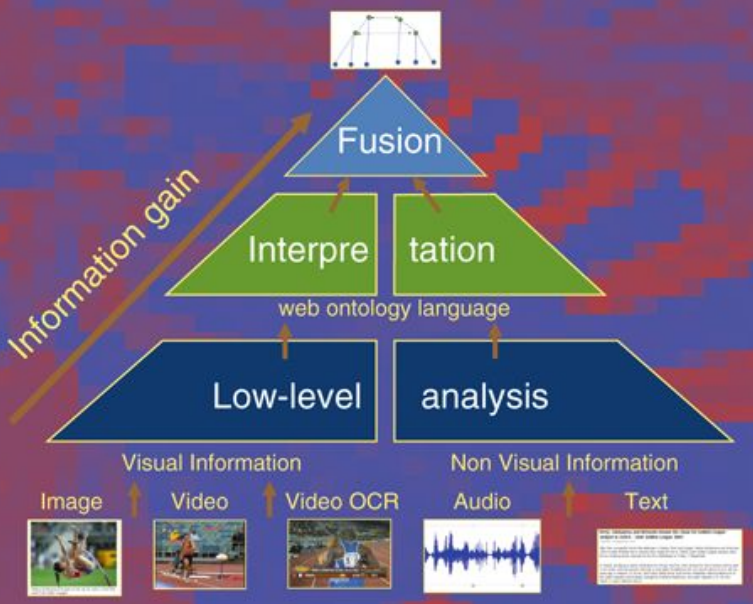
State-of-the-Art Survey

LNAI 6050

Georgios Paliouras
Constantine D. Spyropoulos
George Tsatsaronis (Eds.)

Knowledge-Driven Multimedia Information Extraction and Ontology Evolution

Bridging the Semantic Gap



Springer

Lecture Notes in Artificial Intelligence 6050

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Georgios Paliouras
Constantine D. Spyropoulos
George Tsatsaronis (Eds.)

Knowledge-Driven Multimedia Information Extraction and Ontology Evolution

Bridging the Semantic Gap

Springer

Volume Editors

Georgios Paliouras
Constantine D. Spyropoulos
National Centre for Scientific Research "Demokritos"
Institute of Informatics and Telecommunications
P.O. Box 60228, Ag. Paraskevi, 15310 Athens, Greece
E-mail: {paliourg, costass}@iit.demokritos.gr

George Tsatsaronis
TU Dresden, Biotechnology Center (BIOTEC)
Tatzberg 47-51, 01307 Dresden, Germany
E-mail: george.tsatsaronis@biotec.tu-dresden.de

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-20794-5 e-ISBN 978-3-642-20795-2
DOI 10.1007/978-3-642-20795-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011926569

CR Subject Classification (1998): I.2.4, I.2, H.3, H.2.8, H.5.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This textbook is intended for use by researchers and practitioners in the field of computer science and more specifically in knowledge representation and management, ontology evolution and information extraction from multimedia data. The reader is presumed to have a basic knowledge of knowledge representation with ontologies.

Being authored by acknowledged researchers who participated in the EC-supported project BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction), the book aims to cover the state of the art in the corresponding fields, while also promoting the synergy between ontology evolution and information extraction from multimedia. BOEMIE has shown that this synergy reveals a new area of research that is of great untapped potential.

The book may also constitute an excellent guide to students attending courses of a computer science study program that address information processing and extraction from any type of media (text, images, and video). For the students of computer science, the concepts introduced in this book will provide a sound theoretical framework for the corresponding fields and will hopefully motivate them to join the research community in the effort of developing software that analyzes and "understands" multimedia content. The book also gives concrete examples of applying several of the discussed methods in the athletics (track and field) sports domain.

The first chapter of the book provides an overview of the BOEMIE project and its main achievements. It illustrates the basic bootstrapping framework on which the evolution of ontologies is fed by the extraction of information from multimedia and vice versa. In doing so, it sets the scene for the rest of the book that describes the state of the art in the corresponding fields. It also aims to guide the reader to the technological choices that support the integration of knowledge technologies with multimedia analysis.

Along these lines, the second chapter presents current approaches to the representation of knowledge about multimedia, using ontologies. This chapter illustrates how the aspects of describing multimedia and providing knowledge about a particular domain of application, e.g., sports, can come together in the context of multimedia ontologies. This is the essential "glue" between the different technologies that are involved in the process.

The following two chapters describe the state of the art in extraction methods for two important types of multimedia content, i.e., image and text. The aim is to show how far we can go in understanding what a human perceives by seeing an image or reading a piece of text. Thus, the emphasis is on extracting the semantics from the content at a level that can be linked to the appropriate ontologies.

Once this link of content to ontologies is established, one can employ reasoning methods, in order to combine and interpret the acquired knowledge and obtain an enhanced view of multimedia content. This automated reasoning process is covered by the chapter that follows, where the authors attempt to bridge content and knowledge, in a process inspired by human reasoning, based on perception.

Having reached an adequate level of interpretation, the next step, as proposed by the BOEMIE project, is to attempt to improve the knowledge structures themselves, i.e., evolve the ontologies, based on the extracted information. This is the task covered by the next two chapters, which present the state of the art in ontology learning, population and matching. In other words, they describe methods that can automate the modification of the domain knowledge, which can in turn be used to extract more knowledge from the multimedia content. In this manner the iterative process, proposed by BOEMIE, starts a new cycle of processing.

The last chapter of the book provides a survey of tools that are useful for the annotation of multimedia content with semantics, i.e., concepts and relations that have a particular meaning in the application domain, e.g., sports. Such tools are useful for the manual annotation of training data for the methods presented in the other chapters. Furthermore, they can benefit from the automation proposed by those methods, while combining it with an interactive annotation experience.

January 2011

Georgios Paliouras
Constantine D. Spyropoulos
George Tsatsaronis

Organization

This volume is designed to provide researchers, practitioners, and students with the basic knowledge and skills needed to appreciate the full range of information extraction from multimedia content. The volume contents stemmed largely from the research work conducted over the period of three years under the framework of the BOEMIE research project (IST 6th Framework Programme - FP6-027538 Project).

Editorial Team

Georgios Paliouras	Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos," Greece
Constantine D. Spyropoulos	Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos," Greece
George Tsatsaronis	Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos," Greece Biotechnology Center (BIOTEC), Technische Universität Dresden, Germany

Reviewers

The Editors wish to express their gratitude to the following 'anonymous' reviewers who gave their time and energy during the last 12 months, ensuring the high quality of the published chapters.

Enrique Alfonseca
Werner Bailer
Roberto Basili
Tobias Buerger
Fabio Ciravegna
Marc Ehrig
Jerome Euzenat
George Flouris
Patrick Gross
Paola Hobson
Jose Iria
Michael Kipp
Ferran Marques
Claire Nedellec
Jeff Pan
Marta Sabou
Hichem Sahli
Alan Smeaton
Heiko Stoermer
Umberto Straccia
Sofia Tsekeridou
Chrysa Tsinaraki
Paola Velardi
George Vouros

Table of Contents

Bootstrapping Ontology Evolution with Multimedia Information Extraction

Georgios Paliouras, Constantine D. Spyropoulos, and George Tsatsaronis

Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos",
15310, Ag. Paraskevi, Attiki, Greece
{paliourg,costass,gbt}@iit.demokritos.gr

Abstract. This chapter summarises the approach and main achievements of the research project BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction). BOEMIE introduced a new approach towards the automation of knowledge acquisition from multimedia content. In particular, it developed and demonstrated the notion of evolving multimedia ontologies, which is used for the extraction, fusion and interpretation of information from content of various media types (audio, video, images and text). BOEMIE adopted a synergistic approach that combines multimedia extraction and ontology evolution in a bootstrapping process. This process involves, 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 robustness of the extraction system. Thus, in addition to annotating multimedia content with semantics, the extracted knowledge is used to expand our understanding of the domain and extract even more useful knowledge. The methods and technologies developed in BOEMIE were tested in the domain of athletics, using large sets of annotated content and evaluation by domain experts. The evaluation has proved the value of the technology, which is applicable in a wide spectrum of domains that are based on multimedia content.

1 Motivation and Objectives of the BOEMIE Project

BOEMIE[1] aimed towards the automation of the knowledge acquisition process from multimedia content, which nowadays grows with increasing rates in both public and proprietary webs. Towards this end, it introduced the concept of *evolving multimedia ontologies*. The project was unique in that it linked multimedia extraction with ontology evolution, creating a synergy of great potential.

In recent years, significant advances have been made in the area of automated extraction of low-level features from audio and visual content. However, little progress has been achieved in the identification of high-level semantic features

[1] http://www.boemie.org/

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 1–17, 2011.

or the effective combination of semantic features derived from various modalities. Driven by domain-specific multimedia ontologies, BOEMIE information extraction systems are able to identify high-level semantic features in image, video, audio and text and fuse these features for improved extraction. The ontologies are continuously populated and enriched using the extracted semantic content. This is a bootstrapping process, since the enriched ontologies in turn drive the multimedia information extraction system. Figure 1 provides a graphical illustration of this iterative bootstrapping process, that is implemented in the BOEMIE prototype. The main proposal of the project is illustrated by the continuous iteration that resides at the heart of the process. Information extraction is driven by semantic knowledge, while feeding at the same time the evolution of the ontologies.

Through the proposed synergistic approach, BOEMIE aimed at large-scale and precise knowledge acquisition from multimedia content. More specifically, the objectives of the project were:

Unifying representation for domain and multimedia knowledge. This multimedia semantic model follows modular knowledge engineering principles and captures the different types of knowledge involved in knowledge acquisition from multimedia. It realises the linking of domain-specific ontologies, which model salient subject matter entities, and multimedia ontologies, which capture structural and low-level content descriptions.

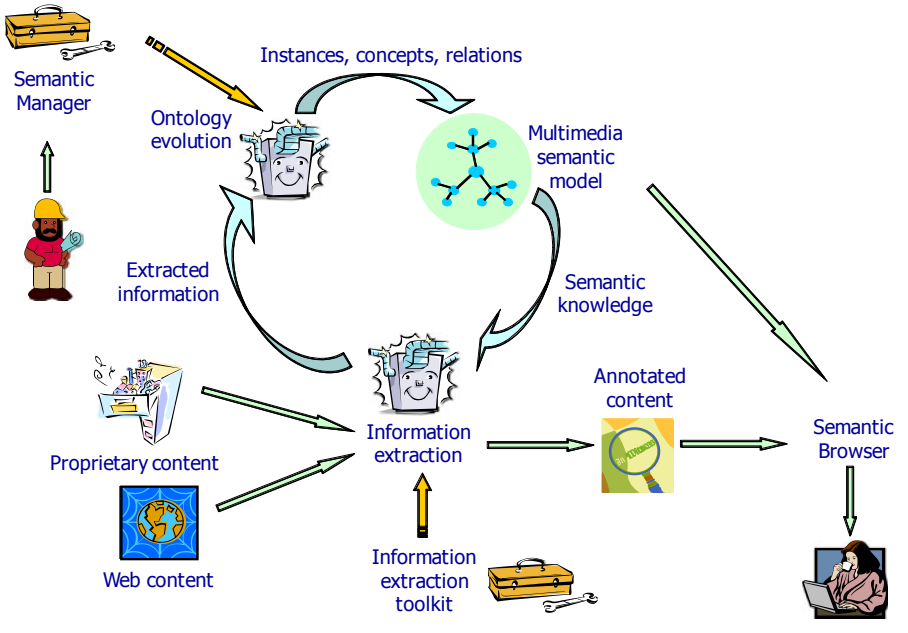


Fig. 1. The BOEMIE bootstrapping process

Methodology and toolkit for ontology evolution. The proposed methodology coordinates the various tools that use the extracted data to populate and enrich the ontologies. The toolkit provides tools to support ontology learning, ontology merging and alignment, semantic inference and ontology management.

Methodology and toolkit for information extraction. The methodology specifies how information from the multimedia semantic model can be used to achieve extraction from various media. Additionally, it fuses information extracted from multiple media to improve the extraction performance. The toolkit comprises tools to support extraction from image, audio, video and text, as well as information fusion.

The resulting technology has a wide range of applications in commerce, tourism, e-science, etc. One of the goals of the project was to evaluate the technology, through the development of an automated content collection and annotation service for athletics events in a number of major European cities. The extracted semantic information enriches a digital map, which provides an innovative and friendly way for the end user to access the multimedia content. Figure 2 illustrates this interaction of the end user with the system, which is provided by a specialised Web application, called the *BOEMIE semantic browser*. Points of interest that are associated with interesting multimedia content are highlighted on the map. The geo-referencing of the content is facilitated by the information extraction process of BOEMIE.

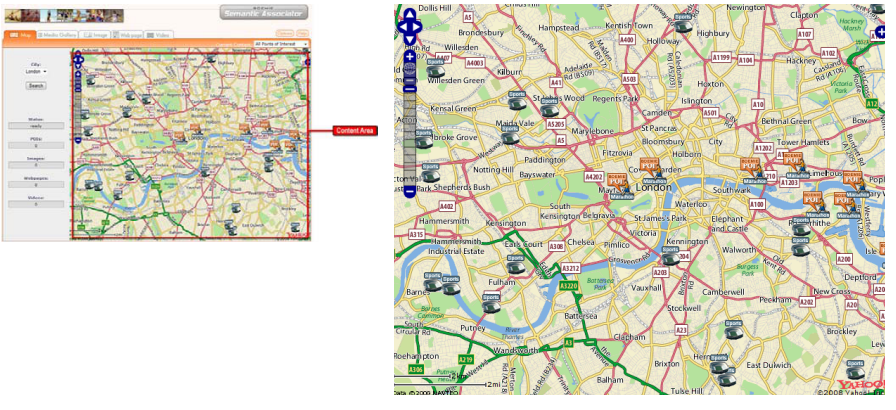


Fig. 2. The map-based interface to multimedia content in the semantic browser

The rest of this chapter is structured as follows. Section 2 presents briefly the main modules of the prototype system that was developed in BOEMIE. Section 3 compares BOEMIE to related projects that took place either before or in parallel with it. Finally, section 4 summarizes the main achievements of the project and proposes interesting paths for further research.

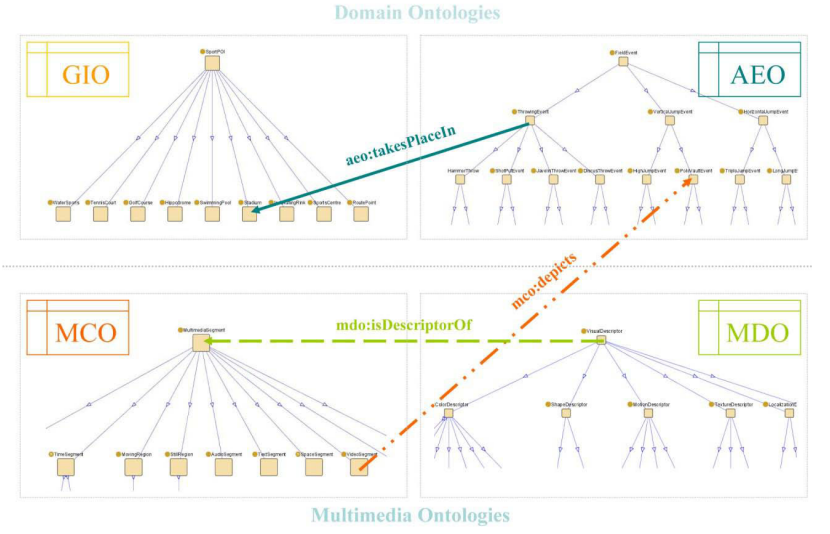


Fig. 3. The Multimedia Semantic Model. AEO (Athletics Event Ontology) models the scenario domain of interest, i.e. public athletics events. GIO (Geographic Information Ontology) models information relevant to geographic objects. MCO (Multimedia Content Ontology) models content structure descriptions, based on MPEG-7 MDS definitions. MDO (Multimedia Descriptor Ontology) models the MPEG-7 visual and audio descriptors.

2 The BOEMIE Prototype

More than 100 different modules and components have been produced in the course of the BOEMIE project, some of which have been made available publicly[2]. Most of the components that were produced have been incorporated in the integrated prototype that was delivered and evaluated at the end of the project. The BOEMIE integrated prototype implements the bootstrapping process, as illustrated in Figure 1. This sketch shows also the main components of the prototype, which are described in the remaining of this section.

2.1 Multimedia Semantic Model

The BOEMIE Multimedia Semantic Model (MSM) [12,11] integrates ontologies that capture our knowledge about a particular domain, e.g. athletics, with ontologies that model knowledge about the structure and low-level descriptors pertaining to multimedia documents (Figure 3).

Besides addressing the interlinking of multimedia document segments with the corresponding domain entities, MSM further enhances the engineering of

[2] `http://www.boemie.org/software`

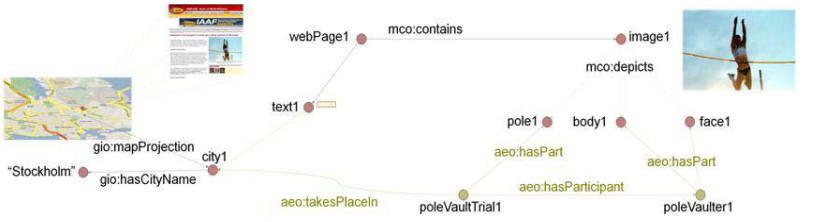


Fig. 4. Linking knowledge between ontologies in the semantic model

subject matter descriptions by distinguishing between *mid-level* (MLC) and *high-level* (HLC) domain concepts and properties, a feature unique to the BOEMIE project. Instances of MLCs represent information that is directly extracted from the multimedia content, using the various analysis tools, e.g. the name of an athlete or her body in a picture. On the other hand, instances of HLCs are generated through reasoning-based interpretation of the multimedia content, using the domain ontology. Such engineering allows incorporating the analysis perspective into the domain conceptualisation, which in turn supports effective logic-grounded interpretation. The developed ontologies allow the utilisation of precise formal semantics throughout the chain of tasks involved in the acquisition and deployment of multimedia content semantics.

In MSM, four OWL DL ontologies are linked in a way that supports the purposes of BOEMIE for semantics extraction, interpretation, evolution, as well as retrieval and representation of the acquired semantics. Figure 4 presents a simple example of this interlinking between the ontologies. This is also the main novelty of the BOEMIE Multimedia Semantic Model.

2.2 Recursive Media Decomposition and Fusion

The information extraction toolkit of BOEMIE integrates a number of tools for content analysis and interpretation, using a recursive media decomposition and fusion framework. In the course of the project, innovative methods for the analysis of single-modality content were developed, going in most cases beyond the state-of-the-art. As Figure 5 illustrates, these methods cover most of the currently available types of media. Most importantly, they support the bootstrapping process through an evolving cycle of analysis of new content, learning of improved analysis models and discovering interesting objects and entities to add to the domain knowledge.

The coordination of the evolving extraction process is achieved by a new method that was developed in BOEMIE and is called Recursive Media Decomposition and Fusion (RMDF) [21]. The method decomposes a multimedia document into its constituent parts, including embedded text in images and speech. It then relies on single-modality modules, the results of which are fused together in a common graph that complies with the domain ontology. In a final step, graph techniques are used to provide a consistent overall analysis of the multimedia

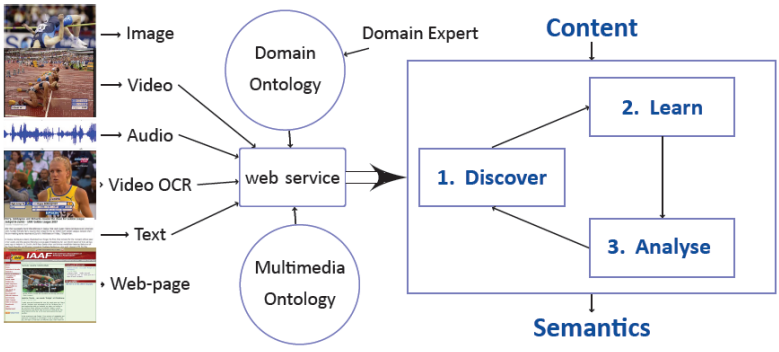


Fig. 5. The Information Extraction Toolkit

document. For instance, a Web page may be decomposed into several parts, one of which containing a video, which may in turn contain a static overlaid image, that embeds text, which refers to a person. This example shows the importance of recursive decomposition and corresponding fusion of results that come from video and text analysis.

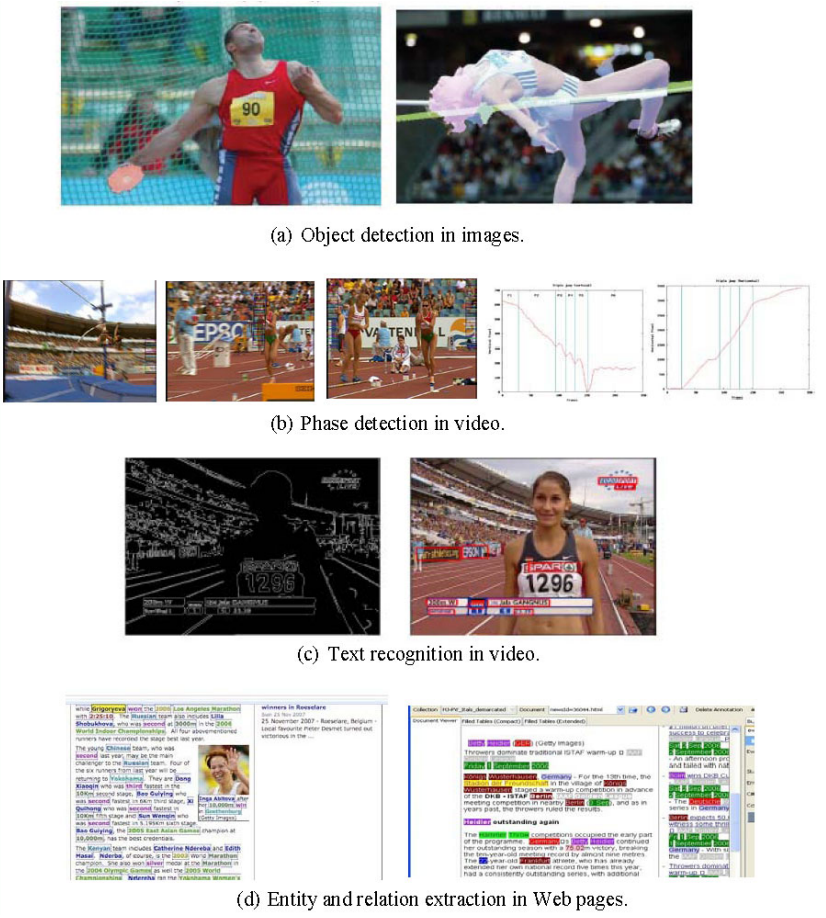
Regarding the single-modality modules, BOEMIE has developed innovative methods to:

- detect and discover objects of various shapes and sizes in images [22],
- track moving objects in video and classify movement phases,
- identify and discover entities in text and relations amongst them [14,20],
- detect overlay and scene text in video and perform optical character recognition on it [23,1],
- recognise and discover audio events and interesting keywords in audio [4,5].

Figure 6 provides examples of such results. Most importantly, however, through interpretation and fusion, the RMDF is able to improve significantly the precision of multimedia analysis, be it in Web pages containing HTML text and images or video footage with audio commentary and overlay text. In addition to the novel decomposition and fusion approach, the single-modality tools support customization to any domain, by allowing the discovery of new semantics in content and learning to identify known objects and entities. Furthermore, the extraction toolkit is easily distributable and scalable, by dynamically integrating per media analysis techniques in an unrestricted number of servers, communicating through a computer network.

2.3 Abductive Multimedia Interpretation

The interpretation of multimedia content in BOEMIE goes well beyond the usual extraction of semantics from individual media. Domain knowledge, in the form of ontologies, is being exploited by a reasoning-based interpretation service that



(a) Object detection in images.

(b) Phase detection in video.

(c) Text recognition in video.

(d) Entity and relation extraction in Web pages.

Fig. 6. Sample results of single-media analysis tools

operates in two levels: single-media interpretation and fusion. The interlinking of domain and multimedia ontologies in the semantic model (Fig. 4) support this process. Figure 7 illustrates the multi-level analysis and interpretation process. Both the single-media and the fusion services are supported by the same reasoning apparatus.

Reasoning for multimedia interpretation is based on the RacerPro reasoning engine[3], which has been extended with many novel methods for the purposes of BOEMIE [3,17,16]. One of the main extensions is the use of abduction to generate interpretation hypotheses for what has been "observed" by the extraction tools. The new abductive query answering service of RacerPro is able, during query evaluation, to recognize non-entailed query atoms and hypothesize them. Since there might be more than one hypothesis (i.e. explanations), a set of

[3] `http://www.racer-systems.com/products/racerpro/`

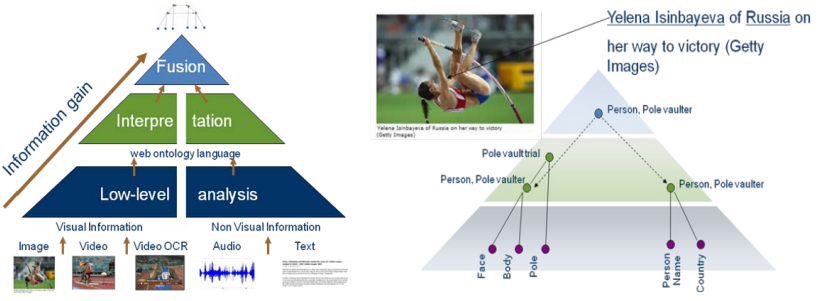


Fig. 7. Multimedia analysis and interpretation process

scoring functions has been designed and implemented in order to prefer certain hypotheses over others. Given the complexity of the interpretation hypothesis (a.k.a. explanation) space, important optimizations have been developed in the reasoner, in order to cut down on the number of consistent and useful interpretations that are produced by the system.

The novel abductive multimedia interpretation machinery of BOEMIE combines Description Logics, as a representation formalism for ontologies, with DL-safe rules that guide the search for interpretations. In the context of BOEMIE, methods to learn these rules have also been developed.

2.4 Pattern-Based Ontology Evolution

The ontology evolution toolkit (OET) implements a pattern-based approach to the population and enrichment of the ontology, which is unique to BOEMIE [10]. In particular, two different cases have been identified for the population process, one in which a single interpretation is produced for a document and one in which more than one candidate interpretation is provided. Furthermore, two cases are defined for the enrichment process, one in which a high-level concept (HLC) and one in which a mid-level concept (MLC) is added. Each of those cases requires different handling in terms of the interaction with the domain expert and the modules that are employed for the semi-automated generation of new knowledge, e.g. concept enhancement, generation of relations and interpretation rules, etc. Figure 8 provides a high-level overview of these four cases (patterns P1 to P4) and the modules that are involved.

The first two cases (P1 and P2), which are responsible for the population of the ontology with new instances, are primarily based on instance matching and grouping methods[7,9]. Novel methods have been developed for this purpose, in order to take advantage of the rich semantics of the BOEMIE semantic model and scale efficiently to large document sets. These methods have been incorporated in the HMatch ontology matching software [6], which is publicly available[4].

A number of innovations have been made also in the area of ontology enrichment (patterns P3 and P4) [18,15]. The discovery of new concepts and properties

[4] tt http://islab.dico.unimi.it/hmatch/

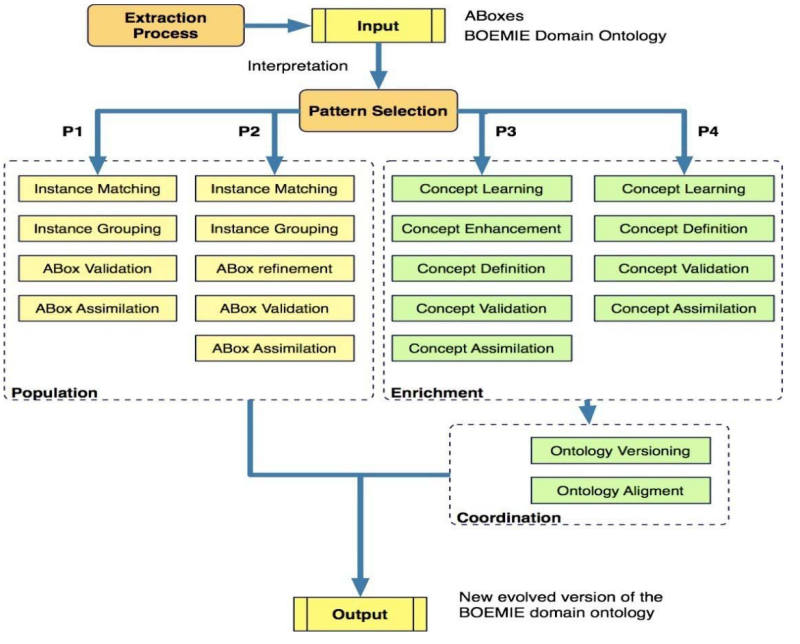


Fig. 8. Pattern-based ontology evolution

is based on a new methodology that incorporates a set of ontology modification operators. Logical and statistical criteria are introduced for the choice of the most appropriate modifications to the ontology, given the observed data. Further to this data-driven enrichment, a concept enhancement method has been developed, matching new constructs to knowledge in external resources, e.g. on the Web.

2.5 Interface Components

In addition to the core processing components, the BOEMIE prototype includes a number of interface components that facilitate the interaction of the users with the prototype, as well as the interaction among the components. Three of these components, which introduce a number of novel features are the *Semantic Browser*, the *Semantic Manager* and the *Bootstrapping Controller*.

The BOEMIE Semantic Browser (BSB) provides an innovative interaction experience with multimedia content. It does so by supporting three modes of interaction with the multimedia content:

- Interactive maps for multimedia retrieval.
- Interactive content of media objects.
- Dynamic suggestion of related information.

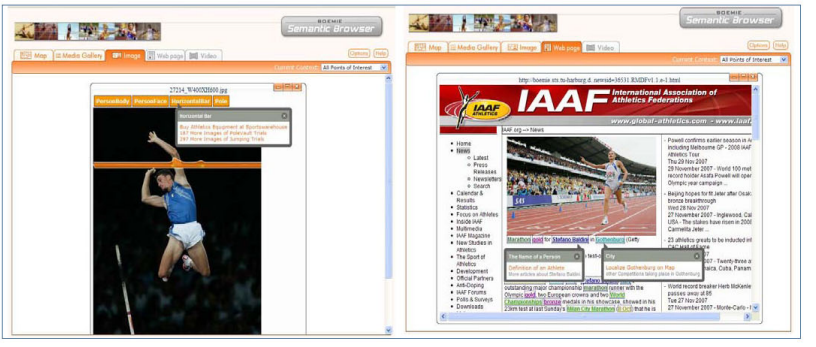


Fig. 9. Suggesting information related to active media objects

A screenshot of the map interface of BSB is shown in Fig. 2. Based on the information extraction technologies of BOEMIE, BSB can associate multimedia content with geopolitical areas and specific Points of Interest (PoIs) on digital maps. In this manner it provides direct access to the multimedia, through what we call "BOEMIE PoIs". Furthermore, BSB uses the semantic annotations generated automatically by BOEMIE to make media objects interactive. More specifically, it automatically highlights relevant content of a specific domain on top of text or images to prepare the interface for further interaction possibilities. Finally, through interpretation, BOEMIE is able to generated deeper semantic information, e.g. the type of sport that an image depicts. Using this implicit knowledge, BSB provides context-sensitive advertisement and suggests related information. This is realized by the idea of context menus, illustrated in Fig. 9.

The BOEMIE Semantic Manager (BSM) [8] is unique in its simplification of a complex and demanding process, i.e., that of adding semantics to multimedia content and maintaining the associated domain knowledge. As an interface to the OET, BSM provides three primary functionalities:

- Population of the ontology with semantically annotated multimedia content.
- Enrichment of the ontology with new knowledge that has been learned from data.
- User-friendly interactive enhancement of new knowledge by the domain expert.

BSM provides interactive selection/approval/rejection of the multimedia content interpretations automatically produced by the BOEMIE system, as well as (similarity-based) document browsing facilities. In order to make the process accessible to the non-skilled in knowledge engineering, it creates a natural language description of the underlying logic representation of ontology instances. Additionally, BSM provides terminological and structural suggestions to support the domain expert in performing ontology enrichment. Suggestions are dynamically extracted from knowledge chunks similar to a given concept proposal by

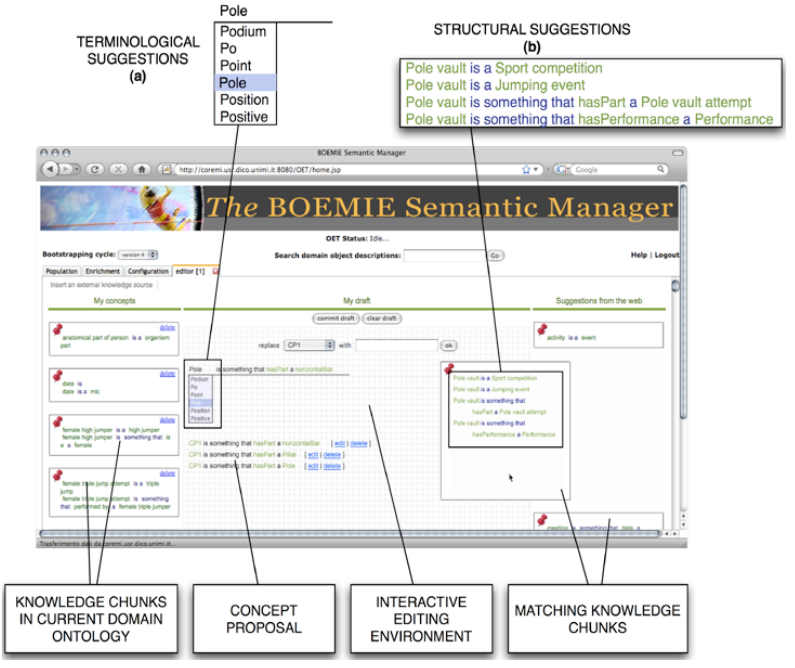


Fig. 10. Concept definition and enhancement, using the semantic manager

exploiting ontology matching techniques. A repository of knowledge chunks is created and maintained through a knowledge harvesting process that periodically searches knowledge of interest from other ontologies, web directories, and, in general, external knowledge repositories. Finally, BSM incorporates a simple ontology editor [13], which uses natural language patterns and autocompletion techniques to facilitate the incorporation of new knowledge to the domain ontology. Figure 10 illustrates this editor.

The Bootstrapping Controller (BSC) is the main application logic component that implements the iterative extraction and evolution process. Using the BSC, the content owner can add content to the Multimedia Repository (MMR) and then send it for processing through predefined workflows. The content is added to the repository, either by uploading specific files or by crawling the Web. Typically, a new document will be sent to RMDF for extraction and the results of its interpretation will be populated into the ontology. When sufficient evidence is accumulated, the OET will generate proposals for changes to the ontology. The domain expert will use these recommendations to change the ontology and the content will be sent again for processing by the RMDF. In some cases, new mid-level concepts (MLCs) will be generated based on the analysis of the multimedia content so far. In these cases, in addition to the extension of the ontologies, the BSC will send sufficient training data to the RMDF, asking for the re-training of the analysis modules.

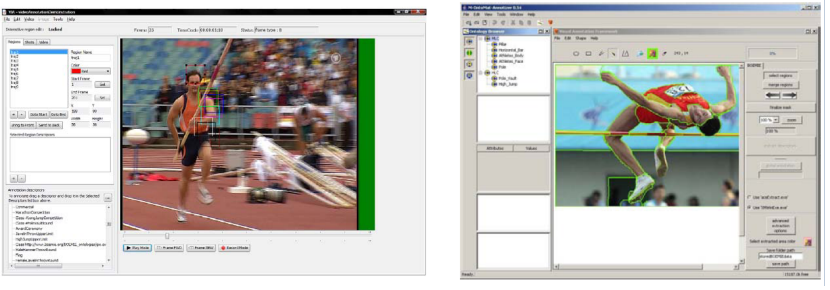


Fig. 11. Video and image annotation with the VIA tool

2.6 Manual Annotation Tools

The information extraction methods developed in BOEMIE are trainable and therefore require training material, in order to learn to identify interesting entities, objects and relations among them in multimedia content. The BOEMIE bootstrapping process generates semi-automatically such training data. However, for the purposes of training and evaluating the initial extractors, we generated significant quantities of training data for all types of media: image, video, audio, text. For these data we used interactive tools for manual annotation. Most of these tools were also developed in BOEMIE and improve significantly the state of the art in the field.

The VIA tool[5] can be used for high-level and low-level video and image annotation. In both cases, annotation is aligned with concepts of the domain ontologies. In the case of image annotation, either image regions and complete images are linked with concepts (high-level annotation) or visual descriptors are extracted per annotated region and associated with the corresponding concept (low-level annotation). To reduce the manual annotation burden, VIA supports the automatic segmentation of a still image into regions and region-merging. Regarding video annotation, VIA supports input in MPEG1/2 video format and frame accurate video playback and navigation. Video annotation can take place either in a frame-by-frame style or as live annotation during playback. Figure 11 illustrates the use of VIA.

The text and HTML annotation tool BTAT[6] [19] has been developed over the Ellogon open-source text engineering platform[7]. It supports the annotation of named entities, the mid-level concepts (MLCs), as well as relations between those named entities. The relations are grouped in tables of specific types. Tables correspond to high-level concepts (HLCs). Furthermore, the tool enables the annotation of relations between HLC instances by creating links between tables in an effective and easy way. One of the innovations in BTAT is its dual manual and automated annotation functionality. Manual annotation is facilitated by a smart

[5] http://mklab.iti.gr/project/via
[6] http://www.boemie.org/btat
[7] http://www.ellogon.org/

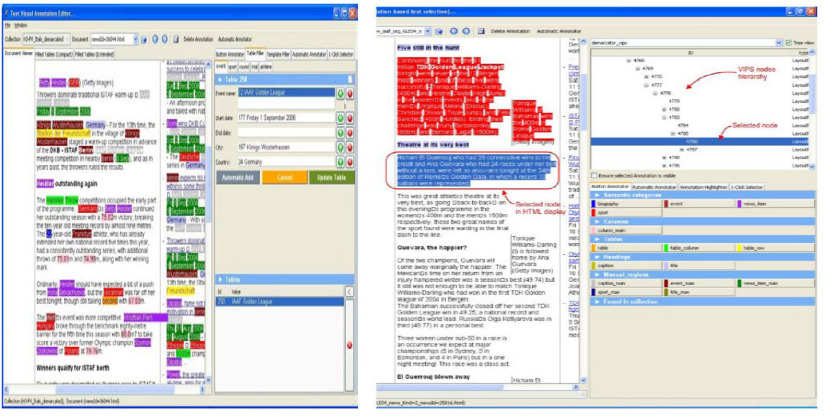


Fig. 12. Text and HTML annotation with the BTAT tool

text-marking system, where the user selects with a mouse click words, instead of single characters. Automatic annotation works by matching user-defined regular expressions. Figure 12 illustrates the use of BTAT.

3 Related Projects

BOEMIE is part of a larger research effort to provide easier access to the ever increasing quantities of multimedia content, particularly on the Web. As such it was preceded and followed by a number of related efforts that influenced or were influenced by it. An extensive list of related projects provided on the BOEMIE Web site[8]. In this section, we refer to a selection of these projects that we consider most relevant.

BOEMIE brought together experts from a number of research fields, who have worked on the analysis of textual, visual and audio content in the past. Even before the beginning of BOEMIE, some of these efforts aimed at the extraction of high-level information, i.e. semantics, from such content. One of the early efforts in this direction was the SCHEMA project[9], which developed technology for content-based analysis and retrieval of multimedia, using ontologies. SCHEMA was followed by the AceMedia project[10], which went further into defining ontologies for multimedia and developing an initial version of a manual annotation tool for images. In parallel, the CROSSMARC project[11] developed ontology-based information extraction technology for cross-lingual text documents, including initial attempts to introduce ontology evolution methods into the process.

[8] http://www.boemie.org/sites/default/files/BOEMIE_related_projects_with_contacts.pdf
[9] http://www.iti.gr/SCHEMA/
[10] http://cordis.europa.eu/ist/kct/acemedia_synopsis.htm
[11] http://labs-repos.iit.demokritos.gr/skel/crossmarc/

Apart from the projects in which BOEMIE partners participated actively, there was a range of other projects on which BOEMIE was based. European networks of excellence, focusing primarily on infrastructure work, were among the major sources of knowledge for BOEMIE. In particular, the OntoWeb project[12] and its follow-up KnowledgeWeb[13] provided interesting survey reports and resources on semantic Web research, which were valuable in the early stages of BOEMIE. Furthermore, the MUSCLE project[14] was a useful source of information about multimedia semantics extraction technology. Apart from the networks of excellence, two European projects that run in parallel to BOEMIE are worth-mentioning: the SEKT[15] and the ALVIS[16] projects. SEKT emphasized on semantic Web ontologies, including ontology matching and enrichment, while ALVIS focussed on semantic search. Beyond European research, there were a number of other projects, particularly in the US, that have provided useful input to BOEMIE. Among these, the projects Marvel[17] and IMKA are particularly worth-mentioning. MARVEL was an IBM research project that aimed at large-scale classification of visual content, while IMKA [2] was a project by the University of Columbia, emphasizing knowledge representation for multimedia.

A number of other interesting projects either run in parallel or followed BOEMIE. Members of the BOEMIE consortium contributed to some them. The Mesh project[18] developed ontology-based extraction technology for multimedia news content, while the LIVE project[19] focused on real-time video search and editing for news broadcasting. In the same industry, CASAM[20] develops technology for computer-assisted annotation of video content by news editors. On the other hand, the WeKnowIT project[21] advocates the importance of social generation of multimedia content, which is centered around important events. Finally, following the series of networks of excellence, the K-Space project[22] produces useful infrastructure for ontology-based analysis of multimedia content.

Apart from the projects in which BOEMIE partners participate, a range of other projects, related to BOEMIE have started. Among these, the projects X-media[23], Vidi-Video[24] and Vitalas[25] aimed to move into large-scale semantic indexing and retrieval of multimedia content, in particular image and video.

[12] http://www.ontoweb.org/
[13] http://cordis.europa.eu/ist/kct/knowledgeweb_synopsis.htm
[14] http://cordis.europa.eu/ist/kct/muscle_synopsis.htm
[15] http://cordis.europa.eu/ist/kct/sekt_synopsis.htm
[16] http://cordis.europa.eu/ist/kct/alvis_synopsis.htm
[17] http://www.research.ibm.com/marvel/
[18] http://www.mesh-ip.eu/?Page=Project
[19] http://www.ist-live.org/
[20] http://www.casam-project.eu/
[21] http://www.weknowit.eu/
[22] http://cordis.europa.eu/ist/kct/kspace-synopsis.htm
[23] http://cordis.europa.eu/ist/kct/x-media_synopsis.htm
[24] http://cordis.europa.eu/ist/kct/vidivideo_synopsis.htm
[25] http://vitalas.ercim.org/

On the other hand, the Bootstrep project[26] studies new methods for the continuous evolution of ontologies and LarKC[27] aims to provide reasoning technology that could be used with large knowledge bases.

4 Summary and Open Issues

The BOEMIE project has brought tightly together two complementary technologies, namely information extraction from multimedia and ontology evolution. It has done so, by introducing a bootstrapping framework, within which each of the two technologies iteratively feeds the other. In this framework, multimedia content of all known types is analysed and interpreted, using multimedia ontologies. The results of multimedia interpretation are used to populate the ontologies and initiate the enrichment of domain knowledge. The improved ontologies can then be used to re-analyse the multimedia content and extract additional information that is translated into new knowledge.

We consider the methods and technology developed in BOEMIE an important contribution to the field of semantic analysis of multimedia. More recent work in this field has emphasized the scalability of the process, aiming to make the corresponding technologies applicable to as large sets of content as the Web itself. The move from the old Web to its social counterpart has intensified the need for such technology, due to the unprecedented volume of media generation by non-expert users. It is thus to be expected that adaptive solutions, such as those proposed by BOEMIE, will play an increasingly important role in the future of multimedia analysis and search.

Acknowledgements

BOEMIE was a collaborative project that was partially funded by the European Commission, under contract number FP6-IST-027538. The work presented here is the result of the common effort of more than 30 people from the following six research teams: Institute of Informatics and Telecommunications (National Centre for Scientific Research "Demokritos", Greece), Institute for Intelligent Analysis and Information Systems (Fraunhofer Gesellschaft, Germany), Institute of Software Systems (Hamburg University of Technology, Germany), Department of Informatics and Communication (Università degli Studi di Milano, Italy), Informatics and Telematics Institute (Centre for Research and Technology - Hellas, Greece), Tele Atlas N.V. (The Netherlands).

References

1. Anthimopoulos, M., Gatos, B., Pratikakis, I.: Multiresolution text detection in video frames. In: Proceedings of the 2nd International Conference on Computer Vision Theory and Applications (VISAPP), Barcelona, Spain, pp. 161–166 (March 2007)

[26] http://cordis.europa.eu/ist/kct/bootstrep-synopsis.htm

[27] http://www.larkc.eu/

2. Benitez, A.B., Chang, S.-F., Smith, J.R.: Imka: A multimedia organization system combining perceptual and semantic knowledge. In: Proceedings of the Ninth ACM International Conference on Multimedia (MULTIMEDIA), pp. 630–631 (2001)
3. Berger, T., Kaplunova, A., Kaya, A., Möller, R.: Towards a scalable and efficient middleware for instance retrieval inference services. In: Proceedings of the 3rd International Workshop on OWL: Experiences and Directions (OWLED), co-located with the 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria (June 2007)
4. Biatov, K.: Two level discriminative training for audio event recognition in sport broadcasts. In: Proceedings of the 12th International Conference on Speech and Computer (SPECOM), Moscow, Russia (October 2007)
5. Biatov, K., Hesseler, W., Kohler, J.: Audio data retrieval and recognition using model selection criterion. In: Proceedings of the 2nd International Conference on Signal Processing and Communication Systems, Gold Coast, Australia (December 2008)
6. Castano, S., Ferrara, A., Lorusso, D., Montanelli, S.: The hmatch 2.0 suite for ontology matchmaking. In: Proceedings of the 4th Italian Workshop on Semantic Web Applications and Perspectives (SWAP), Bari, Italy (December 2007)
7. Castano, S., Ferrara, A., Lorusso, D., Montanelli, S.: On the ontology instance matching problem. In: Proceedings of the 19th International Conference on Database and Expert Systems Applications (DEXA), pp. 180–184 (2008)
8. Castano, S., Ferrara, A., Montanelli, S.: Evolving multimedia ontologies: the bsm tool environment. In: Proceedings of 17th Italian Symposium on Advanced Database Systems (SEBD), Camogli, Italy (June 2009)
9. Castano, S., Ferrara, A., Montanelli, S., Lorusso, D.: Instance matching for ontology population. In: Proceedings of the Sixteenth Italian Symposium on Advanced Database Systems (SEBD), Mondello, Italy (June 2008)
10. Castano, S., Peraldi, S.E., Ferrara, A., Karkaletsis, V., Kaya, A.: Multimedia interpretation for dynamic ontology evolution. Journal of Logic and Computation 19(5), 859–897 (2009)
11. Dasiopoulou, S., Tzouvaras, V., Kompatsiaris, I., Strintzis, M.G.: Capturing mpeg-7 semantics. In: International Conference on Metadata and Semantics Research (MTSR), Corfu, Greece, pp. 113–122 (2007)
12. Dasiopoulou, S., Tzouvaras, V., Kompatsiaris, I., Strintzis, M.G.: Enquiring mpeg-7 based ontologies. Multimedia Tools and Applications 46(2), 331–370 (2010)
13. Ferrara, A., Montanelli, S., Varese, S.C.G.: Ontology knowledge authoring by natural language empowerment. In: IEEE Proceedings of the 1st DEXA Int. Workshop on Modelling and Visualization of XML and Semantic Web Data (MoViX), Linz, Austria (September 2009)
14. Fragkou, P., Petasis, G., Theodorakos, A., Karkaletsis, V., Spyropoulos, C.D.: Boemie ontology-based text annotation tool. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco (May 2008)
15. Peraldi, S.E., Kaya, A.: Ontology and rule design patterns for multimedia interpretation. In: Yamaguchi, T. (ed.) PAKM 2008. LNCS (LNAI), vol. 5345. Springer, Heidelberg (2008)
16. Peraldi, S.E., Kaya, A., Melzer, S., Möller, R.: On ontology based abduction for text interpretation. In: Gelbukh, A. (ed.) CICLing 2008. LNCS, vol. 4919, pp. 194–205. Springer, Heidelberg (2008)

17. Peraldi, S.E., Kaya, A., Melzer, S., Möller, R., Wessel, M.: Towards a media interpretation framework for the semantic web. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI), Silicon Valley, USA, pp. 374–380 (November 2007)
18. Petasis, G., Karkaletsis, V., Krithara, A., Paliouras, G., Spyropoulos, C.D.: Semi-automated ontology learning: the boemie approach. In: Proceedings of the First ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web, Heraklion, Greece (June 2009)
19. Petasis, G., Fragkou, P., Theodorakos, A., Karkaletsis, V., Spyropoulos, C.D.: Segmenting html pages using visual and semantic information. In: Proceedings of the 4th Web as Corpus Workshop: Can we do better than Google?, at the 6th International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco (May 2008)
20. Petasis, G., Karkaletsis, V., Paliouras, G., Spyropoulos, C.D.: Learning context-free grammars to extract relations from text. In: Proceedings of the 18th European Conference on Artificial Intelligence (ECAI), Patras, Greece, pp. 303–307 (July 2008)
21. Petridis, S., Perantonis, S.J.: Semantics extraction from multimedia data: an ontology-based machine learning approach. In: Cutsuridis, A.H.V., Taylor, J.G. (eds.) Perception-action Cycle: Models, Architectures and Hardware (2011)
22. Tsapatsoulis, N., Petridis, S., Perantonis, S.J.: On the use of spatial relations between objects for image classification. In: Proceedings of the 4th IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI), Athens, Greece (September 2007)
23. Vamvakas, G., Gatos, B., Pratikakis, I., Stamatopoulos, N., Roniotis, A., Perantonis, S.J.: Hybrid off-line ocr for isolated handwritten greek characters. In: Proceedings of the 4th IASTED International Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA), Innsbruck, Austria (February 2007)

Semantic Representation of Multimedia Content

Kalliopi Dalakleidi[1], Stamatia Dasiopoulou[2], Giorgos Stoilos[1], Vassilis Tzouvaras[1], Giorgos Stamou[1], and Yiannis Kompatsiaris[2]

[1] School of Electrical and Computer Engineering,
National Technical University of Athens,
Zographou 15780, Athens, Greece
gstoil@image.ece.ntua.gr
[2] Informatics and Telematics Institute
Centre for Research and Technology Hellas,
Thermi-Thessaloniki, Greece
dasiop@iti.gr

Abstract. Multimedia documents constitute extremely rich information resources, whose efficacious management is intertwined with the effective capturing of the underlying semantics. The conveyed meaning may span along multiple levels and relates to search and retrieval tasks, much as to the very extraction and interpretation of content descriptions. In this chapter we consider the formal representation of multimedia semantics that pertain to media and domain specific descriptions, for the purpose of supporting both the extraction and subsequent semantic management of such descriptions. To this end, firstly, we present first an overview of existing approaches to the representation of multimedia content and discuss open issues. Subsequently, we present the ontology infrastructure developed in the context of the BOEMIE project tailored towards the formal representation of multimedia content. Concluding, we present what can non-standard formal representation technologies, such as fuzzy knowledge representation formalisms bring to multimedia document processing and management.

