Knowledge Acquisition from Multimedia Content using an Evolution Framework

D. Kosmopoulos, S. Petridis, I. Pratikakis, V. Gatos, S. Perantonis, V. Karkaletsis, G. Paliouras
Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos",
15310 Aghia Paraskevi Attikis, Athens, Greece
{dkosmo,petridis,ipratika,bgat,sper,vangelis,paliourg}@iit.demokritos.gr

Abstract. We propose an approach to knowledge acquisition, which uses multimedia ontologies for fused extraction of semantics from multiple modalities, and feeds back the extracted information, aiming to evolve knowledge representation. This paper presents the basic components of the proposed approach and discusses the open research issues focusing on the fused information extraction that will enable the development of scalable and precise knowledge acquisition technology.

1 Introduction

The main goal of multimedia content analysis is the automated extraction of indices describing the document content. The high complexity that characterises the multimedia content along with the currently prevailing dearth of precise modelling for multimedia concepts makes automatic semantics extraction a very difficult and challenging task. Although latest advances in multimedia content analysis have improved capabilities for effective searching and filtering, a gap still remains between low-level feature descriptions that can be automatically extracted such as colours, textures, shapes, motions, and so forth, and high-level semantic descriptions of concepts like objects, scenes and events that set the basis for meaningful multimedia content description. A suitable approach to bridge 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 effort required to provide problem-specific extraction tools makes single-media solutions non-scalable, while their precision is also rarely adequate.

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, creating a synergy of enormous yet unrealized potential. 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 these features 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 from multimodal features. Section 2 highlights the related research. Section 3 presents the main aspects of the proposed approach, the architecture designed for its implementation and the basic components of the architecture. Section 4 outlines the semantic extraction approach and section 5 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 and need to be searched. The paper concludes by presenting our next steps.

2 State of the art

The proposed approach towards the automation of knowledge acquisition from multimedia content, through ontology evolution, is based on the synergy of various technologies. This section highlights the state of the art of the technologies involved.

2.1 Semantics extraction from multimedia content

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 three different levels:

- Layout: the syntactic structure an author uses for multimedia documents (camera shots, audio segments, text syntax).
- Content: relates layout segments to elements that an author uses to create in a multimedia document (e.g. setting, objects, humans …).
- Semantics: expresses the 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 [5]. 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 multimedia extraction approaches encountered in the literature are based on learning methods, e.g. naive Bayes classifiers, decision tree induction, k-Nearest neighbor, Hidden Markov model [9], [12].

2.2 Multimedia Ontologies

Ontologies can play a major role in multimedia content interpretation because they can provide high-level semantic information that helps disambiguating the

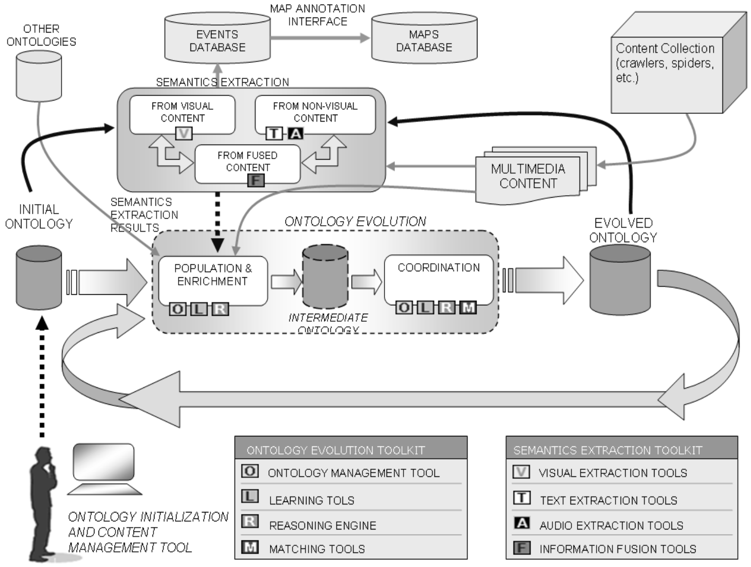

Figure 1. Architecture of the integrated system

labels assigned to multimedia objects. Indicative approaches for constructing multimedia ontologies are the ones presented in [7], [11] [13]. 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 [4], [8], [6] have to be employed.

