Introduction

Raphaël Troncy,¹ Benoit Huet² and Simon Schenk³
¹Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
²EURECOM, Sophia Antipolis, France
³University of Koblenz-Landau, Koblenz, Germany

Digital multimedia items can be found on most electronic equipment ranging from mobile phones and portable audiovisual devices to desktop computers. Users are able to acquire, create, store, send, edit, browse, and render through such content at an increasingly fast rate. While it becomes easier to generate and store data, it also becomes more difficult to access and locate specific or relevant information. This book addresses directly and in considerable depth the issues related to representing and managing such multimedia items.

The major objective of this book is to gather together and report on recent work that aims to extract and represent the semantics of multimedia items. There has been significant work by the research community aimed at narrowing the large disparity between the low-level descriptors that can be computed automatically from multimedia content and the richness and subjectivity of semantics in user queries and human interpretations of audiovisual media – the so-called semantic gap.

Research in this area is important because the amount of information available as multimedia for the purposes of entertainment, security, teaching or technical documentation is overwhelming but the understanding of the semantics of such data sources is very limited. This means that the ways in which it can be accessed by users are also severely limited and so the full social or economic potential of this content cannot be realized.

Addressing the grand challenge posed by the semantic gap requires a multi-disciplinary approach and this is reflected in recent research in this area. In particular, this book is closely tied to a recent Network of Excellence funded by the Sixth Framework Programme of the European Commission named ‘K-Space’ (Knowledge Space of Semantic Inference for Automatic Annotation and Retrieval of Multimedia Content).

By its very nature, this book is targeted at an interdisciplinary community which incorporates many research communities, ranging from signal processing to knowledge
representation and reasoning. For example, multimedia researchers who deal with signal processing, computer vision, pattern recognition, multimedia analysis, indexing, retrieval and management of ‘raw’ multimedia data are increasingly leveraging methods and tools from the Semantic Web field by considering how to enrich their methods with explicit semantics. Conversely, Semantic Web researchers consider multimedia as an extremely fruitful area of application for their methods and technologies and are actively investigating how to enhance their techniques with results from the multimedia analysis community. A growing community of researchers is now pursuing both approaches in various high-profile projects across the globe. However, it remains difficult for both sides of the divide to communicate with and learn from each other. It is our hope that this book will go some way toward easing this difficulty by presenting recent state-of-the-art results from both communities.

Whenever possible, the approaches presented in this book will be motivated and illustrated by three selected use cases. The use cases have been selected to cover a broad range of multimedia types and real-world scenarios that are relevant to many users on the Web: photos on the Web, music on the Web, and professional audiovisual media production process. The use cases introduce the challenges of media semantics in three different areas: personal photo collection, music consumption, and audiovisual media production as representatives of image, audio, and video content. The use cases, detailed in Chapter 2, motivate the challenges in the field and illustrate the kind of media semantics needed for future use of such content on the Web, and where we have just begun to solve the problem.

Nowadays it is common to associate semantic annotations with media assets. However, there is no agreed way of sharing such information among systems. In Chapter 3 a small number of fundamental processes for media production are presented. The so-called canonical processes are described in the context of two existing systems, related to the personal photo use case: CeWe Color Photo Book and SenseCam.

Feature extraction is the initial step toward multimedia content semantic processing. There has been a lot of work in the signal processing research community over the last two decades toward identifying the most appropriate feature for understanding multimedia content. Chapter 4 provides an overview of some of the most frequently used low-level features, including some from the MPEG-7 standard, to describe audiovisual content. A succinct description of the methodologies employed is also provided. For each of the features relevant to the video use case, a discussion will take place and provide the reader with the essential information about its strengths and weaknesses. The plethora of low-level features available today led the research community to study multi-feature and multi-modal fusion. A brief but concise overview is provided in Chapter 4. Some feature fusion approaches are presented and discussed, highlighting the need for the different features to be studied in a joint fashion.

Machine learning is a field of active research that has applications in a broad range of domains. While humans are able to categorize objects, images or sounds and to place them in specific classes according to some common characteristic or semantics, computers are having difficulties in achieving similar classification. Machine learning can be useful, for example, in learning models for very well-known objects or settings.
Chapter 5 presents some of the main machine learning approaches for setting up an automatic multimedia analysis system. Continuing the information processing flow described in the previous chapter, feature dimensionality reduction methods, supervised and unsupervised classification techniques, and late fusion approaches are described.

The Internet and the Web have become an important communication channel. The Semantic Web improves the Web infrastructure with formal semantics and interlinked data, enabling flexible, reusable, and open knowledge management systems. In Chapter 6, the Semantic Web basics are introduced: the RDF(S) model for knowledge representation, and the existing web infrastructure composed of URIs identifying resources and representations served over the HTTP protocol. The chapter details the importance of open and interlinked Semantic Web datasets, outlines the principles for publishing such linked data on the Web, and discuss some prominent openly available linked data collections. In addition, it shows how RDF(S) can be used to capture and describe domain knowledge in shared ontologies, and how logical inferencing can be used to deduce implicit information based on such domain ontologies.