1 Introduction

Multimedia content made available nowadays on the Web and in digital archives amount to a striking volume, intensifying further the urge to process and manage the available content in a semantically rich way. As a multimedia document may convey a wealth of information ranging among others from thematic descriptions addressing scenes, objects and events (e.g. a landscape, a jet engine, scoring, running, etc.), to structural and signal level descriptions (e.g. blue/textured region, linear motion, etc.), the effective representation of such information becomes a critical requirement. This criticality relates not only to the consequent of enabling end-users to efficient query and retrieve multimedia content, but also to the intertwinement with the extraction of content semantics and the intricacies pertaining to automated multimedia interpretation.

During the last decade there have been intense research efforts aiming at developing a proper language by which one would be able to represent and query

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 18–49, 2011.

(at semantic level) multimedia information. Such efforts gave rise to the MPEG-7 standard [32]. Through XML-Schema based definitions, MPEG-7 provides a rich set of tools for the description of multimedia content at different granularities and abstraction levels, including structural, low-level descriptors (e.g. colour, shape) and semantic descriptions, as well as aspects pertaining to authoring, user preferences, and so forth. Although the development of MPEG-7 has been a great advancement towards the systematic description of multimedia documents, significant deficiencies pertain to the means for and the axiomatisation of semantics representation [36, 57]. To a large extent, these deficiencies issue from the use of XML as the underpinning definition language, the flexibility allowed in the structuring of equivalent descriptions, as well as the restricted, and rather rigorous, model provided for the definition of domain specific semantic descriptions. As a result, the descriptions of multimedia information in a machine understandable way that would enable their sharing, reuse and interoperability has been hindered.

Towards this direction the approach of the Semantic Web [3] has proven to be the most promising way to achieve such goals, as ontologies can support a semantically rich, unambiguous and interoperable way of representing semantics, while additionally providing support for reasoning services that allow to extract further knowledge [48]. In addition to the well known paradigm of ontology based multimedia annotation, where domain specific ontologies are used to capture the semantics of subject matter descriptions associated to the multimedia content [27, 44], significant efforts have been undertaken in the last couples of years towards a more substantial deployment of ontologies in the management of multimedia semantics. Specifically, so called multimedia ontologies have been proposed to capture multimedia semantics through the formalisation and extension of the MPEG-7 modelling [1, 25, 41], while appropriately defined ontologies have been used to support tasks such as scene interpretation, object detection and retrieval [10, 33, 34, 45].

However, the aforementioned ontologies are intended for specific applications and tasks, and as a result tend to address the issues involved with respect to the modelling and representation of multimedia semantics in a fragmented fashion. On the contrary, in the BOEMIE[1] project, the formal representation of multimedia semantics has been the subject of research within an integrated application scenario that includes knowledge acquisition and representation, reasoning, multimedia ontology evolution, retrieval and presentation. As such, the proposed representation of multimedia semantics addresses media (content structure and low-level descriptors) and domain specific aspects, and is tailored to the analysis, interpretation and retrieval tasks that constitute the aforementioned chain of semantic content management.

Aiming to provide a systematic view of the aspects involved in the representation of multimedia content semantics within the context of semantics modelling and extraction, we provide in this chapter, on one hand an overview of the relevant literature and its weaknesses, and on the other hand, the ontology-based

[1] http://www.boemie.org/

representation model that has been developed within the BOEMIE project, towards the integrated confrontation of the issues involved. Nevertheless, classical ontology languages are often not capable of handling the type of information that results from multimedia processing tasks, which in many cases is imperfect (vague and/or uncertain). For example, an image analysis algorithm is not always able to assess to 100% accuracy the existence of an object. To account for the handling of such imperfect knowledge in multimedia interpretation and management tasks, non-standard technologies, which extend the proposed ontology infrastructure are also presented.

The rest of the chapter is organised as follows. Sections 2 and 3 provide an overview of relevant approaches towards the representation of multimedia content semantics. Specifically, Section 2 presents the different multimedia ontologies that have been proposed to formally capture the semantics of content structure and of the applicable low-level descriptors, while Section 3 considers the representation of content from the perspective of knowledge-based extraction and interpretation of the underlying semantics. Section 4 describes the semantic model developed within the BOEMIE project for the representation of multimedia content, while Section 5 presents some novel and non-standard ontology languages, which can be used to extend the expressive power of the proposed semantic model in order to handle imperfect information. Finally, Section 6 concludes the chapter and discusses open problems.

2 Multimedia Semantics Representation in Content Management

Multimedia assets form extremely rich sources of information. The conveyed meaning is communicated not only through intertwined multimodal information channels, but also through implicit connotations, narrative and discourse relations that create new levels of meaning. To be able to develop applications and services that are aware of the semantics, both the content and the context of multimedia need to be made explicit. Aiming at interoperable multimedia content description, a variety of multimedia metadata standards have been proposed addressing different levels of the conveyed information. However, in the developed multimedia standards and vocabularies, the semantics are rendered mostly in the form of syntactic norms with respect to corresponding XML Schema definitions, rather than the attachment of formal meaning.

The Semantic Web initiative induced efforts further, pushing towards machine understandable rather than machine readable semantics through the use of ontologies, i.e. explicit specifications of conceptualizations [17]. Ontologies are used to make meaning explicit contributing to the communication, exchange, reuse and sharing of knowledge across heterogeneous agents and applications. A number of ontology languages with varying expressivity have been proposed but the currently most prevalent standard by the World Wide Web Consortium[2](W3C) is the Web Ontology language (OWL) [6]. Building on the Semantic

[2] http://www.w3.org/

Web paradigm, a number of multimedia ontologies have been proposed aiming to attach formal semantics to multimedia content representation and allow for more intelligent content management.

In the following, we describe the proposed multimedia ontologies and discuss the encountered weaknesses. For reasons of completeness, a brief account of the most popular, yet lacking formal semantics, multimedia standards and vocabularies is also given.

2.1 Non-formal Representations

Within the Moving Pictures Expert Group, two relevant multimedia description standards have been developed, namely the Multimedia Content Description Interface (MPEG-7) and the MPEG-21 Multimedia framework.

The goal of MPEG-7 [32] is to provide a rich set of standardised tools for the description of multimedia content, and in addition to support some degree of interpretation of the meaning of information so as to enable the exchange of multimedia metadata across applications as well as their efficient management, e.g. in terms of search and retrieval. It offers a set of audiovisual description tools in the form of Descriptors (Ds) and Description Schemata (DSs), describing the structure of the metadata, their relationships and the constraints to which a valid MPEG-7 description should adhere. MPEG-7 is organised in 8 parts: *Systems*, the *Description Definition Language* (DDL), *Visual*, *Audio*, *Multimedia Description Schemes (MDS)*, *Reference Software*, *Conformance*, and *Extraction and Use*. The DDL consists the standard's core part, specifying the language for the definition of the description tools. The Visual and Audio parts consist respectively of structures and low-level descriptors that cover basic visual and audio features, while the MDS part specifies generic description tools pertaining to multimedia.

The MPEG-21 [35] activities address the definition of an open framework that allows the integration of all components of a delivery chain necessary to generate, use, manipulate, and deliver multimedia content across heterogeneous networks and devices. The key elements of MPEG-21 are: *Digital Item Declaration*, *Digital Item Identification and Description*, *Content Handling and Usage*, *Intellectual Property Management and Protection*, *Terminals and Networks*, *Content Representation*, and *Event Reporting*. From the aforementioned, content handling and usage, addressing the provision of interfaces and protocols to enable creation, search, access, delivery and reuse of content across the content distribution and consumption value chain is specifically interesting for multimedia content description. The same holds for the aspects addressed in the content representation, digital item identification and description elements, etc.

In addition to the MPEG activities, a number of multimedia metadata vocabularies emerged as the outcome of efforts undertaken by individual communities towards shared multimedia content descriptions. We refer indicatively, the Visual Resource Association (VRA) Core that specifies a small and commonly used vocabulary targeted especially at visual resources, and the Exchangeable Image File Format (EXIF), which specifies the formats to be used for images, sound, and

tags, in digital still cameras. Finally, the Synchronised Multimedia Integration Language (SMIL), that is an XML-based two dimensional graphic language that enables simple authoring of interactive audiovisual presentations, while Scalable Vector Graphics (SVG) allows describing scenes with vector shapes, text, and multimedia.

For a more thorough presentation of multimedia related metadata specifications the reader is referred to [42]. As outlined previously, a common characteristic shared among these multimedia representation schemes is that the intended semantics remain implicit in the syntax and the accompanying normative specifications.

2.2 Formal Representations

To enable multimedia on the Semantic Web and alleviate interoperability issues, a number of initiatives engaged in building multimedia ontologies by attaching formal semantics to multimedia content representations. The relevant activities are distinguished in two categories: those building on the MPEG-7 specification, and those following ad hoc modelling choices that are customised to specific application contexts.

Chronologically, the first initiative to make MPEG-7 semantics explicit was taken by Hunter [25] in 2001. The RDF Schema (RDFS) language was proposed to formalise the decomposition patterns of the Multimedia Description Scheme (MDS), the descriptors included in the Visual part, and some additional descriptors representing information about production, creation, usage and media features. The developed ontology has been ported to DAML and eventually to OWL Full [26], while later, extensions that address image analysis terms of the MATLAB Image Processing Toolbox have been also included [19]. The translation approach taken follows rigorously the standard specifications, hence, preserving in this way the intended flexibility of usage. This flexibility however comes with the cost of the inherited ambiguities present in MPEG-7 [36, 57], resulting in descriptions with multiple possible interpretations and ambiguous meaning [11].

Two MPEG-7 based RDFS multimedia ontologies, namely the Multimedia Structure Ontology (MSO) and the Visual Descriptor Ontology (VDO), have been developed within the aceMedia[3] project. MSO covers the complete set of decomposition tools from the MDS, while VDO addresses the Visual Part. The use of RDFS restricts the captured semantics to subclass and domain/range relations [5]. Both these approaches still suffer from the ambiguities that are also observed in the case of the Hunter ontology.

Another effort towards an MPEG-7 based multimedia ontology has been reported within the context of the SMARTWeb[4] project [38]. The developed ontology focuses on the Content Description and Content Management DSs. The respective multimedia content and segment classes along with a set of properties

[3] http://www.acemedia.org/aceMedia
[4] http://www.smartweb-projekt.de/

representing the decomposition tools specified in MPEG-7 enable the implementation of the intrinsic recursive nature of multimedia content decomposition. Although in this approach, axioms have been used to make intended semantics of MPEG-7 explicit, ambiguities are still present due to the fact that the corresponding MPEG-7 normative descriptions have been directly translated into concepts and properties whose semantics lie again mostly in linguistic terms.

Based on the work within the ReDeFer[5] project, the Rhizomik approach proposes a fully automatic translation of the complete MPEG-7 Schema to OWL [41], by mapping the XML schema of MPEG-7 to OWL. Human intervention is required only to resolve name conflicts stemming from the independent name domains for complex types and elements in XML. The resulting MPEG-7 ontology is in OWL DL and has been validated through its comparison against the manual translation of [26], which showed their semantic equivalence. The obvious advantage of the Rhizomik is the automatic translation of the complete MPEG-7 Schemata. However, when it comes to integration with domain specific ontologies, the Rhizomik approach is applicable only under the presumption that these domain ontologies have been re-engineered beforehand so that they extend the classes resulting from the corresponding Semantic DS structures.

An alternative approach has been adopted by the DS-MIRF framework [56]. Exploiting the MPEG-7 semantic description capabilities provided by the SemanticBaseType DS, the resulting ontology intends to serve as an upper multimedia ontology. A systematic methodology has been presented for the integration of domain specific semantics with the general-purpose semantic entities of MPEG-7 [55]. The developed ontology has been conceptualised manually and is in OWL DL. Transformation from XML to OWL, and conversely, is supported through a separate OWL DL ontology that holds the mappings between the original XML Schema and the corresponding OWL entities. Although sharing the same goal with Rhizomik, in terms of using MPEG-7 as a core multimedia content representation ontology, DS-MIRF does not require for the MPEG-7 Schema to be extended, allowing for efficient translation of MPEG-7 metadata to OWL assertions, and inversely.

The most recent approach to the formalisation of MPEG-7 semantics is the Core Ontology for MultiMedia (COMM) initiative [1] developed within the K-Space[6] and X-media[7] projects. Aiming to serve as a core ontology for multimedia, COMM utilises DOLCE [14] to provide a common foundational framework for the description of multimedia documents. COMM is in OWL DL and covers selected descriptors from the media, location and decomposition patterns of MDS, as well as the visual part. COMM extends the design patterns of *Descriptions & Situations (D&S)* [15] and *Ontology of Information Objects (OIO)* [13] in order to axiomatise the description at structural (content decomposition), algorithmic (functionality and parameters), and conceptual (semantics annotation) level. Thereby, COMM underpins at semiotic level the process of integrating

[5] http://rhizomik.net/redefer
[6] http://kspace.qmul.net
[7] http://www.x-media-project.org/

multimedia and domain ontologies for the description of various aspects of content, reinforcing conceptual clarity in the descriptions per se.

As aforementioned, besides the multimedia ontologies that have been developed based on MPEG-7, a number of customised multimedia ontologies have been proposed within specific applications. Thonnat et. al [24, 31], proposed a visual ontology that provides qualitative descriptions with respect to color, texture, and spatial aspects of the characterised content. Analogous qualitative visual descriptors have been also employed in the Breast Cancer Imaging Ontology (BCIO) [23]. In SCULPTEUR [30], an ontology for the museum domain has been combined with a graphical concept browser interface that allows navigation through the domain ontology semantic layer, as well as display of the different content types in appropriate viewers. In [4], a so called pictorially enriched ontology is proposed that uses visual prototypes to represent semantic concepts instead of linguistic concepts. In [20], a visual ontology (VO) is described, which combining MPEG-7 and WordNet descriptions, allows the representation of visual attributes, such as shape, colour, visibility, etc.

Despite sharing a common vision, the aforementioned approaches present substantial conceptual differences, reflected both in the modelling of content semantics as well as in the linking with domain ontologies. The various customised multimedia ontologies, adhering to application specific requirements, are hardly concerned with interoperability issues, while the MPEG-7 based multimedia ontologies, although aiming to alleviate interoperability issues, have introduced new ones, this time at a semantic level [11, 54].

The COMM ontology addresses the axiomatisation of multimedia description patterns, but does not confront the semantic ambiguities that relate to the extensions of the provided definitions through more specialised descriptions as those provided by the rest MPEG-7 based multimedia ontologies. The latter demonstrate a tendency for continually higher utilisation of the expressiveness provided by the ontology languages, yet they all suffer, to a lesser or greater degree, from ambiguous semantics. As a consequence, one ends up with descriptions that have multiple interpretations, even when construed with respect to the reference ontology, thus hindering not only their management but as well their linking with descriptions pertaining to different multimedia ontologies.

We note that in the case of MPEG-7 based multimedia ontologies, the observed semantic ambiguities refer in principle to the representation of the content structure information and of the applicable decomposition schemes, and not to the modelling of the MPEG-7 low-level description tools, since the latter comprise rigid numerical structures rather than conceptual notions. This is no longer the case for the customised multimedia ontologies though, where the different application contexts induce additional discrepancies. Moreover, since correspondence to the MPEG-7 structural and low-level descriptors cannot be always guaranteed, further questions are raised regarding the reuse and linking with existing MPEG-7 based descriptions. Consequently, a critical requirement for enabling the effective extraction and subsequent handling of multimedia semantics is the construction of multimedia ontologies with well-defined semantics.

3 Knowledge-Based Interpretation of Multimedia Content

Multimedia interpretation constitutes a particularly challenging problem that has engaged strong and continuous research interest. It refers to the lack of coincidence between the descriptions that can be extracted automatically from multimedia content at signal level, and the corresponding interpretations as acquired by a human [47]. In this endeavour, the use of background knowledge holds a central role, as the complexity of the problem renders purely data-driven approaches, severely inadequate to approximate what would consist a human like perception of the conveyed meaning.

This background knowledge is usually structured at levels of increasing abstraction, ranging from perceptual representations to logical interrelations that define the entities and notions of interest. Different perspectives on what constitutes multimedia semantics have resulted in the development of knowledge models that address different levels and types of knowledge, and define different interrelations between the employed abstraction levels. These differences affect in turn the espoused knowledge representation formalism as well as the configuration of the multimedia interpretation process as an inferencing task. In each case, the adopted representation formalism determines the degree at which explicit and formal semantics are supported.

In the following, we outline the effects pertaining to the representation of multimedia semantics from the perspective of content interpretation. First, general considerations that apply in the use of knowledge and reasoning in the extraction of multimedia semantics are discussed, and in the sequel characteristic examples of existing works are discussed.

3.1 Knowledge-Based Multimedia Semantics Extraction

The development of knowledge-based approaches to multimedia semantics extraction confronts two crucial questions: i) which representation formalism is suitable for capturing the semantics at hand, and ii) what pieces of information constitute the knowledge that is required for solving the addressed problem.

Regarding the first, and bearing in mind that the focus of this chapter is on formal semantics, the various alternatives, as suggested by the existing literature, have been largely influenced by the Semantic Web initiative. Ontology languages such as OWL [6] and their logical underpinnings, Description Logics [21] have become prevailing choices. The popularity of DLs issues not only from the direct relation with OWL, but also from the fact that they constitute expressive fragments of first order logic, for which decidable reasoning algorithms exist [2].

The expressivity provided by the different representation formalisms, determines the appropriate choice in accordance with the types of knowledge and reasoning tasks that comprise the extraction of content semantics. As will be described in the subsection 3.2, there are approaches that employ hybrid schemes, combining more than one representation formalisms. Given the differences in the

provided expressivity such observations constitute important issues with respect to the kind of expressivity required for supporting multimedia interpretation. It should be noted though that in some cases, the more intuitive formalisms are the ones that finally prevail.

The second question is intertwined to the espoused perspective on what multimedia semantics consists in. An aspect shared among the different approaches, is that the employed knowledge, in addition to providing support for the analysis and extraction processes, it also provides the vocabulary and the semantics of the produced content annotations. This enables content management services, such as search, retrieval, filtering, etc, at a semantic level. This vocabulary is not necessarily restricted to domain specific descriptions, but may include other aspects as well, such as content structure. The latter is a prerequisite in order to provide finer indexing and retrieval services, and support transcoding applications.

Regarding the extraction per se, the tasks for which knowledge and reasoning have been utilised fall roughly into three categories: i) the translation of automatically extracted features to semantic entities, ii) the extraction of descriptions of higher abstraction based on the logical associations that underly the semantic entities that are directly detectable by means of analysis, and iii) the specification of the control strategy, i.e. of the steps and parameters comprising the analysis process itself. Plausibly, the tasks at hand have a strong interrelation with the types of knowledge captured. For example, in approaches tackling the first task, there exists representations of features pertaining to audiovisual manifestations as well as corresponding domain concept definitions with respect to the constraints and range values that apply with respect to the modelled audiovisual features (e.g. colour, texture, motion, etc.). Approaches addressing the second task on the other hand, focus more on the capturing of semantic interrelations and attributes between domain entities. Hence, the background knowledge is populated mostly with concept definitions that reflect complex notions whose meaning lies in logical aggregations, rather than audiovisual manifestations.

It is interesting to note that although multimedia semantics extraction aims at educing descriptions close to what a human interpretation would be, the overview of the state of the art reveals that the majority of the approaches considers mostly the first task. This means that the employed knowledge, even when adequately capturing the specific domain semantics, is mostly utilised for the purposes of annotation, while in the extraction only semantics relevant to audiovisual features are used. Adding to this the fact that axioms defining concept with respect to audiovisual features entail numerical computations rather than logical inference, shows that despite using very expressive knowledge representation languages, with powerful inference services, their potential is poorly exploited.

Another issue relevant to multimedia semantics extraction is the handling of uncertainty, a feature inherent in multimedia analysis and understanding. The numerical nature of segmentation and the incompleteness, to a large extend due to the inability to capture semantics only by means of audiovisual manifestations, of the perceptual models describing semantic entities, allow only for partial

matching against these models. As a result the extracted analysis representations cannot be interpreted as indisputable evidences. From the aforementioned representations, none provides directly the means to handle this uncertainty. As will be described in the next subsection, most approaches handle the uncertainty indirectly, by defining thresholds with respect to the degree of similarity against the defined audiovisual features' models that is acceptable. However, once the similarity is evaluated and the decision is taken, the uncertainty information is usually dismissed, i.e. in the resulting assertions (facts) that comprise the content annotation, there are no degrees. This also means that whichever reasoning is applied afterwards, is performed over crisp terms.

The aforementioned aspects lie in the core of the development of knowledge-based approaches for the extraction of semantic descriptions from multimedia; however, these are not the only dimensions involved. Knowledge acquisition, supported media type, and sequential vs interactive extraction, are indicative examples of relevant issues.

3.2 Related Work

In the following, we briefly summarise different approaches of knowledge-based multimedia systems.

In the series of works presented in [24, 31], an ontology-based approach is followed for the representation of knowledge. The employed knowledge builds upon the premise of addressing separately the three abstraction levels as defined by Marr. A domain ontology provides the corresponding conceptualization for the various domains of images considered, i.e., pollen grain, galaxies, rose diseases, transport vehicles, etc., while a visual concept ontology is employed to provide symbolic, intermediate level definitions related to color, texture and spatial information, that allow linking the domain concepts with the raw image data. The extraction of semantic description for images is realised in the form of rule-based reasoning, performed in a linear fashion in order to derive descriptions of successively higher-abstraction in a stepwise fashion, starting from the available at visual level information.

A similar approach is taken in [26], where rule-based reasoning is employed in an non-iterative manner to derive semantic annotations based on the manually defined mappings between domain concepts and visual characteristics. Three OWL ontologies capture the different knowledge components involved, i.e., low-level visual features, microscope information, and domain specific knowledge (fuel and pancreatic cells). Contrary to the customised visual descriptions of the adopted in [24, 31], the low-level visual features ontology builds on the corresponding MPEG-7 visual descriptors [25].

The ontology-based framework proposed in [5] adopts a similar perspective. A domain ontology captures the logical associations that define the relevant concepts and relations, while two MPEG-7 based ontologies model low-level visual descriptors and content structure, as described in Section 2. The linking of domain concepts with prototypical low-level descriptors' values is realised through

M-Ontomat-Annotizer [39], which formalises the interconnection between the two ontologies.

In [10], semantic concepts in the context of the examined domain are defined in an ontology, enriched with qualitative attributes (e.g., color homogeneity), low-level features (e.g., color model components distribution), object spatial relations, and multimedia processing methods (e.g., color clustering). The RDF(S) language has been used for the representation of the developed domain and analysis ontologies, while for the rules that determine how tools for multimedia analysis should be applied depending on concept attributes and low-level features, are expressed in F-Logic. Compared to the previous approaches, [10] brings in the modelling of analysis new dimensions as well, while for the linking of visual descriptions with domain concepts a similar rationale is followed.

OntoPic [43], is a supervised learning system that utilises DL-based reasoning, treating concept recognition as a classification problem. An appropriately constructed TBox provides the hierarchy of the domain concepts and their spatial topology. The initial definitions are extended during the learning phase with feature roles that associate domain concepts to the features and feature value constraints that resulted from the training. A pseudo-extension to fuzzy DL is introduced to avoid overspecification. During a postprocessing step, the resulted membership values can be re-adjusted according to feature weights reflecting their discriminative power. Finally, the classified regions are checked in terms of spatial consistency, utilising once again the DL inference services. To avoid ending up with inconsistent ABoxes, the violations of spatial constraints are treated as *non-concept* definitions which OntoPic removes successively, starting from the one with the lowest degree of membership, until a consistent ABox, i.e., image description, is reached.

Hence, as in the previously described approaches, two abstraction levels are employed for the representation of content semantics, i.e. domain specific descriptions and low-level visual descriptions. However, contrary to the previous approaches, OntoPic utilises the axiomatic definitions that link the descriptions of the two levels in a more semantically rich way. Specifically, the linking axioms are not used simply as the means to realise the transition from visual descriptions to semantic domain specific notions in the form of "IF" 'THEN" production rules, but support the construction of semantically constituent, logical models.

In [37], Description Logics are used for acquiring scene interpretations. The notion of *aggregate concept* is introduced for realising scene interpretation as a stepwise process utilising taxonomical and compositional relations. The interpretation process works on top of primitive descriptions derived directly from visual evidence, and further contextual information is introduced in the form of spatial and temporal constraints. Four kinds of steps, namely aggregate instantiation, instance specialization, instance expansion and instance merging, are used to realise scene interpretation as model construction. In addition, coupling with a probabilistic framework is proposed in order to provide guidance among the different plausible interpretations.

Rule-based reasoning is employed in the approach to video understanding presented in [28]. Visual, auditory and textual aspects of the video are taken into consideration to semi-automatically construct multimedia ontologies that will provide the definitions required in the sequel for the extraction of video semantics. After automatic speech recognition (ASR) and alignment to video shots, the produced textual data along with the available text annotations are processed using *KAON*[8] and exploiting Wordnet[9], in order to select relevant concepts included in the employed TGM I vocabulary. Similarly, visual detectors based on low-level content features (color, texture, etc.) are used and associated with corresponding terms, while reasoning concerns the application of context rules to adjust the confidence values of the visual detectors.

In [8], an approach to fuzzy reasoning is proposed in order to integrate image annotations at scene and region level, into a semantically consistent final description, further enhanced by means of inference. An ontology is used to capture the underlying domain semantics and allow the detection of incoherences, while rules are used to allow the effective representation of spatial related axioms. The assimilation of fuzzy semantics allows to handle the uncertainty that charasterises multimedia analysis and understanding, while the use of DLs allows to benefit from the high expressivity and the efficient reasoning algorithms in the management of the domain specific semantics. The initial annotations forming the input may come from different modalities and analysis implementations, and their degrees can be re-adjusted using weights to specify the reliability of the corresponding analysis technique or modality.

The aforementioned approaches constitute characteristic examples, where the representation of content semantics not only serves in the semantic structure and management of multimedia descriptions, but in addition underpins the extraction of such descriptions. In their most straightforward form, the proposed approaches involve the representation of some types of perceptual features (often in the form of MPEG-7 descriptors) and the definition of axioms that link domain specific concepts with combinations of valid feature values. In this manner though, reasoning is employed in a rather trivial fashion as it assimilates more the functionality of production rules rather than the construction of logical models. Reasoning as logical entailment is investigated more thoroughly in [8, 37, 43], where the captured semantics are used in order to ensure the construction of consistent content interpretations.

Furthermore, the proposed approaches are tailored to the adopted interpretation perspective, and as such they address only selected content representation aspects. As a result, there exists a lack of an integrated representation framework that would enable to address the formal modelling of the different types and abstraction levels of the relevant information, including the different modalities, as well as the dynamic nature of the knowledge involved. In the following, we present the ontology infrastructure developed in the context of the BOEMIE

[8] `http://kaon.semanticweb.org/`
[9] `http://wordnet.princeton.edu/`

project in order to address such issues and provide support for enriched content interpretation as well as management services.

4 Representation of Multimedia Semantics in BOEMIE

In the current section we present the architecture and design choices followed in the context of the BOEMIE project in order to construct an ontology infrastructure. This infrastructure is developed in such a way that it can provide the means to manage and combine multimedia specific information and domain-specific one in order to enable:

- The semantic labelling of multimedia documents after the detection of concepts and relations from low-level analysis modules.
- The enrichment of the annotation of multimedia documents by providing definitions for complex (high-level) concepts utilised by reasoning services.
- Presentation and retrieval of multimedia documents w.r.t. the information that they convey.
- The evolution and learning process by providing a modular and pattern based ontology infrastructure which can be (semi)automatically evolved.

In order to account for the different types of knowledge involved and meet the different requirements imposed by the different modules which use the ontology infrastructure, the developed ontology model consists in practice of several interrelated and interlinked ontologies that can be divided into two categories. The first category consists of the multimedia ontologies, while the second one of the so called domain ontologies. Each of these two categories further contains two also independent ontologies. More precisely, the domain ontologies include the *Athletics Events Ontology* (AEO), describing our domain of interest which is public athletics events, and the *Geographic Information Ontology* (GIO), describing geographic information. On the other hand, the multimedia ontologies consist of the *Multimedia Content Ontology* (MCO), representing content structure information, and the *Multimedia Descriptors Ontology* (MDO), representing low-level numerical information extracted by analysis modules. An advantage of the proposed architecture is that it is highly modular, as the multimedia structure-related information is independent of the content and common for all multimedia documents, whereas the information about the content of a multimedia document depends totally on its subject. Furthermore, this discrimination can significantly improve the response time of the system to content related end-user queries, since the multimedia structure-related information is usually larger than the domain specific one, but also much less interesting for the end-user.

The four individual ontologies are interconnected and therefore can be used by applications that need to combine information and knowledge from different resources. Thus, the developed ontologies do not stand alone but are interlinked through proper structural, spatial, temporal, or any other kind of relations, of

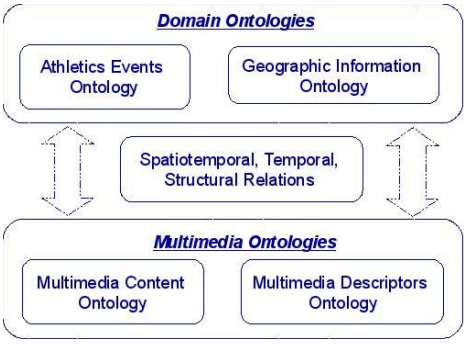


Fig. 1. Architecture of the Multimedia Semantic Model

which the domain or range might be defined in different ontologies. This interconnection finally provides a global and modular ontology infrastructure which is called *Multimedia Semantic Model* (MSM). Figure 1 depicts the overall architecture of the MSM model with the various ontologies and their interconnections. As we can see besides the interconnections between ontologies of the same categories there are also interconnections between ontologies of the multimedia and the domain category.

The knowledge representation formalism that we adopted for the construction of the ontologies of the MSM is Description Logics (DLs) [2]. DLs belong to the family of concept-based representation formalisms and actually consist of expressive fragments of First Order Logic (FOL), providing decidable and empirically tractable reasoning services, like logical consequence (entailment) and concept subsumption, i.e. checking if a concept (class) is a sub-concept (sub-class) of another one.

In the following, an overview of the ontologies of the MSM is provided.

4.1 Domain Knowledge Representation

Athletics Events Ontology. The *Athletics Events Ontology* (AEO) is a formal conceptualization of the domain of interest of the BOEMIE use case scenario which is public athletics events, i.e. jumping, running and throwing events held in European cities. The concepts and relations of the AEO are used for annotation and retrieval of multimedia documents on the subject of athletics events, i.e. on information relevant to athletics competitions and their constituents events as well as information about athletes and performances gained in such events.

During the knowledge acquisition phase of the ontology development process, and taking into consideration the results of analysis, a discrimination has been established between the representation of concepts (semantic entities) that can be immediately instantiated by analysis modules, such as concrete objects, or names of athletes and locations, also called *Mid Level Concepts* (MLCs) in the framework of BOEMIE, and the representation of more abstract concepts that cannot be detected automatically by analysis, also called *High Level Concepts*

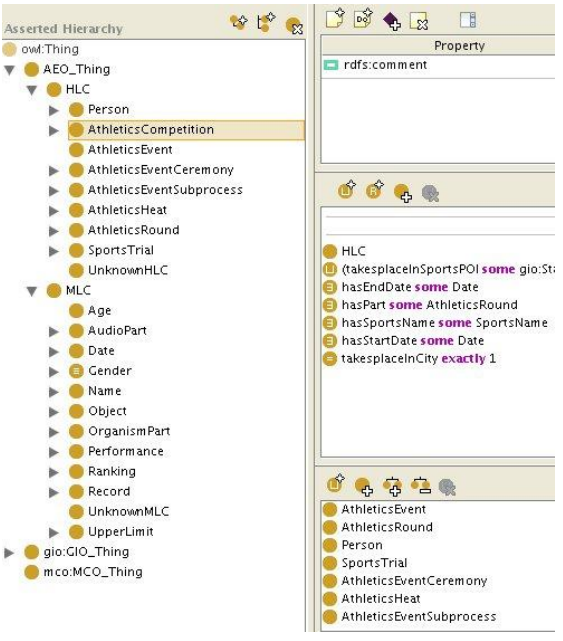


Fig. 2. The root concepts of the Athletics Events Ontology

(HLCs), such as composite events. This discrimination is required both in the ontology evolution process, in terms of applying different patterns for the definition of a new concept, accordingly to its substance, i.e. whether it is a MLC or a HLC as well as in image interpretation and reasoning. As a consequence, the root classes of the AEO hierarchy are the MLC concept and the HLC concept, as shown in Figure 2.

Two different design patterns have been implemented, one for the definition of MLCs and one for the definition of HLCs. MLCs are formalised as atomic concepts, subclasses of the MLC root concept of the AEO hierarchy (e.g. Object $\sqsubseteq$ MLC). Every modality provides its own MLCs. Thus, the subclasses of the MLC concept are Age, Date, Gender, Audiopart, Performance, Ranking, Name, OrganismPart, etc. Among these, the concepts Age, Date, Gender, Performance, Ranking and Name can be instantiated by text analysis whenever a relevant string is detected. On the other hand, image analysis instantiates mainly concepts that are subclasses of the concepts Object and OrganismPart, whenever a relevant image region is detected.

HLCs are formalised as complex concepts that appear in the left-hand side of terminological axioms built using DL concept constructors such as $\exists, \forall, \sqcup, \sqcap$. HLCs are designed as aggregates, which consist of multiple parts that can be either MLCs or other HLCs, and are constrained by relations representing spatial, temporal and other kinds of logical relations between these multiple parts, based on the approach described in [34].

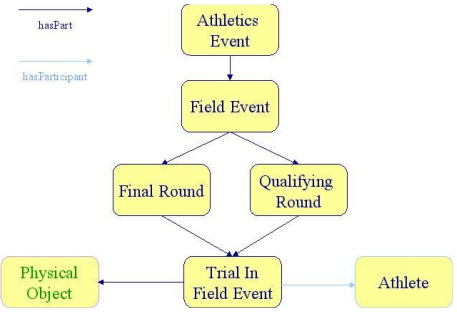


Fig. 3. Conceptualisation of field athletic events and their partonomical relations

The subconcepts of the HLC concept conceptualise the complex concepts of the domain of athletics based on descriptions provided by IAAF Competition Rules and IAAF Technical Regulations[10]. Thus, the most important subconcepts of the HLC concept are the following:

- The concept AthleticsCompetition, which conceptualises series of events held over one or more days, i.e it conceptualises whole athletics competitions, such as the Olympic Games are, which are composed of different kinds of events.
- The concept AthleticsEvent, which conceptualises a single race or contest in a competition that takes place in a specific point of space and time. An athletics event might be a track, a field, a roadrace, a racewalking, a cross country or a combined event. Moreover, track events and field events consist of either one final round or more qualifying rounds.
- The concept AthleticsRound, which conceptualises a single round, final or qualifying, in an event that takes place in a specific point of space and time. A qualifying round consists of more thatn one athletics heats.
- The concept AthleticsHeat, which conceptualises a single heat held in a track or field event that takes place in a specific point of space and time, whenever the number of athletes is too large to allow the event to be conducted satisfactorily in a single round (final).
- The concept AthleticsTrial, which conceptualises a single trial in a field event that takes place in a specific point of space and time.
- The concept Person, which conceptualises persons that participate in very different ways in an athletics competitions. Therefore, its subclasses are not only Athlete but also TechnicalPersonnel, Judge, Coach and Referee.

The partonomical relation that we have used in order to represent the fact that competitions are composed of events, and events are composed of rounds, etc., is the transitive relation hasPart, as can be seen in Figure 3.

[10] `http://www.iaaf.org`

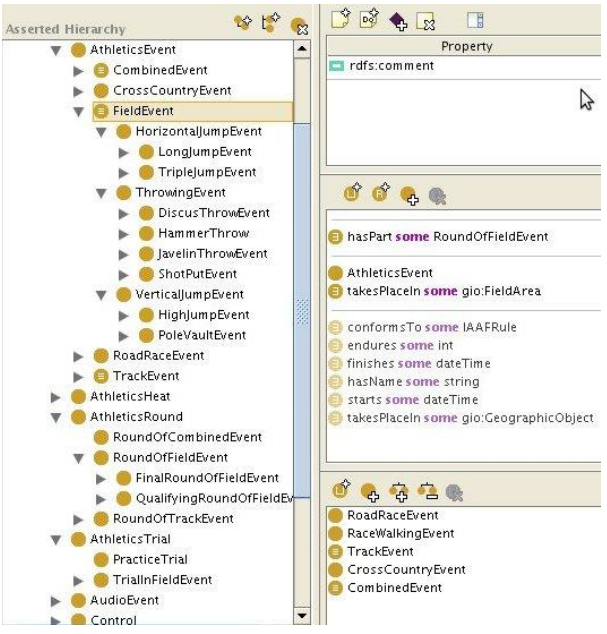


Fig. 4. Conceptualisation of the concept FieldEvent

Finally, in order to address the several characteristic aspects of athletics events, corresponding specialisation concepts have been introduced. In Figure 4 the subclasses of the AthleticsEvent concept are illustrated as well as the definition of the specialisation concept FieldEvent. We can observe that the necessary conditions for an instance of the FieldEvent concept are that it is composed of rounds and that it takes place in the field area of a stadium. In addition, it inherits necessary conditions by its superconcept AthleticsEvent, i.e. it must start and finish on a specific date, it must have a specific duration, it must conform to a specific IAAF rule and it must have a specific name. In the same way, all events are defined with repsect to their specific attributes.

Geographic Information Ontology. The context of usage of the *Geographic Information Ontology* (GIO) within BOEMIE consists in providing the representation of the relevant geographic information in order to associate events/objects from the annotated multimedia content to the respective place/location they take place in (e.g., the stadium and city in which a given athletics competition takes place). In this way, the GIO enables visualisation and navigation on enriched with domain specific information maps (e.g., visualisation of a marathon route on a city map). Moreover, the GIO can provide assistance in the interpretation process through the exploitation of geographic information. To accomplish the aforementioned, the GIO needs to provide support for the representation of the following types of information:

- Geopolitical information, i.e. information about geographic areas, which are associated with some sort of political structure, such as continents, countries and cities.
- Geographic information regarding places and locations of interest.
- Position related information, i.e. coordinates and respective coordinate systems, so that the considered objects can be linked/projected to corresponding map positions.
- Spatial relations, so that from an initial set of geometry-based calculated relations, further ones may be obtained automatically through inference services.

For the development of the GIO, the TeleAtlas database schema model[11] has been used as a guideline, especially for the identification of the types of information that should be covered. TeleAtlas database provides extremely rich, hierarchically structured, thematic information in the form of *Points Of Interest* (POI) and an underlying geometry features' model that enables equally rich functionalities in terms of calculating spatial relations holding among the given geographic objects. Considering the purely geographic information, such as coordinate systems and units of measures, this choice is also justified by the fact that TeleAtlas has followed the corresponding OpenGIS standard specifications. With respect to the thematic information, we observed again compliance to a high degree with the ontologies and vocabularies employed in the relevant literature, so we used the TeleAtlas taxonomy as the basis and applied modifications and further enrichments where necessary. The top level concepts of the developed GIO, illustrated in Figure 5, are the following:

- GeographicObject: The GeographicObject concept is used to represent any type of object used for referring to geographically related information. Each geographic object is associated with some map, on which it is projected, and some coordinates that identify its position within this map. In addition, it is related to other geographic objects through spatial relations, it belongs to a specific timezone and is located in some location. Moreover, the GeographicObject class comprises the GeopoliticalArea, Landform, ManMadeFeature, POI (Point of Interest), Route and the SpecialPurposeArea classes. The GeographicArea concept accounts for the different categories of geographic areas, such as countries and cities. The POI concept models in a hierarchical manner locations / places of general interest. Some indicative subclasses of the POI concept are SportPOI, LeisurePOI and TransportPOI. Subclasses of the SportPOI that are mainly used for representing the locations where athletic competitions take place are the concepts Stadium, SwimmingPool, TennisCourt, etc. In addition, although not included in the Teleatlas database schema, we have defined the concept Route, as a subclass of the concept GeographicObject to represent geographic information relevant to the route of road race events.

[11] http://www.spatialinsights.com/catalog/product.aspx?product=95

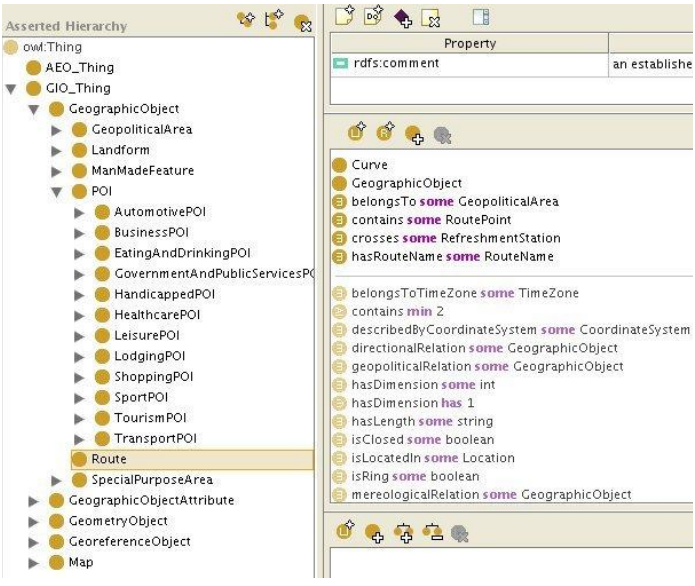


Fig. 5. A part of the GIO hierarchy

- Map: The Map concept is a symbolised depiction of a space which highlights relations between components of that space. To identify the referred map, a string denoting its location (file, url, etc.) is associated with it.
- GeoreferenceObject: The GeoreferenceObject concept is used to represent information for reffering to the location of a specific geographic object by means of coordinates. The subconcepts Coordinate, CoordinateSystem and CoordinateValue are used to represent coordinate related information.
- GeometryObject: It is used to provide geometry-dependent information about geographic objects. The GeometryObject class has subclasses the following concepts Point, Curve, Surface and the concept GeometryCollection. Each geometric object has specific important features, which can be inherited by geographic objects and provide important information about them. For example, since a route of a race event is a curve, and a curve has a certain length, a starting and an ending point, then a route should also have a certain lengh and a starting and ending point.
- GeographicObjectAttribute: This concept represents important attributes of geographic objects, such as their address, their official name, etc.

Additionally, with respect to the different types of geographic areas included, corresponding sets of spatial relations have been defined. More specifically, the properties geopoliticalRelation, topologicalRelation, directionalRelation and mereologicalRelation have been introduced and appropriate sub-properties have been defined.

4.2 Structure and Low-Level Descriptor Representation

Multimedia Content Ontology. The Multimedia Content Ontology (MCO) addresses structural aspects (i.e. decomposition semantics) pertaining to the different multimedia content types. Such knowledge is required to enable attaching annotations to the corresponding content parts (e.g. to annotate a specific still region of an image as depicting an athlete or a video segment as depicting a pole vault trial) and handle part-whole semantics (e.g. an image is comprised of the set of its constituent still regions to which it is segmented, thus if one still region depicts an athlete, the image itself depicts this athlete as well). Providing the means to capture and represent such knowledge, the MCO aims to support for unambiguous multimedia annotation, retrieval, exchange, and sharing of metadata addressing media related aspects, as well as the application of inference. Therefore, its construction is based on the distinct representation of:

- the different types of multimedia content (e.g. images, captioned images, web pages and video),
- the possible logical relations among them (e.g. a web page may consist of a text extract, two images, and an audio sample),
- the semantics of the decomposition of the corresponding media types into their constituent parts according to the level of the produced annotations, e.g. a video can be decomposed into video segments based on shots, each of those segments further decomposed into constituent frames or moving regions when more detail with respect to localization is required,
- and the relations that associate multimedia content to the semantic entities conveyed (e.g. a still region depicts a person face).

As such, the MCO is strongly related to semantics extraction task, since during fusion information, about the provenance of the annotations extracted by the individual modalities is utilised. Furthermore, providing the means to represent the decomposition of multimedia documents into constituent parts, it supports the information retrieval and presentation tasks. The main top level classes include the mco : MultimediaContent class, which captures through its specialisation the various single and multiple modality content types of interest, the mco : MultimediaSegment class, which comprises the different segment types to which the various media items can be (spatially, temporally or spatiotemporally) decomposed to, and the mco : SegmentLocator class, (see Figure 6) which includes information about the various ways for identifying and designating a particular segment. The implemented MCO follows to a large extent the guidelines

$$\begin{aligned}
\mathsf{SingleMediaItem} &\sqsubseteq \exists \mathsf{hasMediaDecomposition.MultimediaSegment} \\
\mathsf{Image} &\sqsubseteq \exists \mathsf{SingleMediaItem} \\
\mathsf{Image} &\equiv \forall \mathsf{mediaHasDecomposition.StillRegion} \\
\mathsf{StillRegion} &\equiv \forall \mathsf{segmentHasDecomposition.StillRegion} \sqcap \\
&\quad \forall\ \mathrm{hasSegmentLocator.VisualLocator}
\end{aligned}$$

Fig. 6. Part of StillImage definition in the MCO

specified in the MPEG-7 structure of content Multimedia Description Scheme, while enhancing it in order to avoid its inherent ambiguities. To accomplish this, the definition of the various content and segment types is logically grounded on the applicable decomposition schemes and the localisation information required for the identifications; thereby, and contrary to the respective definitions in the relevant literature, MCO models unambiguously the semantics of the notions involved.

Multimedia Descriptors Ontology. The Multimedia Descriptor Ontology (MDO) captures knowledge related to low-level representation of multimedia content, i.e. information about the descriptors employed by the different modalities to characterise content at feature (signal) level. The MDO is strongly related to the semantics extraction task, since it supports the individual modalities analysis in the detection of mid-level concepts (MLCs) through the linking of descriptors to domain specific concepts, as well as in the enhancement of their performance, enabling clustering of feature-level similar objects, and thus supporting the handling of unknown MLCs. The MDO has been designed based on two principles:

1. compliance with the respective MPEG-7 Visual and Audio parts to ensure wide coverage and interoperability in case of modalities processing enrichment with additional analysis modules, and
2. support for the requirements specific in the BOEMIE project with respect to the addressed modalities and the used tools.

As a result of the latter for example, since analysis focuses on quantitative descriptions, i.e. numerical representations of the analysed visual properties, quantitative descriptors (e.g. such as bright/dark, smooth/coarse) have not been addressed. The top level concept of MDO is the mdo : MultimediaDescriptor concept which is subclassed with respect to the different modalities into the concepts mdo : VisualDescriptor, mdo : AudioDescriptor, and mdo : TextualDescriptor. In addition, the Adds concept, also subclassed with respect to the different modalities, has been introduced to provide the means to capture information required for representing the corresponding modality descriptors. Each of the latter serves as the root of the ontology component representing the respective

$$\begin{aligned}
\mathsf{DominantColorDescriptor} \sqsubseteq\ & \forall \mathsf{hasDominantColor.DominantColorComboValue} \\
& \sqcap \geq 1 \mathsf{hasDominantColor} \\
\mathsf{DominantColorComboValue} \sqsubseteq\ & \\
& \forall \mathsf{hasColorQuantizationComponent.ColorQuantizationDescriptor} \\
& \sqcap \geq 1 \mathsf{hasColorQuantizationComponent} \\
& \sqcap \forall \mathsf{hasColorSpaceComponent.ColorSpaceDescriptor} \\
& \sqcap \forall \mathsf{hasColorValuesComponent.ColorValuesElement} \\
& \sqcap \leq 8 \mathsf{hasColorValuesComponent} \\
& \sqcap \forall \mathsf{hasSpatialCoherencyComponent.SpatialCoherencyElement}
\end{aligned}$$

Fig. 7. The definition of the Dominant Color Descriptor in the MDO

modality descriptors. Visual descriptors include color, texture, shape, motion and localization descriptors as for example the concepts: mdo : DominantColor, mdo : HomogeneousTexture, mdo : TrajectoryType, etc., while auditory descriptors address basic audio signal features as for example the following descriptors: mdo : FundamentalFrequency, mdo : ZeroCrossingRate, etc. Similarly, the defined properties are organised in a hierarchical way. For example, the relation mdo : hasDominantColorDescriptor is subsumed by mdo : hasColorDescriptor which in turn is subsumed by mdo : hasVisualDescriptor.

4.3 The Multimedia Semantic Model

Although that for the sake of ontology design we have considered the four ontologies as separate ontological modules, their borders are in fact vague. While developing an ontology, we confronted often the situation in which we needed to define a new relation the domain of which belonged to the ontology that we were developing at that time but the range belonged to another ontology of our framework. Thus, and through the definition of appropriate relations spanning across multiple ontologies, a network of structural, spatial and temporal relations, of which the domain and range belonged to different ontologies, emerged gradually. This network of relations comprises the so called Multimedia Semantic Model (MSM) that realises the integration of the different ontological modules into an interlinked and interconnected ontology infrastructure.

We note again, that all four ontologies, as well as the MSM model of interrelations have been manually engineered, while the specifications and requirements for new relations and concepts, as well as for the revision and enhancement of existing definitions, have issued from the feedback received regarding the use of the ontologies in the tasks of multimedia analysis, interpretation, management, and ontology evolution addressed within the BOEMIE project.

