include team names, team order in the score board, player names etc, which are associated with a certain football stadium. We assume that the number of football stadiums is limited in a city. The information is localized in the map based on the known locations of the football stadiums and the teams that are associated with them.

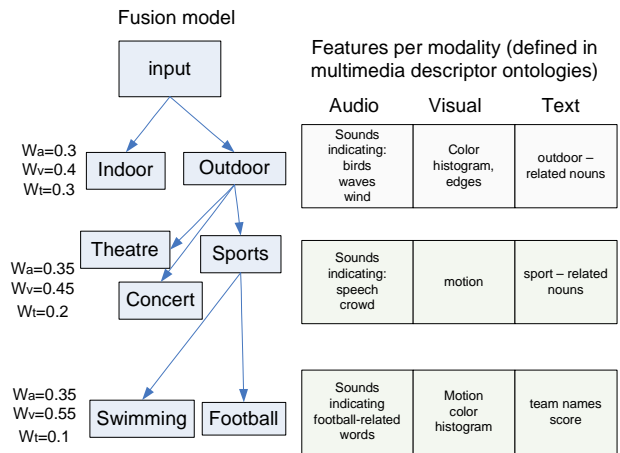


Figure 2: The fusion model for the described scenario, which defines the concept hierarchy and the modality weights per concept and the related multimedia features used.

4.5. Ontology evolution

The former, extraction task will populate the ontologies with instances of the various concepts, together with their properties. This process will also be accompanied by the appropriate annotation of content in the server, in order to provide semantic access to the content by the end-user. The latter, concept modelling task, performed by the extraction methods, will lead to suggestions for the enrichment of the ontologies, through novelty detection. The evolution can be performed through clustering with respect to specific features. As

new sport events may emerge (e.g., football on beaches) the knowledge representation has to evolve accordingly.

5. Discussion

In terms of semantics extraction from multimedia content, we propose the integration of an ontology-based approach with a probabilistic inference scheme. We need to examine carefully the role of the ontology in fusing information extracted from multiple media. We also have to examine new ways to fuse features derived from multimedia content.

Ontologies must be sufficiently expressive to describe the construction space for possible interpretations in general and for specific interpretation results in terms of a particular piece of media. Multimedia applications have highlighted the need to extend representation languages with capabilities which allow for the treatment of the inherent imprecision in multimedia object representation, matching, detection and retrieval. Existing standard web languages do not provide such capabilities. Therefore, considerable research effort needs to be directed towards representation and management of uncertainty, imprecision and vague knowledge in real life applications.

In terms of ontology population and enrichment, we will exploit the multimedia semantic model as well as current research on learning and inference techniques aiming to develop a generic framework for ontology learning and inference from multimedia content, due to the complexities introduced by the multimedia context. Addition of instances in the multimedia descriptor ontology may also require updating the corresponding link with the domain-specific ontology. The semantics extraction process will provide unclassified entities extracted from the multimedia content which may lead to the enrichment of the ontologies with new concepts and relations based on information from all different types of media being processed. Concerning inference techniques for ontology population and enrichment, we need to optimize and enhance description logic inference technology to support learning and retrieval requirements.

We also propose the use of machine learning techniques to assist ontology coordination in this context and we need to investigate the appropriate methods. This depends very much on the type of training data that is available. Supervised learning of complex representations requires data that may not be possible to acquire manually. Unsupervised or partially supervised methods may prove more useful in these cases.

Concerning semantic consistency checking in ontology evolution, there are two main problems. The first occurs at the instance level and requires techniques for efficiently handling incremental additions of instances, while checking integrity constraints. A second one occurs at the concept level and requires techniques for checking the consistency of new concepts against the current ontology, to choose a valid and consistent enrichment solution among a set of possible alternatives.

6. Concluding remarks

We propose a new approach towards automation of knowledge acquisition from multimedia content, by introducing the notion of evolving multimedia ontologies which will be used for the extraction of information from multimedia content. We have outlined the approach

through a sports scenario. This is a synergistic approach since it combines multimedia extraction and ontology evolution in a bootstrapping process involving, on the one hand, the continuous extraction of semantic information from multimedia content in order to populate and enrich the ontologies and, on the other hand, the deployment of these ontologies to enhance the extraction robustness.

The main measurable objective of this initiative is to improve significantly the performance of existing single-modality approaches in terms of scalability and precision. Towards that goal, our aim is to develop a new methodology for extraction and evolution, using a rich multimedia semantic model, and realize it as an open architecture. The architecture will be coupled with the appropriate set of tools.

7. Acknowledgements

BOEMIE is an FP6-IST-Call 4 project (027538), to begin in March 2006. BOEMIE consortium consists of NCSR "Demokritos" (GR), Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung (DE), University of Milano (IT), Centre for Research and Technology Hellas (GR), Hamburg University of Technology (DE), TeleAtlas (BE).

8. References

- Cees G.M. Snoek, M. Worring (2005). Multimodal Video Indexing: A Review of the State-of-the-art, *Multimedia Tools and Applications*, 25, pp. 5–35.
- Manning C.D. and Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, USA.
- Rabiner L.R. (1989). A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286.
- Hunter J. (2001). "Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology", *International Semantic Web Working Symposium (SWWS)*, Stanford, July 30 - August 1,
- Mezaris V., Kompatsiaris I., Boulgouris N.V. and Strintzis M.G. (2004). "Real-time compressed domain spatiotemporal segmentation and ontologies for video indexing and retrieval", *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Audio and Video Analysis for Multimedia Interactive Services*, vol. 14, no. 5, pp. 606-621.
- Troncy R. (2003). "Integrating Structure and Semantics into Audio-Visual Documents", In the 2nd *International Semantic Web Conference, ISWC 2003*, LNCS 2870, pp. 566-581.
- Castano S., Ferrara A., Montanelli S., and Racca G., (2004). "Matching Techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions", *IEEE Proc. of the International Conference on Coding and Computing (ITCC04)*, Las Vegas, Nevada, USA
- Kotis K., Vouros G. (2004). HCONE approach to Ontology Merging. *ESWS'04: The Semantic Web: Research and Applications*, LNCS, Vol. 3053, Springer-Verlag.
- OntoWeb. Deliverable D1.3. (2002). A survey on ontology tools, May (ed. A. Gómez Pérez)

- Maedche A. and Staab S. (2000). Mining ontologies from text. In R.Dieng and O.Corby, editors, Proceedings of EKAW-2000, LNCS, v.1937, pp. 189–202. Springer.
- Brewster, Ciravegna F., and Wilks Y. (2002). User-centred ontology learning for knowledge management. In B. Andersson, M. Bergholtz, and P. Johannesson, editors, NLDB, volume 2553 of LNCS, pages 203–207. Springer.
- Stojanovic L. (2004). Methods and Tools for Ontology Evolution. PhD thesis, University of Karlsruhe.