Having defined the Semantic Web infrastructure, Chapter 7 addresses two questions concerning rich semantics: How can the conceptual knowledge useful for a range of applications be successfully ported to and exploited on the Semantic Web? And how can one access efficiently the information that is represented on these large RDF graphs that constitute the Semantic Web information sphere? Those two issues are addressed through the presentation of SPARQL, the recently standardized Semantic Web Query language, with an emphasis on aspects relevant to querying multimedia metadata represented using RDF in the running examples of COMM annotations.

Chapter 8 presents and discusses a number of commonly used multimedia metadata standards. These standards are compared with respect to a list of assessment criteria using the use cases listed in Chapter 2 as a basis. Through these examples the limitations of the currents standards are exposed. Some initial solutions provided by COMM for automatically converting and mapping between metadata standards are presented and discussed.

A multimedia ontology framework, COMM, that provides a formal semantics for multimedia annotations to enable interoperability of multimedia metadata among media tools is presented in Chapter 9. COMM maps the core functionalities of the MPEG-7 standard to a formal ontology, following an ontology design approach that utilizes the foundational ontology DOLCE to safeguard conceptual clarity and soundness as well as extensibility towards new annotation requirements.

Previous chapters having described multimedia processing and knowledge representation techniques, Chapter 10 examines how their coupling can improve analysis. The algorithms presented in this chapter address the photo use case scenario from two perspectives. The first is a segmentation perspective, using similarity measures and merging criteria defined at a semantic level for refining an initial data-driven segmentation. The second is a classification perspective, where two knowledge-driven approaches are presented. One deals with visual context and treats it as interaction between global classification and local region labels. The other deals with spatial context and formulates the exploitation of it as a global optimization problem.
Chapter 11 demonstrates how different reasoning algorithms upon previously extracted knowledge can be applied to multimedia analysis in order to extract semantics from images and videos. The rich theoretical background, the formality and the soundness of reasoning algorithms can provide a very powerful framework for multimedia analysis. The fuzzy extension of the expressive DL language $SHI^N$, $f$-$SHI^N$, together with the fuzzy reasoning engine, FiRE, that supports it, are presented here. Then, a model using explicitly represented knowledge about the typical spatial arrangements of objects is presented. Fuzzy constraint reasoning is used to represent the problem and to find a solution that provides an optimal labeling with respect to both low-level and spatial features. Finally, the NEST expert system, used for estimating image regions dissimilarity is described.

Multimedia content structuring is to multimedia documents what tables of contents and indexes are to written documents, an efficient way to access relevant information. Chapter 12 shows how combined audio and visual (and sometimes textual) analysis can assist high-level metadata extraction from video content in terms of content structuring and in detection of key events depicted by the content. This is validated through two case studies targeting different kinds of content. A quasi-generic event-level content structuring approach using combined audiovisual analysis and a suitable machine learning paradigm is described. It is also shown that higher-level metadata can be obtained using complementary temporally aligned textual sources.

Chapter 13 reviews several multimedia annotation tools and presents two of them in detail. The Semantic Video Annotation Tool (SVAT) targets professional users in audiovisual media production and archiving and provides an MPEG-7 based framework for annotating audiovisual media. It integrates different methods for automatic structuring of content and provides the means to semantically annotate the content. The K-Space Annotation Tool is a framework for semi-automatic semantic annotation of multimedia content based on COMM. The annotation tools are compared and issues are identified.

Searching large multimedia collection is the topic covered in Chapter 14. Due to the inherently multi-modal nature of multimedia documents there are two major challenges in the development of an efficient multimedia index structure: the extremely high-dimensional feature space representing the content, on the one hand, and the variable types of feature dimensions, on the other hand. The first index function presented here divides a feature space into disjoint subspaces by using a pyramid tree. An index function is proposed for efficient document access. The second one exploits the discrimination ability of a media collection to partition the document set. A new feature space, the feature term, is proposed to facilitate the identification of effective features as well as the development of retrieval models.

In recent years several Semantic Web applications have been developed that support some form of search. Chapter 15 analyzes the state of the art in that domain. The various roles played by semantics in query construction, the core search algorithm and presentation of search results are investigated. The focus is on queries based on simple textual entry forms and queries constructed by navigation (e.g. faceted browsing). A systematic understanding of the different design dimensions that play a role in supporting search on Semantic Web data is provided. The study is conducted in the context of image search and depicts two use cases: one highlights the use of semantic functionalities to support the search, while the other exposes the use of faceted navigation to explore the image collection.
In conclusion, we trust that this book goes some way toward illuminating some recent exciting results in the field of semantic multimedia. From the wide spectrum of topics covered, it is clear that significant effort is being invested by both the Semantic Web and multimedia analysis research communities. We believe that a key objective of both communities should be to continue and broaden interdisciplinary efforts in this field with a view to extending the significant progress made to date.