The Multimedia Semantic Model is illustrated in Figure 8, where we can observe examples of these interlinking relations, which can be divided in the following three categories according to our ontology architecture:

- Relations among concepts of the multimedia ontologies: These relations combine information about structural aspects of multimedia documents with information about low-level features of multimedia objects and can be helpful for the presentation of multimedia objects as well as learning algorithms of new concepts from unknown objects. An indicative example of this kind of relations is the mdo : isDescriptorOf relation which connects instances of descriptors, defined in the MDO, with instances of the multimedia segments that they describe, defined in MCO. For example, in order to represent the fact that an instance of a still region has a certain color descriptor we would use the following assertion:

$$\mathsf{mdo : isDescriptorOf}(\mathsf{mdo : ColorDescriptor1}, \mathsf{mco : StillRegion1})$$

- Relations among concepts of the domain ontologies: These relations connect information about events of the domain of interest with map data and are

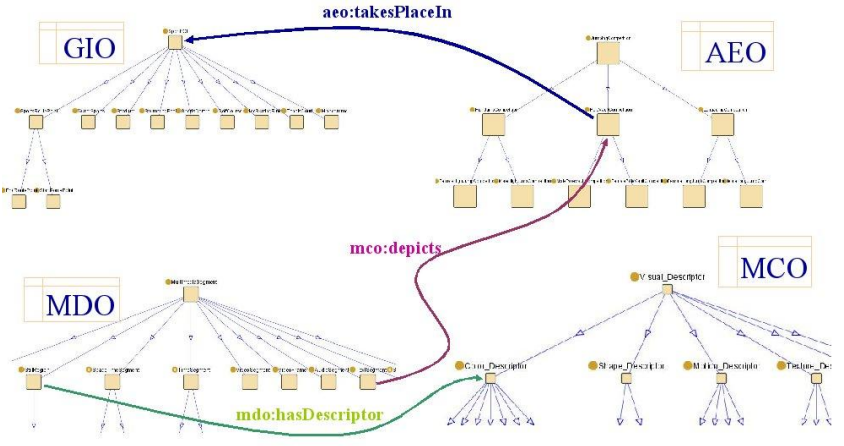


Fig. 8. Interconnections of the ontologies of the MSM

extremely helpful for presentation and retrieval of multimedia documents with respect to the geographic information that they convey by linking the annotated parts of the multimedia documents to geographical map data. In particular, they combine information about athletics events with information about the geographic/geopolitical area that they have taken place. A characteristic example of this category of relations is the aeo : takesPlaceIn relation which connects instances of concepts like aeo : AthleticsEvent, aeo : Athletics *Round*, aeo : AthleticsTrial, defined in AEO, to the location that they have taken place, e.g. to instances of concepts gio : Stadium, gio : StadiumArea, gio : City, gio : Country of the GIO. For example, in order to represent the fact that an instance of a Marathon event has taken place in a specific city, we would use the following assertion:

$$\text{aeo : takesPlaceIn}(\text{aeo : MarathonEvent1}, \text{gio : City1})$$

– Relations among concepts of the multimedia and the domain ontologies: These relations connect structural aspects of multimedia objects with their domain specific content and are really indispensable for presentation and retrieval purposes of multimedia objects or entire documents with respect to end-user queries on the domain of interest. One characteristic relation of this kind is the mco : depicts relation which connects instances of multimedia segments, defined in the MCO, with instances of concepts defined in AEO or GIO. For example, we could use the relation mco : depicts to declare that a specific segment of a text denotes an instance of a stadium, or that a specific region of a still image denotes an instance of a person's face, using the following assertions:

$$\text{mco : depicts}(\text{mco : TextSegment2}, \text{gio : Stadium1})$$
$$\text{mco : depicts}(\text{mco : StillRegion2}, \text{aeo : PersonFace1})$$

5 Representation of Uncertainty

In the previous section, we have shown how to provide a formal representation of multimedia semantics using ontology languages, and more precisely OWL and its underlying technology of Description Logics (DLs). Although DLs are significantly expressive, they feature limitations when it comes to modelling domains where *imperfect*, like *uncertain* or *vague/fuzzy* information is apparent. This is often the case with the task of knowledge-based multimedia processing and interpretation. More precisely, image and video analysis algorithms are usually based on statistical criteria, thus the results they provide also contain *confidence degrees*. Moreover, it is also usual that the information that exists in a multimedia document is inherently vague, like for example the color (red, very red, blue, etc.), the size (large, small, etc) or the shape (long, circular, rectangular, etc.) of a specific object.

The representation and management of imperfect, uncertain and/or vague knowledge, is a huge topic that has received tremendous interest in AI (expert systems, natural language processing and understanding, etc.), in database management systems (relational schemata, deductive databases, etc.), in the field of knowledge representation and reasoning in general (probabilistic logic, Dempster-Shafer theory, Bayesian inference, subjective logic, etc.), and so forth; see [29] for a list of applications of fuzzy sets and fuzzy logic. Corresponding approaches have been developed in the context of ontology languages that extend the underlying mathematical frameworks so as to allow the formal handling of imperfect knowledge. Relevant proposals in the literature, include probabilistic DLs [16], probabilistic OWL [12], possibilistic DLs [40], as well as fuzzy DLs and fuzzy OWL [50–52].

As the aforementioned extensions model different types of imprecision, their appropriateness for a given application depends on the particular semantics involved. In the case of confidence degrees encountered in image and video analysis, the imprecision semantics lie in the nature of "confidence" captured in the computed degrees. Approaches where concepts are detected on the grounds of perceptual similarity, imply a prototypical set of feature values that constitute a visual/perceptual definition of the concept. As the presence of a concept is determined based on the similarity of those values, concepts can be considered as fuzzy sets, where the similarity (distance) function plays the role of the membership function. Contrariwise, approaches that utilise concepts' co-occurrence and correlation, pertain to a probabilistic/possibilistic interpretation of the associations between visual features and semantic concepts. Support Vector Machines [7] constitute a popular example of the former category, while Bayesian Nets [18] and Hidden Markov models fall in the latter.

Apparently, both types of imperfection pertain to the case of multimedia processing and interpretation, while the complementary aspects addressed, render each of them a crucial component towards complete and robust solutions. In this chapter though, we focus solely on handling the vagueness encompassed in the processing of multimedia content. Specifically, in the following, we go through the theory of fuzzy Description Logics, in order to provide an insight on how such

extended theories could be used to represent and reason with the imperfection of the processed multimedia documents We will provide examples on how fuzzy DLs can be used and a short overview of tools that can be used in practical applications.

5.1 Fuzzy Extensions of OWL and DLs

As is the case with classical OWL and Description Logics, fuzzy Description Logics provide the notions of concepts ($\mathbf{C}$), roles ($\mathbf{R}$) and individuals ($\mathbf{I}$) in order to represent the primitive concepts of our domain knowledge. So for example one can use the atomic (primitive) concepts $\mathsf{Blue}, \mathsf{Large}, \mathsf{Arm}, \mathsf{Person}, \mathsf{Car}$ in order to represent entities that are depicted in an image or video, primitive roles $\mathsf{hasColor}, \mathsf{hasPart}$ to describe binary relations or individuals car_1, $person_2$ in order to represent the specific objects of a specific image. Then concepts, roles and individuals are used together with the constructors in order to devise more complex concepts. For example using the construction of conjunction ($\sqcap$) we can describe the concept of blue cars by writing $\mathsf{Car} \sqcap \mathsf{BlueColored}$, or we can use the constructor of existential restrictions ($\exists$) together with the conjunction constructor to describe the notion of a clouded sky as $\mathsf{ClearSky} \sqcap \exists \mathsf{contains.Cloud}$. More formally, fuzzy-$\mathcal{SHOIN}$-concepts and roles are defined as follows.

Definition 1. *Let $RN \in \mathbf{R}$ be a role name and R be an f-$\mathcal{SHOIN}$-role. f-$\mathcal{SHOIN}$-roles are defined by the abstract syntax: $R ::= RN \mid R^-$, where R^- denotes the inverse of the role R. The inverse relation of roles is symmetric, and to avoid considering roles such as R^{--}, we define a function* Inv *which returns the inverse of a role, more precisely $\mathsf{Inv}(RN) := RN^-$ and $\mathsf{Inv}(RN^-) := RN$. The set of f-$\mathcal{SHOIN}$-concepts is the smallest set such that*

1. *every concept name $CN \in \mathbf{C}$ is an f-$\mathcal{SHOIN}$-concept,*
2. *if $o \in \mathbf{I}$ then $\{o\}$ is an f-$\mathcal{SHOIN}$-concept,*
3. *if C and D are f-$\mathcal{SHOIN}$-concepts, R an f-$\mathcal{SHOIN}$-role, S a* simple[12] *f-$\mathcal{SHOIN}$-role and $p \in \mathbb{N}$, then $(C \sqcup D)$, $(C \sqcap D)$, $(\neg C)$, $(\forall R.C)$, $(\exists R.C)$, $(\geq pS)$ and $(\leq pS)$ are also f-$\mathcal{SHOIN}$-concepts.*

As we can see, f-$\mathcal{SHOIN}$-concepts are fairly standard with respect to classical $\mathcal{SHOIN}$-concepts and roles [2].

Similarly to classical DLs, in fuzzy DLs one can also define new concepts using the notion of concept axioms. Let C and D be f-$\mathcal{SHOIN}$-concepts. Concept axioms of the form $C \sqsubseteq D$ are called *inclusion axioms*, while concept axioms of the form $C \equiv D$ are called *equivalence axioms*. Thus, we can describe intentional knowledge in the same way as the standard OWL language. For example we can provide the axiom:

$$\mathsf{CloudedSky} \equiv \mathsf{ClearSky} \sqcap \exists \mathsf{contains.Cloud}$$

[12] A role is called *simple* if it is neither transitive nor has any transitive sub-roles. Allowing only simple roles to participate in number restrictions is crucial in order to get a decidable logic [22].

that defines the new concept of clouded sky. A similar case can be made about roles, where we can capture partonomic relations with the aid of inverse roles, transitive role axioms, and role inclusion axioms.

The power of fuzzy Description Logics comes into play when one wants to represent instance assertions (individual axioms). More precisely, fuzzy ontology languages allow one to represent the degree to which an individual belongs to a concept. For example we could state that object obj_1 is Blue to a degree 0.9, or that it is Large to a degree 0.7. For these reasons in fuzzy ontologies, the notion of an assertion (or fact) is extended to that of a *fuzzy assertion* (or fuzzy fact) [52]. Fuzzy assertions are of the form $(a : C) \geq n_1$, $(a : D) = n_2$ $((a, b) : R) \geq n_3$ and so on, where C, D are concepts (classes) and n_1, n_2, n_3 are degrees from the unit interval ([0,1]).

A *fuzzy ontology* $\mathcal{O}$ consists of a set of the above axioms.

As with classical DLs, fuzzy-DLs provide for a formal meaning to their building blocks, thus they constitute a well-defined and semantic way of representing (vague) knowledge. Such fuzzy semantics are provided with the aid of the (relatively) standard notion of *fuzzy interpretation* introduced in [52]. Roughly speaking, concepts are interpreted as fuzzy sets and roles as fuzzy relations [29]. For example, considering the object $\texttt{Rome}^{\mathcal{I}}$, that denotes the city, and the fuzzy set $\mathsf{HotPlace}^{\mathcal{I}}$ that denotes hot places, a fuzzy set has the form $\mathsf{HotPlace}^{\mathcal{I}}(\texttt{Rome}^{\mathcal{I}}) = 0.7$, meaning that rome is a hot place to a degree equal to 0.7. Fuzzy interpretations can be extended to interpret complex f-$\mathcal{SHOIN}$-concepts and roles, with the aid of the fuzzy set theoretic operations defined and investigated in the area of fuzzy set theory [29]. The interested reader can refer to the wealth of fuzzy DL literature for the complete set of semantics [49, 51–53].

As with classical DLs, fuzzy DLs provide a set of inference services which can be used to query fuzzy ontologies. Interestingly, today there exist reasoning algorithms [50, 52] as well as practical reasoning systems. One such a system is FiRE (Fuzzy Reasoning Engine) which can be found at `http://www.image.ece.ntua.gr/~nsimou/FiRE` together with installation instructions and examples. FiRE currently supports f_{KD}-$\mathcal{SHIN}$, i.e. fuzzy-$\mathcal{SHOIN}$ without the nominal constructor.

Let us now see a specific example of the use of fuzzy DLs in the task of knowledge based multimedia processing. Consider for example pictures that depict athletics, like athletes performing high jump, pole vault, discus throw attempts etc. A segmentation algorithm is applied on such images to identify the different objects that are depicted as image segments. For each segment we can then extract their MPEG-7 visual descriptors. These are numerical values which provide information about the texture, shape and color of a region. One could use such values in order to *move* from low-level descriptions to more high-level ones. For example, if the green component in the RGB color model of region 1 (reg_1) is equal to 243, we can be based on a mapping (fuzzy partition) function [29] and deduce that reg_1 is GreenColored to a degree at least 0.8. On the other hand another region with a green component of 200 could be GreenColored to a degree 0.77. Similarly, we can extract additional fuzzy assertions using other MPEG-7

descriptors, like texture or shape. Subsequently, we can construct an ontology which could be used to provide semantic descriptions (definitions) of the optical objects that exist in our image. A sample ontology could be the following:

$$\begin{aligned} \mathsf{HorizontalBar} &\equiv \mathsf{RectangularShaped} \sqcap \mathsf{Elongated} \sqcap \mathsf{HorizontallyDirected}, \\ \mathsf{LandingPit} &\equiv \mathsf{BrownColored} \sqcap \mathsf{CoarseTextured} \sqcap \mathsf{RectangularShaped}, \\ \mathsf{PoleVault} &\equiv \mathsf{AthleticEvent} \sqcap \exists \mathsf{hasPart.HorizontalBar} \sqcap \exists \mathsf{hasPart.Pole} \end{aligned}$$

Finally, using concept axioms such as the above ones together with fuzzy assertions created by mapping MPEG-7 features to fuzzy concepts and inference services of fuzzy DLs, we can extract all the implied knowledge for a specific image. The following table provides a few examples of initially extracted concepts from MPEG-7 descriptors and inferred concepts using fuzzy-DL reasoning.

Table 1. Semantic labelling

Region	Extracted Concept	Degree	Inferred Concept	Degree
$region_1$	RectangularShaped	0.69	HorizontalBar	0.69
	Elongated	0.85		
	HorizontallyDirected	0.80		
$region_2$	BrownColored	0.85	LandingPit	0.73
	CoarseTextured	0.73		
	RectangularShaped	0.91		

More extended examples on the use of fuzzy-DLs in the context of multimedia processing and interpretation can be found in [9, 46].

6 Conclusions and Open Issues

Today a vast amount of multimedia documents exist in multimedia databases of TV channels, production companies, museums, film companies, sports federations, etc. But all this cultural heritage is almost completely lost or never reused since accessing them is highly inflexible, inefficient and extremely expensive. In most cases these multimedia documents lay in legacy systems free of content descriptions and searching for documents which depict particular content may take hours or even days. To solve this problem one has to provide appropriate ways to represent the multimedia content in a semantically rich and machine understandable way.

Representation of multimedia content semantics is one of the most important issues in the multimedia research community. Firstly, having the description of the content in a semantically rich form enables us to provide semantic access to multimedia documents. Moreover, with the advent of the semantic web publishing such content on the web enables interoperability and reuse of multimedia information. Additionally, the use of semantic technologies gives new possibilities in using inference and reasoning services for the tasks of assisting several

multimedia related tasks, like multimedia analysis. Several proposals for representing the semantics of multimedia documents or for using semantic technologies for performing knowledge-based multimedia processing and interpretation have been proposed in the literature. All these approaches have followed different modelling choices due to the fact that the resulting ontologies were used in different application scenarios or domains.

In the current chapter we have reported on our results of developing ways to represent multimedia content semantics within the BOEMIE project. We have presented four, interconnected ontologies, namely the Athletics Events Ontology (AEO), the Geographic Information Ontology (GIO), the Multimedia Content Ontology (MCO) and the Multimedia Descriptor Ontology (MDO). These ontologies are purposed to capture and represent the information that exists in different parts of multimedia documents. More precisely, the MCO ontology is purposed to represent the structural information of multimedia documents, the MDO ontology the low-level numerical information that is extracted by multimedia analysis modules, while AEO and GIO high-level knowledge about the domain that the specific multimedia documents depict. All aforementioned ontologies, although independently developed, are interlinked using several spatiotemporal relations in order to provide a global framework for representing the semantics of multimedia content. Furthermore, given the imprecision inherent both in the information conveyed by multimedia content and in multimedia analysis and processing, non-standard technologies based on fuzzy extensions to DLs, have been presented as possible means to represent and manage such type of information.

Compared with the relevant literature, the proposed Multimedia Semantic Model, and the opportunities for its extension through the use of fuzzy DLs for the formal handling of uncertainty, brings a number of additional advantages. First, the proposed framework addresses in an integral manner the core issues involved in the interpretation and semantic management of multimedia content, namely the representation and linking of domain with media specific notions in a manner that enables the utilisation of reasoning in a semantically rich way, the handling of imperfect knowledge in terms of vagueness, and the seamless interchange, sharing and reuse of both the background knowledge as well as the resulting semantic interpretations. The specialised ontology patterns proposed for the representation of primitive concepts extracted through analysis and of more complex ones, derivable by means of reasoning, constitute a significant contribution towards the first issue. The clean modelling and axiomatised media specific ontologies, especially with respect to the representation of content structure, constitute the main contribution compared to the existing MPEG-7 multimedia ontologies. Moreover, the advantages from the integral, multiple modalities, view taken on the issues involved, is further strengthened by the modular architecture and the extensible design followed.

Finally, based on the experiences drawn, future research directions and open issues may be summarised in the following.

- The multimedia ontologies have been developed with the aim to live in an evolving environment where apart from representation and reasoning, they will be used for the tasks of presentation, retrieval, learning and evolution. Thus, it remains to evaluate if the proposed architecture is sufficient to support also such tasks.
- First results have shown that DL based ontologies together with rule language, like DL-safe rules are expressive enough to be used for the task of multimedia interpretation and reasoning. On the other hand more extensive evaluation has to be performed in order to estimate the deficiencies and assess the value of DLs for such tasks.
- Currently, although a number of spatiotemporal relations have been used inference services do not go beyond traditional DLs. In order words true spatiotemporal reasoning is not supported. Obviously, such services are important for video analysis and representation as well as for representing image relations. It is an open issue on how existing spatiotemporal extensions to DL languages can be used for representing such multimedia content.

References

1. Arndt, R., Troncy, R., Staab, S., Hardman, L., Vacura, M.: COMM: Designing a Well-Founded Multimedia Ontology for the Web. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 30–43. Springer, Heidelberg (2007)
2. Baader, F., Nutt, W.: Basic description logics. In: Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (eds.) The Description Logic Handbook: Theory, Implementation, and Applications, pp. 43–95. Cambridge University Press, Cambridge (2003)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American (2001)
4. Bertini, M., Del Bimbo, A., Torniai, C.: Enhanced ontologies for video annotation and retrieval. In: Multimedia Information Retrieval, pp. 89–96 (2005)
5. Bloehdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, S., Strintzis, M.G.: Semantic annotation of images and videos for multimedia analysis. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 592–607. Springer, Heidelberg (2005)
6. Brickley, D., Guha, R.V.: OWL Web Ontology Language Overview, W3C Recommendation February 10 (2004), `http://www.w3.org/TR/owl-features/`
7. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
8. Dasiopoulou, S., Heinecke, J., Saathoff, C., Strintzis, M.G.: Multimedia reasoning with natural language support. In: 1st IEEE International Conference on Semantic Computing (ICSC), Irvine, CA, USA (2007)
9. Dasiopoulou, S., Kompatsiaris, I., Strintzis, M.G.: Investigating fuzzy DLs-based reasoning in semantic image analysis. In: Multimedia Tools Appl., S.I. Semantic Multimedia (2009), doi:10.1007/s11042-009-0393-6

10. Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V.K., Strintzis, M.G.: Knowledge-assisted semantic video object detection. IEEE Trans. Circuits Syst. Video Techn. 15(10), 1210–1224 (2005)
11. Dasiopoulou, S., Tzouvaras, V., Kompatsiaris, I., Strintzis, M.G.: Enquiring MPEG-7 based multimedia ontologies. Multimedia Tools and Applications 46(2-3), 331–370 (2010)
12. Ding, Z., Peng, Y.: A Probabilistic Extension to Ontology Language OWL. In: Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37), Big Island, Hawaii, p. 10 (January 2004)
13. Gangemi, A.: Ontology design patterns for semantic web content. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 262–276. Springer, Heidelberg (2005)
14. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002)
15. Gangemi, A., Mika, P.: Understanding the Semantic Web through descriptions and situations. In: Chung, S., Schmidt, D.C. (eds.) CoopIS 2003, DOA 2003, and ODBASE 2003. LNCS, vol. 2888, pp. 689–706. Springer, Heidelberg (2003)
16. Giugno, R., Lukasiewicz, T.: P-$\mathcal{SHOQ}(\mathbf{D})$: A probabilistic extension of $\mathcal{SHOQ}(\mathbf{D})$ for probabilistic ontologies in the semantic web. In: Flesca, S., Greco, S., Leone, N., Ianni, G. (eds.) JELIA 2002. LNCS (LNAI), vol. 2424, pp. 86–97. Springer, Heidelberg (2002)
17. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. In: Inter. Workshop on Formal Ontology, Padua, Italy (March 1993)
18. Heckerman, D.: A tutorial on learning with bayesian networks. Learning in Graphical Models, 301–354 (1998)
19. Hollink, L., Little, S., Hunter, J.: Evaluating the application of semantic inferencing rules to image annotation. In: 3rd International Conference on Knowledge Capture (K-CAP), Banff, Alberta, Canada, pp. 91–98 (2005)
20. Hollink, L., Worring, M.: Building a visual ontology for video retrieval. In: Proc. 13th ACM International Conference on Multimedia, Singapore, November 6-11, pp. 479–482 (2005)
21. Horrocks, I., Patel-Schneider, P.F.: Reducing OWL entailment to Description Logic satisfiability. J. Web Sem. 1(4), 345–357 (2004)
22. Horrocks, I., Sattler, U., Tobies, S.: Practical reasoning for expressive description logics. In: Ganzinger, H., McAllester, D., Voronkov, A. (eds.) LPAR 1999. LNCS (LNAI), vol. 1705, pp. 161–180. Springer, Heidelberg (1999)
23. Hu, B., Dasmahapatra, S., Lewis, P.H., Shadbolt, N.: Ontology-based medical image annotation with Description Logics. In: Proc. Inter. Conference on Tools with Artifical Intelligence (ICTAI), Sacramento, California, November 3-5 (2003)
24. Hudelot, C., Thonnat, M.: A cognitive vision platform for automatic recognition of natural complex objects. In: ICTAI, pp. 398–405 (2003)
25. Hunter, J.: Adding Multimedia to the Semantic Web: Building an MPEG-7 Ontology. In: Proc. The First Semantic Web Working Symposium, SWWS 2001. Stanford University, California (July 2001)
26. Hunter, J., Drennan, J., Little, S.: Realizing the hydrogen economy through Semantic Web technologies. IEEE Intelligent Systems 19(1), 40–47 (2004)
27. Hyvonen, E., Styrman, A., Saarela, S.: Ontology-based image retrieval. In: XML Finland Conference, October 21-22, pp. 15–27 (2002)

28. Jaimes, A., Tseng, B.L., Smith, J.R.: Modal keywords, ontologies, and reasoning for video understanding. In: Bakker, E.M., Lew, M., Huang, T.S., Sebe, N., Zhou, X.S. (eds.) CIVR 2003. LNCS, vol. 2728, pp. 248–259. Springer, Heidelberg (2003)
29. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice-Hall, Englewood Cliffs (1995)
30. Goodall, S., Grimwood, P., Kim, S., Lewis, P., Martinez, K., Stevenson, A., Addis, M., Boniface, M.: Sculpteur: Towards a new paradigm for multimedia museum information handling. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 582–596. Springer, Heidelberg (2003)
31. Maillot, N., Thonnat, M.: A weakly supervised approach for semantic image indexing and retrieval. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 629–638. Springer, Heidelberg (2005)
32. Martínez, J.M.: MPEG-7: Overview of MPEG-7 Description Tools, Part 2. IEEE MultiMedia 9(3), 83–93 (2002)
33. Meghini, C., Sebastiani, F., Straccia, U.: A model of multimedia information retrieval. Journal of the ACM 48(5), 909–970 (2001)
34. Möller, R., Neumann, B.: Ontology-based reasoning techniques for multimedia interpretation and retrieval. In: Semantic Multimedia and Ontologies: Theory and Applications (2008) (to appear)
35. MPEG-21. Multimedia Framework (MPEG-21) - ISO/IEC TR 21000-1:2004 (2002)
36. Nack, F., van Ossenbruggen, J., Hardman, L.: That obscure object of desire: Multimedia metadata on the Web, Part 2. IEEE MultiMedia 12(1), 54–63 (2005)
37. Neumann, B., Möller, R.: On scene interpretation with description logics. Technical Report FBI-B-257/04 (2004)
38. Oberle, D., Ankolekar, A., Hitzler, P., Cimiano, P., Sintek, M., Kiesel, M., Mougouie, B., Baumann, S., Vembu, S., Romanelli, M.: DOLCE ergo SUMO: On foundational and domain models in the SmartWeb Integrated Ontology (SWIntO). J. Web Sem. 5(3), 156–174 (2007)
39. Petridis, K., Anastasopoulos, D., Saathoff, C., Timmermann, N., Kompatsiaris, Y., Staab, S.: M-ontoMat-annotizer: Image annotation linking ontologies and multimedia low-level features. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4253, pp. 633–640. Springer, Heidelberg (2006)
40. Qi, G., Pan, J.Z., Ji, Q.: Extending description logics with uncertainty reasoning in possibilistic logic. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 828–839. Springer, Heidelberg (2007)
41. Celma, O., García, R.: Semantic Integration and Retrieval of Multimedia Metadata. In: Proc. International Semantic Web Conference (ISWC), Galway, Ireland, November 6-10 (2005)
42. Celma, O., Halaschek-Wiener, C., Mannens, E., Troncy, R., Boll, S., Burger, T.: Multimedia vocabularies on the Semantic Web. In: W3Cu Incubator Group Report, July 24 (2007)
43. Schober, J.P., Hermes, T., Herzog, O.: Content-based image retrieval by ontology-based object recognition. In: Haarslev, V., Lutz, C., Möller, R. (eds.) Proc. Workshop on Applications of Description Logics, Ulm, Germany (2004)
44. Schreiber, A.T., Dubbeldam, B., Wielemaker, J., Wielinga, B.J.: Ontology-based photo annotation. IEEE Intelligent Systems 16(3), 66–74 (2001)
45. Di Sciascio, E., Donini, F.: Description logics for image recognition: a preliminary proposal. In: Proceedings of the International Workshop on Description Logics (DL 1999) (1999)

46. Simou, N., Athanasiadis, T., Tzouvaras, V., Kollias, S.: Multimedia reasoning with f-$\mathcal{SHIN}$. In: 2nd International Workshop on Semantic Media Adaptation and Personalization, London, United Kingdom, December 17-18 (2007)
47. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22(12), 1349–1380 (2000)
48. Staab, S., Studer, R. (eds.): Handbook on Ontologies, 2nd edn. Springer, Heidelberg (2009)
49. Stoilos, G., Stamou, G.: Extending fuzzy description logics for the semantic web. In: Proceedings of the 3rd International Workshop on OWL Experiences and Direction (OWL ED 2007) (2007)
50. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrocks, I.: Reasoning with very expressive fuzzy description logics. Journal of Artificial Intelligence Research 30(5), 273–320 (2007)
51. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrocks, I.: Fuzzy OWL: Uncertainty and the semantic web. In: Proc. of the International Workshop on OWL: Experiences and Directions (2005)
52. Straccia, U.: Reasoning within fuzzy description logics. Journal of Artificial Intelligence Research 14, 137–166 (2001)
53. Straccia, U.: Towards a fuzzy description logic for the semantic web. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 167–181. Springer, Heidelberg (2005)
54. Troncy, R., Celma, O., Little, S., GarciaGarcía, R., Tsinaraki, C.: Mpeg-7 based Multimedia Ontologies: Interoperability Support or Interoperability Issue? In: Proc. 1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies (MARESO), Genova, Italy, pp. 2–16 (2007)
55. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Integration of OWL ontologies in MPEG-7 and TV-anytime compliant semantic indexing. In: Persson, A., Stirna, J. (eds.) CAiSE 2004. LNCS, vol. 3084, pp. 398–413. Springer, Heidelberg (2004)
56. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support between mpeg-7/21 and OWL in ds-mirf. IEEE Trans. Knowl. Data Eng. 19(2), 219–232 (2007)
57. van Ossenbruggen, J., Nack, F., Hardman, L.: That obscure object of desire: Multimedia metadata on the web, part 1. IEEE MultiMedia 11(4), 38–48 (2004)

Semantics Extraction from Images

Ioannis Pratikakis[1,2], Anastasia Bolovinou[1],
Bassilios Gatos[1], and Stavros Perantonis[1]

[1] NCSR "Demokritos", Institute of Informatics and Telecommunications,
15310 Ag. Paraskevi, Athens, Greece
[2] Democritus University of Thrace,
Department of Electrical and Computer Engineering,
67100 Xanthi, Greece
ipratika@ee.duth.gr

Abstract. An overview of the state-of-the-art on semantics extraction from images is presented. In this survey, we present the relevant approaches in terms of content representation as well as in terms of knowledge representation. Knowledge can be represented in either implicit or explicit fashion while the image is represented in different levels, namely, low-level, intermediate and semantic level. For each combination of knowledge and image representation, a detailed discussion is addressed that leads to fruitful conclusions for the impact of each approach.

1 Semantics Extraction Basic Pipeline

Semantics extraction refers to digital data interpretation from a human point of view. In the case that the digital data correspond to images, this usually entails an appearance-based inference using color, texture and/or shape information along with a type of context inference (or representation) that can combine and transform these machine-extracted evidence into what we call a scene description. Following Biederman *et al.* [1] definitions of context in a visual scene, we can derive three types of context for real-world scene annotation problems: (i) semantic context which encodes the probability of a certain category to be present in a scene (e.g category "streets" has high probability to coexist with category "building" in the same scene); (ii) spatial context which encodes the spatial relations of categories (e.g sky is usually above grass in a scene) and (iii) scale context which encodes the relative object size (category "human" is expected to occupy a small region in a scene which includes the"building" category).

The research goals in semantics extraction are mostly a function of the granularity of the semantics in question. The goal could be the extraction of a single or multiple semantics of the entire image (e.g. indoor/outdoor setting), or the extraction of the semantics for different objects in an image. In the latter case, the semantics could be generic (e.g. a vehicle) or specific (e.g. a motorbike). Those goals make it clear that semantics extraction is not a new research area. Depending on the goal, the task of semantics extraction can be considered as a categorization, classification, recognition and understanding task that all share in

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 50–88, 2011.

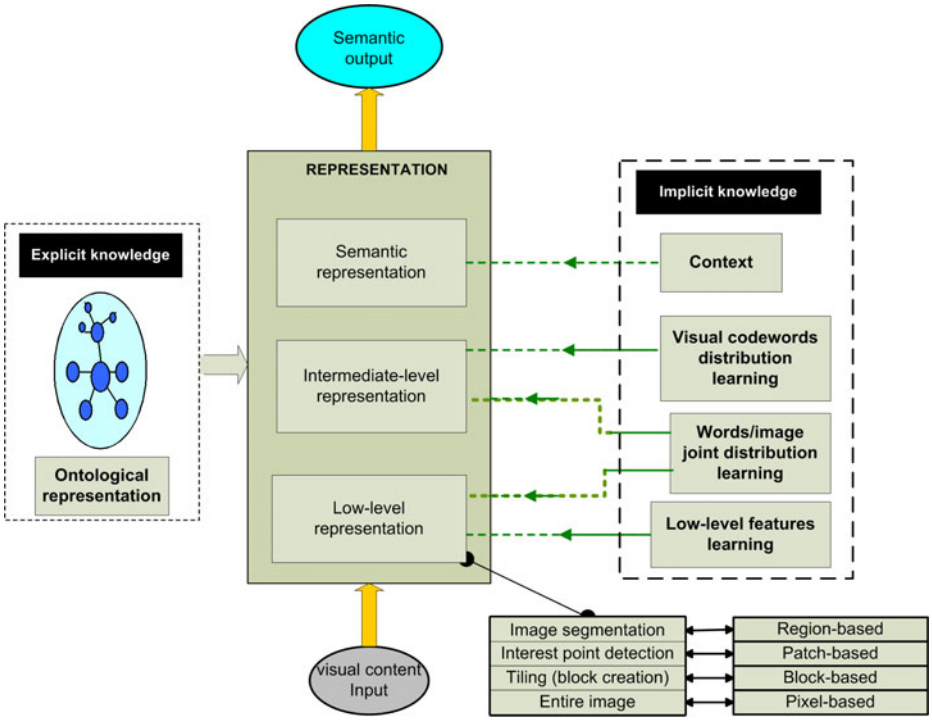


Fig. 1. Schematic diagram of the potential processes for semantics extraction

common the effort for solving the semantic gap. As stated in [2], "Despite intensive recent research, the automatic establishment of a correspondence between the low-level features and the semantic-level information needed to understand the content of the visual medium is a problem still far from being solved or adequately addressed".

In this section, we will report on the existing methodological trends for semantics extraction from images based upon a basic pipeline schema that is shown in Figure 1. Those methodologies are mainly developed by taking into account alternatives from two main axes. The first axis concerns visual content representation, the second axis concerns knowledge representation.

In the case of visual content representation, there exist three possibilities. In the first, we have a low-level representation that can be supported by (i) image segmentation (region-based); (ii) interest point detection (patch-based); (iii) creation of blocks by tiling (block-based) and (iv) the entire image. In the second, we have an intermediate representation that can be supported by visual vocabularies [3]. Finally, the third possibility concerns certain relationship (i.e. spatial) among semantic objects.

In the case of knowledge representation, there are two main trends that depend upon the type of knowledge which is used. Knowledge is involved in a semantics

extraction process either in an implicit or in an explicit way. The former type of knowledge refers to the kind that can be captured from the patterns in data and its validity is of a statistical nature. We assume that machines can analyze dataset's implicit semantics with several, mostly statistical, techniques (see [4]). The latter type of knowledge refers to the kind that is represented in any strict machine processable syntax and is based on a prior domain knowledge. While the former type of knowledge comes with the data and can be pursuit (mathematically) for almost all cases, the latter type depends on prior assumptions which do not always exist.

Based on the image representation level they are built upon, three basic strategies can be found in the literature for implicit knowledge derivation and to which the following three chapters of this work are dedicated. The first strategy (chapter 2) uses low-level features such as color, texture, power spectrum, etc. In this case, implicit knowledge is derived from a low-level image representation. This approach considers i.e. the scene in an image as an individual object [5], [6] and is normally used to classify only a small number of scene categories (indoor versus outdoor, city versus landscape, etc.). The second strategy (chapter 3) uses intermediate representations composed by low-level features' clusters [7] and corresponds to methods built on top of a visual vocabulary. The third strategy (chapter 4) makes use of more complex semantic strategies in order not only to detect objects in an image, but also model the relationships between the detected objects taking into account contextual information. While the methods included in chapter 2 lie on purely appearance-based features, in chapter 3 efforts to couple appearance with spatial or and scale context will be discussed. Finally, in chapter 4, efforts to couple appearance with semantic context will be presented.

For explicit knowledge integration, a prominent position is taken in the literature by ontological knowledge representations due to advantages they exhibit like the provision of a formal framework for supporting explicit, machine-processable semantics definition as well as the ability to derive new knowledge through automated inference. In the following, we will discuss existing approaches that use either implicit (chapters 2,3 and 4) or explicit knowledge (chapter 5) for different levels of representation to address semantics extraction from images. Furthermore, we will also discuss about an important trend which leads to improved semantics extraction that is based upon an interplay between image segmentation and recognition approaches (chapter 6). Let us note here that although we make an effort to report on interesting different learning algorithms applied and features used, our categorization effort is focused in the association of content representation and knowledge inference involved in each work. In this way we strive towards to comprehend the semantic granularity that can be achieved based on a specific image representation and the available inference mechanism. Based on such a deployment of the state of the art techniques, we believe is easier for the reader to form a critical view on new and heterogeneous on their entity methods, depending on the type and requirements of semantic's extraction he/she is interested in.

2 Implicit Knowledge in Low-Level Representation

In this section, the use of implicit knowledge within a low-level image representation will be discussed. According to the type of implicit knowledge used, approaches are distinguished into those which image categorization is based on "low-level features learning" and those which learn the association of keywords and image content, expressed as the joint distribution of words and images. Corresponding architecture diagrams to these two knowledge acquisition methods are included in sections 2.1 and 2.2.

2.1 Low-Level Features Learning

The problem of image categorization is often addressed by computing low-level features (e.g. color and texture), which are processed with a classifier engine for inferring high-level information about the image. These methods consider that an image category can be directly described by the color/texture properties of the image and are mainly suitable for scene-based categorization. For instance, a forest scene presents highly textured regions (trees), a mountain scene is described by an important amount of blue (sky) and white (snow), or the presence of straight horizontal and vertical edges denotes an urban scene. A number of recent studies have presented approaches that classify generic semantics as indoor vs. outdoor, or city vs. landscape, using global cues (e.g. power spectrum, color histogram information). Among them, two trends are mainly distinguished: (i) Global: the object or a scene in an image is described by low-level features computed from the entire image; (ii) Local: the image is first partitioned into several blocks or regions, and then features are extracted from each of those blocks or regions. In both cases, the representation consists in the extraction of ordered features of equal length measured from the image or a set of an image partitions. These are usually low-dimensional features obtained from a filter applied in the image intensity values domain.

In the following, representative discriminative or/and generative approaches will be briefly described. The corresponding semantic output is achieved as shown in Figure 2.

Semantics extraction for scene classification based on low-level global image representation is addressed first, presenting a representative approach where on top of low-level representation a discriminative classifier is applied. Support Vector Machine (SVM) classifiers have shown promising results for visual classification tasks, and the development of specialized kernels suitable for use with local features has emerged as a fruitful line of research [8].

In [9], [10] global features are used to produce a set of semantic labels with a certain belief for each image. They manually label each training image with a semantic label and train k classifiers (one for each semantic label) using SVMs. Each test image is classified by the k classifiers and assigned a confidence score for the label that each classifier is attempting to predict. As a result, a k-nary label-vector consisting of k-class membership is generated for each image.

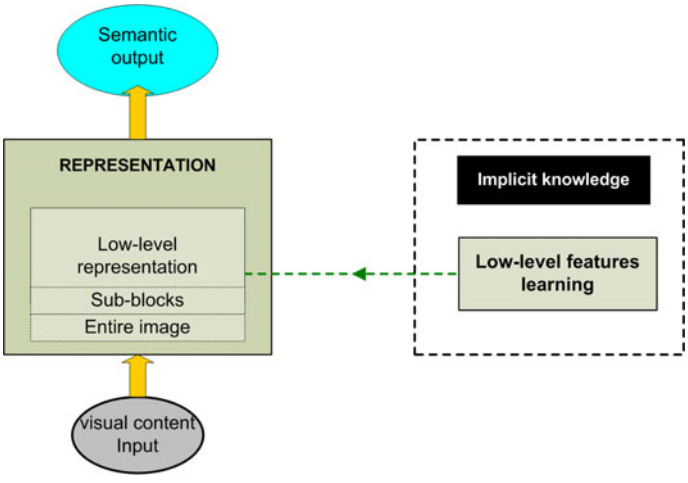


Fig. 2. Semantics extraction based on learning low-level features from the entire image or its block-based representation

This approach is especially useful for Content-Based Image Retrieval (CBIR) combined with Relevance Feedback (RF) systems.

Several attempts that aim to support a local representation, first split the image into a set of sub-regions, which are independently described by their low-level properties. These blocks are then classified, and finally the image is categorized from the individual classification of each block. This approach is originated in 1997, when Szummer and Picard [5] proposed to independently classify image subsections for obtaining a final result using a majority voting classifier. The goal of this work was to classify images as indoor or outdoor. The image is first partitioned into 16 sub-blocks from which Ohta space color histograms and MSAR texture features are extracted. K-NN classifiers are employed to classify each sub-block using the histogram intersection norm, which measures the amount of overlap between corresponding buckets in the two N-dimensional histograms. Finally, the entire image is classified using a majority voting scheme from the sub-block classification results.

The proposal of Serrano *et al.* [11] shares the same philosophy, but using SVM for a reduction in feature dimensionality without compromising classification accuracy. Also, color and texture features are extracted from image sub-blocks and separately classified. Thus, indoor/outdoor labels are obtained for different regions of the scene. The advantage of using SVM instead of K-NN classifier is that the sub-block beliefs can be combined numerically rather than by majority voting, which minimizes the impact of sub-blocks with ambiguous labeling. Both nearest-neighbor or voting-based classifiers followed by an alignment step (e.g. [12]) are quite common for image classification. Still, both may be impractical for large training sets, since their classification times increase with the number of training examples. An SVM, on the other hand, identifies a sparse

subset of the training examples (the support vectors) to delineate a decision boundary.

Global features enjoy the merits of low-dimensionality and fast extraction, they are able to offer a representation of the scene's gist [7] but they can't cope with scene's uneven illumination or affine transformations often present in real-world images. Hence,following evidence that humans seem to integrate both global and local information in order to visually understand a scene (see [13], [14]), recent approaches use global information only in a complementary to local properties fashion. On the other hand, as already discussed in [15], a problem with the methods using image features directly for scene categorization is that it is often difficult to generalize these methods to additional image data beyond the training set. More importantly, they lack an intermediate semantic image description that can be extremely valuable in determining the scene type. Systems that do attempt to find objects or semantic concepts based on intermediate representations or an hierarchy of concepts are described in section 3 and 4 respectively.

2.2 Words/Image Joint Distribution Learning

In this section, methods that are still based on global or local image representations but a label inference is addressed within a probabilistic framework will be presented. In the class of semantics extraction algorithms that an association between keywords and images is addressed (image annotation applications), the underlying principle is to create models of keywords in terms of visual features that can be extracted from images. These models focus on finding the joint probability of images and concepts (keywords).

Joint density estimation of keywords and visual features is an unsupervised procedure producing annotations following the criterion of the largest joint likelihood under the assumed mixture model. Generally, unsupervised labeling leads to significantly more scalable (in database size and number of concepts of interest) training procedures, places much weaker demands on the quality of the manual annotations required to bootstrap learning, and produces a natural ranking of keywords for each new image to annotate. So far, in existing approaches the following models have been used: (i) single-class model; (ii) translation model and (iii) hierarchical model.In this section, single-class models are addressed as they are built upon a low-level representation of the image. The two other models are built upon intermediate and semantic representations and thus, will be addressed later in sections 3.2 and 4, respectively.

In single-class-model approaches, we can estimate an individual distribution function for each keyword. Basically, the idea behind these approaches is to learn a class-conditional probability distribution of each single keyword w of the semantic vocabulary given its training data x. Equivalently, in this case implicit knowledge takes the form of learning keywords and image joint distribution, where images are represented by global low-level features. The corresponding architecture diagram that shows the image representation level associated with this type of knowledge inference is presented in Figure 3.

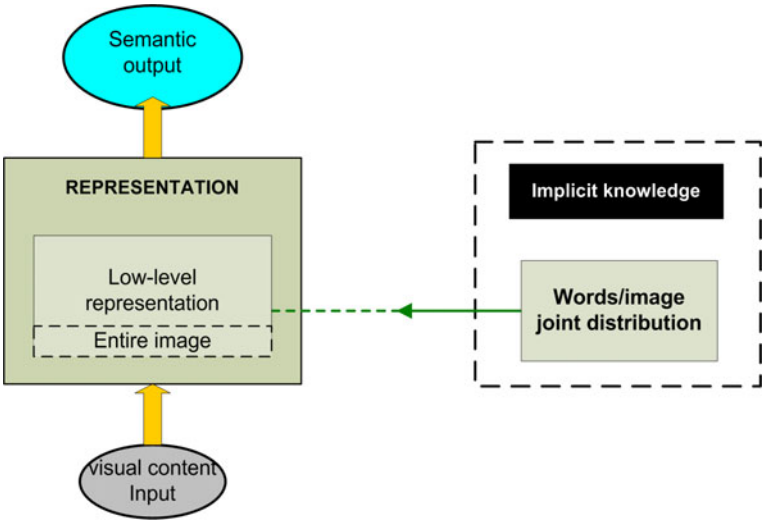


Fig. 3. Semantics extraction based on learning words and images joint distribution using a low-level image representation

Bayes law is used to invert the problem and model p($x|w$) - the features density distribution of a given keyword. Several techniques to model p($x|w$) with a simple density distribution have been proposed: Yavlinsky *et al.* [16] deployed a nonparametric distribution; Carneiro and Vasconcelos [17] a semi-parametric density estimation; Westerveld and de Vries [18] a finite-mixture of Gaussians while Mori *et al.* [19] and Vailaya *et al.* [6] apply different flavors of vector quantization techniques.

In all the above approaches binary Bayesian classifiers are used in an attempt to capture high-level concepts from low-level image features under the constraint that the test image belongs to one of the classes. This approach considers only the class own data ignoring the in-between classes co-occurrence. In most generative model based approaches, the correlations between keywords are ignored to simplify the model calculations. Recently, it has been realized that the correlations between annotated keywords can be used to improve the performance of image annotation [20]. For instance, the keyword set {sky, grass} has a larger probability to be an image caption than {ocean, grass}.

3 Implicit Knowledge in Intermediate Representation

Instead of using directly global image features for object/scene categorization, several approaches make use of an intermediate image description composed by clusters of low-level descriptors (unsupervised learning). In these methods, knowledge acquisition may either involve one step where joint distribution of visual words and keywords is learnt (translation models) or two steps where first

we obtain a global visual vocabulary by clustering descriptors from a training set, and then we represent each image as a histogram of visual words. The remainder of section 3 details the aforementioned alternatives.

3.1 Bag of Words Model and Beyond

Practical needs for robust object recognition under various imaging conditions and recent advances in machine learning algorithms (clustering and classification) opened the road for methods based on a dense local representation of an image. A popular approach is the use of vector quantization on features extracted from image patches to generate codebooks for representation and retrieval. In codebook representation distributions of the feature vectors are estimated and used as signatures. As this model is inspired by former work on data compression and especially text classification [21], the name "bag of words" has been dominated (also known as "bag-of-features" or "bag-of-keypoints" model), while the codebook produced is called visual vocabulary [22]. For reasons of clarity, in our report we will use the term "bag of visual words" (abbreviated to "BoVWs" for practical reasons in the rest of the text) to differentiate from the text-classification related research.

Quantization of robust appearance descriptors extracted from local image patches has been proved an effective means of capturing image statistics for texture analysis and scene classification [7]. Traditional texture models [23], [24] first identify a large vocabulary of useful textons (the codewords). Then, for each category of texture, a model is taught to capture the signature distribution of these textons. We could loosely think of a texture as one particular intermediate representation of a complex scene. After the influential work of [25] on local scale-invariant descriptors making use of a BoVWs model for classification, many studies have followed regarding affine invariant local descriptors that could be used for the construction of a visual vocabulary. The main advantage of such a vocabulary, apart from its simplicity (it uses histograms and no underline data distribution is assumed), is that it inherits the invariance properties of these local descriptors.

In the case of a BoVWs model, as a first step, vocabularies are constructed by using a method such as k-means to cluster the descriptor vectors of patches sampled either densely (on a grid) or sparsely (based on keypoints or salience measures) from a set of training images. After that, each training/test image or region (depending on the annotated data available) has a word distribution h (histogram over the obtained vocabulary) associated with it, thus rendered available for comparison. As stated in [26], the richness in the mathematical formulation of signatures grows together with the invention of new methods for measuring similarity. A similarity metrics' study applied in image multidimensional feature space useful for visual histograms/signatures comparison can be found in [27]. A schematic diagram of the BoVWs model, taken from [15] is presented in Figure 4. In a second step (classification), a discriminative or generative model learns to classify input images based on their visual words histograms.

An overall architecture diagram showing the implicit knowledge inference for semantics extraction in the case of bag of words model is presented in Figure 5.

Apart from texture, other features as color and shape have already been used for visual vocabulary creation [28]. The visual vocabulary construction determines the expressiveness and the discriminatory power of the method [28], [29]. Recent related approaches to object classification by building optimal visual vocabularies from local invariant computed descriptors can be found in [30], [31], [32]. In [33], it is proposed to use the similarity to all codebook vocabulary elements (multi-assignment), retaining expressiveness and discriminatory power, while in [34], different meaningful assignments techniques are studied.

An inherent problem of the BoVWs model is the choice of the resolution of the visual vocabulary. An excessively fine quantization causes features from two images to never match (overfitting), while an excessively coarse quantization yields non-discriminative histograms (bias). Grauman *et al.* [35] proposed Pyramid Matching Kernel to overcome this issue. The idea is to work with a sequence of R progressively coarser vocabularies $B_0, B_1, .., B_{R-1}$ and to define a similarity measure as a positive combination of the BoVWs similarities at the various levels. The formulation yields a proper (positive definite) Mercer kernel.

Whether BoVWs can naturally meet the challenges such as generalization and scalability of visual classification, remains to be proved due to various implementation choices and the absence of any geometric information inherent in the baseline model. Note that, although the generation of the visual vocabulary is performed off-line, it is time consuming and becomes intractable as the number of features increases (>200k). For that reason, hierarchical clustering schemes have recently been developed to cope with large datasets [36]. Even further, promising results were presented when linear k-means algorithm is replaced by its Histogram Intersection kernel counterpart [37] allowing for the clustering scheme to use the non-linear "min" distance (instead of the eulidean distance), making use of the popular kernel trick.

Many researchers explored the strength of BoVWs representation and tried to extend it. In the following, one generative approach, three discriminative approaches (the first using a binary classification set-up while the two others using a multi-class classification set-up), and one combined generative/discriminative approach, applied on top of a BoVWs model are described for object categorization. Finally, the chapter concludes by reporting on methods which aim to integrate spatial and scale context within the traditional BoVWs model.