2.3 Synergy between information extraction and ontologies

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, but it is limited to textual content so far and is not fully automated [10], [3].

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. This method involving on the one hand, the continuous extraction of knowledge from multimedia content sources in order to populate and enrich the ontologies and, on the other hand, the deployment of these ontologies to enhance the robustness of the multimedia information extraction system.

On the side of ontology evolution, we propose (a) a unified representation for multimedia ontologies and related knowledge, which will link domain-specific concepts with low-level features and structural descriptions and (b) a methodology and a toolkit for ontology evolution to support ontology learning, ontology merging and alignment, semantic inference for consistency maintenance, and ontology management. On the side of information extraction, we propose (a) a methodology and an open architecture for information extraction from multimedia content using data fusion techniques and (b) A toolkit for semantic extraction from multimedia content. Within the extraction architecture, tools will be developed to support extraction from image, audio, video and text, as well as information fusion.

We also propose an open architecture of a system that integrates the components for ontology evolution and semantics extraction in order to realise the synergistic bootstrapping approach. As depicted in Figure 1, the major components are:

- The multimedia ontology which links domain-specific ontologies with multimedia content and descriptor ontologies. This will be evolving through the ontology evolution component. An ontology initialization tool will be developed to provide a friendly user interface for the creation of the initial ontology.
- The semantics extraction component which will provide tools for the analysis of single modalities (visual, text and audio extraction tools) as well as tools for fusing information from multiple media sources (information fusion tools). The whole extraction process will be ontology driven in the sense that the ontology will provide the initial knowledge to the extraction process and will also be used to disambiguate the extraction.
- The ontology evolution component which will use the results of the extraction process to populate the multimedia ontology with instances of the various concepts, to enrich the ontology with new concepts and relations, as well as to coordinate the ontologies composing the multimedia ontology.

4 Semantics extraction toolkit

The semantics extraction toolkit is composed of subsystems that process separate modalities and namely visual (still images or image sequences) and non-visual content (audio and text). The results are fused using ontology-based or probabilistic framework and the results are used for content annotation and ontology evolution.

4.1 Semantics extraction from visual content

Our effort concentrates on the development of a semantics extraction toolkit from visual content including tools for:

- *Scene categorisation*, i.e., to categorize the depicted content into various high level classes, e.g. indoor/outdoor city/landscape, office, corridor, street, etc.). Features to be used are color histograms, color coherence vectors, DCT coefficients, edge direction histograms, edge direction coherence vectors, and motion vectors. Sub-blocks analysis can be employed for independent initial classification of blocks, which will be combined next using reasoning.