In [38], primitive geometrical structures as scale normalized straight edges, ridges and blobs have been used for object intermediate-level representation; this has enabled the construction of simple templates for the related structures and the efficient estimation of the likelihood of arbitrary feature poses. The incorporation of top-down information is thus enabled, based on an efficient method for the evaluation of the likelihood of hypothesized part locations. The decisions of the individual detectors are combined using a probabilistic model that simultaneously accounts for the distributions of all codebook entries. This allows them to use graphical modeling techniques to complement bottom-up

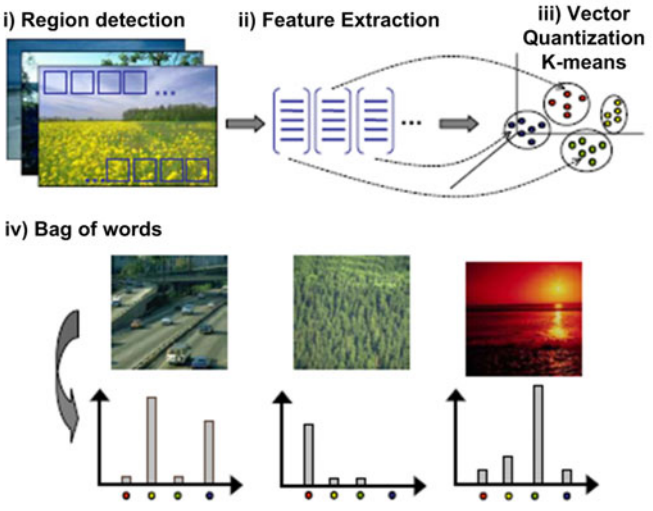


Fig. 4. Four steps to compute the "bag of words": (i-iii) obtain the visual vocabulary by vector quantizing the feature vectors and (iv) compute the image histograms "bag-of-words" for test images according to the obtained vocabulary. Figure is taken from [15].

detection, by proposing and finding the parts of the object that were missed by the front-end feature detection stage. Excellent detection results are obtained by combining bottom-up with top-down without complex appearance descriptors, using a small codebook representation and efficient algorithms.

A quite common and popular approach is the one presented in Csurka *et al.* [39], where the BoVWs model appears as"bag-of-keypoints" model. The approach is based on vector quantization of scale invariant feature transform (SIFT) [40]descriptors of image patches. After a k-means clustering that permits to assign descriptors to clusters in order to form image feature vectors, an SVM classifier is applied to perform the desired categorization and subsequent semantics extraction. In this approach, it is assumed that local patches of an image are independent of each other. A similar approach is followed by [32]. For a comparative study on image classification approaches combining a BoVWs representation with a kernel-based learning method (SVM classifier) which tests the limits of its performance on the most challenging databases available today the interested reader should refer to [41].

In the case of multi-class object recognition, the discriminative approaches to the multi-class setting of [42] and [43] are presented in the sequel. A hybrid of a Nearest Neighbor (NN) coupled with an SVM classifier is proposed in [42] that can deal with multi-class settings with reasonable computational complexity both at training and at run time. The basic idea is to find close neighbors to a query sample and train a local SVM that preserves the distance function on the collection of neighbors. In feature space, the histogram of textons is employed for the texture [24] and the geometric blur descriptor for shape [44]. Following a

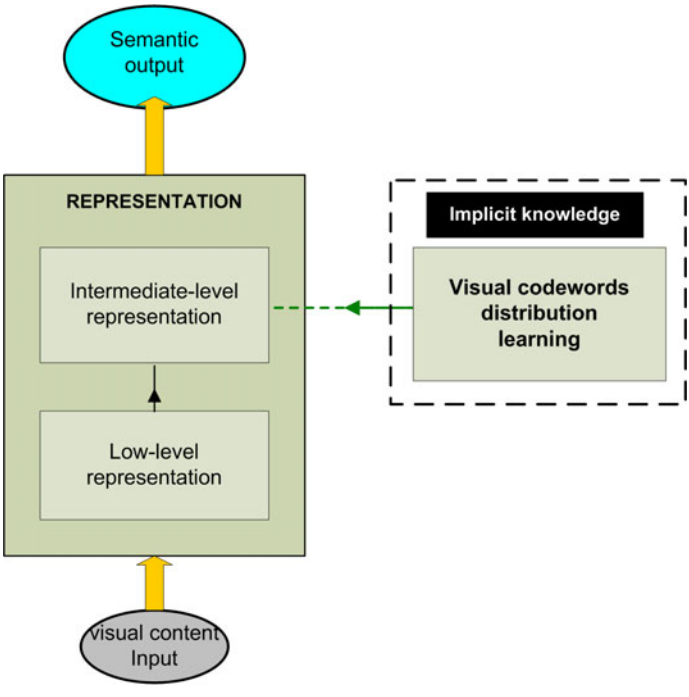


Fig. 5. Semantics extraction in which the domain knowledge is acquired from visual codewords' distribution learning

discriminative approach to classification, NN is used as an initial pruning stage and SVM is used on the smaller but more relevant set of examples that require careful discrimination. Their method is called "SVM-KNN" (where K signifies the method's dependence on the selected number of neighbors).

Semantics extraction is not a traditional supervised learning problem because the training image set does not provide explicit correspondence between keywords and regions. In particular, keywords are associated with images instead of individual regions. To circumvent this problem, in [43], it is proposed to learn an explicit correspondence between image regions and keywords through a Multiple-Instance Learning (MIL) algorithm, namely ASVM (Asymmetrical Support Vector Machine-based MIL), which constitutes a variation of supervised learning as the task is to learn a concept given positive and negative bags of instances. Image segmentation is derived using Normalized-cuts and then a 33 dimensional low-level feature vector is extracted from each region, which includes region color and standard deviation, region average orientation energy (12 filters), region size, location, convexity, first moment, and ratio of region area to boundary length squared. In region-based image annotation, each region is an instance, and the set of regions that comes from the same image can be treated as a 'bag'. From a collection of labeled bags (images), the learner tries to induce a concept that will label individual instances (regions) correctly. In the proposed MIL

framework, an image is annotated by keyword w_i if at least one region in the image has the semantic meaning of w_i. Given an image labeled by keyword w_i, we can expect that at least one region will correspond to w_i even if the segmentation is imperfect.

In [45], a unifying view of state of-the-art techniques for semantic-based image annotation and retrieval is presented in order to explain the motivation for the introduction of a new technique. Therein, limitations of the two basic approaches identified, namely "supervised one vs. all labeling" and "unsupervised labeling", are reported. The first corresponds to the case in which for each concept a binary classifier is trained to detect positive and negative examples of concept's presence or absence (discriminative approach). The second corresponds to the case in which the joint distribution of semantic labels and visual features are modeled in an unsupervised manner using latent variables (generative approach). To address limitations of these two approaches, Supervised Multiclass Labeling (SML) is introduced by [45] which follows a probabilistic formulation for semantic image annotation and retrieval.

In SML formulation, images are represented as bags of localized feature vectors, a mixture density estimated for each image, and the mixtures associated with all images annotated with a common semantic label pooled into a density estimate for the corresponding semantic class. This pooling is justified by a multiple instance learning argument and performed efficiently with a hierarchical extension of expectation-maximization. In this multiple instance learning perspective the question of whether the densities of a semantic class can be estimated without a prior semantic segmentation of the data base is addressed. The benefits of the supervised formulation over the more complex, and currently popular, joint modeling of semantic label and visual feature distributions (unsupervised techniques), are illustrated through theoretical arguments and extensive experiments. The supervised formulation is shown to achieve higher accuracy than various previously published methods at a fraction of their computational cost.

Although "bag of visual words" models have recently demonstrated impressive levels of performance, especially for whole-image categorization tasks [35], they suffer from clutter and occlusions as they are based on orderless local features representation. Context information based on the interaction among possibly existent objects in the scene or on global scene statistics, can provide an essential aid to disambiguate appearance inputs in recognition tasks. In the following, we focus on efforts which aim to incorporate scale and spatial context into a baseline BoVWs model. Note that semantic context as defined in the introduction is discussed separately in ch. 4.

Deriving spatial context from the local features appearance in an image is a tempting research area as the use of spatial relations usually leads to much higher costs in the learning and matching procedure. Many works incorporating geometric information to the BoVWs model follow a graph-based representation approach, which will usually result in a computational complexity increase, which is exponential or polynomial to the number of features. Constellation

models [46], [47] represent the objects with a fixed number of parts which are composed with the local features, and capture the geometry information by modeling the spatial layout of the parts, usually with a joint Gaussian. This type of models is computationally expensive since it requires searching an exponentially large number of hypothesis which give different part assignments to the features. The second type is star shaped models [48], [49] which exploits geometry information by modeling the locations of the local features relative to the center of the object. These models can be easily trained, while usually require searching for an optimal object center in the image during testing. Both constellation models and star shaped models require the training images with bounding boxes. Despite its flexibility, the problem with such an approach is that learning and matching graph representations is known to be very expensive, even if we use fast optimization procedures. Moreover, in these models, geometric variability of objects is modeled explicitly and hence are better suited for describing the spatial layout of structured features (e.g. an articulated human body).

A powerful alternative to graph-based representations is the (model-free) semi-local proximity distribution descriptors which encode the distribution of features in multiple spatial regions of the image. In [50], they propose a new descriptor the Generalized Correlogram that encodes in the same feature vector the local information describing what are the local features in the image and, at the same time, the geometrical information describing the mutual position of these features, using a log-polar (r, θ) quantization of image coordinates' domain (like in Mori's *et al.* [51] Shape Context). The advantage of such a representation is that, by simply comparing two feature vectors we take into account simultaneously the similarity of local features and their spatial distribution. This allows us to employ fast matching techniques that quickly consider the relevant information. The disadvantage is that the dimensionality of the new feature is by a factor of the number of spatial bins bigger than the baseline feature. The implementation of this new spatial feature within the framework of bag of word's models appears in their following work in [52], where now the local features are replaced by the vocabulary label at the corresponding patch. An example of Amores *et al.* descriptor is shown in Figure 6.

Correlogram-based features centred on visual words are translation-invariant as they encode mutual information among visual words. Moreover, they can be made scale and rotation invariant by scale-dependent normalization of the radius and by unifying theta bins to one bin (circular kernel) [53]. A similar approach using polar quantization of the image space was used in [54], where the new features were named "local relational features". Working in a slightly different line, Savarese *et al.* [55] have suggested the usage of correlograms for capturing the spatial arrangement of pairs of codewords. Furthermore, the authors achieve compact spatial modeling without loss of discrimination through the introduction of adaptive vector quantized correlograms, which they call correlatons. Then, object models are obtained as histogram of codewords and correlatons. More recently, this work has been used in [56] to encode visual words co-occurrence in a generative framework for action recognition in video.

A different approach appears in the the work of Lazebnik *et al.* [57], who propose semi-local arrangements of affine features for object detection. Their method builds directly on features, without vector quantization, and starts by detecting geometrically stable triples of regions in pairs of images. The candidate pairs are summarized by a description which averages over their geometric arrangement. This description is validated on other examples and, if found repeatedly, used for recognition. The approach of Quack *et al.* [58] instead, builds on vector-quantized features, defines a scale invariant tiled neighborhood, and employs established data mining techniques to find recurring neighborhoods. Their approach though computationally very efficient, finds class-specific features but the mined configurations can not be directly exploited within a classification problem. In [59], they propose a beyond bag-of-features model which is able to take into account spatial relationships of visual words by analyzing the feature space in different grids through scale.

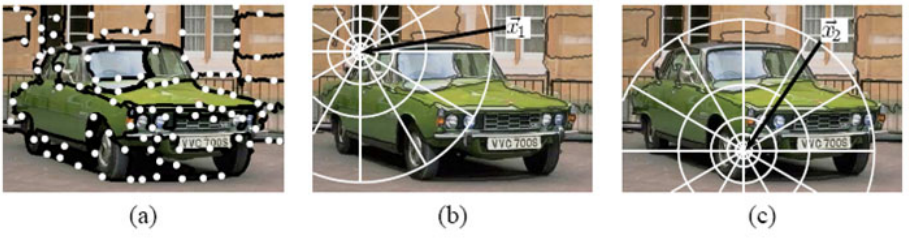


Fig. 6. (a) Sampled set of points taken as reference. (b)-(c) Log-polar spatial quantization of [52] descriptor given two different reference points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (figure taken from [52]).

An important family of methods designed for the above spatial descriptors, is the one which uses Mercer kernels (suitable for image recognition tasks) in order to compare two image representations, based on k-order spatial relationships of visual words. Higher order spatial features, such as doublets or triplets have been used to incorporate spatial information into the BoVWs model as in [53]. However, as the order k increases, the number of features will immediately reach an intractable amount. Therefore, most previous works use 2nd order features [55], or at most 3rd order [53], [60].

Based on the influential work of [35] (see section 3.1), many other kernels have been proposed for object categorization [61], [62], [60]. However, these kernels are either (i) not designed to capture spatial information [35], (ii) not translation invariant [61] since they use absolute coordinates to capture the spatial information, or (iii) computationally expensive [60]. Besides, none of these kernels are designed to calculate higher order (> 2) features. The work of [63] on unbounded-order spatial features addresses all these issues and enables us to compute the kernel in time that is linear to the number of local features in an image (same as the BoVWs approach). High order kernels have been designed for many other applications, such as the string kernel [64] for document classification.

An alternative approach is to augment a BoVWs representation with a region-based approach (an initial segmentation step is applied) in order to exploit regions' spatial relationships. A pixel-wise object class detection and localization by incorporation of local patch-based features into a region-based approach (encapsulate spatial information for deformable regions) and their combination with more common texture-based features is proposed in [65]. This allows to identify regions which may not have a discriminative texture themselves, but which have close proximity to object features. For this purpose, the concept of Region-based Context Features is introduced. A multi-scale representation is achieved by performing image segmentation at three image scales (as in [66]), assigning three different regions to each pixel. In a more demanding task of multiple class segmentation [67] makes use of a semi-local spatial descriptor similar to [55] but using mean shift patches derived from an image segmentation instead of interest-point patches. Other more sophisticated region-based approaches, including spatial modeling, can be found in ch.4.

3.2 Words/Image Joint Distribution Learning

In this section, we focus on models which find the joint probability of images and concepts (keywords) based on an intermediate image description derived from features' clustering. These models are also known as translation models. The corresponding architecture diagram is shown in Figure 7.

In translation model approaches, association between keywords and images may occur either on a global level, either on tiles or on regions of the images. Inspired by machine translation research, Duygulu *et al.* [68] developed a method of annotating image regions with words. First, regions are created using 'normalized cuts' segmentation algorithm. For each region, features are computed and then blobs are generated by clustering the image features for these regions across an image collection. The problem is then formulated as learning the correspondence between the discrete vocabulary of blobs and the image keywords. Following a translation model Jeon *et al.* [69], Lavrenko *et al.* [70] and Feng *et al.* [71] studied a model where blob features of an image are conditionally independent of keywords. Jeon *et al.* [69] recast the image annotation as a cross-lingual information retrieval problem applying a cross-media relevance model (CMRM) based on a discrete codebook of regions. This work was extended by Lavrenko *et al.* [70] using a description of the process about generating blob features with continuous probability density functions (Continuous Relevance Model (CRM)) to avoid the loss of information related to the generation of the codebook.

Instead of coupling the CMRM model with an image description that uses regions produced by an image segmentation method as in [69], Tang *et al.* [72] suggest a coupling with salient regions by using the method proposed in [73], wherein scale-space peaks are detected in a multi-scale difference-of-Gaussian pyramid. This approach is suitable in the case that the aim of the image annotation is to attach words to the entire image instead of objects (regions) in an image. Extending their previous work, Feng *et al.* [71] replace blobs with tiles

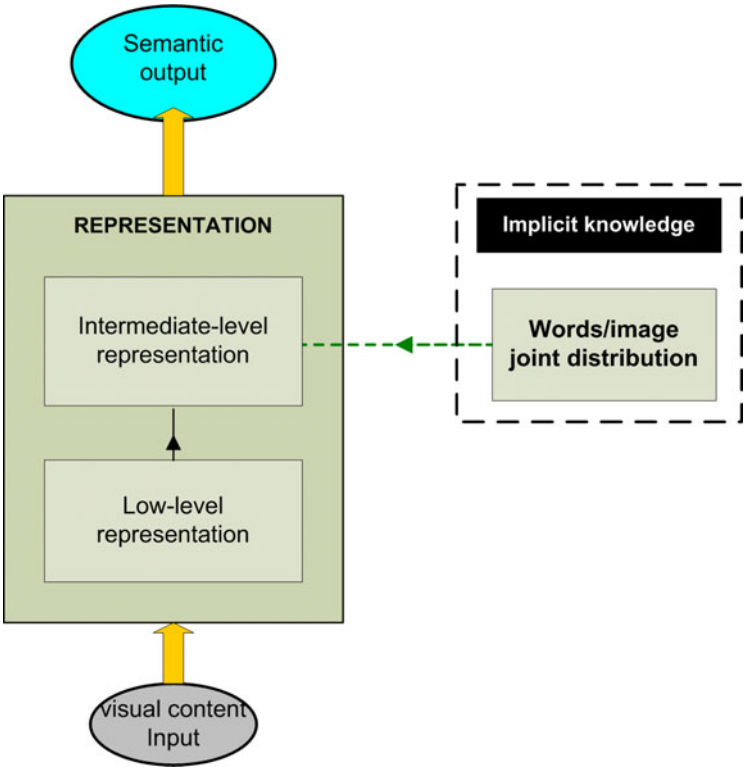


Fig. 7. Semantics extraction wherein the domain knowledge is acquired from words/images joint distribution learning

and model image keywords with a Bernoulli distribution. These methods have the mathematical form of kernel density estimation - the model corresponds to the entire training data - making them computationally very demanding. As in 2.2 these methods are best suited for image annotation and retrieval applications where an unsupervised automatic method is preferred.

4 Implicit Knowledge in Semantic Representation

In general, the constituent parts of a scene in an image do not exist in isolation, and the visual context - the spatial dependencies between scene parts - can be used to improve semantics extraction for the corresponding regions [74], [75]. Two regions, indistinguishable from each other when analyzed independently, might be discriminated as belonging to the correct class with the aid of contextual knowledge. In this section, we focus on methods for semantics extraction based on context-based knowledge inference.

The above-mentioned methods, use the highest level of visual content representation which is the semantic representation. To achieve this representation,

it is required to model the relationships between labeled parts of the image. The corresponding architecture diagram that shows the representation level associated with this type of knowledge inference (context-based) is presented in Figure 8. As one may notice, it is only discussed the use of implicit knowledge since a discussion on the use of explicit knowledge will be added in section 5.

The first part of this section refers to methods that infer a scene label based on inferred scene latent themes, thus they model the joint distribution of words, latent themes and image keywords. The second part of the section refers to methods that infer a scene label based on learnt relations of detected semantic objects usually referred as semantic features or semantic objects.

At first, latent themes discovery models will be addressed, where learning of scene categories is achieved through a generative model built on top of a visual vocabulary method which can incorporate explicit scene configuration models (context). In these models, the algorithms learn both the probability distributions of the

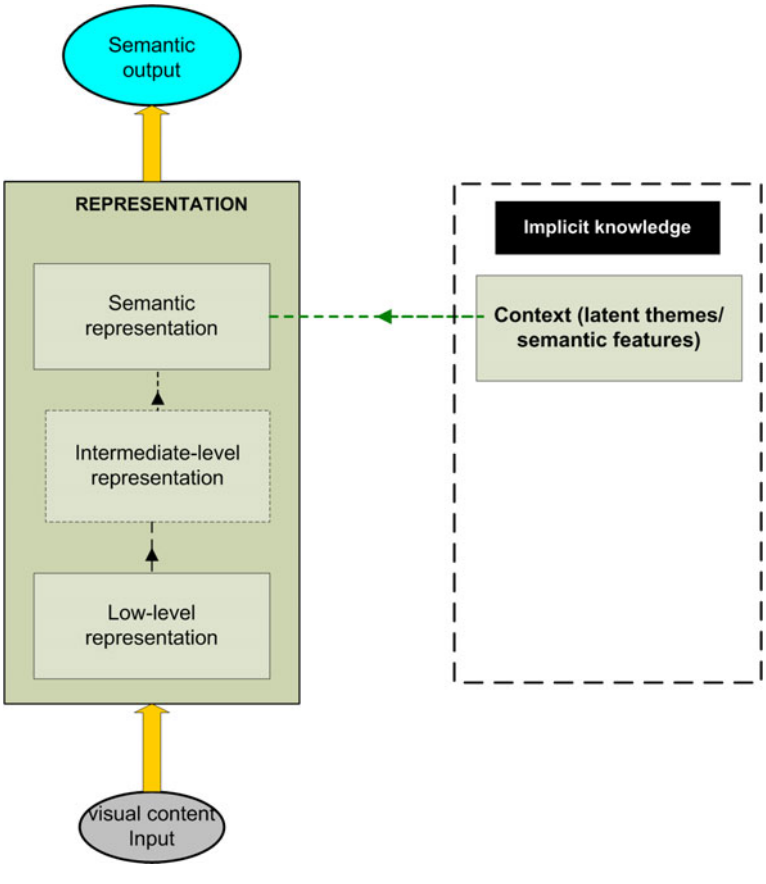


Fig. 8. Semantics extraction for which the implicit knowledge is acquired from context which is identified by latent themes variables

codewords as well as the intermediate themes, see [76], [77], [78], [3], [22], [79]. These models may be considered as an extension of the hierarchical model approaches of words and image association.

In the hierarchical model approaches, the hierarchical relation or the interdependence relation between the elements of an image (words and blobs or tiles) is considered and subsequently reflected in the statistical model, Barnard *et al.* [80] studied a generative hierarchical aspect model, inspired by a hierarchical clustering/aspect model. The data are assumed to be generated by a fixed hierarchy of nodes with the leaves of the hierarchy corresponding to soft clusters. Blei and Jordan [81] describe three hierarchical mixture models to annotate image data, as an extension of the Latent Dirichlet Allocation (LDA) model [82] which assumes that a mixture of latent factors are used to generate words and blob features. It combines the advantages of probabilistic clustering for dimensionality reduction with an explicit model of the conditional distribution from which image keywords are generated.

Magalhaes and Ruger [83] propose a hierarchical model in which each specific keyword-model considers not only its own training data but also the whole training set by utilizing correlations of visual features to refine its own model. More specifically, their model uses a linear combination of a common codebook for all classes, which is by its very nature computationally much simpler. In the proposed generalized linear model the corresponding link function allows to model non-linear relations between the features. The used function is a logit function that is proper for logistic regression.

As an extension of the hierarchical model approach, three representative recent works are presented in the following. The first two of them [76], [78] belong in the latent theme learning class of methods while the third one [84] is placed within the second class of methods dealing with semantic objects. In the latter class, two non-hierarchical methods will also be presented [85], [86]. An equivalent effort integrating co-occurrence and spatial context of words is introduced in [87] but not within a hierarchical framework.

In [76], a generative Bayesian hierarchical model is introduced in order to learn probability distributions of both visual codewords and intermediate "themes" in an unsupervised manner. The goal of learning is therefore to achieve a model that best represents the distribution of these codewords in each category. As one can observe in Figure 9, the proposed process can be split into two phases: (i) learning and (ii) recognition. During learning, a codebook of codewords is learnt from patches drawn from a random half of the entire training set (clustering patches' descriptors). A model for each category of scenes is obtained from the training images. Probability distributions of the local patches as well as the intermediate themes are both learnt in an automatic way; no supervision is required apart from a single category label to the training image. During recognition, local patches are extracted from each test image. Each patch is represented by a codeword from a large vocabulary of codewords (see Figure 9). Then, they find the category model that fits best the distribution of the codewords of the particular image (the category label that gives the highest likelihood probability).

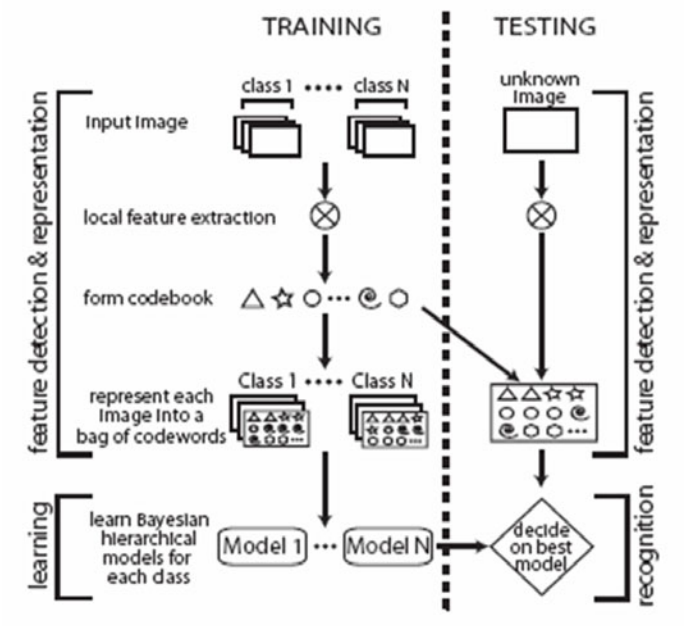


Fig. 9. Flowchart of a learning/recognition process using a generative Bayesian hierarchical model (taken from [76])

An evolution of the above proposed generative framework for object categorization follows with the work of Wang and Fei-Fei [88]. In most previous works using "bag of visual words" models (e.g. [76], [77]), the local patches are assumed to be independent with each other. In this effort, they relax the independence assumption and model explicitly the inter-dependency of the local regions. Certain approaches compute descriptors over regions that are computed with a segmentation algorithm and then perform feature extraction from these regions to feed a BoVWs model.

There is much criticism for the effectiveness of those approaches due to the difficulty with a single segmentation approach to achieve all constituent objects in an image. To overcome this, Russell *et al.* [78] propose a methodology that does not use a single segmentation but rather uses multiple segmentations. The main insight of the approach is that "segments corresponding to objects will be exactly the ones represented by coherent groups (topics), whereas segments overlapping object boundaries will need to be explained by a mixture of several groups (topics)". The multiple segmentation is achieved by changing the parameter values which concern the number of the final segments and the size of the input image using a Normalized Cuts framework [89]. After segmentation is achieved, they perform "topic discovery" on the set of all segments in the image collection using Latent Dirichlet Allocation (LDA) [82], treating each segment as a document. For each discovered topic, they sort all segments by how well

they are explained by the topic. The final result is a set of discovered topics where the top-ranked discovered segments correspond to the objects within that topic. In [90], they improve on the previous work by introducing a model that generates the topic distribution at the region level instead of the world level and by imposing spatial coherency in topics' labels which belong to the same region. Their model is called spatially coherent latent topic model (Spatial-LTM). Spatial-LTM represents an image containing objects in a hierarchical way by over-segmented image regions of homogeneous appearances and the salient image patches within the regions (labeled patches by visual words). Only one single latent topic is assigned to the image patches within each region, enforcing the spatial coherency of the model.

In the following, methods that aim to detect objects of the image in order to describe the scene are discussed. These methods are mainly based on first segmenting the image in order to deal with different regions. Subsequently local classifiers are used labeling the regions as belonging to a known object (semantic objects e.g. sky, people, cars, grass, etc.). Finally, using this local information, the global scene can be classified.

In the framework of the hierarchical approaches, in [84], region-based whole outdoor scene classification is pursued. The work's main novelty is its explicit use of spatial relations [91] in building a generative model to parse a scene, distinguishing it from other work using semantic features. They define semantic, or high-level, features to be labeled regions. A region with ambiguous identity usually has a low belief value, and may also have multiple labels. In this study, high-level features generated from 3 types of detectors are used: (i) output from actual object and material detectors (based on color and texture low-level features); (ii) output from simulated detectors, and (iii) output from best-case detectors (hand-labeled regions). Their generative model is based on the concept of scene configurations, consisting of two parts: the actual spatial arrangement of regions (edge labels in the graph of Figure 10(c) and the material configuration (e.g grass, sand, foliage, beach), the identities of those regions (node labels in Figure 10(c)). A factor graph [92] is used so that we can model interactions between the scene type and various region configurations. The factors in the graph encode the compatibilities between the scene type, the scene configurations, and the detector evidence.

Fan *et al.* [85] used concept sensitive salient objects as the dominant image components to achieve automatic image annotation at the content level. To detect the concept- sensitive salient objects, a set of detection functions is learned from the labeled image regions and each function is able to detect a specific type of these salient objects. Each detection function consists of three parts: (i) automatic image segmentation by using the mean shift technique; (ii) binary image region classification by using the SVM classifiers with an automatic scheme for searching the optimal model parameters and (iii) label-based aggregation of the connected similar image regions for salient object generation. To generate the semantic image concepts, the finite mixture models are used to approximate the class distributions of the relevant objects. After detecting the semantic

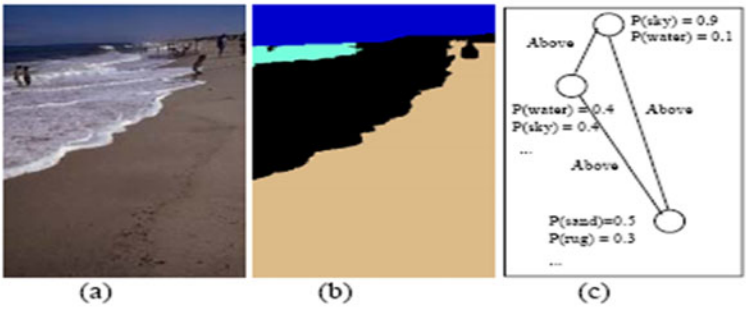


Fig. 10. (a) A 'beach' scene. (b) Its manually-labeled materials. The true configuration includes 'sky above water', 'water above sand', and 'sky above sand'. (c) The underlying graph showing detector results and spatial relations. Figure is taken from [84].

salient objects they carry out the semantic image classification. An adaptive EM algorithm has been proposed to determine the optimal model structure and model parameters simultaneously. In addition, a large number of unlabeled samples are integrated with a limited number of labeled samples to achieve more effective classifier training and knowledge discovery.

In [86], a hybrid approach is proposed: low-level and semantic features are integrated into a general-purpose knowledge framework that employs a Bayesian Network (BN). BN are directed, acyclic graphs that encode the cause-effect and conditional independence relationships among variables in the probabilistic reasoning system. The directions of the links between the nodes (variables) represent causality in the sense that those links express the conditional probabilities of inferring the existence of one variable given the existence of the other variable. Each node can have many such directed inputs and outputs, each specifying its dependence relationship to the nodes from which the inputs originate and nodes where the outputs go.

All the methods presented above make use of an intermediate hierarchy of subconcepts being present in a labeled scene and thus, the system employing them can learn to recognize many categories which include these subconcepts. Despite the fact that they are computationally more demanding and not always theoretically intuitive, they are the only one which can take into account categories co-occurrence within a certain category, thus being the most suitable in a multi-class generic categorization task. Moreover, when combined with region-based representations they can also locate the constituent objects or parts of the scene. Such trends are explored in a dedicated chapter, chapter 6.

5 Explicit Knowledge (Ontology-Driven Approaches)

Until now, we have discussed the use of implicit knowledge for semantics extraction from images. This section will be dedicated to the description of

methodologies that use explicit knowledge. The major bottleneck regarding the construction of new knowledge bases in the case of new application domains was treated by the ARPA Knowledge Sharing Effort [93] that envisioned the construction of new knowledge based systems by assembling reusable components. The envisaged systems would require only specific knowledge and reasoners that did not exist before. The means to fulfill this expectation was ontologies.

The term Ontology refers to the philosophical discipline which deals with the nature and the organization of reality [94]. In this sense, Ontology aims to answer the question: "what is being?", or "what are the features common to all beings?". The meaning of this term has slightly evolved in the artificial intelligence community. Gruber defines the notion of ontology in [95]: An ontology is an explicit specification of a conceptualization. In [96], several complementary definitions are given. A conceptualization is defined as an intensional semantic structure which encodes the rules constraining the structure of a piece of reality. In AI, the term ontology has largely come to mean one of two related things [94]. First of all, an ontology is a representation vocabulary, often specialized to some domain or subject matter. More precisely, it is not the vocabulary as such that qualifies as an ontology, but the conceptualizations that the terms in the vocabulary are intended to capture. In its second sense, the term ontology is sometimes used to refer to a body of knowledge describing some domain, typically a commonsense knowledge domain, using a representation vocabulary. In other words, the representation vocabulary provides a set of terms with which to describe the facts in some domain, while the body of knowledge using that vocabulary is a collection of facts about a domain.

An ontology is composed of several entities as (i) a set of concepts (C) (e.g. geometric concepts); (ii) a set of relations (R) (e.g. spatial relations); (iii) a set of axioms (e.g. transitivity, reflexivity, symmetry of relations). A concept is defined as a notion, usually expressed by a term (or more generally by a sign). A concept represents a group of objects or beings sharing characteristics that enable us to recognize them as forming and belonging to this group. The aim of ontologies is to define which primitives, with their associated semantics, are necessary for knowledge representation in a given context. An ontology is supposed to be the support of reasoning mechanisms.

Using an ontology for multimedia information processing offers several advantages [97]: (a) The ontology provides a source of shared and precisely defined terms that is used to (i) index the metadata describing the semantic content of the document; (ii) express the queries and (iii) describe the content of each source of documents (also called views). (b) An ontology-based approach allows more precise queries on metadata. For example, it is possible to ask for all the documents containing "the successful efforts of a particular player to score in the 2002 World Football Cup". (c) An ontology-based search is more powerful than a keyword search. The inference that is drawn from the ontology enable to derive information that was not explicitly stated in the metadata, and thus to provide documents that would have been missed otherwise.

Ontologies are characterized by the following features: knowledge sharing, machine interoperability and intercommunication, extensibility, scalability, and inferencing. Due to the broad power of the capabilities of ontologies, it was a natural side-effect to put forward efforts for the revision of tools and languages that were available for constructing knowledge-based systems. In the case of knowledge representation language selection, it has to be done by taking into account the expressiveness of the language, the efficiency of the reasoning mechanisms supported by the knowledge representation language, and the ease of use of the language. In the ontology-driven approaches, semantics extraction underlies upon ontologies that can enable machines to both generate and interpret visual descriptions which can be used for multimedia reasoning.

In [98], an ontology-driven methodology suited for search in large collections of heterogeneous images is presented. The proposed approach employs a fully unsupervised segmentation algorithm to divide images into regions wherein low-level descriptors for the color, position, size and shape of each region are extracted. The computed descriptors are automatically associated with appropriate qualitative intermediate-level descriptors, which form a simple vocabulary termed object ontology (Figure 11). The object ontology is used to allow the qualitative definition of the high-level concepts the user queries for (semantic objects, each represented by a keyword) and their relations in a human-centered fashion. When querying for a specific semantic object (or objects), the intermediate-level descriptor values associated with both the semantic object and all image regions in the collection are initially compared, resulting in the rejection of most image regions as irrelevant. Following that, a relevance feedback mechanism, based on support vector machines and using the low-level descriptors, is invoked to rank the remaining, potentially relevant image regions and produce the final query results. As it can be seen in Figure 12, each concept has a particular representation that has to be unique in the complete list of concepts.

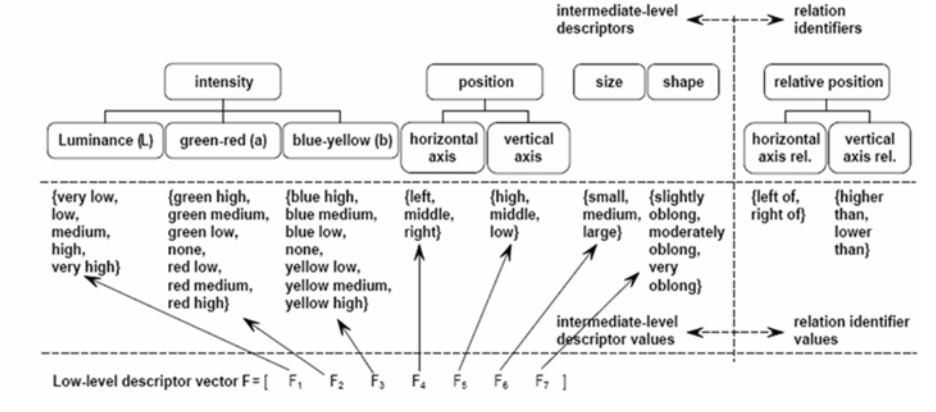


Fig. 11. Object ontology (taken from [98])

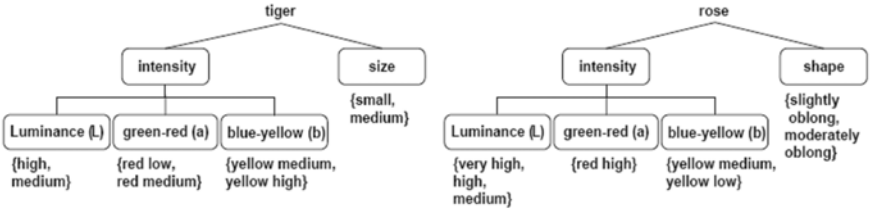


Fig. 12. Exemplary concepts using the object ontology (taken from [98])

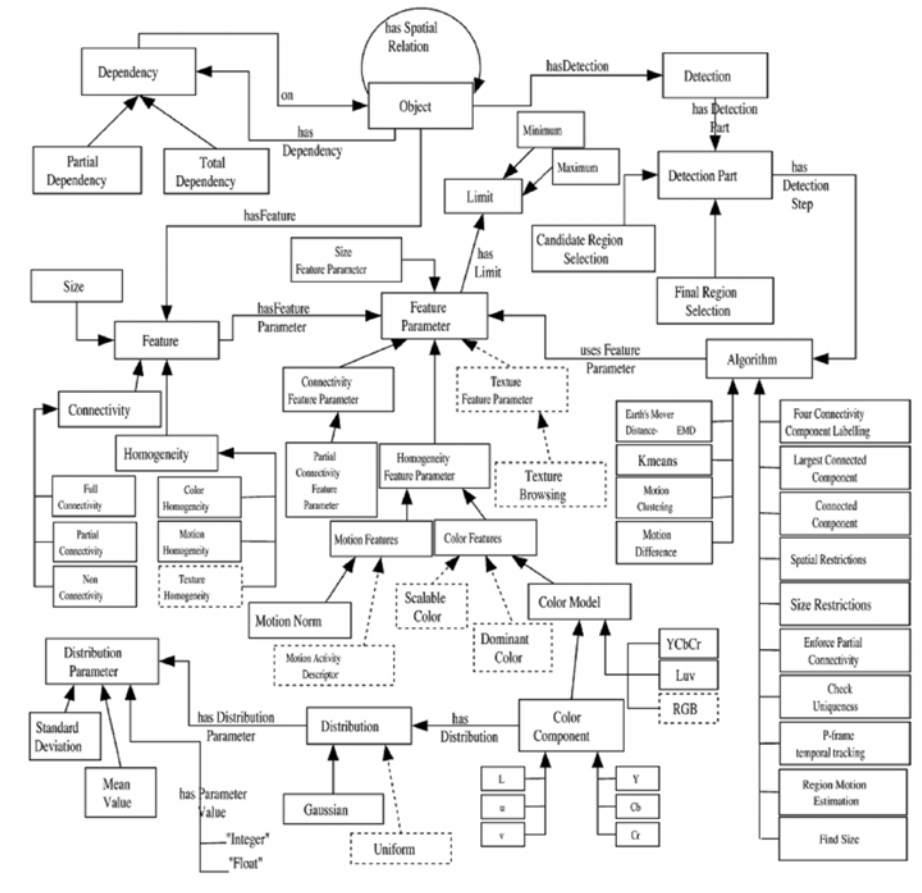


Fig. 13. Multimedia Analysis ontology (taken from [99])

In [99], a multimedia ontology infrastructure is presented to support knowledge-assisted semantic video object detection. The semantic concepts defined in the ontology are enriched with qualitative attributes (e.g., color homogeneity),

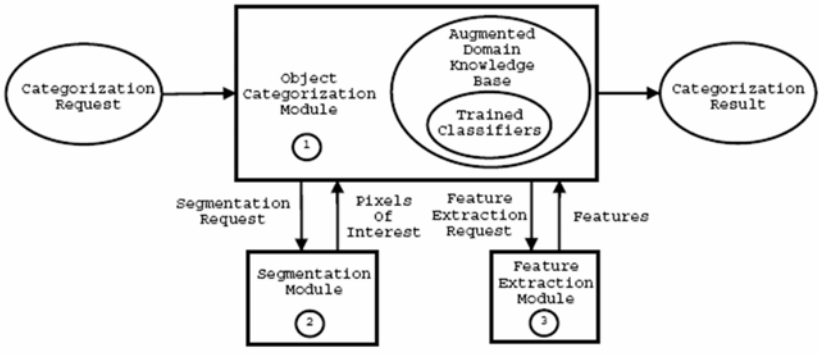


Fig. 14. The Object categorization phase (taken from [100])

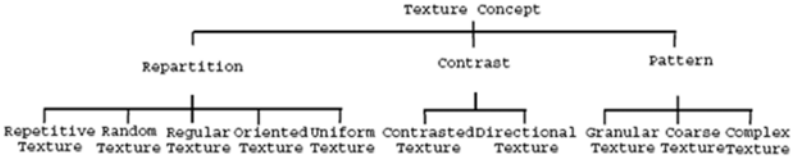


Fig. 15. Texture concept hierarchy (taken from [100])

low-level features (e.g., color model components distribution), object spatial relations, and multimedia processing methods (e.g., color clustering) (Figure 13). For knowledge representation in the RDFs metadata standard, they use Semantic Web technologies. A construction of rules in F-logic is tackled to describe how tools for multimedia analysis should be applied, depending on concept attributes and low-level features, for the detection of video objects corresponding to the semantic concepts defined in the ontology. This aims to support flexible and managed execution of various application and domain independent multimedia analysis tasks. Furthermore, this semantic analysis approach can be used in semantic annotation and transcoding systems, which take into consideration the users environment including preferences, devices used, available network bandwidth and content identity. The proposed approach was tested for the detection of semantic objects on video data of three different domains.

In [100], an object categorization method is presented (Figure 14). The proposed approach involves machine learning and knowledge representation. A major element of their approach is a visual concept ontology composed of several types of concepts (spatial concepts and relations, color concepts (Figure 16) and texture concepts (Figure 15). Visual concepts contained in this ontology can be seen as an intermediate layer between domain knowledge and image processing procedures (Figure 17). The proposed approach is composed of three phases: a knowledge acquisition phase, a learning phase and a categorization phase.

Red	Purple
Reddish Orange	Reddish Purple
Orange	Purplish Red
Orange Yellow	Purplish Pink
Yellow	Pink
Greenish Yellow	Yellowish Pink
Yellow Green	Brownish Pink
Yellowish Green	Brownish Orange
Green	Reddish Brown
Bluish Green	Brown
Greenish Blue	Yellowish Brown
Blue	Olive Brown
Purplish Blue	Olive
Violet	Olive Green

Fig. 16. Set of Hue concepts (taken from [100])

A major issue is the symbol grounding problem (symbol grounding consists in linking meaningfully symbols to sensory information). They propose a solution to this difficult issue by showing how learning techniques can map numerical features to visual concepts.

Hudelot *et al.* [101] propose a solution to the grounding problem (the mapping between the numerical image data and the high level representations of semantic concepts). To establish the correspondence links for the mapping they present two approaches: (i) a learning approach, in which links between low-level image data features and visual concepts are learned from image samples and (ii) an a priori knowledge-based approach, in which links between low-level image data features and visual concepts are built explicitly. In the case of approach (ii), the symbol grounding knowledge base encodes the corresponding expertise in a declarative manner and depends on a visual concept ontology as well as on an image processing ontology. In the proposed approach, linking between visual concepts and image data is based on the modeling of each low level feature as a fuzzy linguistic variable with a domain, a possible set of linguistic values and their associated fuzzy sets. To take into account the spatial structure of semantic concepts, the proposed framework supports spatial relation representation via the visual concept ontology. Furthermore, using an image processing

Class	POACEAE
{	
SuperClass:	POLLENWITHPORI
SubParts:	
PORI pori1	[PORIWITHANULUS]
SpatialAttributes :	
GeometricConcept **geometry :**	[*CircularSurface EllipticalSurface*]
SizeConcept **size :**	[*ImportantSize*]
ColorAttributes :	
HueConcept **hue:**	[*Pink*]
BrightnessConcept **brightness:**	[*Dark*]
TextureAttributes :	
TexturePatternConcept **pattern:**	[*GranulatedTexture*]
TextureContrastConcept **contrast:**	[*Slight*]
SpatialRelations :	
SpatialRelation **r1:**	[*NTTP*(POACEAE,pori1) *TTP*(POACEAE,pori1)]
}	

Fig. 17. High level description of domain class "Poaceae" (taken from [100])

ontology they use object extraction criteria for the decision on the application of constraints in subsequent image processing requests.

Town and Sinclair [102] use an extensible ontology to support a language-based querying of image collections. Querying is achieved by combining ontological concepts constrained by a grammar. The proposed query language OQUEL features a generic base vocabulary built on extracted image features and intermediate level content that correspond to segmented image regions. The mapping between image data and concepts is performed by supervised machine learning techniques. More specifically, multi-layer perceptrons (MLP) and radial basis functions (RBF) networks were used, whose topology was optimized to yield best generalization performance for each particular visual category.

6 Explicit / Implicit Knowledge during a Segmentation / Recognition Interplay

As discussed in the previous sections, the methods which solve the problem of semantics extraction (filling the semantic gap) use either implicit knowledge in a bottom-up manner or explicit knowledge in a top-down fashion. In this

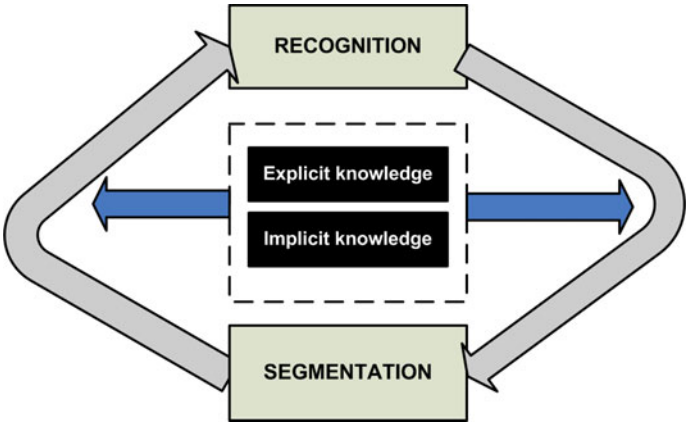


Fig. 18. A schematic diagram of the segmentation / recognition interplay for semantics extraction

section, there will be a discussion about methodologies which address a coupling of bottom-up and top-down approaches that is translated to an interplay between image segmentation and recognition stages. A schematic diagram of the complete process can be seen in Figure 18.

One of the first efforts found in the literature investigating this interplay is that of [103] in 2004. In this work, they study how figure-ground segmentation may be achieved as a result of object recognition without putting the requirement of an initial segmentation step. The proposed algorithm achieves this by learning a codebook of local appearances of a particular object category. Starting with each patch (size 25x25 pixels extracted with the Harris interest point detector [104]) a visual vocabulary is obtained by agglomerative clustering based on the Normalized Greyscale Correlation (NGC) (patches' clusters corresponding to codebook entries/words). Rather than to use the codebook directly to train a classifier as in [105], they propose to use a probabilistic voting scheme. Given a test image, they extract image patches and match them to the codebook to activate codebook entries. For every codebook entry, all the positions it was activated in are stored, relative to the object center. Each activated entry then, casts votes for possible positions of the object center in a probabilistic framework. Figure 19 illustrates this procedure. Moreover, they can refine the hypothesis by sampling all the image patches in its surroundings, not just those locations returned by the interest point detector. As a result, they get a representation of the object including a certain border area.

Based on the refined object hypotheses obtained from recognition part (a probability of both object identity and position given an extracted patch), we now want to know whether a certain image pixel is figure or ground (segmentation). Given the patches' votes, we can obtain information about a specific pixel by summing over all the patches that contain it. Instead of keeping a fixed segmentation mask (as in [106]), a separate mask for every stored occurrence

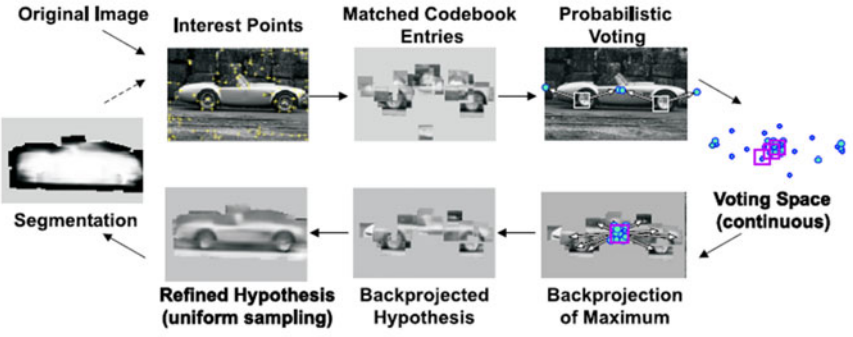


Fig. 19. The recognition procedure which leads to a segmentation mask generation (taken from [103])

position of each codebook entry is stored. Whenever a codebook entry is matched to the image using this approach, a separate segmentation mask is associated with every object position it votes for. The final selection of an option depends on the winning hypothesis and its accumulated support from other patches. The derived probabilistic formulation of the problem allows them to incorporate knowledge about the recognized category as well as the supporting information in the image. As a result, they obtain a segmentation mask of the object together with a per-pixel confidence estimate specifying how much this segmentation can be trusted. The resulting images show that for more accurate segmentation results, the combination with traditional contour or region based segmentation algorithms is required while the probabilistic formulation lends itself to an easy integration with other segmentation methods.

More recently, accurate segmentation of (one) object of interest (for real-time embedded implementations) based on object's region's shape is achieved in [107]. A classification-segmentation wrapper approach is proposed, where the image classification is used to assemble regions derived from the traditional segmentation algorithms, and then to further direct these algorithms to modify their segmentation parameters if the match is inadequate. In this framework, the probability of correct classification is used as a metric to determine the quality of the segmentation. The selected feature for object classification concerns the shape of the object of interest. Region-based moment representation of the images is used for the description of the object shape. For image segmentation and binary labeling (figure/background) EM algorithm is selected where the mixture component with the highest probability becomes the label for each pixel.

There is one common issue associated with using the EM algorithm and that is the selection of the number of components in the mixture. In the former work they rely on the classification accuracy in order to determine if more components are required. For example, beginning with a fixed number of components and finding that the "classification distances" for the corresponding pattern classes are too high, then they assume the image may be under-segmented, and they increase the number of mixture components. Multiple combinations of regions are evaluated based

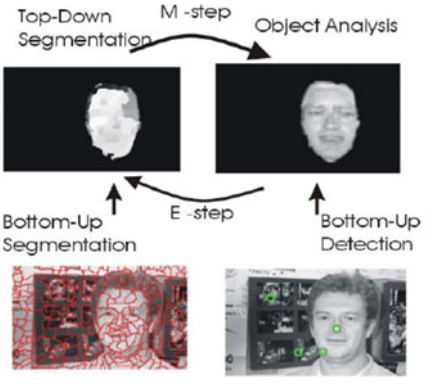


Fig. 20. Overview of the EM approach to cooperative image segmentation and object categorization (taken from [108])

on the probability of correct classification for a given class. The classification is assumed to be correct when a minimum distance between the segmented test sample (manually segmented image into figure/background) and the candidate segmented image is attained. One apparent advantage of this work is that existing image segmentation techniques may be executed within the proposed wrapper-framework. Moreover, by using shape as the classification feature (semantics), they are able to develop a segmentation algorithm that relaxes the requirement that the object of interest to be segmented must be homogeneous in some low-level image parameter, such as texture, color, or greyscale intensity.

In [108], they deal with the problem of modeling and exploiting the interaction between the processes of image segmentation and object categorization. A probabilistic framework to address this problem is proposed that is based on the combination of the Expectation Maximization (EM) algorithm and generative models for object categories. The system accepts as its bottom-up input an over-segmentation of the image (morphological watershed segmentation algorithm) and a set of locations proposed by an object detection system. They experimented initially with the part-based detection system of [109], which relies on a Markov Random Field (MRF) formulation for object detection and requires training with hand-labeled keypoint data. To overcome this constraint, they have experimented next with a simplification of the bottom-up part of [110], which generates an object hypotheses based on a compact vocabulary of local appearance.

In the EM formulation of the interaction between segmentation and object categorization algorithm, segmentation is interpreted as the E step, assigning observations (segments' content) to models (object hypotheses), whereas object detection/analysis is modeled as the M-step, fitting models to observations (see Figure 20). In the E-step (object-based segmentation) the object and background hypotheses compete for the occupancy of image regions. The content of each region is modeled by generative models. Generative models for objects are based

on Active Appearance/Morphable Models [111] [112] which have been successful in high level tasks as object recognition, pose estimation etc. In the M-step the parameters of the morphable models are updated in order to model the areas of the image assigned to them during the E-step. Results on the joint detection and segmentation of the object categories of faces and cars are demonstrated.

In [113], a probabilistic model that makes use of top-down shape templates (such as limbs, head and shoulders) is constructed to guide the grouping of homogenous bottom-up regions produced by a multiscale hierarchical segmentation graph [114]. The hierarchy of bottom-up segments in multiple scales is used to construct a prior on all possible figure-ground segmentations of the image. This prior increases as more pixels that are strongly connected (weight in the graph) to the same salient segments (boundary properties and homogeneity of the segment) are grouped together as figure or ground. It decreases as more salient regions are split to figure and background parts. We then apply a top-down process to guide the segmentation in forming a specific desired shape while maintaining a high prior for it. Initially, shape templates representing object parts are locally detected in the image. The detected parts are integrated to produce a global approximation for the object's shape, which is then used by an inference algorithm to produce the final segmentation.

With respect to the former method, experiments with a large sample of images demonstrate strong figure-ground segmentation despite high object and background variability. The segmentations are robust to changes in appearance since the matching component depends on shape criteria alone. The model could potentially learn and apply other top-down knowledge than shape. For instance, it could be used to"supervise" the learning of color or texture characteristics of a specific shape (i.e. regions grouped to form a horse head are more likely to be brown than green). Remaining difficulties include addressing the estimation of object scale and incomplete figures, occurring in conditions where both segmentation processes are challenged in the same region - such as the top-down missing a body part while the bottom-up merges it with its background.

In [115], two enhancements for the creation and analysis of Binary Partition Trees (BPTs) are proposed towards bridging the gap between visual content and semantics cast in a segmentation framework: (i) Bottom-up BPT construction: Introducing and combining multiple and generic homogeneity criteria based on low- and middle-level features. Such features are referred to as syntactic features, since they are defined by the relative positions of the regions they represent; (ii) Top-down BPT analysis: The problem of detections of a single instance of the same object is assessed. To do so, a model for semantic classes and its application on BPTs is presented.

The framework presented in [115] for BPT creation is an extension of the segmentation algorithm presented in [116], which performs an iterative region-merging trying to optimize an initial partition based on regions' visual (color, size, partial inclusion) and statistical homogeneity criteria. The visual descriptors are generated by associating a dissimilarity measure to each pair of neighboring regions given certain rules. Even though the rules are computed locally over a

pair of regions, an estimate of the distribution of the dissimilarity measures for each rule is computed by means of the histogram over the whole image. In the case of color homogeneity, the histogram counts the number of times that the same color difference occurs among the whole set of regions. The distributions for size and partial inclusion are similarly estimated, extracting features which are both local and global. During a decision step, for each pair of connected regions from the current partition a dissimilarity measure that combines local and global descriptors is computed and the pair of regions with the lowest value is selected for merging. Regarding the top-down approach, an improvement on BPT Semantic Neighborhoods (a subset of connected BPT nodes that represent instances of the same semantic class) is introduced in order to choose among all the BPT nodes which are candidate to contain a semantic instance of a class. The proposed technique makes use of simple and composite semantic classes allowing for a semantic decomposition of the observed scene and imposes a novel specific rule to the process of adding a new node into the Semantic Tree.

While the aforementioned methodologies used implicit knowledge as a common framework, in the sequel, explicit knowledge in the form of ontologies will be used.

In [117], a semantic image segmentation approach is proposed that combines two types of learning algorithms, namely SVMs and Genetic Algorithm (GA), with explicitly defined knowledge in the form of an ontology that specifies domain objects and fuzzy spatial relations. SVMs are employed (an individual SVM for each semantic object) for performing an initial mapping between low-level visual features and the domain objects in the ontology (i.e. generating an initial hypothesis set for every image region) at a region level, whereas a GA is subsequently used to optimize this mapping over the entire image, taking into account the spatial context. Representation of the latter relies on fuzzy spatial relations extraction which builds on the principles of projection and angle-based methodologies inspired by [118] and [119]. The resulting learnt fuzzy spatial relations serve as constraints denoting the "allowed" domain objects spatial topology. After the initial set of hypotheses is generated, based solely on visual features and the fuzzy spatial relations are computed for every pair of image segments, the genetic algorithm (GA) is introduced to decide on the optimal image interpretation using the spatial-related domain knowledge as produced by the particular training process. An extension of the Recursive Shortest Spanning Tree (RSST) algorithm has been used for segmenting the image, while regions' description utilizes MPEG7 visual descriptors (see [120]).

In [121], they present a framework for simultaneous image segmentation and region labeling leading to automatic image annotation. The proposed framework operates at semantic level using possible semantic labels to make decisions on handling image regions instead of visual features used traditionally. In order to stress its independence of a specific image segmentation approach they applied their idea on two region growing algorithms, i.e. watershed and recursive shortest spanning tree. Techniques are modified to operate on the fuzzy sets stored in the ARG in a similar way as if they worked on low-level features. Additionally they exploit the

notion of visual context by employing fuzzy algebra with characteristics derived from the Semantic Web. For this purpose ontological taxonomic knowledge representation is implemented (discussed in depth in their previous work [122]), incorporating in this way global information and improving region interpretation. In this process, semantic region growing labeling results are being re-adjusted appropriately, utilizing contextual knowledge in the form of domain-specific semantic concepts and relations. The performance of the overall methodology is demonstrated on a real-life still image dataset from the popular domains of beach holidays and motor sports. Testing of this method revealed that semantic region growing algorithms did not perform well when the corresponding segments differed visually and the possible detected object was a composite one - in contradiction to other encountered material objects - and was composed by regions of completely different characteristics.

7 Concluding Remarks

In this chapter, we reviewed the state-of-the-art methodologies that aim to extract semantics of different granularity from images. In the case of methodologies that use visual content to directly extract semantics of different granularity, we have set-up an overall framework for semantics extraction to facilitate the understanding of methodologies categorization during the discussion of all available trends. From our discussion, it is obvious that there is a bulk of methodologies for semantics extraction that are developed upon two axes, the axis of knowledge and the content representation, respectively. There are methods that achieve a good performance for a few number of semantics but there is much more way to cover till a generic approach is achieved. The task at hand (object, part or scene-oriented), the nature of the dataset (level of invariance required) and the availability of training data should always guide the choice of the method. Although no method can be considered as the gold standard for any semantics, those methods which strive towards the integration of context (scale, spatial or semantic) into semantics extraction inference mechanism should be given a special attention and are considered very promising. The optimal use of those methods can be addressed by the combination of generative and discriminative models as well as by methodologies that involve a segmentation/recognition interplay, using either explicit of implicit knowledge.

References

1. Biederman, I., Mezzanotte, R.J., Rabinowitz, J.C.: Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology 14(2), 143–177 (1982)
2. Hobson, P., Kompatsiaris, Y.: Advances in semantic multimedia analysis for personalised content access. In: ISCAS (2006)
3. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part IV. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)

4. Sheth, A., Ramakrishnan, C., Christopher, T.: Semantics for the semantic web: The implicit, the formal and the powerful. Int. Journal on Semantic Web and Information Systems 1(1), 1–18 (2005)
5. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: CAIVD, pp. 42–51 (1998)
6. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.: Image classification for content-based indexing. IEEE Transactions on Image Processing 10(1), 117–130 (2001)
7. Oliva, A., Torralba, A.B.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
8. Huang, F.J., LeCun, Y.: Large-scale learning with svm and convolutional nets for generic object categorization. In: Proc. Computer Vision and Pattern Recognition Conference (CVPR 2006). IEEE Press, Los Alamitos (2006)
9. Chang, E., Goh, K., Sychay, G., Wu, G.: Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on Circuits and Systems for Video Technology 13(1), 26–38 (2003)
10. Pratikakis, I., Gatos, B., Thomopoulos, S.C.: Scene categorisation using low-level visual features. In: VISAPP 2006, vol. 3309, pp. 155–160 (2006)
11. Serrano, N., Savakis, A.E., Luo, J.: Improved scene classification using efficient low-level features and semantic cues. Pattern Recognition 37(9), 1773–1784 (2004)
12. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: International Conference on Computer Vision, vol. 1, pp. 525–531 (2001)
13. Vogel, J., Schiele, B.: Semantic scene modeling and retrieval for content-based image retrieval. Int. Journal of Computer Vision 72(2), 133–157 (2007)
14. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in Cognitive Sciences 11(12), 520–527 (2007)
15. Bosch, A., Muñoz, X., Marti, R.: Which is the best way to organize/classify images by content? Image Vision Comput. 25(6), 778–791 (2007)
16. Yavlinsky, A., Schofield, E., Rüger, S.M.: Automated image annotation using global features and robust nonparametric density estimation. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 507–517. Springer, Heidelberg (2005)
17. Carneiro, G., Vasconcelos, N.: Formulating semantic image annotation as a supervised learning problem. Computer Vision and Pattern Recognition, 163–168 (2005)
18. Westerveld, T., de Vries, A.P.: Experimental result analysis for a generative probabilistic image retrieval model. In: SIGIR, pp. 135–142 (2003)
19. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: MISRM 1999 First International Workshop on Multimedia Intelligent Storage and Retrieval Management, Orlando, FL, USA (1999)
20. Zhou, X., Wang, M., Zhang, Q., Zhang, J., Shi, B.: Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In: CIVR, pp. 25–32 (2007)
21. Lodhi, H., Shawe-Taylor, J., Cristianini, N., Watkins, C.J.C.H.: Text classification using string kernels. In: NIPS, pp. 563–569 (2000)
22. Liu, D., Tsuhan, C.: Semantic-shift for unsupervised object detection. In: Workshop on beyond patches in conjunction with CVPR, pp. 16–16 (2006)
23. Malik, J., Belongie, S., Shi, J., Leung, T.K.: Textons, contours and regions: Cue integration in image segmentation. In: ICCV, pp. 918–925 (1999)

24. Leung, T.K., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision 43(1), 29–44 (2001)
25. Schmid, C., Mohr, R.: Combining greyvalue invariants with local constraints for object recognition. In: CVPR, pp. 872–877 (1996)
26. Datta, R., Li, J., Wang, J.Z.: Content-based image retrieval: approaches and trends of the new age. In: MIR 2005: Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 253–262. ACM, New York (2005)
27. Yu, J., Tian, Q., Amores, J., Sebe, N.: Toward robust distance metric analysis for similarity estimation. In: CVPR (1), pp. 316–322 (2006)
28. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR (2), pp. 1447–1454 (2006)
29. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: ICCV (2005)
30. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: *textonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
31. Winn, J.M., Criminisi, A., Minka, T.P.: Object categorization by learned universal visual dictionary. In: ICCV, pp. 1800–1807 (2005)
32. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: CVPR (2), pp. 994–1000 (2005)
33. van Gemert, J.C., Geusebroek, J.M., Veenman, C.J., Snoek, C.G., Smeulders, A.W.: Robust scene categorization by learning image statistics in context. cvprw 0, 105 (2006)
34. Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: Conference on Computer Vision & Pattern Recognition (June 2009)
35. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: ICCV, pp. 1458–1465 (2005)
36. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: NIPS, pp. 985–992 (2006)
37. Wu, J., Rehg, J.M.: Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In: IEEE International Conference on Computer Vision (ICCV) (2009)
38. Kokkinos, I., Maragos, P., Yuille, A.L.: Bottom-up & top-down object detection using primal sketch features and graphical models. In: CVPR (2), pp. 1893–1900 (2006)
39. Csurka, G., Dance, C., Willamowski, J., Fan, L., Bray, C.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision (2004)
40. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV, pp. 1150–1157 (1999)
41. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classifcation of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA, Antipolis (2005) Technical report
42. Zhang, H., Berg, A., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: CVPR, pp. 2126–2136 (2006)
43. Yang, C., Dong, M., Hua, J.: Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: CVPR (2), pp. 2057–2063 (2006)

44. Berg, A.C., Malik, J.: Geometric blur for template matching. In: CVPR (1), pp. 607–614 (2001)
45. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 29(3), 394–410 (2007)
46. Niebles, J.C.: A hierarchical model of shape and appearance for human action classification. In: CVPR (2007)
47. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 264–271 (June 2003)
48. Sudderth, E.B., Torralba, A.B., Freeman, W.T., Willsky, A.S.: Learning hierarchical models of scenes, objects, and parts. In: ICCV, pp. 1331–1338 (2005)
49. Mikolajczyk, K., Leibe, B., Schiele, B.: Multiple object class detection with a generative model. In: CVPR (1), pp. 26–36 (2006)
50. Amores, J., Sebe, N., Radeva, P.: Class-specific binary correlograms for object recognition. In: BMVC (2007)
51. Mori, G., Belongie, S., Malik, J.: Efficient shape matching using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(11), 1832–1837 (2005)
52. Amores, J., Sebe, N., Radeva, P.: Context-based object-class recognition and retrieval by generalized correlograms. IEEE Trans. Pattern Anal. Mach. Intell. 29(10), 1818–1833 (2007)
53. Liu, D., Hua, G., Viola, P., Chen, T.: Integrated feature selection and higher-order spatial feature extraction for object categorization. In: CVPR 2008, pp. 1–8 (2008)
54. Setia, L., Teynor, A., Halawani, A., Burkhardt, H.: Grayscale medical image annotation using local relational features. Pattern Recognition Letters 29(15), 2039–2045 (2008); Image CLEF 2007 - Automatic annotation of medical images for image retrieval
55. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: CVPR 2006, vol. II, pp. 2033–2040 (2006)
56. Savarese, S., DelPozo, A., Niebles, J., Fei-Fei, L.: Spatial-Temporal correlatons for unsupervised action classification. In: IEEE Workshop on Motion and Video Computing, WMVC 2008, pp. 1–8 (2008)
57. Lazebnik, S., Schmid, C., Ponce, J.: Semi-local affine parts for object recognition. In: BMVC, pp. 959–968 (2004)
58. Quack, T., Ferrari, V., Leibe, B., Gool, L.J.V.: Efficient mining of frequent and distinctive feature configurations. In: ICCV, pp. 1–8 (2007)
59. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2), pp. 2169–2178 (2006)
60. Ling, H., Soatto, S.: Proximity distribution kernels for geometric context in category recognition. In: ICC 2007, pp. 1–8 (2007)
61. Bosch, A., Zisserman, A., Muñoz, X.: Representing shape with a spatial pyramid kernel. In: CIVR, pp. 401–408 (2007)
62. Vedaldi, A., Soatto, S.: Relaxed matching kernels for object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)
63. Zhang, Y., Chen, T.: Efficient kernels for identifying unbounded-order spatial features, pp. 1762–1769 (2009)
64. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.J.C.H.: Text classification using string kernels. Journal of Machine Learning Research 2, 419–444 (2002)

65. Pantofaru, C., Dorko, G., Schmid, C., Hebert, M.: Combining regions and patches for object class localization. In: The Beyond Patches Workshop in conjunction with the CVPR, pp. 23–30 (June 2006)
66. Martin, D.R., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV, pp. 416–425 (2001)
67. Yang, L., Meer, P., Foran, D.: Multiple class segmentation using a unified framework over mean-shift patches. In: CVPR 2007, pp. 1–8 (2007)
68. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
69. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models, pp. 119–126 (2003)
70. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: NIPS (2003)
71. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: CVPR (2), pp. 1002–1009 (2004)
72. Tang, J., Hare, J.S., Lewis, P.H.: Image auto-annotation using a statistical model with salient regions. In: ICME, pp. 525–528 (2006)
73. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
74. Murphy, K., Torralba, A., Freeman, W.: Using the forest to see the trees: a graphical model relating features, objects and scenes. In: NIPS (2003)
75. Kumar, S., Hebert, M.: Discriminative random fields: A discriminative framework for contextual interaction in classification. In: ICCV, vol. 2, pp. 1150–1157 (2003)
76. Fei-Fei, L., Fergus, R., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, June 20-25, vol. 2, pp. 524–531 (2005)
77. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. PAMI 28(4), 594–611 (2006)
78. Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR (2), pp. 1605–1614 (2006)
79. Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T., Gool, L.J.V.: Modeling scenes with local descriptors and latent aspects. In: ICCV, pp. 883–890 (2005)
80. Barnard, K., Duygulu, P., Forsyth, D.A., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. Journal of Machine Learning Research 3, 1107–1135 (2003)
81. Blei, D., Jordan, M.: Modeling annotated data. In: SIGIR 2003. ACM, New York (2003)
82. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
83. Magalhães, J., Rüger, S.M.: Logistic regression of generic codebooks for semantic image retrieval. In: Sundaram, H., Naphade, M., Smith, J.R., Rui, Y. (eds.) CIVR 2006. LNCS, vol. 4071, pp. 41–50. Springer, Heidelberg (2006)
84. Boutell, M.R., Luo, J., Brown, C.M.: Factor graphs for region-based whole-scene classification. In: CVPR Workshop on Semantic Learning Applications in Multimedia, p. 104 (2006)

85. Fan, J., Gao, Y., Luo, H., Xu, G.: Statistical modeling and conceptualization of natural images. Pattern Recognition 38(6), 865–885 (2005)
86. Luo, J., Savakis, A.E., Singhal, A.: A bayesian network-based framework for semantic image understanding. Pattern Recognition 38(6), 919–934 (2005)
87. Monay, F., Quelhas, P., Odobez, J.M., Gatica-Perez, D.: Integrating co-occurrence and spatial contexts on patchbased scene segmentation. In: CVPR Workshop on Beyond patches. IEEE Computer Society, Los Alamitos (2006)
88. Wang, G., Zhang, Y., Fei-Fei, L.: Using dependent regions for object categorization in a generative framework. In: CVPR 2006, New York, NY, USA, June 17-22, vol. 2, pp. 1597–1604 (2006)
89. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: CVPR, pp. 731–737 (1997)
90. Cao, L., Fei-Fei, L.: Spatially coherent latent topic model for concurrent object segmentation and classification. In: Proceedings of IEEE Intern. Conf. in Computer Vision (ICCV) (2007)
91. Singhal, A., Luo, J., Zhu, W.: Probabilistic spatial context models for scene content understanding. In: CVPR (1), pp. 235–241 (2003)
92. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory 47(2), 498–519 (2001)
93. Neches, R., Fikes, R., Finin, T.W., Gruber, T.R., Patil, R.S., Senator, T.E., Swartout, W.R.: Enabling technology for knowledge sharing. AI Magazine 12(3), 36–56 (1991)
94. Chandrasekaran, B., Josephson, J.R., Benjamins, V.R.: What are ontologies, and why do we need them? IEEE Intelligent Systems 14(1), 20–26 (1999)
95. Gruber, T.: Towards principles for the design of ontologies used for knowledge sharing. International Journal for Human-Computer Studies 43, 907–928 (1995)
96. Guarino, N., Giaretta, P.: Ontologies and knowledge bases: Towards a terminological clarification. In: Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, pp. 25–32 (1995)
97. Towards text recognition in natural scene images. In: SWAMM 2006, collocated with WWW 2006, Edinburgh, Scotland (2006)
98. Mezaris, V., Kompatsiaris, I., Strintzis, M.G.: Region-based image retrieval using an object ontology and relevance feedback. EURASIP J. Appl. Signal Process. 2004(1), 886–901 (2004)
99. Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V.K., Strintzis, M.G.: Knowledge-assisted semantic video object detection. IEEE Trans. Circuits Syst. Video Techn. 15(10), 1210–1224 (2005)
100. Maillot, N.: Ontology-based object learning and recognition. PhD thesis, ORION / INRIA Sophia-Antipolis (December 2005)
101. Hudelot, C., Maillot, N., Thonnat, M.: Symbol grounding for semantic image interpretation: from image data to semantics. In: ICCV, Workshop on Semantic Knowledge in Computer Vision (2005)
102. Town, C., Sinclair, D.: Language-based querying of image collections on the basis of an extensible ontology. Image Vision Comput. 22(3), 251–267 (2004)
103. Leibe, B., Schiele, B.: Interleaved object categorization and segmentation. In: BMVC (September 2003)
104. Harris, C., Stephens, M.: A combined corner and edge detector. In: 4th ALVEY Vision Conference, pp. 147–151 (1988)
105. Agarwal, S., Roth, D.: Learning a sparse representation for object detection. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 113–127. Springer, Heidelberg (2002)

106. Borenstein, E., Ullman, S.: Class-specific, top-down segmentation. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2351, pp. 109–122. Springer, Heidelberg (2002)
107. Farmer, M.E., Jain, A.K.: A wrapper-based approach to image segmentation and classification. In: ICPR (2), pp. 106–109 (2004)
108. Kokkinos, I., Maragos, P.: An expectation maximization approach to the synergy between image segmentation and object categorization. In: ICCV, pp. 617–624 (2005)
109. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient matching of pictorial structures. In: CVPR, p. 2066 (2000)
110. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV 2004 Workshop on Statistical Learning in Computer Vision (May 2004)
111. Jones, M., Poggio, T.: Multidimensional morphable models: A framework for representing and matching object classes. International Journal of Computer Vision 29, 107–131(25) (1998)
112. Cootes, T., Taylor, C.: Statistical models of appearance for computer vision. Technical report, University of Manchester, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, Manchester M13 9PT, United Kingdom (September 1999), http://www.wiau.man.ac.uk
113. Borenstein, E., Malik, J.: Shape guided object segmentation. In: CVPR 2006, pp. 969–976. IEEE Computer Society, Los Alamitos (2006)
114. Galun, M., Sharon, E., Basri, R., Brandt, A.: Texture segmentation by multiscale aggregation of filter responses and shape elements. In: ICCV, pp. 716–723 (2003)
115. Ferran, C., Giró, X., Marqués, F., Casas, J.R.: BPT enhancement based on syntactic and semantic criteria. In: Avrithis, Y., Kompatsiaris, Y., Staab, S., O'Connor, N.E. (eds.) SAMT 2006. LNCS, vol. 4306, pp. 184–198. Springer, Heidelberg (2006)
116. Salembier, P., Garrido, L.: Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. IEEE Transactions on Image Processing 9(4), 561–576 (2000)
117. Papadopoulos, G.T., Mezaris, V., Dasiopoulou, S., Kompatsiaris, I.: Semantic image analysis using a learning approach and spatial context. In: Avrithis, Y., Kompatsiaris, Y., Staab, S., O'Connor, N.E. (eds.) SAMT 2006. LNCS, vol. 4306, pp. 199–211. Springer, Heidelberg (2006)
118. Skiadopoulos, S., Giannoukos, C., Sarkas, N., Vassiliadis, P., Sellis, T., Koubarakis, M.: 2d topological and direction relations in the world of minimum bounding circles. IEEE Transactions on Knowledge and Data Engineering 17(12), 1610–1623 (2005)
119. Wang, Y., Makedon, F., Ford, J., Shen, L., Goldin, D.Q.: Generating fuzzy semantic metadata describing spatial relations from images using the r-histogram. In: JCDL 2004, Tuscon, AZ, USA, June 7-11, pp. 202–211 (2004)
120. Sikora, T.: The mpeg-7 visual standard for content description-an overview. IEEE Trans. Circuits Syst. Video Techn. 11(6), 696–702 (2001)
121. Athanasiadis, T., Mylonas, P., Avrithis, Y.S.: A context-based region labeling approach for semantic image segmentation. In: Avrithis, Y., Kompatsiaris, Y., Staab, S., O'Connor, N.E. (eds.) SAMT 2006. LNCS, vol. 4306, pp. 212–225. Springer, Heidelberg (2006)
122. Athanasiadis, T., Tzouvaras, V., Petridis, K., Precioso, F., Avrithis, Y., Kompatsiaris, Y.: Using a multimedia ontology infrastructure for semantic annotation of multimedia content. In: Proc. of 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2005), Galway, Ireland (November 2005)

Ontology Based Information Extraction from Text

Vangelis Karkaletsis, Pavlina Fragkou, Georgios Petasis, and Elias Iosif

Institute of Informatics and Telecommunications
National Center for Scientific Research (N.C.S.R.) "Demokritos"
15310 Aghia Paraskevi Attikis, Greece
{vangelis,fragou,petasis,iosife}@iit.demokritos.gr

Abstract. Information extraction systems employ ontologies as a means to describe formally the domain knowledge exploited by these systems for their operation. The aim of this survey is to study the contribution of ontologies to information extraction systems. We believe that this will help towards specifying a concrete methodology for ontology based information extraction exploiting all levels of ontological knowledge, from domain entities for named entity recognition, to the use of conceptual hierarchies for pattern generalization, to the use of properties and non-taxonomic relations for pattern acquisition, and finally to the use of the domain model itself for integrating extracted entities and instances of relations, as well as for discovering implicit information and detecting inconsistencies.

1 Introduction

Information extraction from textual content is situated between information retrieval and text understanding. Unlike information retrieval where the aim is to locate passages of text relevant to a domain-specific topic or a user's query (e.g. news on pole-vault events), information extraction aims to locate inside a text passage domain-specific and pre-specified facts (e.g., facts about the athlete participating in a pole-vault event, such as his/her name, nationality, performance, as well as facts about the specific event, such as the event name, location). Unlike text understanding, only a small portion of a text is typically relevant to an extraction task.

Information extraction (IE) can be defined as the automatic identification of selected types of entities, relations or events in free text (see [16] for an introduction to the approaches used). More specifically, IE is about extracting from texts the following different types of information:

1. *Entities*: textual fragments of particular interest, such as persons, places, organizations, dates, etc.
2. *Mentions*: the identification of all lexicalisations of an entity in texts. For example, the name of a particular person can be mentioned in different ways inside a single document, such as "Tatiana Lebedeva", "T. Lebedeva", or "Lebedeva".
3. *Relations between entities*: the identification of the relations holding between extracted entities, according to an existing specification (domain knowledge). Usually these relations are triggered by linguistic evidence found between mentions of entities, such as lexico-syntactic patterns. For example, a lexico-syntactic pattern of the

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 89–109, 2011.

form "<location name>, <country name>" can suggest that a particular place is related to a country through the "located_in" relation.

4. *Events involving the entities*: the identification of all the entities involved in an event described in a document, as well as the identification of other related events to it. For example, the identification of an entity representing an athlete, which is composed of different related entities, such as his/her name, age, nationality, performance, as well as the identification that this specific athlete's entity has participated in a specific entity representing a pole-vault event.

The processing steps usually followed to find the aforementioned types of information, are [1]:

1. *Named Entity Recognition (NERC):* entity mentions are recognized and classified into proper types for the thematic domain.
2. *Co-reference:* all the mentions that represent the same entity are identified and grouped together according to the entity they refer to.
3. *Template Element Task:* all mentions of an entity (representing a person, object or event) are interpreted in order to create the entity that represents this person, object or event.
4. *Template Relation Task*: relations between recognized entities are identified.
5. *Scenario Template Task:* a pre-specified template for a specific event is filled organizing the information extracted from all the previous processing steps.

The task of information extraction from text has been the subject of significant research in the past two decades. Research was influenced by the Message Understanding Conferences (MUCs) [12, 13], a series of evaluations of IE technology that helped to establish common evaluation measures. The shift in the latest MUCs from black-box evaluations to glass-box evaluations led to the establishment of a common decomposition of IE into the processing steps mentioned above.

Robustness and fast adaptation to new domains are key issues in IE systems. In the first MUCs, IE was tackled as a full natural language understanding (NLU) problem that required complete syntactic and semantic analysis, resulting in systems with limited computational efficiency. After the 3rd Message Understanding Conference in 1991, it became clear that IE systems differ significantly from traditional NLU systems [15]. IE systems based on simple pattern matching techniques [23] were reported to achieve better results than systems that attempted to perform "deep" syntactic and semantic analysis, e.g., [20]. Also, they were faster and easier to debug and adapt to new domains. Furthermore, several systems that employ machine learning techniques, e.g., [4, 31], have been proved easier and faster to port to new domains, mainly compared to systems that use hand-crafted patterns and rules. Hybrid approaches that combine knowledge-based techniques with machine learning have been presented, in an attempt to exploit the advantages of both worlds, usually leading to more accurate systems as demonstrated by the top ranked system for the Named Entity task at MUC-7 [26] and more recent approaches in the area.

Despite the advances introduced by the use of machine learning, portability to new thematic domains still remains an open issue. Many of the tasks performed by a traditional IE system (especially the ones that relate to templates) have a strong dependency on knowledge about the thematic domain, which is very frequently scattered

among the various tasks. Ontology based IE systems try to alleviate this problem through the use of ontologies, which provide the means to disassociate an IE system from the domain knowledge required for its operation. Making domain knowledge explicit through an ontology, not only enhances portability, but also provides new opportunities for IE systems, ranging from using the ontology for storing the extracted information to using reasoning for implementing various IE tasks. For example, the BOEMIE IE system[1] maintains the traditional NERC and co-reference steps, whose results are used to populate an ontology, and substitutes all the template-related steps with reasoning over this ontology, driven by a set of inference rules stored explicitly, along with the ontology. In addition, the fact that domain knowledge is explicitly described by an ontology allows the adaptation of the system's behaviour through changes in its ontology, usually in a synergistic approach where extracted information is used to enhance the ontology, which in return affects the performance of the IE system.

This survey provides first a classification of IE systems in four different groups according to the level of ontological knowledge they exploit (Section 2.1). The most important features that characterise an IE system for the purposes of this survey are presented in Section 2.2. Based on these features, Section 3 describes representative systems belonging in each of the fours groups (Sections 3.1 – 3.4) and discusses the characteristics of each group. Section 4 concludes this survey, summarising the current trends in ontology-based information extraction research.

2 Semantics Extraction from Textual Content

In this section we first classify OBIE systems in four different groups according to the level of ontological knowledge they exploit. We then describe the basic features of OBIE systems. These features are used in section 3 for the description of representative systems in each of the four different groups.

2.1 Classification of Ontology-Based Information Extraction Methods

Ontologies in OBIE systems provide the domain knowledge model required for the systems' operation. This model can be a rather poor one (e.g., a flat list of athlete names, location names, etc., the so-called gazetteer lists) or a rich one (e.g., a model built using an ontology language like OWL, which enables the representation of complex entities or events as well as the reasoning over them) enabling the categorization of information extraction systems according to the level of ontological knowledge they use.

In order to classify ontology-based information extraction systems we follow the classification proposed in [27], according to which four different levels of ontological knowledge can be exploited by an IE system.

The first level includes the domain entities (e.g., person, location) and their variations (synonyms, co-referents), as well as word classes (i.e., keywords/terms and their variations, specifiers/descriptors of entities). These are mainly used in the IE process for named entity recognition and classification, for named entity normalisation where

[1] http://www.boemie.org

the various forms of a name can be annotated with a value corresponding to their normalised form, as well as for co-reference resolution (e.g., that the phrases "she", "this athlete" co-refer to the person name "Tatiana Lebedeva").

At a second level, domain entities or word classes are organized in conceptual hierarchies. For instance, an ontology for athletics may include a concept (class) "person" with sub-concepts for "athlete", "trainer", etc., whereas in WordNet [14], word classes (synsets) are structured via the hypernym/hyponym relation into a hierarchy. Such conceptual hierarchies can be exploited by an IE system for generalizing/specializing its extraction rules (either in a rule-based or a machine learning based system). In the case of the athletic domain, extraction rules for recognizing person names can be specialized in order to recognize those names that correspond to athletes exploiting some other features that are derived from the ontology. For example, since the concept of athlete requires among others, a name and an age as attributes, an athlete's name it is expected to be in close textual proximity with numerical values that denote age.

The third level of ontological knowledge that can be exploited by IE systems concerns the concepts' properties and/or the relations between concepts. For example, the "athlete" concept can be defined by the "name" property filled by a string value, the "nationality" relation filled by values of the concept "country", the property "age" filled by a numeric value, etc. These properties and relations guide the information extraction system in various processing stages, from named entity recognition to template relation extraction, independently of the techniques used. In a rule-based approach, the knowledge engineer writes rules for detecting specific types of relations inside a text (e.g., "nationality" associates a person-name with a country-name, therefore the corresponding rule looks for instances of such named entities inside a sentence associated in different ways, such as "… the Russian T.Lebedeva…", "T.Lebedeva, the Rusian athlete"). In a machine learning based approach such extraction rules are acquired from corpora that had previously been annotated according to the ontology.

The fourth level of ontological knowledge is the domain model itself. This knowledge is exploited at the final processing stage of IE, that of template filling. It is not enough to detect named entities inside the text, associate them with properties, and relate them with other named entities, according to the entity types (concepts), properties and relations types defined in the ontology. These extracted facts must be combined according to the domain model in order, for instance, to detect an athlete's instance and associate it with a sport in which the specific athlete participated. For example consider the following sentence: "*At last night's IAAF World Athletics Tour meeting in New York, Jamaica's Usain Bolt set a new World record for the men's 100m in a time of 9.72 seconds*". Assume that the tokens, "New York"', `"Jamaica's", "Usain Bolt", "men's", and "100m", are recognized as named entities, being instances of the concepts: "City", "Country", "Name", "Gender", and "Sport", respectively. Next, the following relations are generated in the form of tuples: "(Usain Bolt, Jamaica)", "(Usain Bolt, men)", "(Usain Bolt, 9.72 seconds)" and "(100m, New York)". At this point the aforementioned instances and relations can be furthermore exploited according to the domain model in order to build instances of higher level: *"Athlete= (Athlete_Name= "Usain Bolt", Athlete_Gender= "men", Athlete_Nationality = "Jamaica")", and "Sport= (Sport_Name= "100m", Sport_City= "New York")".*

Finally, a binary relation, encoding the participation of the athlete to the sport, connects the two high-level instances. In this example, the domain model enables different structures to be merged, and also, several constraints can be set, e.g., the identification of athlete's name and nationality is required for detecting an instance of an athlete's concept. Moreover, additional actions can be performed, such as consistency checking, adoption of assumptions in case of missing information, as well as discovery of implicit information.

2.2 Descriptive Features of Information Extraction Systems

A set of features that we consider as important ones for the clear and comparative description of IE systems, for the purposes of this survey, are presented below:

1. Initial requirements in terms of the type of input documents (raw texts, semi-structured such as web pages, fully structured content in the form of tables or databases) and the preprocessing required (lexical analysis, syntactic analysis, semantic analysis).
2. Extraction process: rule-based, machine learning-based, hybrid. The features on the extraction process help identifying the current trends, for instance, with respect to the infrastructure used.
3. Use of the ontology in the IE process, that is, which level/levels of the ontological knowledge are exploited. It refers to the four classes of IE systems distinguished according to the level of ontological knowledge they exploit.
4. Ontology features: technical and other related aspects regarding the use of ontology, such as the knowledge representation formalism, the ontology implementation language, the inference mechanism employed.
5. Output (e.g., annotated corpus, filled templates, populated/enriched ontology). This is of greater interest in the cases where a bootstrapping approach is involved according to which the IE result populates the ontology and the populated version is then employed to improve IE performance
6. Portability, which examines the role of the ontology in the porting to new domains.

3 Information Extraction (IE) Systems

In this section we describe several representative information extraction systems classified to the four categories defined in the previous section. A comparative analysis is included in order to identify the trends that exist in this research field, as well as to highlight a number of different approaches that are followed in certain cases. Overall, the aim is to make clearer the contribution of ontological knowledge to the process of information extraction.

3.1 IE Systems Exploiting Domain Entities, Word Classes

As noted in Section 2.1, the first level of ontological knowledge includes the domain entities and their variations (synonyms, co-referents), as well as word classes

(i.e., keywords/terms and their variations, specifiers/descriptors of entities). Consider for example the sentence describing an athletic event:

"New York, **USA** – At last night's **IAAF World Athletics Tour meeting** in **New York**, **Jamaica**'s **Usain Bolt** set a new World record for the **men**'s **100m** in a time of **9.72 seconds**."

IE systems belonging to this category may exploit lists of known entities and their variations (these can also be encoded using rules). For instance, in the athletic domain these may be lists of athletes, countries, cities, sport names, along with rules denoting performance types, variations of athletes' and events' names, etc., based on an athletic domain ontology. Such systems may recognize the **IAAF World Athletics Tour meeting** as an already known event name, and the **100m** as an expression corresponding to the sport of 100 meters. In this section, we present indicative systems that exploit these types of ontological knowledge.

The knowledge about the ontological entities of the domain of interest can be exploited through a wide range of approaches, which span from entity semantic annotation to normalization, and posing of constraints over conceptual properties. The semantic annotation of entities is mainly performed for the construction of a training corpus. The entities may be stored either in flat repositories, e.g., gazetteers, or in more structured resources, such as ontologies. For example, a gazetteer can contain several distinct entries under the entity of city, while an ontology can relate the entity of city with other entities, such as country. The use of ontology for information extraction goes one step further in the case of normalization. For example, "United States of America", "USA, and "US" are lexical variations of the same entity which can be represented by a single lexical pattern. Furthermore, the ontology can be used for identifying text of interest according to constraints posed on conceptual properties. For instance, if we are interested in recognizing the age of an athlete, a number greater than a reasonable value, cannot be accepted as a valid athlete's age.

There are several systems that can be classified under this category. However, we focus our description in three of them that we consider as representative ones: Learning Pinocchio [10], CROSSMARC [21] and 2PP [24].

In [10], domain entities are manually annotated in a training corpus. For these tagged examples, features derived from morphological analysis, POS tagging, semantic labeling by gazetteer consulting, are taken into consideration, for learning extraction rules. Furthermore, the system generalizes the learnt rules by exploiting contextual dependencies.

Various approached are examined in the CROSSMARC project, [21], where entities are directly retrieved from the domain ontology. The process of multilingual information extraction consists of two major phases: named entity recognition (NERC) and fact extraction. Different approaches are used by the monolingual NERC modules: machine learning-based, rule-based and hybrid. Fact extraction modules either employ machine learning techniques or apply a hybrid approach in order to learn an editable model. CROSSMARC ontology is used for the storing and normalization of named entities. Fact extraction deals with the filling of information templates consisting of slots, by assigning domain specific roles to identified entities. This extraction step takes into account some ontological constraints regarding the properties of entities.

A different approach is proposed by [24], where patterns of orthographic and semantic type are discovered and the ontological knowledge is used in the final phase of the extraction process. For the case of orthographic patterns, features such as delimiters and capitalization are taken into account. In addition, dictionaries are employed for the identification of certain entities. The results from the application of the orthographic patterns may be exploited by the semantic patterns according to the underlying ontological scheme. For example, given the sequence "Athens, Greece, December 15, 2010", the orthographic procedure gives "<Entity1>, <Entity2>, <Date>", which then becomes "<City>, <Country>, <EventDate>", exploiting the corresponding (or the most similar) transformation encoded in the ontology.

According to the above descriptions, the ontology can be exploited in different levels and phases during the process of information extraction. In [10], concept instances are annotated in the corpus and are then exploited during the training phase. Concepts' relations are not encoded in the training corpus; hence, they do not affect explicitly the process of information extraction, although the contextual dependencies of annotated instances are exploited during training. Using an example about a lecture event, it is obvious that the concepts of "speaker" and "lecture time" are strongly related, but this kind of knowledge is not encoded in the domain ontology. The ontology is exploited in a similar way in [24]. In [21] the ontology consists of three layers: meta-conceptual, conceptual and instances layer. The first layer defines the language of the subsumed layers. Also, the structure of the templates used in the phase of fact extraction is defined in the meta-conceptual layer. The conceptual layer includes the domain concepts and the relations between them. The instances layer contains instances of concepts, as well as lexical instantiations of them. For example, "Pentium 3" is an instance of the concept "processor name", while "Pentium III", "P3", "PIII", etc., are equivalent lexical instantiations of "Pentium 3". Overall, it is interesting to note that the ontological knowledge in [10, 21] is exploited throughout the process of information extraction, while in [24] has a secondary role in the final phase of the process. Regarding evaluation process, all of the presented systems use precision and recall (or F-measure), in order to measure the correctness of assigning an extracted instance to an ontological concept. The following table summarizes the main features of the aforementioned systems that belong to the first category.

It is interesting to note that in addition to the typical pre-processing operations for extracting linguistic features, some systems put effort in identifying document regions of semantic coherence [21, 24]. The machine learning-based approach appears to facilitate porting to new domains. In practice there are numerous ways about the incorporation of the domain ontological knowledge in the context of a machine learning approach. The most straightforward way is the annotation of conceptual instances in the training corpus [10], while the ontology can be further exploited, as in [21], where instances are encoded reflecting the ontological structure.

In this category, the ontological knowledge is not always represented in a full scale, because only the individual conceptual instances are mainly exploited. For example, dictionaries and gazetteers are used for the instantiation of ontological concepts, for 2PP and Learning Pinocchio systems, respectively. In contrast, in the CROSSMARC system a multi-layered ontology is used.

Table 1. Representative IE systems of the 1st category

System	Input, Preprocessing	IE approach	Use of ontology	Ontology features	Output
Learning Pinocchio [10]	-Web pages -Tokenization, lemmatization, POS tagging	Machine learning	Instance annotation in training corpus	Gazetteer for ontological concepts	Annotated corpus
CROSSMARC [21]	-Web pages -Tokenization, POS tagging, NER, identification of document regions	Combination of machine learning and rule-based techniques	- Instantiation of concepts - Template definition	Multilayered ontology	Filled templates
2PP [24]	-Web pages - Identification of document regions	Machine learning	Disambiguation	Dictionary for ontological concepts	Mapping to ontological concepts

The output seems to have several levels, some of which are common across different systems, while others are dependent on the goals of the specific application. For example, all three systems are able to identify in a given text instances of the domain concepts. This can be used for corpus annotation [10], and for the mapping of the identified information to the conceptual schema of the underlying ontology. Moreover, this can be extended according to the specifications of a certain application, as in the case of [21], where entity constraints have an important role, such as a computer processor of a certain speed.

3.2 IE Systems Exploiting Conceptual Hierarchies

As noted in Section 2.1, at a second level, domain entities or word classes are organized in conceptual hierarchies. Such hierarchies can be exploited by an IE system for generalizing/specializing its extraction rules.

Using the same example as before, systems belonging to this category go a step further. More specifically, they recognize that the expression **Usain Bolt** corresponds to an athlete name belonging to the ***Athlete*** concept which is a sub-concept of a ***Person.*** In the same way, the expression **100m** is recognized as instance of the ontology concept ***100 meters*** and a sub-concept of ***Running Sports***. As noted in [27], IE systems do not focus on the use of the hierarchical organisation, or they do not adequately report on this, and for this reason the contribution of conceptual hierarchies in the IE process is unclear.

Representative systems are presented under this category, giving emphasis on the use of conceptual hierarchies (taxonomic relations), although they also employ non-taxonomic relations (see Section 3.3). A taxonomic relation between two concepts reflects the sub- or super-ordination between them [11].

The taxonomic relations that are used during the information extraction process can be classified into general and domain-specific ones. Some widely used general

relations are “is-a” and “part-of”, which can be exploited in a large variety of domains. On the other hand, domain-specific relations, as for example, the relation “located-in” from a river to a region in the sentence “Amazon river is located in South America”, can be exploited only within a specific domain. Despite their differences these types of relations encode valuable ontological knowledge, which can be efficiently exploited by information extraction systems [3, 8, 9].

The system proposed in [8] is an interesting example of using general taxonomic relationships in order to classify instances with respect to a given ontology. The key idea is simple and utilizes specific linguistic patterns that imply such relationships. This approach is fully unsupervised and each pattern is used “as is”. Hence, the existence of a training corpus is not pre-requisite in order to learn extraction rules. On the other hand, the goal of discovering such relationships requires an adequately large corpus. To tackle this, they used Web as a corpus. The procedure of instance classification is summarized as follows. First, the system takes as input a web page and identifies proper nouns. Next, the identified proper nouns and the ontology concepts are connected via the linguistic patterns that denote the above relationships. An example of such association is “<CONCEPT> such as <INSTANCE>”, where “<CONCEPT>” denotes a concept stored in ontology, “<INSTANCE>” is an identified proper noun, and “such as” is a linguistic pattern which implies the relation of hyponymy. For each possible combination between instances, concepts and patterns, a web search engine is used in order to find the frequencies of all combinations, retaining just the number of the returned hits. An instance is categorized to an ontological concept according to the ranking of the previously computed list of frequencies. Finally, the web page is annotated, using the conceptual classification of the proper nouns that appear in it. An extension of the described system is proposed in [9]. Instead of taking into account the number of document hits, an abstract of each document is obtained, as this capability is provided by the search engine. Then, linguistic patterns are identified in the downloaded abstracts and their context is explored in order to overcome ambiguity problems that occurred in the initial system [8].

In [3] the main exploited relations for information extraction are not domain-general, but they are oriented to certain events encoded in the domain model. The events are described by verbs, involving a pair of actors. The subject and object of the verb - reflecting the event of interest - determine the actors that participate in the event. For example, in “CompanyA was sold to CompanyB” the event of “Selling” is present, while “CompanyA” and “CompanyB” are the participating actors. The extraction of such an event is accomplished through the activation of rules, which pose certain semantic constraints. For instance, the actors should be “companies”, which ontologically belong to the concept of “social group”. This taxonomic relationship is used during the information extraction process that validates a set of constraints, while traversing the hierarchical structure of the participating actors. That is, any identified named entity inherits the properties of its ontological ancestors and these properties are validated during the activation of rules. Also, it should be noted that for the instantiation of objects, subjects, and also verbs, the corresponding EuroWordNet synsets are used, as they are retrieved from the ontology. For example, the “Selling” event can be expressed by a set of verbs, such as {“acquire”, “buy”, etc.}. The rules are triggered by the presence of a verb, while they are induced by a machine learning approach, which uses a flat consideration of the ontology as this is performed by the systems of the previous category.

According to the above descriptions, the use of the ontological knowledge during the process of information extraction must be investigated according to the nature of the relations that are exploited. For domain-general relations that are expressed through typical linguistic patterns, the main concerns of the extraction algorithm are the definition of such patterns and the use of sufficient number of examples that tackle the data sparseness problem. The use of such patterns for information extraction purposes is a well-studied problem with numerous approaches, inspired by the work of Hearst [19], for identifying and extracting relations between the entities of interest. In the framework of this approach the Web is a very popular knowledge resource, due to its size and the semantic diversity. On the other hand, approaches like [8] consider search engines as "black boxes", without having an insight into the semantics of the returned hits. A deeper understanding of the underlying semantics is attempted by the successor of [8], which uses contextual information in order to alleviate semantic ambiguities. Concerning the ontological use the systems presented in [8, 9], requires a specific ontology to retrieve the concepts, and a set of linguistic patterns that imply general taxonomic relations. The interesting aspect is the independence of the linguistic patterns from the ontology used. In contrast, the use of domain-specific relations in [3] depends on the domain ontology. This happens because the actors of an event are associated through a verb, whose semantics is domain-dependent. The ontological relation of actors is then traversed for consistency checking. The use of linguistic patterns for classification of candidate instances to ontological concepts is a straightforward and powerful method, due to its unsupervised and intuitive nature [8]. However, it lacks additional features, such as posing constraints on conceptual attributes. This is addressed by systems like [3] in which the relations are exploited for the needs of constraint validation throughout the process of information extraction.

For evaluation purposes, NAMIC (News Agencies Multilingual Information Categorization) system [3], employs two precision measurements; one regarding the quality of slot filling and another for the assignment of extracted information to event types. In PANKOW (Pattern-based Annotation through Knowledge On the Web) system [8], the traditional metrics of precision and recall are applied. However, in the

Table 2. Representative IE systems of the 2nd category

System	Input, Preprocessing	IE approach	Use of ontology	Ontology features	Output
PANKOW C-PANKOW [8, 9]	-Web pages -POS tagging	Pattern-based	Provide concepts incorporated in patterns	Ontological representation with KAON tool	Annotated corpus
NAMIC [3]	-Semi-structured -Lemmatization, POS tagging, NER, chunking, clause boundary detection	Machine learning	Use taxonomic relations for checking constraints	Ontological representation with XI language	Fill templates for event recognition

extended system C-PANKOW (Context-driven PANKOW) [9], the evaluation measurement of Learning Accuracy is adopted, which acknowledges that there is not always a unique correct assignment for an identified instance. The following table summarizes the main features of the discussed systems.

Systems like [8, 9] seem to be portable in new domains, under the assumption that the linguistic patterns that indicate certain taxonomic and domain-general relations are available. In the case of systems like [3] the used ontological relations seem to contribute less to domain portability, since they are used partially in the process of information extraction. Furthermore, the exploitation of these relations is strongly dependent to the design of the domain ontology. Despite the simplicity of the method proposed in [8], the final result is a promising step towards the automatic semantic annotation of a given document. On the other hand, the use of domain-specific relations in systems like [3] provides a methodology for building inference mechanisms. The systems of this category seem to need a formal ontological representation. PANKOW and C-PANKOW use KAON ontology management infrastructure. NAMIC system uses XI, a Prolog-based language that is able to represent knowledge about individuals, classes, and their relations.

3.3 IE Systems Exploiting Conceptual Properties and Relations

The third level of ontological knowledge concerns the concepts' properties and/or the relations between concepts. These properties and relations guide the information extraction system in various processing stages, from named entity recognition to template relation extraction, independently of the techniques used.

Thus, regarding our example, appropriate patterns may be defined or machine learning approaches may be used in order to capture relation instances as follows:

1) **Event name** with **city and country,** (in our case the tuples (**IAAF World Athletics Tour meeting**, **New York**) and (**IAAF World Athletics Tour meeting**, USA))
2) **Athlete name** with **Nationality, Gender** and **Performance** (in our case the tuples (**Jamaica**, **Usain Bolt),** (**Usain Bolt, men)** and **(Usain Bolt, 9.72 seconds))**

In this section, we present representative systems that exploit this level of ontological knowledge.