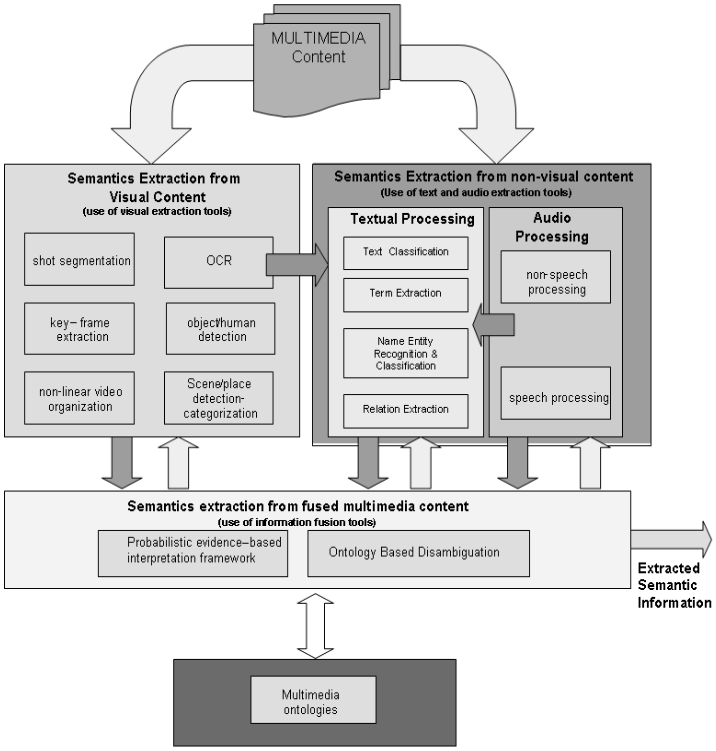
MULTIMEDIA Content
Semantics Extraction from Visual Content
(use of visual extraction tools)
shot segmentation
OCR
key–frame extraction
object/human detection
non-linear video organization
Scene/place detection-categorization
Semantics Extraction from non-visual content
(Use of text and audio extraction tools)
Textual Processing
Text Classification
Term Extraction
Name Entity Recognition & Classification
Relation Extraction
Audio Processing
non-speech processing
speech processing
Semantics extraction from fused multimedia content
(use of information fusion tools)
Probabilistic evidence–based interpretation framework
Ontology Based Disambiguation
Extracted Semantic Information
Multimedia ontologies

Figure 2 Architecture of the semantics extraction subsystem

- *Video-OCR*, in order to extract as much information as possible about the depicted settings, objects, persons through appropriate name/noun/verb identification.
- *Object detection, recognition and tracking*, mainly through motion analysis and matching with related patterns to differentiate solid objects from deformable ones, e.g., vehicles from humans. Other features used may include skin color or facial features for humans. For tracking, the particle filtering technique may be employed or the MPEG moving object descriptors may be exploited if available. The expected found objects can be significantly reduced by the previous steps.
- *Place recognition*, i.e., recognition of known places based on features that are unique for each place (e.g., landmarks) and based on the results of the previous processing steps.

It is critical to be able to spot independent visual entities either in single images or image sequences. Furthermore, we provide support to the ontological description which will require expressing relations (structural, spatial, temporal) between the visual objects. In light of this need, our methodology focuses on research of tools for automatic image segmentation and video segmentation, which will enable us to partition an image or a sequence of images into meaningful spatial or spatiotemporal objects. In this perspective, all mid-level feature identification are treated in a region-based fashion. Moreover, we use advanced machine learning techniques for recognition and categorization, in order to address various problems that are related to the use case. Semantics are extracted by matching/linking with visual information included in the multimedia ontologies.

Some additional tools perform video segmentation to isolate the video shots, key-frame extraction to limit processing to those frames that describe adequately the content and video organization (in a non-linear fashion) to find similar shots through clustering and to limit processing to shot representatives and to assist user browsing.

To provide a qualitative measure for the visual detectors involved, confidence measures are employed. These measures will be taken into account in BOEMIE's reasoning engine, which will be able to modify the visual detector's confidence scores according to a set of contextual rules and supplementary rules expressed by the corresponding semantic model that determines how likely it is for the given object (or scene) to appear in the given visual content.

4.2 Semantics Extraction from non-visual content

Non-visual content can be textual or audio/speech, which will be used to construct the basis for a meaningful contextual knowledge culminating at a toolkit for semantics extraction from non-visual content.

The text is provided by unstructured (e.g. raw text) or semi-structured documents (e.g. HTML pages) or is included in image/video data as raw data (e.g., OCR text, speech transcriptions) or as annotations (e.g. textual descriptions of images). The effort is concentrated on:

- using available text processing tools and resources in combination with the domain-specific and geographic ontologies of the multimedia ontology to extract information that "localises" the content directly or indirectly, e.g., names of exhibition places, street names;
- extracting semantic information, e.g., in the vehicle exhibition case, extract information such as the event name, event dates, event organisers, exhibitors (vehicle brand names), exhibits (vehicle models) and multimedia related to it.