The ontological relationships of a domain can also be viewed according to the semantic complexity that they reflect. The taxonomic hierarchy of concepts captures a portion of the existing domain semantics, involving fundamental relations, such as "is-a". The domain-specific relationships, on the other hand, can give a better understanding of the domain. The semantic complexity of such relationships is theoretically unbounded, since in general there is not a unique perception of domain knowledge. In essence, the relation between any two concepts through domain-specific relationship is an issue limited by the assumptions adopted during the process of the ontology design. However, this degree of semantic freedom decreases as the knowledge diversity becomes narrower. For example, in a relatively generic ontology design, various domain-specific relationships can hold among the encoded concepts [17, 18]. On the other hand, the domain-specific relationships of a more focused design are more

restricted and expert consultancy is of greater need [29]. The domain semantics can be used, either individually [30] or synergistically with linguistic features for the recognition of domain-specific relationships [17, 18, 32]. Moreover, lower levels of ontological knowledge can be exploited, including the instantiation of domain concepts [29], up to taxonomic relationships [17, 18, 30, 32]. An additional source for information extraction employing domain-specific relationships relies on the use of lexical patterns [29] and grammar rules [30].

The system proposed in [32] follows a machine learning-based approach for relation extraction, in which numerous linguistic and semantic features are used. The linguistic features are derived through a preprocessing phase, including part of speech tags, NP/VP chunks, etc. WordNet synsets are used in order to provide the appropriate senses for the words of interest. Heuristics are used in order to avoid a complicated word disambiguation procedure. The system uses an annotated corpus for training purposes. The annotation covers all the aforementioned features, from linguistic and semantic information, to taxonomic and non-taxonomic relations. The task of information extraction is considered as a multi-class classification problem, using Support Vector Machines, where each class represents a relationship. Every mention in the corpus that refers to two entities is regarded as an evidence for their relation and is classified to one of the candidate classes.

In [17, 18], two types of knowledge are considered during the process of information extraction, grammatical and conceptual. Grammatical knowledge imposes syntactic constraints. For example, consider the sentence "The engine of the manufacturer …" in which the word "engine" is unknown. Regarding the use of ontology, initially all the top-level concepts are candidates for the word "engine". But, when the fragment "of the manufacturer" is parsed the word "engine" is hypothesized to be associated with the word "manufacturer" that is related to the concept of "Manufacturer", which already exists in the domain ontology. The concept of "Manufacturer" is connected with other concepts, e.g., "Address" with non-taxonomic relations, e.g., "HasAddress". During the parse of a sentence the system checks the fulfillment of such relations in order to justify these hypotheses. Also, new concepts are learnt by exploiting patterns of apposition and exemplification, e.g., "the X engine" and "the X is an engine", respectively.

An interesting approach of combining statistical and knowledge-based techniques is proposed in [30], using stochastic context-free grammars (SCFG). The terminal symbols of grammar consist of instances of the conceptual classes and several generic patterns like "and" that increase the coverage of rules. The non-terminal grammar symbols represent domain concepts, while rules express taxonomic and non-taxonomic relations. The system aims to relation extraction by finding that parse which has the maximum probability.

In [29], a pattern-based approach is proposed in order to populate non-taxonomic relations, in the field of philosophy. The non-taxonomic relations are stored in the domain ontology while the taxonomic organization of the ontology is populated mainly using manual effort. For the population of the non-taxonomic relations apart from the stored entity instances, Wikipedia is also exploited along with other domain-specific resources. The use of Wikipedia has a significant contribution to the extraction process. Finally, it is important to note that experts verify the extracted non-taxonomic relations.

The exploitation of linguistic features during the extraction of non-taxonomic relations is performed by [17, 18, 32], however, their particular use differs. In the case of [32] the linguistic features add another annotation parameter in the training corpus and we can say that they exist in parallel with the taxonomic and non-taxonomic relations that are also annotated in the same corpus. In the case of [17, 18] grammatical knowledge encoded as lexicalized dependency grammar plays the role of a model, which syntactically constraints the extraction process driven by the conceptual ontology. Despite these differences, both systems can be considered as examples where the linguistic information captures the underlying domain semantics. These semantics are useful during the process of information extraction, even if the types of the implied semantic relations are not always known. A suggestion for reducing the manual effort of annotating a training corpus is proposed in [30]. This is attempted by writing SCFG rules, which are trained over a corpus. The generalization power of the rules and thus, the required amount of the training corpus are issues that are dependent to the rule design. It is important that a grammar can be easily changed without demanding a new, large-scale annotation, compared to the systems that fully rely on annotated training corpora. The use of non-taxonomic relations during the process of information extraction is situated in the framework of grammar rules, instead of the encoding of them into an ontological scheme. This feature enhances the generalization ability of exploiting ontological relations, due to the probabilistic nature of the system. Moreover, the efficiency of this system, given limited annotated training data, is greater compared to typical machine learning-based methodologies of information extraction for which larger training corpora are needed. A different use of non-taxonomic relationships is followed in [29], according to specific patterns that are dependent to a particular external source (Wikipedia is used for acquiring biographic data). Despite the fact that the non-taxonomic relations are stored into the domain ontology, the structure of information provided by the source must be taken into consideration during the ontology design. That is because, the relationship of interest is explicitly declared within the first sentence of the retrieved passage. Despite this limitation, the exploitation of such external information sources can become a valuable tool for the task of ontology evolution. The major features of the described systems are present in the table that follows

In [32], the preprocessing step extracts various features, but this approach was mainly followed in order to investigate the contribution of each feature to the task of relation extraction. In the case of [17, 18], syntactic information is needed in order to proceed to conceptual hypotheses for an unknown word according to its context. In contrast to [17, 18, 32], the hybrid approach proposed in [30] does not require any syntactic information for the SCFG rules. From development perspective, [30] is more flexible compared to the other systems that employ patterns that require greater amount of training data. On the other hand, the availability of adequate data can enhance the task of learning new concepts as in [17, 18]. As it is observed by the system description, the contribution of non-taxonomic relationships in information extraction process is situated in different levels. In the simplest level, the non-taxonomic relations can be used for annotating a training corpus [32], for constraint checking [17, 18], even in the level of SCFG rule writing [30]. A different use of non-taxonomic relations is proposed in [29], where relations are expected to be found in a particular external source. Of course this dependency decreases the system's portability.

Relation extraction seems to be a common task for systems of this category. It should be noted that in very demanding knowledge fields, such as philosophy, feedback from experts is used regarding the validity of the extracted relations [29].

Table 3. Representative IE systems of the 3rd category

System	Input, Preprocessing	IE approach	Use of ontology	Ontology features	Output
OBIE [32]	-Raw, semi-structured text -Tokenization, sentence splitting, POS tagging, NP/VP chunking, Bu-Chat/MiniPar parsers	Machine learning	Annotation of entities, relations	Annotations as ontology	Relation extraction
SYNDI-CATE [17, 18]	- Raw text - POS tagging	Machine learning	- Linguistic constraints - Validation of relations	KL-ONE	Ontology population, enrichment
TEG [30]	Raw text	Rule-based	Use of entities, relations for training corpus annotation and rule development	Annotations as ontology	- Relation extraction - Ontology population
[29]	-Web pages	Pattern-based	Definition of non-taxonomic relations	OWL in hand-built ontology	- Relation extraction - Ontology population

The majority of systems of this category do not follow the use of a standardized ontology language. In OBIE and TEG systems, the ontological knowledge is in the form of annotations (an annotated corpus is used), rather than in a representation generated by an ontology-authoring tool. In particular, OBIE system uses ACE annotations, and TEG system uses MUC and ACE annotations. SYNDICATE system uses KL-ONE representation, and only the system proposed in [29] uses OWL over a hand-built ontology. In general, we feel that the exploitation of non-taxonomic relations for information extraction is a function of ontological design. The nature of non-taxonomic relations allows multiple points of view for a knowledge domain, in contrast to taxonomic relations. Thus, the design of ontology and extraction algorithms, especially when exploiting non-taxonomic relations, should be considered in a unified development plan.

3.4 IE Systems Exploiting the Domain Model

As noted in Section 2.1, the fourth level of ontological knowledge is the domain model itself. It is not enough to detect named entities inside the text, associate them with properties, and relate them with other named entities, according to the entity types (concepts), properties and relations types defined in the ontology. These extracted facts must

be combined in order to be semantically interpreted according to the domain model. In this case, the domain model enables different structures to be merged, checking consistency, making valid assumptions in case of missing information, as well as discovering implicit information. Thus, regarding our example in question, the instantiated tuples: (**IAAF World Athletics Tour meeting**, **New York**), (**Jamaica**, **Usain Bolt)** and (**men**, **100m)** are further exploited by the extraction mechanism and create the following ontology class instances:

- **Sport_class_instance**= (Sport_name = 100m, Sport_city = New York)
- **Event_class_instance** = (Event_name = IAAF World Athletics Tour meeting, Event_city = New York, Event_country = USA)
- **Athlete_class_instance** = (Athlete_name = Usain Bolt, Athlete_gender = men, Athlete_nationality = Jamaica, Athlete's_performance = 9.72 second)

The extraction mechanism of those systems not only creates those instances and populates accordingly the ontology in question, but also relates those instances by applying the inference rule set. The result of the application of the inference rule set is that the specific athlete participated in the described sport of the specific event and achieved the observed performance.

Although there are several IE systems that employ ontology management systems which can provide such inference services over the domain model, this is not exploited in practice, as we can tell from the descriptions they provide. In this section, we present representative systems that exploit the domain model.

The result of using domain knowledge during the process of information extraction is reflected on different levels of completion. Relations between concepts can be inferred if the intrinsic conceptual characteristics of a domain are taken into consideration. For example, regarding the domain of sports, a relation denoting that a goal is an own goal, is inferred by knowing that a goal achieved by a player against his team is an own goal. Clearly, this sort of specific knowledge is reasonable for certain sports, and reflects the idiomorphic characteristics of the corresponding domains. An important aspect of systems of this category is the synergistic relation between the inference mechanism and the ontology-based information extraction process. In the simplest case there are systems that infer relations among the ontological concepts according to the extracted information [5, 6]. Other systems [7, 25], make a step further, since the inference procedure enhances the information extraction task.

In [5, 6], the SOBA system is presented. SOBA performs information extraction for domain specific question answering. The information extraction process uses generic grammars for the identification of persons, locations, etc., as well as manually developed rules for the recognition of domain-specific entities and events, dealing with the sport of soccer. Once the extraction phase is completed, the extracted information is used for ontology population. Next, discourse analysis is applied in order to infer relations between the extracted events. The use of discourse analysis is motivated by the observation that information expressing relations between events is spread across multiple sentences of a passage [2]. The inference of relations between events does not use any sophisticated reasoning methodology, but is based on the order of events. For example, such rule can infer the relation that a certain action,

e.g., a shot, caused a goal, if this action precedes the event of goal. So, questions like "How many goals did PlayerX score in Champions League" can be answered.

In [25], a machine learning approach is followed for the information extraction task. Using the domain ontology the system deals with the instantiation of the ontological concepts, as well as the identification of relations between them into a corpus. The inference engine of this system is built using F-Logic [22],and allows the description of the ontological scheme and the instantiation of it. For example, the fact that the concept "Building" is a sub-concept of "Accommodation" is denoted as "Building::Accommodation". Continuing with the same example, "Accommodation" is related with concept "Location", i.e., "Accommodation[InArea => Location]". In this manner, an instance of "Building" is defined as "hilton:Building[InArea => athens]". Inference rules are predefined and given as input to the system. Such rules provide relation inference, by stating that a relation is, for example, transitive. In order to illustrate this, assume that in the ontology it is defined that a team of engineers designs a building, and also that this team includes civil engineers. According to the transitive nature of the rules, a direct relation is inferred, associating the concepts of "Building" and "CivilEngineer", denoting that the team of civil engineers had a contribution to the building design. By adding such inferred facts into ontology the information extraction process is bootstrapped. Note that, human subjects review the inference results.

In BOEMIE system [7], the core idea is the bootstrapping of ontology evolution in the framework of ontology-based information extraction. The extraction process is layered, having in the first layer the identification of ontological concepts and their relations that can be attributed to text segments. In the next layer more composite concepts of higher level and relations among them are generated, based on the previously extracted concepts, using an inference mechanism. In contrast to the lower-level concepts of the first level, the higher-level concepts usually cannot be mapped to textual fragments. For example, assume that the instances referring to an athlete, a sport and a tournament were extracted in the first layer. The inference mechanism relates the instantiations of these concepts, generating a higher-level concept according to the domain ontology. The ontology evolution task of BOEMIE system can be roughly distinguished into two branches: ontology population and ontology enrichment. The procedure of ontology population adds new individual entities to the ontology by accounting disambiguation and consistency maintenance issues. The domain knowledge is extended by the addition of new concepts and relations that are obtained through the process of ontology enrichment.

The description of the above systems highlights the different major uses of domain model in the framework of ontology-based information extraction. These differences can be considered from the perspectives of information extraction and ontology usage. However there are cases where these perspectives cannot be studied in complete discrimination [7, 25]. This happens because information extraction and ontologies participate into an iterative procedure [27, 28]. In particular, ontological knowledge is used for information extraction, while the latter enhances the former. In the case of [5, 6] the domain knowledge is exploited through simple, but efficient inference heuristics, which are focused on the task of question answering. These heuristics take advantage of event ordering that is in correspondence with the domain of interest in which this ordering has a meaningful role. That is, a soccer game can be viewed as a

sequence of events, whose appropriate ordering can draw domain interpretations. More sophisticated inference mechanisms are used in the case of SMES and BOEMIE systems [7, 25]. In the case of SMES system [25], a logic-based language, F-Logic, is used to infer relations among the extracted facts. The inference approach in the case of BOEMIE system seems to have a greater ability for generalization, since it is applied during a higher extraction level [7]. As it was mentioned in the previous paragraph the information extraction process in BOEMIE system consists of two phases, identifying low and high-level concepts (and their relations), respectively. In both systems, SMES and BOEMIE, consistency checking is taken into consideration by the inference engines. Regarding the information extraction techniques, the SOBA system uses a rule-based approach to populate the ontology. SMES system annotates a training corpus and machine learning methods are used for bootstrapping, while in BOEMIE both rule-based and machine learning techniques are applied. SMES and BOEMIE systems perform ontology evolution that is used during the information extraction process, following the bootstrapping way. In this framework two main phases are involved at operational level before the enhancement of ontology, information extraction and inference. Although these two phases are tightly connected, they have different design requirements. As we saw, information extraction techniques vary from rule-based to machine learning and hybrid approaches. For all these options well-studied cases are reported to the literature, while mature preprocessing tools are available. On the other hand the development of inference mechanism for such systems is strongly dependent to the specific characteristics of the domain of interest.

Table 4. Representative IE systems of the 4th category

System	Input, Preprocessing	IE approach	Use of ontology	Ontology features	Output
SOBA [5, 6]	- Web pages - Tokenization, POS tagging, morphological analysis, NER	Grammar, rule-based	Rule development	RDF F-logic	Ontology population
SMES [25]	- Raw text, web pages - Tokenization, morphological analysis, POS tagging, NER, chunk parsing	Machine learning	-Instantiation of concepts and their relations -Inference	RDF F-logic	Ontology enrichment
BOEMIE [7]	- Web pages - Tokenization, identification of document regions, sentence splitting, POS tagging, stemming, chunking, NER	- Rule-based - Machine learning	-Instantiation of concepts and their relations - Inference	OWL Description Logics	Ontology enrichment

This demands more elaborative design (in contrast to the phase of information extraction) because the available facilities are of general purpose, thus probably requiring some tuning effort. Regarding evaluation process, systems proposed in [5, 6, 7] use the measurements of precision and recall, as almost all the systems of the previous categories. The following table summarizes the major characteristics of the discussed systems.

It is obvious that the systems of this category, as well as systems of the previous categories, share many steps of preprocessing. The described systems apply their information extraction techniques using the web as a textual resource. The exploitation of web pages requires the manipulation of textual information of different structural types, e.g., typical text segments, tabular text, and image captions. An interesting example of exploiting text of such heterogeneous types is proposed by SOBA system [5, 6]. Regarding text modality, the BOEMIE system [7], apart from ordinary text segments, uses also image captions. Furthermore, it is important to underline that BOEMIE system incorporates two additional sources of different modality for gathering textual information, i.e., video OCR and speech recognition results. The ontological knowledge is used for the mapping of textual segments to the ontological scheme that favors machine learning approaches [7, 25], since the annotated corpus created can be used for training the machine learning algorithms. The usefulness of this knowledge is also true for rule-based systems, in which ontological entities and relations serve as the fundamental development units [5, 6, 7]. Doubtless, the ontological knowledge becomes of greater importance in the case of inference-based systems, where a deep understanding and representation of the domain special characteristics is needed. For the systems of this category we observe that standardized ontology formalisms are used, such as RDF and OWL. This is also true for the inference mechanisms, in which the approaches of F-logic and Descriptions Logics are followed. The gain by establishing an inference mechanism is the ontology enrichment, which in turn enhances the process of information extraction [7, 25]. Since the development of a very large set of inference rules engine for an open-domain is not realistic, the discussed systems are reported with respect to specific domains. However, we estimate that the employed information extraction techniques and general inference principles can be ported to new domains. This is justified by the fact that, the employed information extraction techniques appear to work synergistically with general inference principles in a closed loop [7, 25]. This decreases the manual effort for porting to new domains, compared to other approaches where the information extraction is not bootstrapped by the inference results. This happens because they provide several levels of extraction during a cyclic process, which can be controlled according to the domain characteristics. Such approaches can use a mixture of rule-based and machine learning techniques that can reduce the manual effort during the extraction task. The rules can serve as seeds that the system can learn and generalize. Of course, the development of an inference mechanism is strongly dependent to expert knowledge; however, within the bootstrapping approach enhances the extraction step of the next cycle. Moreover, this cyclic approach can have a bi-mode operation, ranging from semi-automatic up to full automatic functionality, where in the semi-automatic mode an expert can validate the intermediate results of each cycle, and also, he can select the appropriate levels of extraction.

4 Conclusions

Ontology-based information extraction employs ontologies as a means to describe formally the domain knowledge exploited by an information extraction system for its operation. The aim of this survey is to study the contribution of ontologies to information extraction systems.

From this study, it seems that the majority of information extraction approaches follow similar pre-processing steps. They are rather application-oriented ones, although they have functionalities that enable domain adaptation, since the goal is to develop IE systems that get better evaluation results in specific application areas. This is probably justified by the influence of the IE evaluation conferences during the 90's, the so-called Message Understanding Conferences (MUCs), which established a decomposition of IE into standard processing steps, as well as by the influence of machine learning techniques which facilitated experimenting and porting to new application areas. It is interesting to notice that the pre-processing tasks are now more mature, establishing, thus an infrastructure upon which new techniques with stronger involvement of knowledge models (i.e., ontologies) can be exploited.

Also, as the ontological use is increased, more standardized formalisms for ontology representation and inference are followed, in contrast to the majority of systems of the first category in which the ontological representation has the poor form, for example, of a gazetteer.

The use of different ontological aspects during the information extraction process forms a range of systems with different capabilities. These capabilities range from concept instantiation and relation extraction to inference of new concepts and relations. As the output becomes more sophisticated, more complex ontological relationships and deeper understanding of the domain model is required. We believe that the systems belonging to the last category i.e. which exploit the domain model in a framework of inference, give a unified approach regarding the use of ontologies in information extraction. Making domain knowledge explicit through an ontology, it does not only enhance portability, but it also provides new opportunities for information extraction systems, ranging from using the ontology for storing the extracted information to using reasoning for implementing various IE tasks. We consider BOEMIE as the most representative example of this case, since text IE in BOEMIE maintains the traditional NERC and coreference steps, whose results are used to populate an ontology, and substitutes all the template-related steps with reasoning over this ontology, driven by a set of inference rules stored explicitly, along with the ontology. Furthermore, the fact that domain knowledge in an ontology-based information extraction system is explicitly described by an ontology, allows the adaptation of the system's behaviour through changes in its ontology, usually in a synergistic approach where extracted information is used to enhance the ontology, which in return affects the performance of the system, as is the case in BOEMIE's bootstrapping approach.

References

[1] Appelt, D.E., Hobbs, J.R., Bear, J., Israel, D.J., Tyson, M.: Fastus: A finite-state processor for information extraction from real-world text. In: IJCAI, pp. 1172–1178 (1993)

[2] Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press, Cambridge (2003)

[3] Basili, R., Moschitti, A., Pazienza, M.T., Zanzotto, F.M.: Personalizing Web Publishing via Information Extraction. IEEE Intelligent Systems and Their Applications 18(1), 62–70 (2003)
[4] Bikel, D., Miller, S., Schwartz, R., Weischedel, R.: Nymble: A High-Performance Learning NameFinder. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 194–201. Morgan Kaufmann, CA (1997)
[5] Buitelaar, P., Cimiano, P., Racioppa, S., Siegel, M.: Ontology-based Information Extraction with SOBA. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 2321–2324 (2006)
[6] Buitelaar, P., Cimiano, P., Frank, A., Hartung, M., Racioppa, S.: Ontology-based Information Extraction and Integration from Heterogeneous Data Sources. International Journal of Human Computer Studies (JHCS) 66, 759–788 (2008)
[7] Castano, S., Peraldi, I.S.E., Ferrara, A., Karkaletsis, V., Kaya, A., Möller, R., Montanelli, S., Petasis, G., Wessel, M.: Multimedia Interpretation for Dynamic Ontology Evolution. Journal of Logic and Computation (2008)
[8] Cimiano, P., Handschuh, S., Staab, S.: Towards the Self Annotating Web. In: Proceedings of the 13th World Wide Web Conference (2004)
[9] Cimiano, P., Ladwig, G., Staab, S.: Gimme The Context: Context driven Automatic Semantic Annotation with CPANKOW. In: Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, pp. 332–341 (2005)
[10] Ciravegna, F., Lavelli, A.: LearningPinocchio: Adaptive Information Extraction for Real World Applications. Natural Language Engineering 1(1), 1–21 (2003)
[11] Croft, W., Cruse, D.A.: Cognitive Linguistics. Cambridge University Press, Cambridge (2004)
[12] Defence Advanced Research Project Agency: Proc. of the 6th Message Understanding Conference. Morgan Kaufmann, San Francisco (1995)
[13] Defence Advanced Research Project Agency: Proc. of the 7th Message Understanding Conference, http://www.muc.saic.com/proceedings/muc_7_toc.html
[14] Fellbaum, C.: WordNet, an electronic lexical database. MIT Press, Cambridge (1998)
[15] Grishman, R.: Information Extraction: Techniques and Challenges. In: Pazienza, M.T. (ed.) SCIE 1997. LNCS, vol. 1299, pp. 10–27. Springer, Heidelberg (1997)
[16] Grishman, R.: Information Extraction. In: Mitkov, R. (ed.) Handbook of Computational Linguistics Information Extraction. Oxford University Press, Oxford (2003)
[17] Hahn, U., Marko, K.G.: Ontology and Lexicon Evolution by Text Understanding. In: Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT 2002), Lyon, France (2002)
[18] Hahn, U., Romacker, M., Schulz, S.: Creating Knowledge Repositories from Biomedical Repots: MEDSYNDIKATE Text Mining System. In: Proceedings PSB 2002, pp. 338–349 (2002)
[19] Hearst, M.A.: Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of COLING 1992, pp. 539 – 545 (1992)
[20] Hobbs, J.R., Stickel, M., Appelt, D., Martin, P.: Interpretation as Abduction, Technical Note 499, AI Center, SRI International (1990)
[21] Karkaletsis, V., Spyropoulos, C.D., Grover, C., Pazienza, M.T., Coch, J., Souflis, D.: A Platform for Cross-lingual, Domain and User Adaptive Web Information Extraction. In: Proceedings of the European Conference in Artificial Intelligence (ECAI), Valencia, Spain, pp. 725–729 (2004)
[22] Kifer, M., Lausen, G., Wu, J.: Logical Foundations of Object Oriented and Frame Based Languages. Journal of the ACM (1995)

[23] Lehnert, W., Cardie, C., Fisher, D., Riloff, E., Williams, R.: University of Massachusetts: Description of the CIRCUS system as used for MUC-3. In: Proceedings of the Third Message Understanding Conference. Morgan Kaufmann, CA (1991)

[24] Ma, L., Shepherd, J.: Information Extraction Using Two-Phase Pattern Discovery. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, pp. 534–535 (2004)

[25] Maedche, A., Neumann, G., Staab, S.: Bootstrapping an Ontology-Based Information Extraction System. In: Szczepaniak, P.S., Segovia, J., Kacprzyk, J., Zadeh, L.A. (eds.) Intelligent Exploration of the Web Series. Studies in Fuzziness and Soft Computing. Springer, Heidelberg (2002)

[26] Mikheev, A., Grover, C., Moens, M.: Description of the LTG system used for MUC-7 (1998), http://muc.saic.com/proceedings/muc_7_toc.html (last visited October 1999)

[27] Nédellec, C., Nazarenko, A.: Ontologies and Information Extraction (2005), http://arxiv.org/abs/cs.AI/0609137

[28] Nédellec, C., Nazarenko, A., Bossy, R.: Information Extraction. In: Staab, S., Studer, R. (eds.) To appear in Ontology Handbook. Springer, Heidelberg (2008)

[29] Niepert, M., Buckner, C., Allen, C.: A Dynamic Ontology for a Dynamic Reference Work. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2007) (2007)

[30] Rosendfeld, B., Feldman, R., Fresko, M.: TEG—A Hybrid Approach to Information Extraction. Knowledge Information Systems 9(1), 1–18 (2005)

[31] Soderland, S.: Learning text analysis rules for domain-specific natural language processing, PhD thesis. Amherst: University of Massachusetts, Department of Computer Science (1997)

[32] Wang, T., Bontcheva, K., Li, Y., Cunningham, H.: D2.1.2 / Ontology-Based Information Extraction (OBIE) v.2, EU-IST Project IST-2003-506826 SEKT SEKT: Semantically Enabled Knowledge Technologies (2005)

Logical Formalization of Multimedia Interpretation

Sofia Espinosa, Atila Kaya, and Ralf Möller

School of Electrical Engineering and Information Technology,
Hamburg University of Technology, Hamburg, Germany
{sofia.espinosa,at.kaya,r.f.moeller}@tu-harburg.de

Abstract. Nowadays, many documents in local repositories as well as in resources on the web are multimedia documents that contain not only textual but also visual and auditory information. Despite this fact, retrieval techniques that rely only on information from textual sources are still widely used due to the success of current text indexing technology. However, to increase precision and recall of multimedia retrieval, the exploitation of information from all modalities is indispensable, and high-level descriptions of multimedia content are required. These symbolic descriptions, also called deep-level semantic annotations, play a crucial role in facilitating expressive multimedia retrieval. Even for text-based retrieval systems, deep-level descriptions of content are useful (see, e.g., [7]).

1 Introduction

There is a general consensus that manual annotation of multimedia documents is a tedious and expensive task which must be automated in order to obtain annotations for large document repositories. *Multimedia interpretation* is defined here as the process of producing deep-level semantic annotations based on low-level media analysis processes and domain-specific conceptual data models with formal, logical semantics.

The primary goal of this chapter is to present logical formalizations of interpretation. The chapter presents pioneering work on logic-based scene interpretation that has a strong influence on multimedia interpretation. Early approaches are discussed in more detail to analyze the main reasoning techniques. More recent approaches, which are more formal and therefore harder to understand, are referred to by providing references to the literature such that the reader can get an overview over the research field of logic-based media interpretation.

The discussion about scene interpretation is complemented with a presentation of logical approaches to text interpretation. Logical representations for deep-level video interpretation are discussed afterwards. The main goal of the chapter is to investigate the role of logic in the interpretation process. In order to focus on this goal, we neglect probabilistic approaches to this topic (but we give some pointers to the literature).

Logic-based media interpretation builds on initial symbolic descriptions of media content. In the next section, we argue that it is reasonable to expect

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 110–133, 2011.

so-called *multimedia analysis* processes to be able to reliably produce description about information that is, more or less, directly observable.

2 Prerequisites for Interpretation: Media Analysis

The identification of directly observable information in different modalities, also called *surface-level information*, has been studied in the past for at least three decades. In natural language processing, *information extraction* is one of the major tasks that aims to automatically extract structured information such as named entities and certain relations between entities. Evaluations have shown that state-of-the-art information extraction systems are very powerful language analysis tools that can recognize names and noun groups with an accuracy higher than 90% [12]. Different systems exploit various machine-learning techniques such as k-nearest neighbors or Hidden Markov Models. They have been successfully used for solving real-world problems [2]. However, information extraction is a more restricted problem than general language understanding, and language analysis techniques employed in these systems provides for simple, reliable symbolic content descriptions but are not as powerful as full syntactic language analysis. A state of the art system for text analysis is OpenCalais (`http://www.opencalais.com`), which returns its results as annotations to a text in a logic-based language. However, when it comes to extracting more abstract information such as events that require a deep understanding of the domain, information extraction systems are reported not to perform well in general [16].

In computer vision, *object recognition* aims to find objects in images (scenes) or image sequences (videos). Even though object recognition has been successfully applied in specific domains, e.g., for finding faces in images [55], general object recognition is still an unsolved problem. In many approaches, object recognition follows segmentation, where images are partitioned into homogeneous regions, i.e. sets of pixels. The pixels in a region are similar w.r.t. some feature such as color, intensity or texture [53]. The derivation of homogeneous regions is supported by techniques such as color histograms or shape analysis. However, when used without further knowledge resources, these "global" techniques are not appropriate for general-purpose object recognition in images [25]. Therefore, a wide range of local descriptors, such as Harris corners [19], Shape Context [8] and Scale Invariant Transform (SIFT) [29], have been proposed. Nowadays, local descriptors are successfully used for solving practical problems. For example, SIFT has been applied to the problem of robot localization in unknown environments in robotics [51]. Mikolajczyk and Schmid present a comprehensive evaluation of various local descriptors [32]. We would like to point out that logic-based representations have also been used at the analysis level (maybe in combination with probabilistic or fuzzy representations such as, e.g., in [49]).

Recently, Leibe and Schiele presented an approach that considers object recognition and segmentation as intertwined processes and uses top-down knowledge for guiding the segmentation process [28]. The authors reported on experimental results that show the capacity of the approach to categorize and segment

diverse categories such as cars and cows. As a result, even though object and event recognition in the general domain is beyond the capabilities of current technology [26], the identification of observable information in image and video sequences in specific domains can indeed be achieved with state-of-the-art computer vision systems. Information extraction from text and the field of computer vision are related research fields providing the input required for the interpretation process.

Thus we can reasonably assume that the above-mentioned analysis processes can compute symbolic descriptions of media content, and make such descriptions available as input to multimedia interpretation processes. It is also very well possible that media analysis can be influenced by media interpretation. But for the time being we consider analysis and interpretation as sequential steps. In any case, the discussion reveals that recent advances in media analysis provide for a solid foundation to the derivation of deep-level abstract content descriptions based on a logical representation language.

3 Logic-Based Scene Interpretation

In this section we present related work on scene interpretation that has a strong influence on the design of multimedia interpretation processes. In fact, the multimedia interpretation problem, for which also modalities beyond images are relevant, can be considered as a generalization of scene interpretation. Although there exist a substantial number of approaches to high-level scene interpretation in the literature, unfortunately, many of them are not built on representation languages with a formal semantics. In this section we focus on approaches that exploit formal, declarative representations for scene interpretation and that have been implemented as software systems. Our goal is not only to cite relevant work on scene interpretation but also to identify key problems in scene interpretation. We expect the reader to be familiar with first-order logic and, to some extent, with logic programming as well as description logic (see pointers to the literature in the text).

3.1 Scene Interpretation Based on Model Construction

The first formal theory of scene interpretation based on logics was introduced by Reiter and Mackworth [45]. They propose a so-called theory of depiction and interpretation that formalizes image-domain knowledge, scene-domain knowledge and a mapping between the image and scene domains using first-order logic [46]. An interpretation of an image is then defined as a logical model of a set of logical formulae which formalize background knowledge as well as the output of low-level scene analysis processes.

We shortly discuss the main ideas of the approach in [46], and we recapitulate the system *Mapsee*, which has been implemented for the interpretation of hand-drawn sketch maps of geographical regions [34]. Given a sketch map consisting of chains[1], regions and various relations between them, the goal of the system

[1] Chain is the term used in the original paper for polylines.

is to compute an interpretation in terms of roads, rivers, shores, areas of land, and areas of water, etc.

The image-domain knowledge includes general knowledge about maps such as the taxonomy of image-domain objects, which are specified through first-order logic axioms:

$$\forall x: \ \mathit{image\text{-}object}(x) \ \Leftrightarrow \ \mathit{chain}(x) \vee \mathit{region}(x)$$
$$\forall x: \ \neg(\mathit{chain}(x) \wedge \mathit{region}(x))$$

The first axiom states that chains and regions, so-called image primitives, are the only objects that can exist in a map, whereas the latter axiom states that an object cannot be both chain and region at the same time (disjointness of image primitives). Relations between image-domain objects are also part of the image-domain knowledge and are specified using predicates such as $\mathit{tee}(c, c')$, $\mathit{bound}(c, r)$ etc. For example, the predicate $\mathit{tee}(c, c')$ means that chain c meets chain c′ at a T-junction.

The approach assumes a map description to consist of finitely many chains and regions together with finitely many relations between the chains and regions. Therefore, the system makes the *domain closure assumption* by postulating that all map objects are completely known. To this end, closure axioms of the following form are used (i_m and i'_n are constants):

$$\forall x: \ \mathit{chain}(x) \ \Leftrightarrow \ x \ = \ i_1 \vee \cdots \vee x \ = i_m$$
$$\forall x: \ \mathit{region}(x) \ \Leftrightarrow \ x \ = \ i'_1 \vee \cdots \ \vee x \ = i'_n$$
$$\forall x, y: \ \mathit{tee}(x, y) \ \Leftrightarrow \ (x \ = \ i_1 \wedge y \ = \ i'_1) \vee \cdots \vee \ (x \ = \ i_k \wedge y \ = \ i'_k)$$
$$\cdots$$

Furthermore, the system makes the *unique name assumption* by assuming that all constants (e.g., image primitives such as chains and regions) denote different objects. Scene-domain knowledge is represented by axioms for objects such as roads, rivers, shores, or land and water areas. For instance, the following equivalence, coverage and disjointness axioms are used.

$$\forall x: \ \mathit{scene\text{-}object}(x) \ \Leftrightarrow \ \mathit{linear\text{-}scene\text{-}object}(x) \vee \mathit{area}(x)$$
$$\forall x: \ \mathit{linear\text{-}scene\text{-}object}(x) \ \Leftrightarrow \ \mathit{road}(x) \vee \mathit{river}(x), \vee \mathit{shore}(x)$$
$$\forall x: \ \neg(\mathit{road}(x) \wedge \mathit{river}(x))$$
$$\forall x: \ \neg(\mathit{linear\text{-}scene\text{-}object}(x) \wedge \mathit{area}(x)) \ \ldots$$

In addition, the scene-domain knowledge base contains also specific restrictions such as, for instance, rivers do not cross each other:

$$\forall x, y: \ \mathit{river}(x) \wedge \mathit{river}(y) \Rightarrow \ \neg \ \mathit{cross}(x, y)$$

Axioms that restrict the domain and range of relations to scene objects only are also used:

$$\forall x, y: \ \mathit{cross}(x, y) \Rightarrow \ \mathit{scene\text{-}object}(x) \ \wedge \ \mathit{scene\text{-}object}(y)$$

Besides the specification of intra image- and scene-domain axioms, also inter-domain axioms between the image and scene domain are specified (so called

mapping axioms). The mapping axioms are represented using the binary predicate $\Delta(i, s)$ meaning that image object i depicts scene object s. The depiction relation only holds between image and scene objects:

$$\forall i, s: \ \Delta(i, s) \Rightarrow \mathit{image\text{-}object}(i) \wedge \mathit{scene\text{-}object}(s)$$

For specifying image-scene-domain mappings, closure and disjointness axioms are provided.

$$\forall x: \ \mathit{image\text{-}object}(x) \vee \mathit{scene\text{-}object}(x)$$
$$\forall x: \ \neg(\mathit{image\text{-}object}(x) \wedge \mathit{scene\text{-}object}(x))$$

Furthermore, it is assumed that every image object i depicts a unique scene object, which is denoted by $\sigma(i)$:

$$\forall i: \ \mathit{image\text{-}object}(i) \Rightarrow \mathit{scene\text{-}object}(\sigma(i)) \wedge \Delta(i, \sigma(i)) \wedge [\forall s: \Delta(i, s) \Rightarrow s = \sigma(i)]$$

and every scene object is depicted by a unique image object:

$$\forall s: \ \mathit{scene\text{-}object}(s) \Rightarrow (\exists_i^1: \ \mathit{image\text{-}object}(i) \wedge \Delta(i, s))$$

The notation $\exists_i^1 : \alpha(x)$ means that there exists exactly one x for which $\alpha(x)$ holds. Finally, mappings between the image- and scene-objects

$$\forall i, s: \ \Delta(i, s) \wedge \mathit{region}(i) \Rightarrow \mathit{area}(s)$$
$$\forall i, s: \ \Delta(i, s) \wedge \mathit{chain}(i) \Rightarrow \mathit{linear\text{-}scene\text{-}object}(s)$$

and mappings between relations of the image and scene domains are specified:

$$\forall i_1, i_2, s_1, s_2: \ \Delta(i_1, s_1) \wedge \Delta(i_2, s_2) \Rightarrow \mathit{tee}(i_1, i_2) \Leftrightarrow \mathit{joins}(s_1, s_2)$$
$$\forall i_1, i_2, s_1, s_2: \ \Delta(i_1, s_1) \wedge \Delta(i_2, s_2) \Rightarrow \mathit{chi}(i_1, i_2) \Leftrightarrow \mathit{cross}(s_1, s_2)$$
$$\ldots$$

The above-mentioned axioms state that *tee*[2] relations in the image depict *joins* relations in the scene and vice versa, whereas *chi*[3] relations in the image depict *cross* relations in the scene.

Given the specification of all relevant image-domain axioms, scene-domain axioms and mapping axioms, Reiter and Mackworth define an *interpretation* of an image, specified as set of logical facts, as a logical model of these facts w.r.t. the axioms in the knowledge base.

The main problem here is that, in principle, a set of first-order formulas may have infinitely many models, which in turn might be infinite, and, therefore, the computation of all models may become impossible. Even worse, it is undecidable in general whether a set of first-order formulas has a model at all. However, Reiter and Mackworth show that as a consequence of the assumptions made in their logical framework, it is possible to enumerate all models. In fact, under the additional closed-world assumption, finite extensions of all predicates can be used

[2] Shorthand for T-junction.
[3] Shorthand for X-junction.

in the models, and therefore quantified formulas can be replaced with quantifier-free formulas. Consequently, first-order formulas can be reduced to propositional formulas, for which the computation of all models is possible [14]. Reiter and Mackworth formulate the problem of determining all models of the resulting propositional formulas as a *constraint satisfaction problem* (CSP). Although, in general, CSPs of this kind are NP-hard, and thus computationally intractable, several efficient approximation algorithms exist, which have also been used in the Mapsee system [34].

Reiter and Mackworth also show that for the computation of the models using CSP algorithms, only scene axioms are relevant and all other axioms can be ignored. This gives rise to the question whether the distinction between image- and scene-domain knowledge is necessary. This distinction makes the formal specification more involved, but at the same time, allows for a separate representation of general knowledge about the image and scene domains. However, in the first-order logical framework it is not possible to check for the consistency of general knowledge bases, for which no domain-closure axioms can be specified. Furthermore, the logical framework presumes the unambiguous acquisition of image objects, scene objects and their relations, as well as the depiction relations such that unique specifications can be obtained. These assumptions are obviously too strict for general purpose scene interpretation and largely neglect issues such as noise and incompleteness (see also the discussion in [46]). Therefore, in [43] Poole, the exploitation of probabilistic knowledge is studied using the Mapsee scenario.

Schröder [50] criticizes that representing interpretation results in terms of logical models (as done in the Mapsee approach) yield interpretations that might be too specific, which, in turn, might cause an over-interpretation of observations. He suggests the notion of a partial model [50], a relational structure detailed enough to represent the commonalities between all models.

3.2 Scene Interpretation Based on Abduction

Inspired by the work of Reiter and Mackworth, Matsuyama and Hwang present a vision system called SIGMA, in which they apply logic to scene interpretation [31]. In contrast to Reiter and Mackworth, they do not assume the availability of an a priori image segmentation, and do not make domain-closure and unique-name assumptions for the image domain. Constant symbols representing image-domain objects are not available in the beginning, but have to be created through an expectation-driven segmentation approach, which is part of the interpretation process. Consequently, also constant symbols representing scene objects are not available in the beginning of the interpretation process and have to be computed through hypotheses. This is why Matsuyama and Hwang call their approach constructive, and we will argue that nowadays it would have been called *abductive*.

Matsuyama and Hwang use aerial images of suburban areas that typically show houses and roads. First-order logic axioms are used to represent general knowledge about the application domain. For example, the fact that every house

is related to exactly one street is represented as follows (for the sake of the example the relation is called *rel*)

$$\forall x : house(x) \Rightarrow (\exists y : \ road(y) \wedge rel(x, y) \wedge \forall z : (road(z) \wedge rel(x, z)) \Rightarrow z = y)$$

This formula can be transformed into *clausal normal form* (with an implicit conjunction operator between the formulas on separate lines).

$$\neg house(x) \vee road(f(x))$$
$$\neg house(x) \vee rel(x, f(x))$$
$$\neg house(x) \vee \neg road(z) \vee \neg rel(x, z) \vee z = f(x)$$

Existential quantification is replaced with terms using so-called *Skolem functions*. A Skolem term replaces an existentially quantified variable and denotes a certain domain object, depending on the universally quantified variable in whose scope the replaced existentially quantified variable is located. As an example, assume an aerial image depicting a house. The house is represented by the constant h_1. Given the above-mentioned axioms representing the general knowledge about the domain and information about the existence of a house in the scene, namely $house(h_1)$, the following information is entailed:

$$road(f(h_1))$$
$$rel(h_1, f(h_1))$$
$$\neg road(z) \vee \neg rel(h_1, z) \vee z = f(h_1)$$

Here, the new domain object $f(h_1)$ denoted using the Skolem term f is called an *expected object*, in this example a road, and has to be identified in the image.

In the SIGMA system, different classes of scene objects and spatial relations are defined through necessary conditions.

$$\forall x : road(x) \Rightarrow greater(width(x), 5) \wedge less(width(x), 100) \wedge ribbon(shape(x))$$
$$\forall x, y : rel(x, y) \Rightarrow parallel(axis(x), axis(y)) \wedge distance(center(x), center(y), 50)$$

Object attributes such as *width*, *shape*, *axis* or *center* are modeled through functions, predicates regarding spatial attributes such as *greater*, *less*, *ribbon*, *parallel* or distance are modeled as constraints. These axioms define conditions that must hold for objects of the scene domain, and thus can eliminate certain models.

Assume that our sample image depicts also a road. Then, the road is represented in the scene domain as well, e.g. by the constant r_1. After adding a new axiom to represent this information, namely $road(r_1)$, the following information is entailed:

$$\neg rel(h_1, r_1) \vee r_1 = f(h_1)$$

In the SIGMA system, spatial relations of the image domain are not mapped to relations whose domain and range are the scene domain. In addition, for spatial relations of the scene domain such as *rel* only necessary conditions are defined but not sufficient ones. Therefore, it cannot be proved logically, whether $rel(h_1, r_1)$ holds or not. To solve this problem, a special equality predicate is used in SIGMA, which reflects two important assumptions about equality of objects:

i) Two scene objects are considered to be identical, if they are of the same type, e.g. road, and have the same shape and position, i.e. occupy the same space. ii) If an existing object in the scene domain fulfills all conditions that an expected object has to fulfill, both objects are considered to be identical.

In our example, if r_1 fulfills all conditions that have to be fulfilled by the expected object $f(h_1)$ then as a result of the equality assumption, the *hypothesis* $r_1 = f(h_1)$ is generated, and later $rel(h_1, r_1)$ is derived. In case no suitable scene object that is identical to the expected object $f(h_1)$ exists, the conditions of the expected object $f(h_1)$ are used for an expectation-driven image analysis process to identify an object in the image. In case an object is identified, a new constant symbol is introduced in the image domain, e.g. r_2, and the hypothesis $road(r_2)$ is created. Afterwards, the hypothesis $r_2 = f(h_1)$ is generated and $rel(h_1, r_2)$ is derived.

In order to guarantee termination, expected objects are not allowed to trigger the derivation of new expected objects, e.g. $g(f(r_1))$. In other words, expectations are not used to derive further expectations. Expectation generation is done solely through the exploitation of constant symbols, which can only be introduced by an expectation-driven image analysis process.

The hypothesis generation process in SIGMA computes so-called *interpretation networks*, i.e., networks consisting of mutually related object instances. Multiple interpretation networks can possibly be constructed for an image. In an interpretation network, multiple objects instances may be located in the same place in the scene. Such instances are called *conflicting instances*, and a so-called *in-conflict-with* relation is established between them. It should be noted that the SIGMA system applies no heuristics to select among the possible sets of networks but delivers the first computed set of networks as the result.

In [31], Matsuyama and Hwang not only present the general approach followed in the SIGMA system, but also discuss the computation of scene interpretations. According to the authors the goal of scene interpretation is to provide for an explanation of the observations, i.e. of the images, through the exploitation of axiomatize general knowledge about the world. They observe that the computation of scene interpretation cannot be achieved through deductive reasoning only: $axioms \not\models observations$.

The axioms representing general knowledge in terms of universally quantified formulas do not entail concrete observations (facts). Instead of a deductive reasoning approach, Matsuyama and Hwang follow the *hypothetical reasoning* approach of Poole et al. [41] where the task is to compute a set of logical hypotheses such that following conditions are fulfilled:

i) $Axioms \cup Logical\_Hypotheses \models Observations$
ii) SAT($Axioms \cup Logical\_Hypotheses$)

Poole's work is the first in which the space of abducibles is declaratively specified. He uses Horn rules and a set of so-called assumables (aka abducibles) in order to specify which predicates are assumed to be true in a backward-chaining inference process over the Horn rules. The set of these hypotheses are returned as part of the result of the reasoning process (see also [43]). This form of

reasoning has been introduced by Peirce [40] under the name *abduction* in the late 19th century. Contrary to deduction where we can reason from causes to effects, in abduction we can reason 'backwards', i.e, from effects (observations) to causes (explanations). Abduction is also often defined as a reasoning process from evidence to explanation, which is a type of reasoning required in several situations where the available information is incomplete [1]. Abduction has been widely used to formalize explanation-based reasoning and plays an important role in intelligent problem solving tasks such as medical diagnosis [41] and plan recognition [11]. Formalizing the interpretation of camera data in a robotics scenario, Shanahan has also argued for an explanation-based (abductive) approach to scene interpretation [52]. As described in [52], logic is used for analyzing the behavior of specific procedural programs developed for scene interpretation.

Despite the fact that logic is a useful tool for analyzing (and describing) the behavior of computational systems, and despite the fact that the retrospective use of logic has its merits, nowadays logical reasoning systems have reached a state of maturity such that declarative reasoning services can be used to directly solve the interpretation problems in an abductive way. As has been said before, [43] uses Horn clauses for generating scene interpretations (in an abductive way) and exploits Bayesian networks for ranking alternatives. Recent developments of this significant theory, for which even a practical reasoner implementation exists, can be found in [44] and [42].

Another logical approach in which scene interpretation is realized by a practical reasoning engine for ontologies (which are, in some sense, more expressive than Horn clauses) is described in [13,10]. This approach has been extended in [15] in terms of control strategies and w.r.t. ranking explanation probabilities using Markov logic networks. [15] is the first approach in which the abduction process is systematically controlled by generating an explanation only if the agent can prove that the probability that the observations are true is substantially increased. This solves the termination problem in explanation generation inherent in early approaches such as, e.g., the one of Matsuyama and Huang.

Besides abduction, in [13] also deduction plays an important role. Something is abduced only if it cannot be proven to hold. We therefore analyze related work on scene interpretation based on deduction. The main question is whether the input (stemming from low-level scene analysis processes) can be made specific enough such that useful conclusions can be computed using deduction principles. Interestingly, somewhat contrary to common expectations, the main message here is, logical deduction is indeed able to compute important results w.r.t. scene interpretation based on sensible expectations w.r.t. analysis results.

3.3 Scene Interpretation Based on Deduction

First Approaches to Logic-based Interpretation using Deduction. The VEIL project (Vision Environment Integrating Loom) [47,48] is a research project that aims to improve computer vision programs by applying formal knowledge representation and reasoning technology. To this end a layered architecture integrating vision processing and knowledge representation has been proposed.

In this architecture a computer vision program operates at the pixel level using specialized data structures to deal with low-level processing, whereas the knowledge representation system Loom uses symbolic structures to represent and reason higher-level knowledge.

One of the major goals of VEIL is to enable the construction of explicit declarative vision models. This is achieved by exploiting the knowledge representation and reasoning facilities provided by the Loom system [30]. The Loom system provides support for an expressive knowledge representation language in the KL-ONE family and reasoning tasks. It supports not only deductive reasoning but provides also facilities to apply production rules. The declarative specification of knowledge offers various benefits: i) It is easier to maintain than a procedurally specified program. ii) It enables the application of automatic validation and verification techniques. iii) Data is represented in a high-level specification instead of application-specific data structures, and thus can easily be shared or reused by other applications.

Similar to the Mapsee and to SIGMA systems, also in the VEIL project domain knowledge is represented in two different models. The *site model* is a geometric model of concrete objects found in a particular image such as runways, markings, buildings and vehicles. The so-called *domain model* contains not only concrete objects such as roads, buildings, and vehicles but also abstract objects such as convoys (groups of vehicles) and field training exercise events.

The VEIL application scenario is the detection and analysis of aerial photographs of airports. Airports are modeled as collections of runways, which are long thin ribbons with markings (smaller ribbons) in certain locations. Aerial images are analyzed by the computer vision system through standard analysis techniques such as the Canny edge detector [9] to produce hypotheses. A sequence of filtering and grouping operations are then applied to reduce the number of hypotheses. In the next step, hypotheses are verified using the information in Loom's site model. For example, the site model describes markings in terms of their sizes, relative positions and position on the runway. The domain knowledge represented using Loom is used to constrain the set of possible hypotheses. For example, descriptions of the size and location of markings are used to rule out some hypotheses generated by the computer vision system. The generation of hypotheses, however, is not declaratively modeled. Logic-based deduction (consistency checking) is used to narrow down the space of possible hypotheses.