On the other hand, audio processing is used to extract features from both speech and non-speech audio. The effort here is concentrated on the detection of:

- Environmental sounds for the detection of setting
- Names in speech data to infer references to places of interest

Methodologically, the audio channel is separated into speech and non-speech segments using BIC segmentation and speaker detection. As regards non-speech environmental sounds, a state-of-the-art classification method, based on machine learning, is to be used for the generic automatic construction of (possibly nested) sequences of data transformations (tree-like features). The sound classifier deploys acoustic models of sounds and permits differentiation of different sound sources (applause, laughter) and environments (traffic vs. quiet museum). The speech segments are recognized using a syllable-based speech recognizer for the English language. Since the ontologies providing the basis for feature derivation are evolved throughout the content extraction process, it is critical that the vocabulary of words which are identified by the speech recognition component is dynamic.

4.3 Semantics Extraction from Multimedia Content

Single modalities are not always powerful enough to encompass all aspects of the content and identify concepts precisely. Fusing information from multiple media sources [1] is expected, on one hand, to improve the validity of indices extracted independently from each modality and, on the other, to provide a unifying framework promoting the complementarity of the modalities in respect to different aspects and/or different levels of granularity of the underlying concepts encompassed in the ontology of a (possibly evolving) domain. At the same time, a cross-modality fusion architecture is also expected to naturally bridge the gap between low level mode-specific features and higher level concepts, by facilitating the identification and separation of mode-independent and mode-specific features necessary to capture the ontology concepts.

In particular, fusion techniques are investigated based on:

- Ontology-based disambiguation: The results of the extraction process are used to build hypotheses. These are then matched to information included in the multimedia ontology to build more precise higher-level hypotheses. During this process, conflicts or inconsistencies may be found, prompting the revision of intermediate results, and, possibly, the adjustment of parameters for low-level processing modules to achieve more precise results at higher levels. This is implemented as a closed-loop extraction process.
- Probabilistic evidence-based interpretation framework: The evidence in this case is provided by low-level features and the extracted semantics may belong to intermediate abstraction levels. This thread of thought has been particularly studied in the context of synchronous or quasi-synchronous information streams for audio-visual speech recognition, modelled as variants of HMM [2].

The semantic data is input into the multimedia ontology, which is divided into (a) Multimedia content ontology, which represents content structure (b) Multimedia descriptor ontology, which models concepts and properties that describe visual characteristics of objects including MPEG-7 standard features and (c) Domain-specific ontologies, which contain concepts and properties related to the knowledge of the domain of interest (related application scenario). The multimedia ontology will be then evolved through the mechanisms described in the previous section.

4. Application scenario

The application concerns the enrichment of digital maps with semantic information. The process involves an automatic collection and annotation service for public events in a number of major cities from the Web and proprietary sources. The domain of public events include commercial exhibitions, sport events, concerts etc. The results of the annotation process, i.e., the identified entities and their properties, will be linked to geographical locations and stored in a content server. The user will be provided with immediate access to the annotated content base, through the user-friendly interface of digital maps, which will also provide immediate navigation guidance to the place of interest. The domain-dependent semantic model will be used by the extraction architecture to identify multimedia information related to the concepts in the ontologies. Further, from the extracted information, new concepts will be generated to extend the ontologies, using the evolution architecture.

As a concrete example of the application scenario, consider the domain of vehicle exhibitions, an application that has significant commercial and social interest, while at the same time it is associated with a wealth of complementary multimedia content that is evolving over time. Given such a domain, the following stages will be followed to customize and use the envisaged system (see Fig. 1):

- Initialization: Forming of the initial multimedia semantic model for the domain by collecting and merging existing ontologies for sub-domains, referring for example to car, and motorcycle exhibitions and linking them to multimedia descriptor ontologies, using the ontology initialization and content annotation tool.
- Training: Training of the various semantics extraction and ontology evolution tools to the domain. To that end, a training dataset containing representative and annotated multimedia content will be constructed using the ontology initialization and content annotation tool.
- Information gathering: After customization, the first step of its run-time use is to collect content from various Web and proprietary sources. In the case of car exhibitions, such sources will include TV and news programmes, on-line magazines, the sites or proprietary databases of car suppliers and dealers, specialized discussion fora and Weblogs, as well as generic content sources. Different sources will provide different types of content, which when fused semantics can lead to a rich description of concepts and their instances.
- Semantics extraction: The trained semantics extraction tools will be applied at regular intervals to the incoming stream of multimedia content, performing two parallel tasks: (a) Extracting the relevant information from each piece of content, such as event venue, event dates, event organisers, exhibits and multimedia related to it, exhibitors, etc. Single modalities will be processed separately and then will be fused (b) associating the information to concepts of the ontologies, by identifying their characteristic elements in multimedia content, e.g., terms in the text and audio, objects in visual content, crowd density, indoor/outdoor, etc.
- Ontology evolution: Population of the ontologies with instances of the various concepts, together with their properties (accompanied by annotation). The concept modelling task, performed by the extraction methods, will lead to suggestions for the enrichment of the ontologies, through novelty detection.
- Information positioning and retrieval: The concept instances annotated on the multimedia content will be linked to the map data in the digital maps server. The user will be able to browse content and issue queries about exhibits, events, etc. The results will always be associated to places on the digital map also considering time.

5 Concluding remarks

We proposed 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. This is a synergistic approach, combining 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 improve significantly the performance of existing single-modality approaches in terms of scalability and precision.

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 will also examine ways to learn optimal multimedia-based feature combinations. Synchronization and alignment of the different modalities is another issue, since all modalities must refer to a common timeline.

Ontologies must be sufficiently expressive in order 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. Since existing standard web languages do not provide such capabilities, research effort needs to be directed towards representation and management of uncertainty, imprecision that exists 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 aiming to develop a generic framework for ontology learning and inference from multimedia content.

References

1. Belur V. Dasarathy, Elucidative fusion systems - an exposition, Information Fusion, Vol 1, pp 5-15, 2000.

2. Samy Bengio, Multimodal speech processing using asynchronous Hidden Markov Models, Information Fusion, Volume 5, pp 81-89, 2004

3. Brewster, F. Ciravegna, and Y. Wilks. User-centred ontology learning for knowledge management. In B. Andersson, M. Bergholtz, and P. Johannesson, editors, NLDB, volume 2553 of Lecture Notes in Computer Science, pages 203–207. Springer, 2002.

4. S. Castano, A. Ferrara, S. Montanelli, and G. Racca, “Matching Techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions”, IEEE Proc. of the International Conference on Coding and Computing, Las Vegas, USA, 2004

5. Cees G.M. Snoek, M. Worring Multimodal Video Indexing: A Review of the State-of-the-art Multimedia Tools and Applications, 25, 5–35, 2005

6. OntoWeb. Deliverable D1.3. A survey on ontology tools, 2002 (ed. Gómez Pérez)

7. J. Hunter, "Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology", International Semantic Web Working Symposium (SWWS), Stanford, 2001

8. K. Kotis, G. Vouros. HCONE approach to Ontology Merging. ESWS'04. The Semantic Web: Research and Applications, LNCS, Vol. 3053, Springer-Verlag, (2004)

9. C.D. Manning and H. Schutze, Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, USA, 1999.

10. A Maedche and S. Staab. Mining ontologies from text. In R.Dieng and O.Corby, editors, EKAW, vol. 1937, Lecture Notes in Computer Science, pp. 189–202. Springer, 2000.

11. 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 Transactions on Circuits and Systems for Video Technology, Special Issue on Audio and Video Analysis for Multimedia Interactive Services, vol. 14, pp. 606-621, May 2004.

12. L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE, Vol. 77, No. 2, pp. 257–286, 1989.

13. R. Troncy, “Integrating Structure and Semantics into Audio-Visual Documents”, In the second International Semantic Web Conference, LNCS 2870, pp. 566-581, 2003.