The work on VEIL shows that declarative representations and deduction as an inference service are useful for scene understanding, although the construction of the space of hypotheses for each scene is not done in terms of logical reasoning in VEIL but using a procedural program. In the VEIL project, deductive reasoning is employed to classify an instance as belonging to a concept. For example, assume that a group of pixels in an image is identified as a vehicle instance v_1 and added to the knowledge base. Further analysis of the same group of pixels might unveil that v_1 has tracks. Adding this information to the knowledge base, Loom classifies v_1 as a *tracked-vehicle* instance, where the concept

tracked-vehicle is defined as a subconcept of the concept vehicle. This is possible, because the concept tracked-vehicle is defined with necessary and sufficient conditions, which are all fulfilled by v_1. Note that instance classification has been used even before VEIL in the context of detecting visual constellations in diagrammatic languages (cf. [17,18]).

Ontology-based Interpretation. The exploitation of the ideas behind VEIL in the much more formal context of ontologies has been investigated by Hummel in [22,23]. In her work, Hummel describes a realistic scenario for logic-based traffic intersection interpretation. Based on a crossing model using carefully selected primitives, ambiguity is reduced by "integrating" cues in a logical framework. It is interesting to see how underspecified information derived by low-level analysis processes can be enriched using logical reasoning. In contrast to VEIL, which is based on incomplete reasoning, the work of Hummel uses a sound and complete reasoner and an expressive description language. Hummel found that soundness and completeness are mandatory in order to effectively reduce ambiguity such that (indefinite) cues from analysis processes are condensed to obtain useful interpretation results by deductive interpretation processes.

The overall goal of the system defined by Hummel is to facilitate autonomous car driving through the interpretation of road intersections. To this end, the system is provided as input with sensor data from a camera and a global positioning system (GPS) mounted on a vehicle, as well as with data from a digital map. For each road intersection the system is then requested to answer questions such as 'Which driving directions are allowed on each lane?', 'Which of the map's lanes is equivalent to the vehicle's ego lane?', etc. Answering such questions requires reasoning since general regulations of roads and intersections as well as partial and non-complementary information from various sensors about the current situation of the car have to be considered together.

In her work, Hummel investigates appropriate ways for representing relevant scene information in description logics (DLs). Being a decidable subset of first-order logic, DLs are are family of logical representation languages for which highly optimized reasoning systems exist. Terminological knowledge is formalized in terms of terminology (concepts and relations) in a so-called Tbox. Assertional knowledge about particular objects is described in a so-called Abox. For an introduction to DLs see [4].

For typical classes of scene information she proposes generic DL representations, which she refers to as design patterns. In particular, she presents design patterns for representing sensor data and qualitative scene geometry models in DLs. In the context of road intersection interpretation, different sensor setups are investigated as well. If a still image from a single sensor is interpreted, the unique-name assumption (UNA) should be imposed such that two individuals in the Abox are always interpreted (in the sense of first-order logic) as different objects. However if data is acquired by multiple, non-complementary sensors, objects are detected multiple times, and hence the UNA need not hold. For the multiple sensor setup, Hummel requires the UNA to hold within data acquired by a single sensor only, which she calls the local UNA. She reports the local UNA

to have been implemented as a procedural extension that enhances a knowledge base through the application of rules in a forward-chaining way.

Furthermore, Hummel investigates scene interpretation tasks with respect to their solvability through standard deductive DL inference services. These tasks are

1. Object detection, i.e., the discovery of new scene objects
2. Object classification, i.e., the assignment of labels to a detected object
3. Link prediction, i.e., predicting the existence and types of relationships between objects
4. Data association, i.e., the identification of a set of measurements as referring to the same object.

For her experiments Hummel develops a sophisticated Tbox for representing a road network ontology (RONNY), in which the qualitative geometry and building regulations of roads and intersections are specified. Building on these grounds, she describes a case study where the logic-enhanced system solves interpretation tasks using RONNY and sensor data from a stereo vision sensor, a global positioning system, and a digital map. The performance of the system in solving object detection, object classification and data association tasks has been evaluated on a sample set of 23 diverse and complex intersections from urban and non-urban roads in Germany.

She shows that in order solve the object classification task with standard DL inference services, the maximum number of individuals in a scene have to be added a priori to the Abox, which describes the scene. A corresponding design pattern has been proposed in [23]. In fact, if this design pattern is applied, the task of object detection can be reduced to the task of object classification, which can be solved using the so-called Abox realization DL inference service. In a nutshell, Abox realization is a deductive DL inference service that computes for all individuals in an Abox A their most-specific concept names w.r.t. a Tbox T. This way, in a sense, objects are "classified", and the classification determines in terms of symbols (names) what the systems knows about domain objects (see the previous subsection on VEIL).

In contrast to object detection and object classification, Hummel identified that the task of link prediction and data association cannot elegantly be solved using DLs.

In [23], it is shown that the system built through the integration of a deductive DL reasoner and a computer vision system can be used to significantly improve recognition rates of the computer vision system.

4 Logic-Based Text Interpretation

In a similar way as for scene interpretation, logic-based approaches have been used for text interpretation. In particular, the work of Hobbs et al. in [20,21] has been influential in conceptualizing text interpretation as a problem that requires

abduction in order to be solved. They developed a linguistic and knowledge-intensive framework to solve the problem of text interpretation starting from the derivation of the so-called logical form of a sentence, a first-order representation capturing its logical structure, together with the constraints that predicates impose on their arguments. The central idea of Hobbs et al. is to show that logical forms of (parts of) sentences can be established as consequences from background knowledge and additional assumptions (formulae to be added). The added formulae provide for a deeper interpretation.

As an example, consider the following sentence, on which a sequence of interpretation steps are applied.

(1) Disengaged compressor after lube-oil alarm.

The *reference resolution* step analyzes the words "compressor" and "alarm" and identifies them as so-called references. To establish the reference of compressor, the following logical form is generated for this part of the sentence

(2) $\exists x : compressor(x)$

Given a background knowledge base containing,

$starting\_air\_compressor(c_1)$
$\forall x : starting\_air\_compressor(x) \Rightarrow compressor(x)$

i.e., an instance of a "starting air compressor", namely c_1, and the definition of starting air compressor as a specific type of compressor, then the logical form (2) extracted from the sentence (1) can be resolved to the instance c_1, i.e.,

$compressor(c_1)$

is derived, and in this sense, the entailment of expression (2) is proved. In this case no additional assumptions are required.

When a reference formula cannot be proved to be entailed (w.r.t. the background knowledge), then it is assumed to be true. Here, we find the principle of abduction be applied. For example, "Lube-oil alarm" is a *compound nominal*, thus composed of two entities which are implicitly related to each other. The problem of determining the implicit relation between the two is called compound nominal resolution. To interpret "lube-oil alarm", a logical form is first extracted, namely

$\exists y, z, nn : lube\_oil(ys) \wedge alarm(z) \wedge nn(y, z),$

and, due to the principle explained above, w.r.t. the background knowledge, it should be possible to find one entity for lube-oil and another for alarm, and there must be some implicit relation (called nn) between them. If the entailment of the above formulae cannot be shown, assumptions are necessary (possible with Skolem terms, see above). Note, however, that assumptions need not be "least-specific". For instance, if the background knowledge contains information about the most common possible relations for an implicit relation, e.g. to denote part-whole relations,

(3) $\forall x, y : part(x, y) \Rightarrow nn(x, y)$

or complex relations that can be explained as a *for* relation,

(4) $\forall x, y : for(x, y) \Rightarrow nn(x, y)$

an assumption using *part* or *for* can in principle be made rather than use the more "abstract" relation directly.

As can be observed, there might exist more than one possibility to make assumptions. To choose between possible candidates, [21] defines a preference strategy to support this decision problem, called weighted abduction which will be explained below. We first continue with the example.

Deciding whether "after lube-oil alarm" modifies the compressor or the disengaging event is the problem of *syntactic ambiguity resolution.* To solve this problem, Hobbs et al. propose the transformation of the problem to a constrained coreference problem, where the first argument of the predicate is considered as existentially quantified. In this sense, the extracted logical expression is:

(5) $\exists e, c, y, a : after(y, a) \wedge y \in \{c, e\}$

where the existentially quantified variable y should be resolved to the compressor c or the disengaging event e. This problem is often solved as a by-product of metonymy resolution. *metonymy resolution* which involves the "coercion" of words such that the constraints that predicates impose on their arguments are fulfilled.

For example, in the above sentence (1), the predicate *after* requires events as arguments:

(6) $\forall e_1, e_2 : after(e_1, e_2) \Rightarrow event(e_1) \wedge event(e_2).$

Therefore, it is necessary to coerce the logical form in (5) such that the requirements of the predicate *after* are fulfilled. For this purpose, coercion variables satisfying the constraints are introduced:

$$(7)\ \exists k_1, k_2, rel_1, rel_2, y, a : after(k_1, k_2) \wedge event(k_1) \wedge rel_1(k_1, y) \wedge event(k_2) \wedge rel_2(k_2, a)$$

in this case k_1 and k_2 are the coercion variables related to *after* instead of y and a as it was before. Also coercion relations (rel_1, rel_2) are introduced. As can be seen from the example, coercion variables and relations are implicit information and are also generic, which suggests that any relation can hold between the implicit and the explicit arguments. If there are axioms in the background knowledge base, expressing the kind of "coercions" that are possible:

$\forall x, y : part(x, y) \Rightarrow rel(x, y)$
$\forall x, e : function(e, x) \Rightarrow rel(e, x)$

then, metonymy resolution is solved by abduction.

The next phase aims at computing the cost of the resulting interpretation. It is anticipated that during the process of proving the entailment of a logical form

(see above) different proofs can be found. In order to find the "less-expensive" proof Hobbs et al. developed a method called weighted abduction, which is characterized by the following three features: First, goal expressions should be assumable at varying costs, second it should be possible to make assumptions at various levels of specificity, and third, natural language redundancy should be exploited to yield more economic proofs. In this method, each atom of the resulting ungrounded logical form is weighted with a cost. For instance, in the formula

$$\exists e, x, c, k_1, k_2, y, a, o : Past(e)^{\$3} \wedge disengage'(e, x, c)^{\$3} \wedge compressor(c)^{\$5} \wedge after(k_1, k_2)^{\$3} \wedge event(k_1)^{\$10} \wedge rel(k_1, y)^{\$20} \wedge y \in \{c, e\} \wedge event(k_2)^{\$10} \wedge rel(k_2, a)^{\$20} \wedge alarm(a)^{\$5} \wedge nn(o, a)^{\$20} \wedge lube\_oil(o)^{\$5}$$

costs are indicated as superscripts with $ signs. Costs indicate different weights. An explanation is preferred if the costs of the things to assume are minimal.

The costs are given according to linguistic characteristics of the sentence, thus if the same sentences is expressed in a different way, the cost might vary accordingly. They have analyzed how likely it is that a linguistic expression conveys new information, and therefore failing to prove the entailment of the construct is not so costly, contrary to other expressions in which no new information is conveyed, and therefore it should be possible to prove the corresponding entailment. For example, the main verb is more likely to convey new information than a definite noun phrase which is generally used referentially. Failing to prove a definite noun phrase is therefore expensive. For a more detailed description of this linguistic characteristics, refer to [21].

Besides these weights, there are other factors used to determine the appropriateness of an interpretation, namely simplicity and consilience. A simple interpretation would be one that exploits redundancy in the discourse, such that the number of assumptions can be reduced, for example by assuming that two atoms are identical due to semantic knowledge. Consilience refers to the relation between the number of atoms that have been proved exploiting redundancy and the total number of atoms to prove. The highest the number of atoms that have been proved with the less number of assumptions, the more the explanation is consilient.

In their approach, less-specific explanations are preferred, due to the fact that the more specific the assumptions are, the more information can be obtained but it is also more likely that they are not correct. This is the so called informativeness-correctness trade-off. However, if there is evidence in the background knowledge that allows for a more specific assumption, then the more specific proof is considered.

As we have argued, the work of Hobbs et al. show us that logic-based interpretation can account for a large number of effects that naturally occur in text interpretation. We are now ready to study another modality, namely the video modality.

5 Logic-Based Video Interpretation

For video interpretation, various ontologies have been used. Whereas in some approaches time points are used (with time points being specified by quantitative numerical values), other approaches use time intervals and qualitative relations between them. What distinguishes the approaches is the level of declarativeness of how events to be recognized are specified.

5.1 Early Approaches

The beginnings of symbolic video interpretation can be dated back to the seminal publication of Tsotsos et al. [54] describing the ALVEN system for automatic heart disease detection. The basic idea of ALVEN is to use a frame-based representation in which each frame can be associated with spatio-temporal constraints describing instantiation restrictions. Spatio-temporal motion phenomena such as heart contractions are described in terms of area changes (the initial area is larger than the resulting area). The change can, for instance, be further characterized using a speed specification, which can be further constrained using additional predicates describing necessary conditions (e.g., the area change must not be too large or too small). A small set of primitive movement descriptors, such as time interval, location change, length change, area change, shape change, and so on are used to describe all higher-level motion concepts. Event frames can be linked to one another using so-called similarity links. Different techniques for event recognition and hypothesis ranking are explored. The description language used in ALVEN is inspired by natural language descriptions for motion events investigated in [6].

Although ALVEN uses a procedural description for the event recognition process, and does not model event recognition as a logical reasoning problem (besides inheritance reasoning), it was one of the first systems to use explicit symbolic representations. ALVEN has influenced the work of Neumann et al. who were among the first to use a logic-based approach for recognizing events in street scenes.

5.2 Quantitative Approaches for Event Definitions

The goal of Neumann and Novak [39,38,36] was to support query answering and the generation of natural language descriptions for street scene events (the system was called NAOS: NAtural language description of Object movements in Street scenes). The basis for the NAOS system is a so-called geometric scene description (GSD): Per timepoint the description consists of detected objects including their types and their positions.

Given a GSD determined by low-level video analysis processes, basic motion event descriptions of single objects are generated. Basic motion events such as move, accelerate, approach, etc. are associated with two timepoints (start point and end point) in such a way that the resulting interval is maximal w.r.t. a sequence of GSD snapshots. Given a set of assertions for basic motion events,

high-level motion events are instantiated based on a set of declarative event models. The following example demonstrates the main ideas behind NAOS[4].

```
(define-event-class ((overtake ?obj1 ?obj2) *t1 *t2)
   (?obj1 vehicle)
   (?obj2 vehicle)
   ((move ?obj1) *t1 *t2)
   ((move ?obj2) *t1 *t2)
   ((approach ?obj1 ?obj2) *t1 *t3)
   ((behind ?obj1 ?obj2) *t1 *t3)
   ((beside ?obj1 ?obj2) *t3 *t4)
   ((in-front-of ?obj1 ?obj2) *t4 *t2)
   ((recede ?obj1 ?obj2) *t4 *t2))
```

Events are specified as Horn rules with object variables (indicated with `?`) and time variables (prefixed with `*`). The first two conditions impose non-temporal static restrictions on the types of `?obj1` and `?obj2`. The temporal relation between subevents are indicated using corresponding time variables. See [33] for a detailed definition of the semantics of event classes in terms of logical rules.

Implicit constraints are established between temporal variables. We give the semantics of the above definition in CLP($\mathfrak{R}$) [24], where *holds*(*Atom*) means that *Atom* can be proven using an external prover, in this case a description logic reasoner.

```
overtake(Obj1, Obj2, T1, T2) :- T1 < T2,
   holds(vehicle(Obj1)),
   holds(vehicle(Obj2)),
   move(Obj1, T1, T2), T1 < T2,
   move(Obj2, T1, T2), T1 < T2,
   approach(Obj1, Obj2, T1, T3), T1 < T3,
   behind((Obj1, Obj2, T1, T3), T1 < T3,
   beside((Obj1, Obj2, T3, T4), T3 < T4,
   in_front_of((Obj1, Obj2, T4, T2), T4 < T2,
   recede(Obj1, Obj2, T4, T2), T4 < T2.
```

An example for a set of basic motion events derived from a GSD is given below (we use constants `vw1` and `vw2`).

```
(define-assertion ((move vw1) 7 80))
(define-assertion ((move vw2) 3 70))
(define-assertion ((approach vw1 vw2) 10 30))
(define-assertion ((behind vw1 vw2) 10 30))
(define-assertion ((beside vw1 vw2) 30 40))
```

[4] The original syntax used in the NAOS system slightly deviates from the example presented here. We describe the syntax used in a reimplementation of the NAOS event recognition engine, which is based on the work described in [33].

```
(define-assertion ((in-front-of vw1 vw2) 40 80))
(define-assertion ((recede vw1 vw2) 40 60))
```

With `(define-assertion ((R X) T1 R2))` a corresponding CLP($\mathfrak{R}$) fact `R(X, T1, T2).` is denoted (analogously for `(R X Y)`).

An example query for our scenario is given as follows (with query results printed below in terms of variable bindings).

```
(?- ((overtake ?obj1 ?obj2) *t1 *t2))
-> OBJ1 = VW1 OBJ2 = VW2 T1 = [10, 29] T2 = [41, 60]
```

The substitutions for object and time variables indicate recognized events. Time intervals indicate the minimum and maximum values for which the event can be instantiated given the assertions for basic motion events specified above. It is also possible to query for events involving specific objects (e.g., `vw2`).

```
(?- ((overtake ?obj1 vw2) *t1 *t2))
-> OBJ1 = VW1 T1 = [10, 29] T2 = [41, 60]
```

Note that in contrast to CLP($\mathfrak{R}$), in NAOS there are actual solutions being generated, and not only consistency checks performed for time variables (or real variables). Given a query (goal specification), NAOS applies some form of backward chaining of event class rules to determine bindings for variables. Backward chaining involves constraint propagation for time variables [38].

In NAOS it is also possible to find instantiations of all declared event models. The rules are applied in a forward chaining way if there is no specific goal.

```
(?-)
Rule OVERTAKE indicates ((OVERTAKE VW1 VW2) 10 60).
```

Based on the bindings found for events, it is possible to explicitly add event assertions to the knowledge base (e.g., `define-assertion ((overtake vw1 vw2) 10 60)`. These assertions can then be used to detect even higher-level events.

As can be seen from the example, the original NAOS system can be used for an a-posterior analysis of a given set of event assertions. In principle, the approach can be extended to support incremental event recognition (see [27] for an early approach based on quantitative data) such that one can also query for events which might be possible at a certain timepoint.

It should also be emphasized that in general there might be multiple possibilities for instantiating events. Thus, a combinatorial space for navigating through logic-based scene models is defined (see [37] for details). Scene models define classes for high-level events using a first-order language. The construction process for valid interpretation hypotheses described in [37] is extra-logical, however.

Horn clauses are not the only logical representation language that has been used in the literature to specify events. In an attempt to formulate scene understanding and event recognition as a (sequence of) logical decision problem(s), the approach described in [35] uses ontologies (aka description logic knowledge bases) as the underlying formalism. In particular, high-level event descriptions

are generated by employing ontology-based *query answering*. The approach in [35] does not specify, however, from which knowledge sources the queries are taken. Along the same lines, a more methodological approach is presented by the same authors in [33]. In this work, event-query generation is formalized as a form of abductive reasoning, and the space of abducibles is defined by rules.

As we have seen, in the NAOS approach, event recognition is based on quantitative information on timepoints, and (simple) constraints over the reals ensure the semantics of time to be represented in NAOS. In addition, event assertion with maximal time intervals must be made available to the NAOS formalism. All assertions are maintained in a large knowledge base.

5.3 Qualitative Approaches for Event Definitions

Another idea to represent events is to subdivide facts into temporally ordered partitions and use qualitative relations between the partitions. This has been explored in the VEIL system (see above) in order to detect event sequences that span multiple images. The goal of this scenario is to process a sequence of images and detect events such as field training exercises. Forty images of a hypothetical armored brigade garrison and exercise area that share a common site model have been used in the experiments reported in [47].

In the VEIL context, an event is a sequence of scenes that satisfy certain criteria. A scene is represented as a set of object descriptions (called a world), which can be associated with a timestamp. Some of the criteria such as the temporal order apply across different scenes, whereas other criteria apply only within a single scene.

A field training exercise is a sequence of scenes showing an armored unit in a garrison, then moving in convoy, then deployed in a training area and finally in a convoy again. In order to extract the scenes that meet the criteria of a field training exercise event, the following query is used:

```
(retrieve (?Y ?S1 ?S2 ?S3 ?S4)
          (and (within-world ?S1 (in-garrison ?Y))
               (within-world ?S2 (convoy ?Y))
               (within-world ?S3 (deployed-unit ?Y))
               (within-world ?S4 (convoy ?Y))
               (before+ ?S1 ?S2)(before+ ?S2 ?S3)(before+ ?S3 ?S4)))
```

Query terms, e.g. *in-garrison* and *deployed-unit*, are defined in the domain model. The result of the query is a set of tuples. Each tuple is a field training exercise event since it satisfies all conditions defined in the query.

It should be pointed out that qualitative relations between states (worlds) are used in the query language. In this context, there are means for adding specification of events to the Tbox (see, e.g., [3]).

In all approaches to image sequence understanding, be they quantitative or qualitative, it is important to understand what is made explicit and what is added by logical reasoning. The corresponding "design space" is investigated in detail in [56]. The main insight is that, given the features of contemporary

reasoning systems, event recognition can be formalized as query answering in the expressive description logic Abox query language nRQL (for an introduction to nRQL see [57]). Building on this query language, [5] formalize event recognition (actually, in [5] event recognition is called situation recognition) by transforming specifications of linear temporal logic into nRQL queries.

6 Summary

We have sketched major logic-based representation languages that formalize interpretation using logical decision processes. The most important insights gained from these works are:

- Existing computer vision systems are well-equipped to process pixel-level data, whereas formal knowledge representation and reasoning systems are more appropriate to process symbolic structures. Therefore it is reasonable to distinguish between surface-level and deep-level information when building a software system for scene interpretation.
- Even though a scalable system for declarative scene interpretation could not be built yet, promising results have been achieved. Various benefits such a system would offer motivate us to develop future logic-based approaches for multimedia interpretation.
- It is hardly possible to compute interpretations of an image through deductive reasoning only. The generation of hypothesis in an abductive way is crucial for scene interpretation, and provides for an appropriate formalization of the generative nature of the interpretation process. Representing interpretation results in terms of logical models (see the Mapsee approach) seems to be too specific, and the specificity of models provides for an over-interpretation of observations.

The goal of scene interpretation is to provide for explanations of the observations made through the analysis of an image. The explanations have to be hypothesized since, in general, observations are not entailed by available background knowledge. In fact, if the available background knowledge would contain explanations of observations, the computation of an scene interpretation would be unnecessary since the scene interpretation would already be part of the background knowledge. Therefore the observations can logically follow from the background knowledge only if appropriate explanations are hypothesized and added to the background knowledge before. Different approaches exist in the literature for specifying the "space of abducibles".

References

1. Aliseda, A.: Abductive Reasoning: Logical Investigations into Discovery and Explanation. Synthese Library, vol. 330. Springer, Heidelberg (2006)
2. Appelt, D.E., Hobbs, J.R., Bear, J., Israel, D.J., Tyson, M.: FASTUS: A Finite-state Processor for Information Extraction from Real-world Text. In: Proceedings of IJCAI, pp. 1172–1178 (1993)

3. Artale, A., Lutz, C., Toman, D.: A description logic of change. In: Veloso, M. (ed.) Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007), pp. 218–223. AAAI Press, Menlo Park (2007)
4. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, Cambridge (2003)
5. Baader, F., Bauer, A., Baumgartner, P., Cregan, A., Gabaldon, A., Ji, K., Lee, K., Rajaratnam, D., Schwitter, R.: A Novel Architecture for Situation Awareness Systems. In: Giese, M., Waaler, A. (eds.) TABLEAUX 2009. LNCS, vol. 5607, pp. 77–92. Springer, Heidelberg (2009)
6. Badler, N.: Temporal scene analysis: Conceptual descriptions of object movements, report tr-80. Technical report, Dept. of CS, University of Toronto (1975)
7. Bast, H., Chitea, A., Suchanek, F., Weber, I.: Ester: efficient search on text, entities, and relations. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), pp. 671–678 (2007)
8. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 24(24), 509–522 (2002)
9. Canny, J.F.: A Computational Approach To Edge Detection. IEEE Transactions on Pattern Recognition and Machine Intelligence 8(6), 679–698 (1986)
10. Castano, S., Espinosa, S., Ferrara, A., Karkaletsis, V., Kaya, A., Möller, R., Montanelli, S., Petasis, G., Wessel, M.: Multimedia Interpretation for Dynamic Ontology Evolution. Journal of Logic and Computation, Advance Access published on September 30 (2008), doi:10.1093/logcom/exn049
11. Charniak, E., Goldman, R.: Probabilistic Abduction For Plan Recognition. Technical report, Brown University, Tulane University (1991)
12. Cucerzan, S., Yarowsky, D.: Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In: Proceedings of Joint SIGDAT Conf. on Emprical Methods in Natural Language Processing and Very Large Corpora (1999)
13. Espinosa, S., Kaya, A., Melzer, S., Möller, R., Wessel, M.: Towards a Media Interpretation Framework for the Semantic Web. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2007), November 2007, pp. 374–380. IEEE Computer Society, Washington, DC, USA (2007)
14. Genesereth, M.R., Nilsson, N.J.: Logical Foundations of Artificial Intelligence. Morgan Kaufmann Publ. Inc., Los Altos (1987)
15. Gries, O., Möller, R., Nafissi, A., Rosenfeld, M., Sokolski, K., Wessel, M.: A probabilistic abduction engine for media interpretation based on ontologies. In: Alferes, J., Hitzler, P., Lukasiewicz, T. (eds.) RR 2010. LNCS, vol. 6333, pp. 182–194. Springer, Heidelberg (2010)
16. Grishman, R.: Information Extraction. In: Handbook of Computational Linguistics Information Extraction (2003)
17. Haarslev, V.: Formal semantics of visual languages using spatial reasoning. In: Proceedings of the 11th IEEE Symposium on Visual Languages, Darmstadt, Germany, September 5-9, pp. 156–163. IEEE Press, Los Alamitos (1995)
18. Haarslev, V.: A fully formalized theory for describing visual notations. In: Proceedings of the AVI 1996 Post-Conference Workshop on Theory of Visual Languages, Gubbio, Italy, May 30 (1996)

19. Harris, C., Stephens, M.: A Combined Corner and Edge Detector. In: Proceedings of 4th Alvey Vision Conference, pp. 147–151 (1988)
20. Hobbs, J., Stickel, M., Martin, P., Edwards, D.: Interpretation as Abduction. In: Proceedings of the Conference on 26th Annual Meeting of the Association for Computational Linguistics (1988)
21. Hobbs, J.R., Stickel, M., Martin, P.: Interpretation as abduction. Artificial Intelligence 63, 69–142 (1993)
22. Hummel, B., Thiemann, W., Lulcheva, I.: Description logic for vision-based intersection understanding. In: Proc. Cognitive Systems with Interactive Sensors (COGIS). Stanford University, CA (2007)
23. Hummel, B.: Description Logic for Scene Understanding at the example of Urban Road Intersections. Südwestdeutscher Verlag für Hochschulschriften (2010)
24. Jaffar, J., Michaylov, S., Stuckey, P.J., Yap, R.H.C.: The CLP(R) language and system. ACM Transactions on Programming Languages and Systems 14(3), 339–395 (1992)
25. Jing, Y., Baluja, S.: PageRank for Product Image Search. In: Proceedings of 17th International World Wide Web Conference WWW 2008 (April 2008)
26. Katz, B., Lin, J., Stauffer, C., Grimson, E.: Answering Questions About Moving Objects in Surveillance Videos. In: Proceedings of AAAI Spring Symposium on New Directions in Question Answering (March 2003)
27. Kockskämper, S., Neumann, B., Schick, M.: Extending process monitoring by event recognition. In: Proc. Second International Conference on Intelligent System Engineering, ISE 1994, pp. 455–460 (1994)
28. Leibe, B., Schiele, B.: Interleaved Object Categorization and Segmentation. In: Proceedings of British Machine Vision Conference (BMVC 2003) (September 2003)
29. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
30. MacGregor, R.M., Bates, R.: The Loom Representation Language. Technical Report ISI/RS-87-188, Information Sciences Institute, University of Southern California (1987)
31. Matsuyama, T., Hwang, V.S.: SIGMA: A Knowledge-Based Aerial Image Understanding System. Perseus Publishing (1990)
32. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Desciptors. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27(10), 1615–1630 (2005)
33. Möller, R., Neumann, B.: Ontology-based Reasoning Techniques for Multimedia Interpretation and Retrieval. In: Semantic Multimedia and Ontologies: Theory and Applications. Springer, Heidelberg (2008)
34. Mulder, J.A., Mackworth, A.K., Havens, W.S.: Knowledge structuring and constrataint satisfaction: The Mapsee approach. IEEE Transactions in Pattern Analysis and Machine Intelligence 10(6), 866–879 (1988)
35. Neumann, B., Möller, R.: On Scene Interpretation with Description Logics. In: Christensen, H.I., Nagel, H.-H. (eds.) Cognitive Vision Systems. LNCS, vol. 3948, pp. 247–275. Springer, Heidelberg (2006)
36. Neumann, B., Novak, H.-J.: NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenenxs. Informatik Forschung und Entwicklung 1, 83–92 (1986)

37. Neumann, B., Weiss, T.: Navigating through logic-based scene models for high-level scene interpretations. In: Crowley, J.L., Piater, J.H., Vincze, M., Paletta, L. (eds.) ICVS 2003. LNCS, vol. 2626, pp. 212–222. Springer, Heidelberg (2003)
38. Neumann, B.: Retrieving events from geometrical descriptions of time-varying scenes. In: Schmidt, J.W., Thanos, C. (eds.) Foundations of Knowledge Base Management – Contributions from Logic, Databases, and Artificial Intelligence, p. 443. Springer, Heidelberg (1985)
39. Neumann, B., Novak, H.-J.: Event models for recognition and natural language description of events in real-world image sequences. In: Proc. International Joint Conference on Artificial Intelligence, IJCAI 1983, pp. 724–726 (1983)
40. Peirce, C.S.: Deduction, Induction and Hypothesis. In: Popular Science Monthly, vol. 13, pp. 470–482 (1878)
41. Poole, D., Goebel, R., Aleliunas, R.: Theorist: A Logical Reasoning System for Defaults and Diagnosis. In: Cercone, N., McCalla, G. (eds.) The Knowledge Frontier: Essays in the Representation of Knowledge, pp. 331–352. Springer, Heidelberg (1987)
42. Poole, D., Mackworth, A.: Artificial Intelligence: foundations of computational agents. Cambridge University Press, New York (2010)
43. Poole, D.: Probabilistic horn abduction and bayesian networks. Artificial Intelligence 64(1), 81–129 (1993)
44. Poole, D.: The independent choice logic and beyond. In: De Raedt, L., Frasconi, P., Kersting, K., Muggleton, S.H. (eds.) Probabilistic Inductive Logic Programming. LNCS (LNAI), vol. 4911, pp. 222–243. Springer, Heidelberg (2008)
45. Reiter, R., Macworth, A.K.: The Logic of Depiction. Technical Report 87-24, Department of Computer Science, University of British Columbia, Vancouver, Canada (1987)
46. Reiter, R., Macworth, A.K.: A Logical Framework for Depiction and Image Interpretation. Artificial Intelligence 41, 125–155 (1989/90)
47. Russ, T., Price, K., MacGregor, R.M., Nevatia, R., Salemi, B.: VEIL: Research in Knowledge Representation for Computer Vision, Final Report. Technical Report A051143, Information Sciences Institute, University of Southern California (February 1998)
48. Russ, T.A., MacGregor, R.M., Salemi, B.: VEIL: Combining Semantic Knowledge with Image Understanding. In: Firschein, O., Strat, T.M. (eds.) Radius: Image Understanding for Imagery Intelligence, pp. 409–418. Morgan Kaufmann, San Francisco (1997)
49. Saathoff, C., Staab, S.: Exploiting spatial context in image region labelling using fuzzy constraint reasoning. In: Ninth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008), pp. 16–19 (2008)
50. Schröder, C.: Bildinterpretation durch Modellkonstruktion: Eine Theorie zur rechnergestützten Analyse von Bildern. PhD thesis, University of Hamburg (1998)
51. Se, S., Lowe, D., Little, J.J.: Global Localization using Distinctive Visual Features. In: Proceedings of International Conference on Intelligent Robots and Systems (IROS 2002), Lausanne, Switzerland, pp. 226–231 (November 2002)
52. Shanahan, M.P.: Perception as Abduction: Turning Sensor Data Into Meaningful Representation. Cognitive Science 1, 103–134 (2005)
53. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. Thomson Learning (April 2007)
54. Tsotsos, J.K., Mylopoulos, J., Covvey, H.D., Zucker, S.W.: A framework for visual motion understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (1980)

55. Viola, P., Jones, M.: Robust Real-time Object Detection. International Journal of Computer Vision (2001)
56. Wessel, M., Luther, M., Möller, R.: What happened to Bob? Semantic data mining of context histories. In: Proc. of the 2009 International Workshop on Description Logics DL 2009, Oxford, United Kingdom, July 27-30. CEUR Workshop Proceedings, vol. 477 (2009)
57. Wessel, M., Möller, R.: A high performance semantic web query answering engine. In: Horrocks, I., Sattler, U., Wolter, F. (eds.) Proc. International Workshop on Description Logics (2005)

Ontology Population and Enrichment: State of the Art

Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras,
Anastasia Krithara, and Elias Zavitsanos

Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos",
15310, Ag. Paraskevi, Attiki, Greece
{petasis,vangelis,paliourg,akrithara,izavits}@iit.demokritos.gr

Abstract. Ontology learning is the process of acquiring (constructing or integrating) an ontology (semi-) automatically. Being a knowledge acquisition task, it is a complex activity, which becomes even more complex in the context of the BOEMIE project[1], due to the management of multimedia resources and the multi-modal semantic interpretation that they require. The purpose of this chapter is to present a survey of the most relevant methods, techniques and tools used for the task of ontology learning. Adopting a practical perspective, an overview of the main activities involved in ontology learning is presented. This breakdown of the learning process is used as a basis for the comparative analysis of existing tools and approaches. The comparison is done along dimensions that emphasize the particular interests of the BOEMIE project. In this context, ontology learning in BOEMIE is treated and compared to the state of the art, explaining how BOEMIE addresses problems observed in existing systems and contributes to issues that are not frequently considered by existing approaches.

Keywords: Ontology learning, Ontology population, Ontology enrichment.

1 Introduction

In recent years, ontologies have become extremely popular as a means for representing machine-readable semantic knowledge. The rapid growth of the Web and the information overload problem that it has caused has triggered significant research in the development of practical information extraction solutions that process Web content. However, the difficulty of extracting information from the Web, which was produced mainly for visualising information, has driven the birth of the Semantic Web. The Semantic Web will contain many more resources than the Web and will attach machine-readable semantic information to these resources. The first steps towards that goal, addressed knowledge representation issues for this semantic information, with the development of ontologies. Realizing the difficulty of designing the grant ontology for the world [96], research on the Semantic Web has focused on the development of domain or task-specific ontologies which have started making their appearance in fairly large numbers. Having provided an ontology for a specific domain, the next step is to annotate semantically related Web resources. If done manually, this

[1] The BOEMIE project is presented in chapter 1.

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 134–166, 2011.

process is very time-consuming and error-prone. Information extraction is a promising solution for automating the annotation process. However, it comes along with the aforementioned knowledge acquisition bottleneck and the need for learning.

At the same time, acquiring domain knowledge for ontologies is also a resource demanding and time-consuming task. Thus, the automated or semi-automated construction, enrichment and adaptation of ontologies, is highly desired. The process of automatic or semi-automatic construction, enrichment and adaptation of ontologies is known as ontology learning [79]. From our perspective, ontology learning is a wide research area that includes work on ontology enrichment, inconsistency resolution and ontology population. Ontology enrichment is the task of extending an existing ontology with additional concepts and semantic relations and placing them at the correct position in the ontology. Inconsistency resolution is the task of resolving inconsistencies that appear in an ontology with the view to acquire a consistent (sub)ontology. Ontology population, on the other hand, is the task of adding new instances of concepts to the ontology.

Despite the fact that it is an emerging field, a significant amount of research has been performed already, leading to a large number of proposed approaches and practical systems. A fairly complete overview of the work performed in the field until 2003 is presented in [45], as well as in [99]. An updated overview of the field is also presented in [24]. Ontology learning has also significant presence in the major AI conferences, with workshops such as "Ontologies and Texts" (OLT) (EKAW2000 [8], ECAI2002 [9]), and other important conferences (IJCAI2001 [76], ECAI2000 [105] and workshops ECAI2004-OLP [18], OLP2 [20] and ECAI2008-OLP3 [22]).

The purpose of this chapter is to present the state of the art in ontology learning, by presenting the major approaches and most important practical systems that appear in the literature. The BOEMIE project is compared to these systems throughout this chapter and the solutions it gives to the various problems faced by the others are discussed. Systems and approaches are categorised along significant dimensions, such as the ontology elements learned, the starting point, the learning approach and the final outcome. The task of ontology learning is presented in section 2, covering the most significant approaches found in the literature. In section 3, ontology population is presented, as well as some important ontology population tools, which are also compared. Section 4 discusses ontology enrichment and follows a comparative presentation of ontology enrichment tools. Ontology evaluation is presented in section 5, while section 6 concludes this document.

2 Ontology Learning Foundations

Ontologies are a means for sharing and re-using knowledge, a container for capturing semantic information of a particular domain. A widely accepted definition of ontology in information technology and AI community is that of "a formal explicit specification of a shared conceptualization" [44], where "formal implies that the ontology should be machine-readable and shared that it is accepted by a group or community" [19]. Additionally, in the case of a domain ontology, it is usually assumed that it conveys concepts and relations relevant to a particular task or the application domain, which is the case we are interested in.

Ontology learning is the process of acquiring (constructing or integrating) an ontology (semi-) automatically. The acquisition of ontologies can be performed through three major approaches:

- By integrating existing ontologies. The integration process tries to capture commonalities among ontologies that convey the same or similar domains, in order to derive a new ontology. Several methods have been proposed in the literature, such as:
    - the merging of ontologies to create a single coherent ontology,
    - the alignment of ontologies by establishing links between them and allowing them to reuse information from each another, and
    - the mapping of ontologies by finding correspondence among elements in the ontologies.
- By constructing an ontology from scratch or by extending (populating and enriching) an existing ontology, usually based on information extracted from domain-specific content.
- By specialising a generic ontology, in order to adapt it to a specific domain.

In this chapter we will concentrate on the last two approaches, the construction of new ontologies and the enrichment/specialisation of existing ontologies.

Research in ontology learning studies methods and techniques for the acquisition of an ontology, based on semantic information, extracted from domain-specific content. Being closely related to the field of knowledge acquisition, a significant amount of the work presented in the bibliography concentrates on the task of knowledge acquisition from text, through the re-use of widely adopted natural language processing and machine learning techniques. However, ontology learning is not simply a replication of existing work under a different name, as it adds novel aspects to the problem of knowledge acquisition [19]:

- Ontology learning combines research from knowledge representation, logic, philosophy, databases, machine learning, natural language processing, image/audio/video analysis, etc.
- Ontology learning in the context of the Semantic Web must deal with the massive and heterogeneous data of the World Wide Web and thus improve existing approaches for knowledge acquisition, which target mostly small and homogeneous data collections.
- Substantial effort is being put into the development of extensive and rigorous evaluation methods in order to evaluate ontology learning approaches on well-defined tasks with well-defined evaluation criteria.

Following [19], the ontology learning process can be decomposed into six layers, forming a "layer cake"[2] of increasingly complex subtasks, which can be seen in Fig. 1.

[2] Ontology learning "layer cake" has been originally formulated with terminology originating from the textual modality. However, since the "layer cake" is applicable to multiple modalities, the labels of the layers have been slightly extended to cater for multimodality.

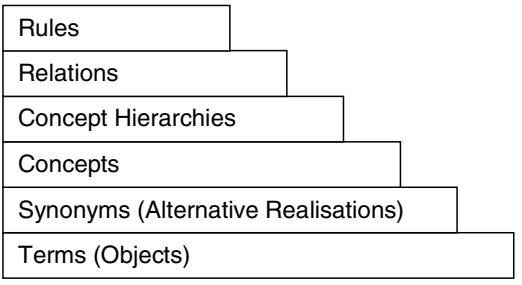


Fig. 1. Ontology learning "layer cake"

The main target of ontology learning is the definition of concepts and the relations between them. However, this implies substantial knowledge about the "symbols" that represent these concepts and relations and "instantiate" these into entities of the real word. We will use the notion of object or term to refer to these instances of concepts and relations, but it should be noted that we do not necessarily refer to the text modality: an object can be an audio, image or video segment that instantiates a concept or a relation in a corpus of the corresponding modality. Thus, in order to define new concepts/relations, the acquisition of knowledge about the objects that instantiate these concepts/relations in content is equally important. In addition to knowledge about objects/terms, object/term synonyms are also important: all terms that are synonyms (alternative realisations) refer to the same real object or event, and thus all materialise a single concept or relation. Failure to identify which terms/objects are synonyms may result in the introduction of redundant concepts or relations in an ontology, which in most cases is undesirable.

Among relations, one type is of particular importance to ontologies, namely hierarchical ones. These are the relations that realise the taxonomy backbone of an ontology, such as the subsumption relation (also referred as "is-a" relation in many cases). On the other hand, non-hierarchical relations are all relations that are not used in the formation of the concept hierarchy. Despite the fact that the relations are categorised into types, no type categorisation is performed at the concept level in the vast majority of the work presented in the literature.

Finally, an important aspect of an ontology is the ability to derive and make explicit facts that are implied by the knowledge in the ontology, mainly through reasoning. But for such derivations to occur, rules must be defined (and possibly acquired) to allow for such derivations. All of these aspects of ontology learning, related to things that can be learned, can be organised into the "layer cake" of Fig. 1 [19]. In the following subsections we are going to briefly present the state of the art for each layer of this "cake".

2.1 Object Identification

Object extraction (or identification) is a prerequisite for all aspects of ontology learning. An object is an instance of a recognisable entity in a multimedia corpus that conveys a single meaning within a domain (concept). A recognizable entity is something that can be recognized in multimedia corpora, such as words or phrases in textual

corpora, or areas in images. Since objects "materialise" a concept, objects found in a corpus usually represent candidate concepts that can enrich an ontology. Thus, the main objective is the identification of objects in a multimedia corpus that possibly convey concepts, which can be used for enriching an ontology. The object identification task can be decomposed into three subtasks [61]:

- Object recognition. This task is responsible for finding recognisable entities in the corpus that are objects.
- Object classification. This task assigns a semantic category to recognised objects. This categorization is important for the task of ontology learning, as these categories are often the concepts of the thematic domain.
- Object mapping. This task tries to link identified objects with relevant entities in other data sources, such as object libraries, vocabularies, lexica, thesauri and databases. A frequent use of this task is for exploiting similarities that potentially exist in the referred data sources, in order to identify clusters of objects that represent the same concept – synonyms/alternative realisations.

As object/term identification is an important task, not only for concept discovery for ontology learning but also for textual information extraction and retrieval, many approaches have been presented in the literature (mainly for the processing of textual corpora). Among the most successful ones are statistical methods, which usually measure the significance of each word with respect to other words in a corpus, based on word occurrence frequencies. TF/IDF [91] is often employed for this task [3, 30], possibly combined with other methods, such as latent semantic indexing [41] or taking into account co-occurrence information among phrases [43].

Clustering techniques also play an important role in object identification: recognizable entities can be clustered into groups based on various similarity measures, with each cluster being a possible object (consisting of synonyms). Approaches like [2, 37, 57] employ clustering techniques and other resources, such as the WWW and WordNet [38], to successfully extract terms. Additionally, both frequency and clustering-based approaches can be substantially enhanced through the use of natural language processing techniques, such as morphological analysis, part-of-speech tagging and syntactic analysis, as terms usually are noun phrases or obey specific part-of-speech patterns [47, 49]. Finally, morphological clues, such as prefixes and suffixes, can be very useful for some thematic domains: suffixes like "-fil" and "-itis" quite often mark terms in medical domains [50, 51].

Other methods use filters and heuristics. For example, Glossex [60] filters terminological candidates using lexical cohesion and a measure of domain relevance. It also uses some additional heuristics for extracting useful terms. TermExtractor [93] extracts a list of "syntactically plausible" terms and uses two entropy-based measures. The first metric, called Domain Consensus, is used to select only the terms which are used consistently throughout the corpus. The second one, Domain Relevance, is used to select only the terms that are relevant to the domain of interest. Finally, extracted terms are further filtered using Lexical Cohesion, which measures the degree of association of all the words in a terminological string.

2.2 Alternative Realization/Synonym Identification

Alternative realisations/synonyms are objects that refer to the same real object or event, variants in a corpus that can be thought to represent the same concept or relation. A significant amount of work has been performed mainly for text corpora, by exploiting resources such as WordNet [38]. Employing standard word sense disambiguation techniques [29, 64, 109] they seek to identify the most appropriate (WordNet) sense of each term, in order to collect synonyms associated with the sense. Other approaches try to locate term synonyms through clustering, mainly based on Harris's distributional hypothesis, according to which similar terms in meaning tend to share syntactic contexts [54, 68, 70, Hindle, 1990]. Related work is also performed in the field of information retrieval for term indexing, such as the family of Latent Semantic Indexing algorithms (LSI, LSA, PLSI, etc.), and the family of probabilistic topic models, e.g. Latent Dirichlet Allocation (LDA [12]). These methods apply dimensionality reduction techniques to reveal inherent relations between terms, in order to form clusters [63, 94]. Finally, more recent approaches extract synonyms by applying statistical approaches over the Web [10, 107]. For more information on such methods, the reader is referred to [19].

2.3 Concept Identification

Despite the fact that concepts are an important part of an ontology, what constitutes a concept is controversial. According to [19], concept formation should provide:

- An intentional definition of the concept.
- A set of concept instances.
- A set of realisations (i.e. terms).

Two types of intentional concept definition can be identified: informal and formal. An informal concept definition does not define a concept in terms of properties and relations between them, but in a more general, descriptive way, like for example a textual description or a concept gloss in a dictionary. Informal concept identification is quite rare, with only one approach appearing in the literature, the OntoLearn system [111], which associates WordNet glosses with domain specific concepts. Formal concept definition, on the other hand, builds on top of object and synonym identification, by formulating concepts as clusters of "related" objects. It exploits relations among objects that are discovered using approaches which will be described in the following two subsections. Basing the definition of a concept on a cluster of objects automatically provides the set of realisations of the new concept. The association of a set of instances with a concept is known as ontology population or ontology tagging, and it will be presented in greater detail in section 3.

2.4 Taxonomy Construction

An important part of an ontology is its taxonomy, or the hierarchy of concepts. Subsumption relations (also known as "is-a" or inclusion relations) provide a tree view of the ontology and determine inheritance between concepts. A popular approach for taxonomy discovery in textual domains is the use of lexico-syntactic patterns (such as

Hearst patterns [53]). According to this approach, syntactic elements (such as noun phrases) are combined with characteristic phrases to identify inclusion relations. Examples of such patterns can be the following ones (NP stands for noun phrase):

- NP such as NP, NP,..., and NP
- such NP as NP, NP,..., or NP
- NP, NP,..., and other NP
- NP, especially NP, NP,..., and NP
- NP is a NP

Several systems have been proposed based on simple variations of the above idea, such as [56, 57, 84]. More recent systems also employ pattern learning algorithms to automate pattern construction [1, 31, 103]. For non-textual domains, machine learning methods, such as hierarchical clustering, can be used. Further details on such approaches can be found in [115] and [19].

Yang and Callan [108], in a metric-based taxonomy induction framework, combine the strengths of pattern-based and clustering-based approaches. The framework incorporates lexico-syntactic patterns as one type of feature in a clustering framework. It integrates contextual, co-occurrence, syntactic dependency, lexical-syntactic patterns, and other features to learn an *ontology metric*, i.e. a score indicating semantic distance, for each pair of terms in a taxonomy; it then incrementally clusters terms based on their ontology metric scores.

Snow et al. [102] have presented an algorithm for inducing semantic taxonomies, which attempts to globally optimize the entire structure of the taxonomy. The model has the ability to integrate heterogeneous evidence from different classifiers, offering a solution to the key problem of choosing the correct word sense for a new hypernym.

A particularly interesting machine learning technique for hierarchy construction is the estimation of Probabilistic Topic Models that produce a hierarchical modelling of a particular collection. Among the well known models of this family is the hierarchical Latent Dirichlet Allocation (hLDA) [13], where each document is modeled as a set of topics across a specific path of the learned hierarchy from the root to a leaf. In addition, the models of the Pachinko Allocation family, like PAM [66], hPAM [83] and non-parametric PAM [67] deal with some of the problems of hLDA, such as the lack of multiple inheritance between topics at different levels of the hierarchy. Among the major benefits of methods that rely on such models is that the identification of topics, which serve as concepts in the ontology, and their taxonomic arrangement is performed simultaneously. In addition, these models do not require an initial ontology to start from. They construct a taxonomic backbone without any prior knowledge, but a collection of documents. In order to learn topic ontologies, probabilistic topic models have been applied in [117,118] and in [114].

2.5 Semantic Relation Extraction

Relations beyond the concept hierarchy (non-taxonomic relations) constitute also an important component of an ontology. Such relations can be extracted with approaches similar to the ones used for extracting taxonomic relations. In textual domains, where most of the existing work has focussed, lexico-syntactic patterns again play an important role. Verbs usually represent actions or relations between recognisable entities in

sentences. As a result, verbs are assumed to express relations between entities, which may be useful for enriching an ontology, provided that the involved entities can be associated with concepts of the ontology. Systems like the RelExt tool [95] use such patterns to identify related pairs of concepts. Additionally, semantic clustering of verbs has been reported to help in situations where extraction of specific relation types is desired [101]. Finally, association rule mining algorithms have been used for the acquisition of non-taxonomic relations for ontology enrichment [74, 75].

2.6 Ontology Rule Acquisition

Ontology rule acquisition is probably the least addressed aspect of ontology learning, as almost no work has been presented that acquires rules. An initial attempt to formulate the problem is presented in [69], where an unsupervised method for discovering inference rules from text is presented. Learned rules are of the following form "*X is author of Y* ≈ *X wrote Y*, *X solved Y* ≈ *X found a solution to Y*, and *X caused Y* ≈ *Y is triggered by X*" [69]. Also, Sangun et al., [92] proposed an ontology rule acquisition procedure using an ontology, which includes information about the rule components and its structure. The procedure comprises rule component identification and rule composition. They use stemming and semantic similarity for the former and a Graph Search method for the latter. Finally, in the field of inductive logic programming (ILP), which deals with the induction of first-order rules, some attempts have been made to address reasoning for the Semantic Web [71].

2.7 Comparative Analysis of Ontology Learning Tools

During the last decade, a large number of approaches and practical systems have been presented that try to automate ontology construction. The presented approaches are so diverse, and thus trying to classify existing systems along a single "dimension" will be at least incomplete. Thus, for this document a comparison framework similar to the one proposed in [99] will be adopted, where some important comparison "dimensions" are defined. Following [99], we will classify existing approaches/practical systems performing both ontology population as well as ontology enrichment, according to the following categorisation criteria:

- **Elements of the "layer cake" learned.** The elements of the "layer cake" that are learned provide a good view of the complexity and capabilities of an ontology learning system, through the ontological aspects learned by the system. It is desirable for a system to provide solutions to as much layers as possible.
- **Initial requirements.** Initial requirements, such as prior knowledge and type of required input for learning an ontology, clarify the starting point of an ontology learning system, the background knowledge and the resources available in order to help knowledge acquisition. In addition, the use of domain-depended resources affect directly the feasibility of a system, as it restricts its portability to new thematic domains.
- **Learning approach.** Of particular interest is also the approach an ontology learning tool adopts in order to extract knowledge, and whether this approach is specialised to the domain, e.g. an extraction engine based on manually

constructed patterns, or a more general one, e.g. based on machine learning or statistical methods. The learning approach adopted by a system usually affects other categorisation criteria, such as the initial requirements and of course the degree of automation, as the usage of machine learning methods usually reduces the degree of manual intervention of the domain expert during knowledge acquisition.

- **Degree of automation.** The degree that a system automates decisions is important, as it contributes to the plausibility of the system. A fully automated system is of course desirable, but it may not be always possible, especially with tasks related to ontology enrichment. But even in the case of semi-automated or cooperative systems, various degrees of automation can be identified. For example, the required knowledge expected by the expert: interaction through a domain expert may be more desirable than interaction through an ontology expert, who is expected to know both the thematic domain in addition to ontology engineering.
- **Consistency maintenance and redundancy elimination.** We are also interested in the outcome of the system and the knowledge representation structures used for storing the acquired information. Systems that do not enhance an ontology usually do not deal with aspects such as consistency maintenance or redundancy elimination. Maintaining the consistency of an ontology is crucial, as an ontology that contains conflicting information is of little use. Redundancy elimination on the other hand is not as crucial as consistency, i.e., redundancy cannot render an ontology useless, unless it also introduces contradictions. However, redundancy elimination can enhance the plausibility of an ontology by facilitating the process of querying the ontology, and at the same time by limiting the size (and complexity) of the ontology.
- **Domain portability.** An important aspect of an ontology learning system is whether it can be ported to other thematic domains or not. Systems that exhibit increased domain portability tend to explicitly define the required domain knowledge, whereas less portable system can contain domain specific knowledge in the internals of the system.
- **Corpora Modality.** It is desirable for a system to be able to process more than one modalities, as it can provide evidence of the ability of a system to accommodate and exploit diverse knowledge sources, fuse the extracted information and provide unified results that are valid across modalities.

2.8 A Procedural View of Ontology Learning

Based on our experience in the area from our involvement in several relevant projects, we consider that the task of ontology learning involves the subtasks of population, enrichment, and inconsistency resolution. Ontology population is the process of adding new instances of concepts/relations into an ontology, usually by locating the corresponding object/terms and synonyms in the corpus. Ontology enrichment is the process of extending an ontology with new concepts, relations and rules. Inconsistency resolution is responsible for remedying problems introduced by population and enrichment. In addition to these subtasks, ontology evaluation is also needed in order

to measure the plausibility of the learned ontology by evaluating the usefulness of the changes. Fig. 2 depicts a typical ontology learning process.

Very often, ontology learning is modelled as a bootstrapping process: an initial ontology is used as a basis for learning a new ontology, which in turn substitutes the initial one and the whole process restarts. In particular, an initial ontology is used to analyze and extract information from a corpus. The extracted information is used to evolve the ontology, and through the evolved ontology the extraction of information is improved. The bootstrapping process continues until no more information can be extracted from the corpus. Here we have to note that in every cycle the consistency of the ontology is checked and in the case of inconsistency, the changes are discarded. In the following section, the steps involved in ontology population will be described in more detail, along with a comparative analysis of the most important approaches and practical systems performing ontology population. The steps of ontology enrichment will be presented in section 4, along with a comparative analysis of the most important approaches and practical systems performing ontology enrichment. Finally, ontology evaluation will be presented in section 5.

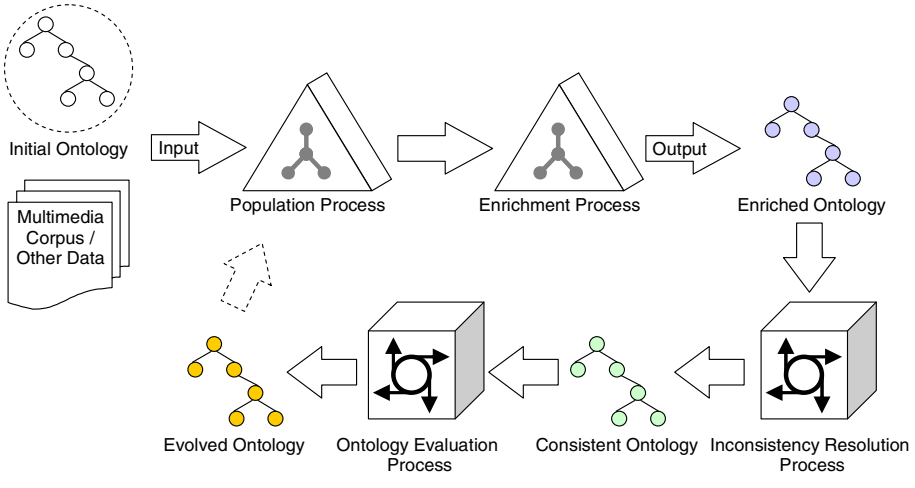


Fig. 2. The process of ontology learning

3 Ontology Population

Ontology population is the process of inserting concept and relation instances into an existing ontology. In a simplified view, an ontology can be thought of as a set of concepts, relations among the concepts and their instances. A concept instance is a realisation of the concept in the domain, e.g. the instantiation of the concept as a phrase in a textual corpus. The process of ontology population does not change the structure of an ontology, i.e., the concept hierarchy and non-taxonomic relations are not modified. What changes is the set of realisation (instances) of concepts and relations in the domain. A typical ontology population methodology is depicted in Fig. 3.

Ontology population requires an initial ontology that will be populated and an instance extraction engine. The extraction engine is responsible for locating instances (realisations) of concepts and relations in a multimedia corpus. A multimedia corpus is processed by the extraction engine, in order to locate concept/relation. The list of extracted concept/relation instances is subsequently used to populate the ontology.

Recalling the "layer cake" idea, the population process involves some of the layers presented in section 2. In particular, it deals with the acquisition of realisations (i.e. objects and alternative realisations/synonyms) of both concepts and relations. A typical approach is to use known realisations associated with concepts/relations which may have been identified during concept/relation formation, to locate the corresponding objects/synonyms in a corpus. This process is also known as lookup text extraction or prototype recognition in image analysis. The result is an annotated corpus, which can be used to construct more general instance extractors, using machine learning.

An interesting aspect of ontology population, which is not addressed adequately in the literature, is the handling of redundancy. The elimination of redundancy in the instance set requires entity disambiguation, which is the process of identifying instances that refer to the same real object or event. If an ontology is populated with an instance without checking if the real object or event represented by the instance already exists in the ontology, then redundant instances will be inserted. A worst case scenario is that redundant instances contain contradicting information, which may lead to an inconsistent ontology.

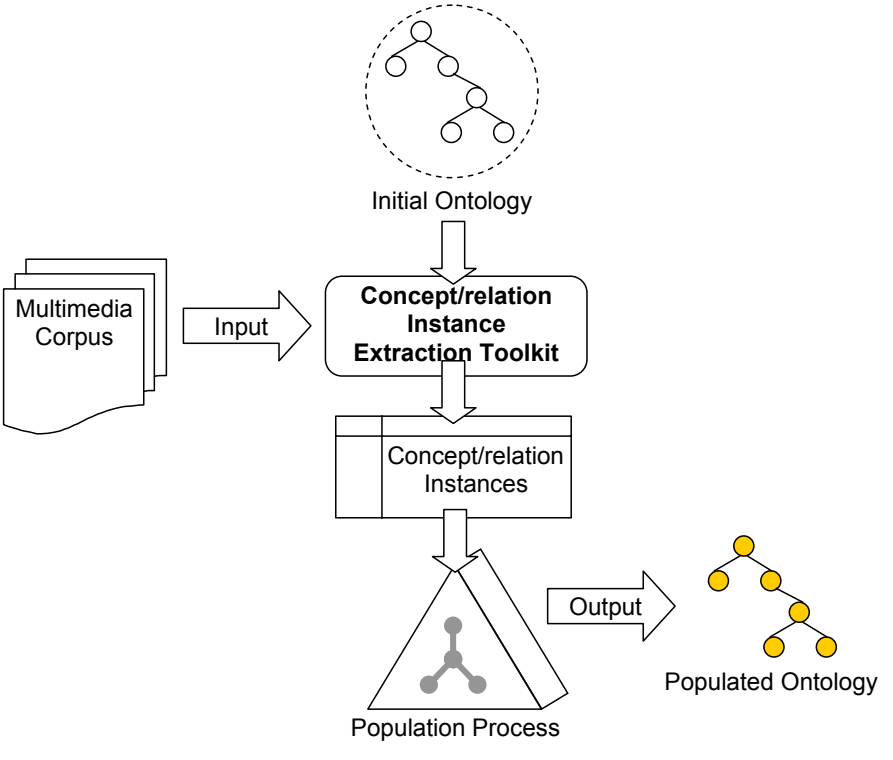


Fig. 3. The ontology population process

To our knowledge, only three approaches address this problem. The Artequakt system [4, 5, 6, 59] applies manually written heuristics, in order to merge instances that refer to the same real object or event. These heuristics are evaluated after a batch of instances has populated the ontology. The SOBA system [21], on the other hand, performs simple checks using special mapping rules, during instance creation (i.e. before the instances populate the ontology), in order to re-use instances that refer to the same real object or event instead of creating new ones. The approach followed by BOEMIE enhances that of Artequakt, through the use of machine learning instead of manually-developed heuristics.

3.1 The BOEMIE Approach to Ontology Population

BOEMIE [23] implements an ontology based information extraction system, that is able to extract objects from a variety of modalities, including texts, images, and videos. Due to its multimodal nature, the BOEMIE system clearly distinguishes entities from their realisations (through properties) in the various modalities. Exploiting the idea that you cannot find entities in corpora but rather their properties, BOEMIE adopts a different approach that separates the concepts into two types: "primitive" concepts that can be easily attributed to objects (i.e. have direct realisations) – mid-level concepts (MLCs) in BOEMIE terminology – and "composite" concepts (that represent real objects or events), usually build on top of primitive ones. These "composite" concepts do not have direct realisations as they cannot be mapped directly to an object and are named high-level concepts (HLCs) in BOEMIE. For example, consider a person that is referenced in a set of textual documents, images and videos. From the text modality BOEMIE can extract a person name, an age, a gender or a profession: this set of properties is considered instances of MLCs for the text modality. In addition, by exploiting linguistic information (such as verbs), relations may be extracted that relate these MLC instances with each other (i.e. suggesting that a specific age, gender and profession are related with a specific person name). Similarly, from an image anatomical parts (i.e. a person face) can be extracted, and possibly a person name from the caption or through OCR. Again, all these are instances of MLCs for this modality, possibly related to each other through spatial and proximity relations.

Despite the fact that instances of properties of a person have been extracted from the involved modalities, a person instance has not yet been identified. This is because "person" is a "composite" concept, an HLC. The identification of entities, and thus the instantiation of HLC instances, is performed as a second processing step: reasoning is employed, where through rules MLC instances (properties) extracted from the various modalities are fused and interpreted. During fusion and interpretation, relations between MLC instances will be examined in order to identify the number of involved entities (i.e. persons) and which properties belong to which person. The result of the interpretation process will be instances of HLC concepts, for all identified entities.

Since the vast majority of work in ontology learning does not discriminate between "primitive" and "composite" concepts, ontology population in these systems is performed as a single step, i.e. the instances that are assimilated into the ontology are identified directly by the instance extraction tool, thus requiring the incorporation of

considerable domain knowledge in the extraction tool. Instance extraction tools typically instantiate complex composite structures with groups of realisations (objects/terms) related to each other through ontology relations.

The population methodology proposed by the BOEMIE project distinguishes between two layers of complexity when populating an ontology with concept instances. Concepts are divided into "primitive", called mid-level concepts, and "composite" ones, called high-level concepts. In contrast to mid-level concepts that are populated by extraction tools as described above, the high-level concepts are populated by reasoning over the mid-level instances, since they are defined in terms of "primitive" concepts. The main differences between the BOEMIE approach and the state of the art are:

- The concept/relation instance extraction engine is not expected to extract instances of "composite" concepts. It is expected to extract only instances of "primitive" concepts. A clear advantage is the fact that the extraction engine becomes immune to changes in the organisation of the ontology, which is a desired property in environments where the ontology evolves over time. The extraction engine needs to adapt only when new "primitive" concepts or relations involving "primitive" concepts are modified.
- The ontology is used to create instances of "composite" concepts from populated "primitive" concept instances and populated relation instances, through non-standard reasoning[3]. The advantages of such an approach are two–fold: a) "composite" concept instances are always in sync with the current formal definition of the relevant concepts, and b) the formation of "composite" instances respects the constraints that may be imposed by the ontology, i.e. through rules, thus helping maintaining the consistency of the ontology.

To our knowledge, there is no method in the bibliography following this two-stage approach to ontology population.

3.2 Comparative Analysis of Ontology Population Tools

The vast majority of the systems found in the literature for ontology population, share the architecture depicted in Fig. 3: an extraction toolkit is used for object/term identification or named-entity recognition, in order to locate instances of concepts and in some cases also instances of relations between concepts, which are then assimilated into the ontology. Ontology population systems are closely related to ontology-based information extraction systems, since the latter provide mechanisms to associate pieces of the data with concepts of an ontology. Thus, every ontology-based information extraction system can be viewed as an ontology population system, as it can be extended to assimilate extracted instances into the ontology.

In the rest of this section we present a comparative analysis of the main approaches and practical systems that have been presented in the literature for ontology population.Table 1 presents a summary of the systems. The comparison is guided by our categorisation criteria described in subsection 2.7, relating also important features of

[3] BOEMIE employed abductive reasoning in order to create "composite" objects from "primitive" ones.

the BOEMIE project, such as portability to other thematic domains, preservation of the ontology consistency and entity disambiguation, as explained in subsection 3.1. Also, due to the focus of BOEMIE on multimedia corpora, we categorize the different systems according to the modality of the data they can handle. This parameter has proved particularly important, as the majority of the systems use textual corpora, and they rely heavily on linguistic processing, such as syntactic analysis, or exploitation of additional resources like thesauri and semantic hierarchies.

Elements extracted. Some systems are more complete in the sense that they populate an ontology with instances of both concepts and relations, such as Artequakt [4, 5, 6, 59], WEB→KB [26], SOBA [21], [85, 86], OPTIMA [58] and ISOLDE [113]. Others, such as Adaptiva [15], LEILA [106] and [7] concentrate only on relation instances. Finally, the KnowItAll system [34, 35] identifies only concept instances, while BOEMIE is able to extract both concept and relation instances in order to populate the ontology.

Table 1. Brief description of the different systems for ontology population

System	Description
Artequakt	Extracts knowledge from the web about artists, populates a knowledge base and uses it to generate personalized biographies. Once instances have been identified, the system uses a domain specific ontology and a generic one in order to extract binary relations between two instances. It uses heuristics to remove redundant instances from the ontology.
WEB→KB	Combines statistical and logical (FOIL rule learning) methods to learn concept instances and relation instances from web documents. The system employs document classification to identify and classify as instances whole pages from the web. Instances of relations are retrieved by examining hyperlink paths that connect web pages.
KnowItAll	Uses domain-independent lexico-syntactic patterns to extract possible instances. It selects the instances by evaluating their plausibility, using a version of the pointwise mutual information statistical measure.
Adaptiva	Employs a bootstrapping approach, extracting instances of relations from a corpus and asking an ontology expert to validate them. The outcome of validation is used by Amilcare [25], functioning as a pattern learner. Once the learning process is completed, the induced patterns are applied to unseen corpora and new examples are returned for further validation by the user.
SOBA	Automatically populates a knowledge base by information extracted from soccer match reports as found on the web. It employs standard rule-based information extraction to extract named entities related to soccer events. The extracted information is converted into semantic structures, as defined by the ontology, with the help of mapping rules.
[85, 86]	A pattern-based system to automatically enrich a core ontology with the definitions of a domain glossary. It uses manually developed lexico-syntactic patterns for extracting instances of concepts. These instances are processed in order to extract relation instances which associate extracted information with concept properties.

Table 1. (*Continued*)

LEILA	A system that learns to extract instances of binary relations from natural language corpora. The system employs statistical techniques to learn the extraction patterns for the relation.
[7]	Automatically learns extraction patterns for finding semantic relations in unrestricted text, based on statistical corpus processing.
OPTIMA	A (semi-)automated system for populating ontologies from unstructured or semi-structured texts. It extracts relational information with natural language processing techniques. It assigns instances to concepts by calculating a fitness value between a candidate instance and each concept in the ontology, using the hierarchical syntactic information of the ontology schema.
ISOLDE	Generates a domain ontology from a seed ontology by exploiting a general purpose NER system and lexico-syntactic patterns to extract concept candidates. Concept candidates are then filtered according to their statistical significance and the knowledge that can be derived from available Web resources.
BOEMIE	Combines an ontology-based information extraction (OBIE) engine based on machine learning, with an inference engine, in order to extract "primitive" concept instances from multiple modalities, which are then fused and interpreted (through abductive reasoning) to form instances of "composite" and more abstract concepts.

Initial requirements. In order to be self-sustained, an ontology population system should have as few initial requirements as possible, in terms of resources or background knowledge. Some systems do not perform object/term and synonym identification, but rather employ publicly available processing resources for this task. Artequakt is based on the information extraction toolkit GATE [27, 28] to perform named entity recognition, syntactic and semantic analysis. SOBA uses a standard rule-based information extraction system, an enhanced version of SProUT – [32], while [7] a part of speech tagger and a module for named entity recognition. Other systems, instead of employing a term/synonym extraction engine, require extraction patterns to be provided by the user. For example, KnowItAll uses domain-independent lexico-syntactic patterns, inspired by Hearst patterns [53]. On the other hand, the system presented in [85, 86] uses manual extraction patterns to populate the CIDOC CRM ontology with terms extracted from glosses of the Art and Architecture Thesaurus (AAT). OPTIMA uses user-defined named entity types, organized in a hierarchy, and user-defined binary relations. A name-entity recogniser based on these particular entity types is used for the extraction of instances. ISOLDE uses a general purpose named entity recogniser to find instances in a base ontology and then uses Hearst patterns to find class candidates. Systems like WEB→KB, Adaptiva and LEILA include an adaptable term/synonym extraction engine which can be taught with the help of concept/relation instance examples. BOEMIE adopts a similar term/synonym extraction approach. An adaptable term/synonym extraction engine is employed using examples of instances that are provided either through manually annotated corpora, or by the previous ontology population steps.

Learning approach. Machine learning seems to be the choice of the majority of systems, as all but three of the examined systems (Artequakt, SOBA, [85, 86]) employ some form of learning. The systems employing machine learning either use statistical methods to identify terms, or perform automated pattern extraction. For example, Adaptiva uses a tool for adaptive Information Extraction from text (IE), to learn patterns. KnowItAll uses an extended version of the pointwise mutual information [107] statistical measure, which selects the instances that will populate the knowledge base, by evaluating their plausibility. OPTIMA uses a trainable named entity recognizer, combining a boundary detector using CRFs [62] and a named-entity classifier using maximum entropy. ISOLDE employs a seed ontology and the general-purpose NER system SProUT [32] to extract instances for concepts in the seed ontology. Then lexico-syntactic patterns [53] are applied to identify possible new concepts, which are then filtered with the help of heuristics and knowledge obtained from online resources, such as Wikipedia[4], Wiktionary[5] and DWDS[6]. Finally, WEB→KB uses both a statistical and a symbolic approach (FOIL [88]) to learn classifiers that can detect instances and relations between instances. The three systems that do not use machine learning either employ an external, publicly available term/synonym extraction engine or require manually-constructed patterns as input, as they seem to rely mostly on linguistic information. The LEILA system also relies on linguistic knowledge, but employs additional filtering based on statistical approaches, such as adaptive k-Nearest-Neighbor-classifiers and Support Vector Machines. BOEMIE also uses machine learning. In particular, the term/synonym extraction engine makes use of both linguistic information (especially shallow syntactic analysis) and machine learning to identify concept instances and relations, while automated pattern extraction is used for relation extraction.

Degree of automation. This criterion examines the extent to which the domain expert needs to intervene during knowledge acquisition. With the exception of Adaptiva, all other systems examined here do not require interaction with the domain/ontology expert. This is an indication that the population process can be fully automated, which is also true for the approach adopted in BOEMIE. BOEMIE directly populates an ontology instead of producing an intermediate representation of instances. In addition, BOEMIE provides a graphical user-interface that enables the domain expert to examine and revise the populated instances, if such a need arises.

Consistency maintenance and redundancy elimination. These issues are only addressed by three systems (Artequakt, SOBA and BOEMIE). The Artequakt system uses manually-written heuristics, in order to merge populated instances that refer to the same real object or event. SOBA, on the other hand, performs simple checks during instance creation, i.e., before the instances populate the ontology, in order to reuse instances that refer to the same real object or event instead of creating new ones. The BOEMIE approach enhances the Artequakt proposal through the use of matching techniques instead of manually developed heuristics. More specifically, BOEMIE instance matching methods try to identify instances that refer to the same real entity or event and group them, rather than merging them into a single instance.

[4] http://en.wikipedia.org/
[5] http://en.wiktionary.org/
[6] http://www.dwds.de/

Domain portability. Some of the systems are domain-independent (KnowItAll, Adaptiva, LEILA, OPTIMA, ISOLDE, BOEMIE), as they do not use any domain-specific resources, while others are domain specific (SOBA, [85, 86] and [7]. There are also some systems that have limited portability, such as Artequakt and WEB→KB. The reason for this is either that they are applicable only to domains with specific characteristics, or that they require adaptation to the new domain, in ways not tested in their current work.

Corpora Modality. All the mentioned systems with the exception of BOEMIE are applied to text. No special effort has been made for other modalities, such as video, images or multimedia. BOEMIE explores this direction, by analysing multimedia corpora. The BOEMIE system supports the identification of objects from multiple modalities (such as text, image, video, audio and text from image/video OCR), which are then fused through reasoning (employing both deduction and abduction) to form instances of modality-independent concepts.

4 Ontology Enrichment

Ontology enrichment is the process of extending an ontology, through the addition of new concepts, relations and rules. It is performed every time that the existing domain knowledge is not sufficient to explain the information extracted from the corpus. Thus, the ontology enrichment activity is expected to extend the background knowledge, in order to better explain extracted information in the future. Since new concepts and relations can be added during enrichment, the structure of the ontology changes. Recalling our discussion of the "layer cake", the enrichment process involves all of the layers presented in section 2, unlike ontology population which is concerned only with the lower layers. The main approach adopted by the state-of-the-art methods starts with the identification of objects and their alternative realisations/synonyms. Each object, along with a possible set of alternative realisations, is a candidate concept to be added to an ontology. Advancing to the third layer of the "cake", each proposed cluster of objects and alternative realisations that possibly represent a concept must be evaluated in order to decide whether it constitutes a concept or not. In case the object represents a concept, the concept must be formulated by creating an intentional definition (section 2.3) and possibly augmented with evidence/instances that justify the addition of the new concept. At the next layer, relations (either taxonomic or non-taxonomic) must be identified between concepts, usually based on spatio-temporal information for modalities like image and video or linguistic information (either syntactic or semantic) for text. Finally, in order to support reasoning and derive facts not explicitly encoded but derivable from the ontology, rules and constraints must be acquired.

4.1 The BOEMIE Approach to Ontology Enrichment

Unlike ontology population which can be fully automated, ontology enrichment remains typically a semi-automated procedure. All systems presented in the literature require the manual intervention of a domain expert, in order to review and accept or reject the system's proposals (Fig. 4). The methodology proposed by the BOEMIE

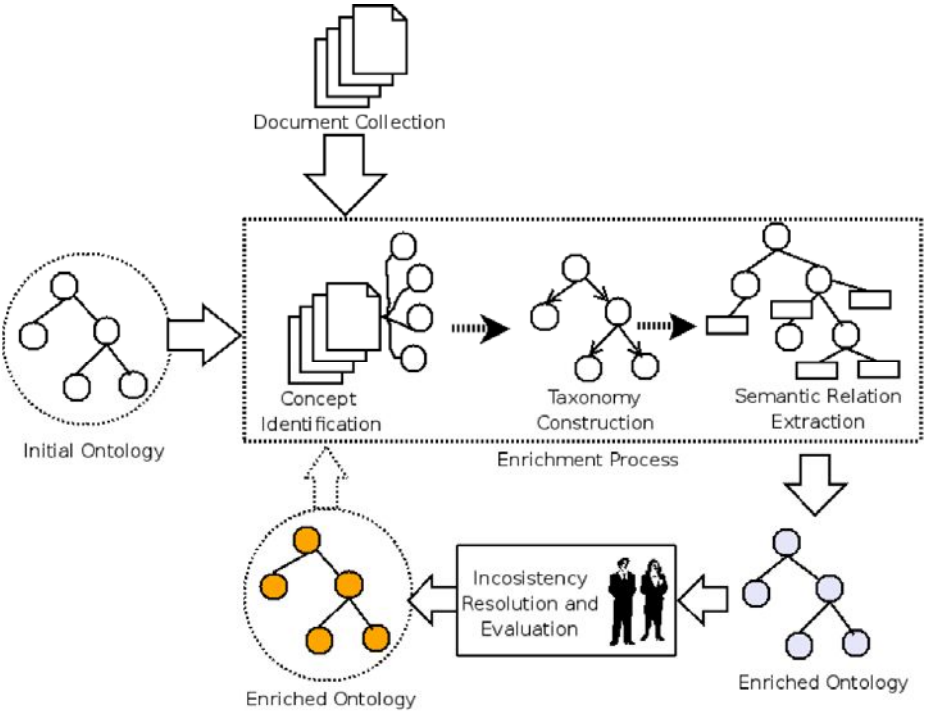


Fig. 4. The ontology enrichment process

project is not an exception. BOEMIE proposes a semi-automated approach which tries to minimise the role of the expert as much as possible.

As in ontology population, a two-stage approach is used. That is, the system distinguishes between high-level and mid-level concepts, as introduced in subsection 3.1. Ontology enrichment in BOEMIE is driven by the quality of the interpretation achieved for a multimedia resource: if a sufficient number of MLCs (properties) have been extracted from the involved modalities, and a large percent of these MLC instances have been successfully interpreted (through their relation to HLC instances), the background knowledge (ontology) is considered as sufficient to describe the multimedia resource. Ontology enrichment is triggered when the background knowledge is not sufficient to interpret adequately a resource: if a significant number of MLC instances are not part of the interpretation (i.e. not related to HLC instances), then the system tries to enrich the ontology through the addition of new HLC concepts. Similarly, if an inadequate number of MLC instances have been identified for one or more modalities, the system tries to enrich the ontology through the addition of new MLC concepts, by triggering the relevant modality-specific enrichment process for the involved modalities. Both enrichment processes rely on clustering techniques to perform proposal of possible new MLCs/HLCs, which are then enhanced with the use of external knowledge sources, through ontology matching techniques, before presented to a domain expert for final verification/approval. Once a concept has been approved

for inclusion into the ontology, the required fusion/interpretation rules used during reasoning are automatically created. Among the innovative aspects of BOEMIE, are the use of non-standard clustering, which tries to cluster ontological fragments, and the use of external knowledge sources aiming to provide the expert additional information during concept and relation definition. More information about this approach can be found in [23].

4.2 Comparative Analysis of Ontology Enrichment Tools

In this subsection we perform a comparative analysis of the most influential ontology enrichment systems. Table 2 presents the systems along with a brief description.

Elements learned. Some of the examined systems are more complete than others, in the sense that they cover several layers of the "cake" presented in section 2. Systems like ASIUM [39, 40], HASTI [97, 98, 100], TEXT-TO-ONTO [77], VIKEF[7] (Virtual Information and Knowledge Environment Framework) and KAON [79] perform learning of new concepts, relations and in some cases even rules. On the other hand, systems like SYNDIKATE [52], ABRAXAS [17, 55], ATRACT [82], and [104] concentrate on concept or relation learning. The BOEMIE ontology enrichment methodology incorporates methods to extract concepts, hierarchical and non-hierarchical relations and rules.

Initial requirements. Almost all systems rely on some form of linguistic analysis, exploiting syntactic relations to identify new concepts, relations or even rules. Besides linguistic knowledge, only a few systems require additional background knowledge, such as a domain ontology, domain specific lexicons or lexicon-syntactic patterns (SYNDIKATE, ABRAXAS, VIKEF, ATRACT). The BOEMIE approach follows a slightly different direction, as it has no initial requirements. Operating solely on the results of information extraction that have been enhanced through reasoning, BOEMIE learns concepts and relations through instance clustering. Furthermore, it tries to associate unknown objects with existing concepts/relations, through the use of external knowledge sources.

Learning approach. Machine learning seems to be the choice of most of the systems, especially in the form of clustering (e.g. ASIUM, HASTI, TEXT-TO-ONTO, KAON and BOEMIE) or lexico-syntactic pattern acquisition (ABRAXAS). BOEMIE also uses clustering on the results of multimedia interpretation through reasoning, rather than at the term/synonym level which is the common approach. As a result, clustering in BOEMIE effectively operates on ontological instances.

Degree of automation. In contrast to ontology population, the enrichment process cannot be fully automated, at least by the existing systems. Most systems interact with an ontology expert who has the final word on the modification of the ontology. Those systems that do not involve the expert either require significant background knowledge and/or support very limited knowledge acquisition (e.g. SYNDIKATE, ABRAXAS, VIKEF, ATRACT, [104]). SYNDIKATE requires an almost completeontology, which can be augmented with new concepts originating from unknown

[7] http://cordis.europa.eu/ist/kct/vikef_synopsis.htm, http://www.vikef.net/

Table 2. Brief description of ontology enrichement systems

System	Description
ASIUM	Learns terms, synonyms, concepts and hierarchical relations from unrestricted text corpora, based on syntactic analysis. It employs machine learning (hierarchical clustering) in order to learn concept hierarchies, with manual supervision by the domain expert.
HASTI	Learns terms, concepts, hierarchical and non-hierarchical relations and axioms in incremental and non-incremental modes. It starts from a small kernel ontology, using a hybrid approach, combining logical, linguistic, template-driven, and heuristic methods.
SYNDIKATE	A system for automatically acquiring knowledge from real-world texts and representing it into formal structures. Through reasoning, an unknown term is either added to an existing concept or creates a new one.
TEXT-TO-ONTO	Learning concepts and relations from unstructured, semi-structured, and structured data, using a multi-strategy method which combines association rules, formal concept analysis and clustering.
ABRAXAS	Performs concept and relation extraction, using automated lexico-syntactic pattern acquisition. This process spots all instances of concepts and relations already in the ontology and acquires extraction patterns using machine learning. These patterns are subsequently applied to the corpus, in order to detect new concepts and relations, the plausibility of which is accessed by a statistical measure.
KAON	Provides components for each subtask of the learning process. It contains an algorithmic library that supports clustering, classification and other techniques. It learns concepts, taxonomic relations and other general binary relations between concepts.
[104]	Learns instances of relations from unstructured corpora. It extracts triples that represent relations between entities/terms. The system employs various metrics for filtering the list of extracted triples in order to decide if a new relation has been discovered.
VIKEF	The system proposes a methodology for extracting information from product catalogues, aimed by an ontology to provide domain knowledge and guide the disambiguation process. The domain ontology can be enriched with parts from other ontologies, selected from a pool of ontologies.
ATRACT	Used for terminology recognition and clustering based on the C/NC-value method (a method for the automatic extraction of multi-word terms, which combines linguistic and statistical information) [43]. It specialises to the domain of molecular biology.
BOEMIE	BOEMIE employs an OBIE extraction engine along with a semantic interpretation engine orchestrated by a bootstrapping approach in order to enrich a seed ontology. The system continuously monitors the quality of interpretations achieved for multimedia resources and performs ontology enrichment when the background knowledge is found inadequate to interpret a set of resources, through a semi-supervised approach. Concept proposals expressed in natural language are automatically generated by exploiting both internal and external knowledge, which must be revised and approved by a domain expert.

terms. However, these concepts can be added mainly near the existing conceptual taxonomy, assuming that there is resemblance in the syntactic usage of the unknown term and concept lexicalisations already in the ontology. ATRACT serves mainly as a workbench for terminology recognition and clustering and is mainly targeting the domain of molecular biology. VIKEF also uses an initial ontology, which is created using a subset of the taxonomical glossary obtained from a product catalogue. This ontology forms the basis for the development of the final ontology about product catalogues. VIKEF applies pattern matching techniques to identify individual product descriptions. For each identified product, its natural language description is processed in order to identify relevant entities and relations between them. The learning process takes advantage of the results of the extraction to enrich the ontology. In addition, similar existing ontologies or parts of them are retrieved from a pool of available ontologies, and they are used to extend the domain ontology. ABRAXAS uses three external resources, namely a corpus of text, some lexico-syntactic textual patterns and an ontology. It considers ontology learning as a process that maintains these resources in some form of equilibrium, as a change in one resource triggers actions in the rest of the resources, in order to reach a consistent overall state. Specia and Motta [104] concentrate mainly on relation identification, thus supporting a very limited type of enrichment. BOEMIE belongs in the family of methods that interact with a domain expert, thus implementing a semi-automated approach to enrichment. However, BOEMIE aims to automate as many tasks as possible, employing also the use of diverse knowledge sources, in order to help the domain expert. It is worth noting that BOEMIE needs a domain expert and not an ontology expert, presenting in a natural-language format only part of the ontology. For example, when a cluster is identified as a candidate concept, a formal definition of the concept is automatically induced along with the required interpretation rules, augmented with its instances. In addition, external knowledge sources, such as other ontologies or Web directories sharing the same or similar thematic domain, are aligned to the concepts of the BOEMIE ontology and used to further enhance the suggested formal definition of a concept. Following the TEXT-TO-ONTO paradigm, BOEMIE provides a natural user interface to the domain expert, who is requested to revise, if needed, and approve the proposed definition. More details about the methodology proposed by BOEMIE can be found in [23].

Consistency maintenance and redundancy elimination. BOEMIE puts significant effort in maintaining the consistency of the ontology while at the same time keeping the ontology clean from redundant information. Consistency maintenance is an automated process performed with the help of reasoning, while redundancy elimination is performed mainly by the domain expert, who is responsible to evaluate whether the supportive information (i.e. clustered instances) for a new concept/relation is enough to justify its addition. Alternatively, this information can be associated with an existing concept/relation.

Domain portability. Most of the presented systems are domain independent, except SYNDIKATE and VIKEF that require significant background knowledge.

Corpora modality. As in the case of ontology population, most of the systems focus on text corpora. Only VIKEF uses both text and images extracted from product catalogues. BOEMIE goes a step further and tries to combine various modalities, such as text, images, video and audio.

5 Evaluation

Evaluation in the context of ontology learning measures the quality of a learned ontology with respect to some particular criteria, in order to determine the plausibility of the learned ontology for the purposes it was built for. Approaches for evaluating learned ontologies can be distinguished into four major categories:

- "Gold standard" evaluation: the learned ontology is compared to a predefined (and usually manually-constructed) "gold standard" ontology.
- Application-based evaluation: the learned ontology is used in an integrated system and is implicitly evaluated through the evaluation of the complete integrated system.
- Data-driven evaluation: the learned ontology is evaluated through comparison with a data source covering the same domain as the learned ontology.
- Human evaluation: the learned ontology is examined/evaluated by domain experts based on predefined criteria, requirements, standards, etc.

An ontology can be evaluated at different layers, such as:

- Lexical, vocabulary or data layer. The evaluation here focuses on which concepts and instances have been included in the ontology and the vocabulary used to identify them.
- Relational layer. The evaluation of this layer deals with the relations between the concepts of the ontology:
  - Hierarchy, taxonomy. An ontology almost always includes hierarchical inclusion relations between its concepts. Thus, the evaluation of these taxonomic relations is very important.
  - Semantic relations. This layer of the ontology concerns other relations besides inclusion and can be evaluated separately.
- Structure, architecture. At this layer we assess whether the design of the ontology has followed some predefined strategies and if it is possible to further develop the ontology easily.
- Philosophical layer. At this level we evaluate the ontology against highly general ontological notions, drawn from the field of philosophical ontology. Thus, we want to decide whether a property of a concept is essential for the specific concept, whether a concept is easily identified among others, etc.

The majority of the evaluation approaches fall into the first category, i.e. gold standard evaluation, and the last category, i.e. evaluation by humans. These categories can also be combined and thus, they are commonly viewed as different sides of the same coin. In what follows, we will discuss these two categories in more detail, while we will give some insights regarding the application-based and the data driven evaluation.

5.1 "Gold Standard" Evaluation

During the "gold standard" evaluation, a learned ontology is compared to a predefined ontology which is considered to be "correct" and which is usually developed by domain

experts. A typical strategy for evaluating against a "gold standard" ontology is as follows: As a first step, the "gold standard" ontology must be created, an action usually performed manually by the domain experts. Then, the "gold standard" ontology is deliberately damaged, usually some concepts, relations and rules are removed from the ontology. At the third step, the pruned ontology is enriched with ontology learning. What is measured is the degree to which learning managed to reconstruct the pruned knowledge.

The comparison can be performed at various levels of the ontology. At the lexical level various string similarity measures can be used, such as the Levenshtein edit distance [65], in order to measure the similarity of concept and relation names. The evaluation at this point is usually performed by measuring Term/Lexical Precision and Term/Lexical Recall [90]. At the relational level, precision and recall can also be used, in order to determine how many identified relations are correct and how many relations of the "golden standard" ontology were found. An interesting approach is presented in [78] based on the notion of *semantic cotopy*. The semantic cotopy of a concept in a given taxonomy is the set of its super and sub-concepts. The overlap of the semantic cotopies of two concepts can be used as a similarity measure between the two concepts. The *taxonomic similarity of concepts* [33, 89] compares the relative placement of concepts in the ontology, based on their distance (shortest path) to other concepts. This set of distances can be used to compare the learned ontology to the "golden standard". Similar ideas have been proposed in [80], where the measures of Augmented Precision and Recall have been used to measure the similarity between two ontologies, taking into account the distance of each concept from the root. Treating the hierarchical backbone as a partition of instances, the evaluation can also be performed using the OntoRand index [14]. This approach measures the similarity between concepts of different hierarchies based either on their common ancestors, their distances in the hierarchy, and the overlap of their sets of instances. Finally, the work in [116] introduces the measures of P-value and R-value, which measure the similarity between ontologies based on the cotopy sets of the concepts and the distance of the concepts, when treated as probability distributions over their instances.

Evaluation against a "gold standard" is an interesting approach but it also has some drawbacks. Besides the obvious problem of constructing manually the "gold" ontology, this approach is somewhat "subjective". The "gold" ontology models a domain in a specific way, chosen by the domain experts that crafted the ontology. Bad evaluation results of a learned ontology do not necessarily mean that the learned ontology is wrong. It is possible that the learned ontology conceptualises the domain with a slightly different model or even captures information not addressed by the domain experts and thus not contained in the "gold" ontology. Thus, the same learned ontology may exhibit different scores with a slightly modified "gold" ontology. Finally, the results of this method are affected by the quality of the matching between the learned and the gold ontology. Thus, a correct ontology matching [36, 81] between the two ontologies is of particular importance, in order to derive meaningful conclusions and penalize accordingly the learned ontology. A combination of matching methods with the measures of P-value and R-value and a relevant discussion can be found in [117, 118].

5.2 Application-Based Evaluation

An important reason for creating an ontology is, among others, to be used in a specific application. Thus, a reasonable approach in evaluating an ontology is to evaluate the performance of the system that uses this ontology, assuming of course that the quality of the ontology plays a role in the performance of the system. Possible measurable objectives in the performance of a system may include low query computation effort, efficient reasoning with the ontology, correctness and completeness of the provided answers. A disadvantage of this evaluation approach is that the results are affected by the dependency of the system on the used ontology. In other words, the evaluation figures depend on the way the ontology is used by the system and the aspects of the ontology that are exploited. As a result, various ontology aspects may not be evaluated.

Although many papers report good results and successful applications of learned ontologies in various tasks, the first experimental conclusions are given in [48]. In this work, the ontology supported a speech recognition task and its role was to determine how closely related the meaning of two concepts was. The task was to assign the correct senses to ambiguous lexical items. These senses were provided by the ontology concepts. The accuracy of the senses assigned to the lexical items was measured against a gold standard.

Similarly, the peculiarities of application-based ontology evaluation are also examined in [87], in the task of tagging the ontological relations that hold between ontologically marked-up entities. This mark-up is obtained from a concept tagging system and constitutes a form of sense disambiguation, whereby the specific senses correspond to items of the ontology's vocabulary. The authors measure the accuracy of the tagging task with respect to ground truth. In addition, they notice various shortcomings of the learned ontology, when comparing the results against those obtained with a gold ontology.

5.3 Data-Driven Evaluation

An ontology may also be evaluated on existing data sources. These are usually collections of text documents, Web pages or dictionaries. The most important requirement for these data sources is to be representative and to cover the domain of the ontology.

Data-driven evaluation has been applied at the lexical [110], and the relational [16] layer of the ontology. This kind of evaluation is particularly suitable for evaluating ontologies learned from textual sources, since we can use a corpus of documents as facts to check whether these facts can be logically derived from the ontology. The metrics of precision and recall are applicable, since they provide an indication of the information that the learning algorithm has captured from the document collection.

Evaluation can also be performed using a set of domain-specific terms or concepts extracted from a corpus, which is compared against the concepts in the ontology. The overlap of the two sets measures the fit between the ontology and the corpus [16]. In the special case that the learned ontology is the result of a document clustering algorithm, it can be evaluated against pre-categorized document collections, such as the Reuters corpus.

Data-driven evaluation requires representative and domain-specific data. Consequently, a question usually arises regarding the choice of the datasets that will be used for the evaluation and how to measure whether they are representative or not.

5.4 Human Evaluation

In human evaluation, the ontology is assessed by human experts, based on desired predefined criteria. The evaluation can be performed by ontology experts, usually the ones that have designed the ontology learning system, users testing the ontology in applications or both. Features evaluated by ontology experts usually include ontology consistency, completeness or conciseness of the model implemented by the ontology. Users on the other hand are interested in the applicability of the ontology to a target task.

The OntoMetric [72, 73] methodology is an example of a principled ontology evaluation by the users of the ontology. A tool is introduced which helps users determine the suitability of an ontology for a particular application, allowing them to compare the importance of the ontology objectives and carefully evaluate its characteristics based on multiple criteria.

A set of ten criteria that can be used for ontology evaluation, are presented in [11]. These criteria cover various ontology aspects like richness, i.e. number of features used, and lawfulness, i.e. frequency of errors, interpretability, clarity, comprehensiveness, accuracy, relevance, authority and history.

A different view to human evaluation focuses on the competence of the ontology [42]. Competence is measured by constructing queries in such a manner that helps the evaluator to check if the ontology meets predefined requirements. A set of generic criteria that are proposed in this work include: (a) efficient reasoning, (b) minimality, i.e. if the ontology contains only the necessary information, (c) functional completeness, i.e. if the ontology can represent the required information to support some task, (d) generality, i.e. if it can be shared among domains, and (e) perspicuity, i.e. if it is easily understood by the users.

From a philosophical point of view, the notion of rigidity, introduced in [46], can be used to check the taxonomical structure of the ontology. Rigidity is based on the more abstract notion of essence. A concept is essential for an instance, if and only if the instance is necessarily an instance of this concept among all universes and at all times. This method is supported by the OntoEdit tool. An important drawback of this approach, though, is that much manual tagging of the concepts participating in the ontology is required. AEON [112] is a tool that aims at enhancing this process by automatically tagging the ontology.

5.5 Comparing the Various Approaches

In the above subsections, various approaches for evaluating a learned ontology have been presented. Each of them has different advantages and disadvantages. First, in order to make data-driven evaluation applicable to a particular domain, a substantial set of data about this domain is required. However, it is not always easy to acquire such data, making the approach difficult to adopt. Similarly, application-based evaluation requires the whole application to be evaluated by humans, which is also a difficult task. In addition, evaluation must be performed by multiple users, in order for the evaluation results to have some statistical significance.

Human-based evaluation is the most complete approach, as all aspects of a learned ontology can be measured and evaluated. However, this evaluation approach is difficult to automate and must be supported by special tools, which help humans in the evaluation. The "gold standard" evaluation is a convenient approach for evaluating ontologies that provides a clear view of the performance of the ontology learning, by comparing the ontology to a predefined gold one in an automated way, using various metrics and measures from the field of information retrieval. To our view, all other approaches evaluate ontologies in an abstract way, which is not always operational and meaningful especially if the ontology is decoupled from the application that uses it. In addition, the fact that the "gold standard" ontology is developed manually provides the ontology engineers the opportunity to develop an ontology that will score well in human-defined criteria and is also suitable for the domain of application. Thus, measuring the closeness of a learned ontology to this "gold" ontology performs also an implicit evaluation according to criteria that are used in human evaluation.

6 Conclusions

In this chapter, we have attempted a detailed presentation of the state-of-the-art on ontology learning, focusing on ontology population and enrichment. A generic framework has been proposed, to facilitate the comparative presentation of the most influential approaches found in the literature.

The comparative presentation of both population and enrichment systems leads to a number of interesting conclusions. The first observation concerns the modality of corpora the systems use to learn ontologies. While a significant amount of work has been performed on text corpora, work on other modalities is practically non-existent. A second observation is that work on learning from text relies heavily on linguistic preprocessing, especially syntactic analysis and exploitation of additional resources like thesauri and semantic hierarchies, such as WordNet. This is due to the fact that many practical systems employ a pattern-based approach, especially for the discovery of relations between concepts. Finally, despite the wide use of machine learning, many systems still require significant manual intervention, usually by ontology experts who make the final decisions for modifying the ontology. Systems that perform ontology population seem to require less manual intervention, effectively automating a large portion of the population process.

In this context, BOEMIE addresses a number of problems identified in the state of the art. In particular, BOEMIE works on multimedia corpora instead of text. The distinction made between "primitive" and "composite" concepts helps in making the information extraction process independent of the ontology structure. Also, BOEMIE puts significant effort in handling redundancy and maintaining the consistency of the ontology. The BOEMIE approach supports interaction with a domain expert rather than an ontology expert, as it presents the discovered knowledge in a natural language format. Finally, as the approach is domain-independent, it is expected to have a wide range of applications in different domains.

References

[1] Agichtein, E., Gravano, L.: Snowball: Extracting Relations from Large Plain-Text Collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries (ACM DL), pp. 85–94 (2000)

[2] Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching Very Large Ontologies Using the WWW. In: Workshop on Ontology Construction of the European Conference of A.I., ECAI 2000 (2000)

[3] Ahmad, K., Davies, A., Fulford, H., Rogers, M.: What is a term? The Semi-Automatic Extraction of Terms from Text. John Benjamins Publishing Company, Amsterdam (1994)

[4] Alani, H., Sanghee, K., Millard, E.D., Weal, J.M., Lewis, P.H., Hall, W., Shadbolt, N.: Automatic Extraction of Knowledge from Web Documents. In: Proceeding of (HLT 2003) (2003)

[5] Alani, H., Sanghee, K., Millard, E.D., Weal, J.M., Lewis, P.H., Hall, W., Shadbolt, N.: Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation. In: Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003), Florida, USA (2003)

[6] Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., Shadbolt, N.R.: Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems 18(1), 14–21 (2003)

[7] Alfonseca, E., Ruiz-Casado, M., Okumura, M., Castells, P.: Towards Large-scale Non-taxonomic Relation Extraction: Estimating the Precision of Rote Extractors. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, Sydney, Australia, pp. 49–56 (July 2006)

[8] Aussenac-Gilles, N., Biebow, B., Szulman, S. (eds.): EKAW 2000 Workshop on Ontologies and Texts (2000), http://CEURWS.org/Vol-51/CEUR

[9] Aussenac-Gilles, N., Maedche, A. (eds.): ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontolology Learning (2002), http://www.inria.fr/acacia/OLT2002

[10] Baroni, M., Bisi, S.: Using cooccurrence statistics & the web to discover synonyms in a technical language. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, vol. 5, pp. 1725–1728 (2004)

[11] Burton Jones, A., Veda Storey, C., Sugumaran, V., Ahluwalia, P.: A Semiotic Suite for Assessing the Quality of Ontologies. Data and Knowledge Engineering (2004)

[12] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)

[13] Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical Topic Models and the Nested Chinese Restaurant Process. Advances in Neural Information Processing Systems 16 (2004)

[14] Brank, J., Mladenic, D., Grobelnik, M.: Gold standard based ontology evaluation using instance assignment. In: Proceedings of the EON Workshop (2006)

[15] Brewster, C., Ciravegna, F., Wilks, Y.: User-Centred Ontology Learning for Knowledge Management. In: Andersson, B., Bergholtz, M., Johannesson, P. (eds.) NLDB 2002. LNCS, vol. 2553, pp. 203–207. Springer, Heidelberg (2002)

[16] Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: Proceedings of the International Conference on Language Resources and Evaluation (2004)

[17] Brewster, C., Iria, J., Zhang, Z., Ciravegna, F., Guthrie, L., Wilks, Y.: Dynamic Iterative Ontology Learning. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2007), Borovets, Bulgaria (2007)
[18] Buitelaar, P., Handschuh, S., Magnini, B. (eds.): Proceedings of the ECAI 2004 Workshop on Ontologies, Learning and Population (2004)
[19] Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam (2005) ISBN: 1-58603-523-1
[20] Buitelaar, P., Cimiano, P., Loos, B.: Bringing the Gap between Text and Knowledge. In: Workshop on Ontology Learning and Population (2006)
[21] Buitelaar, P., Cimiano, P., Racioppa, S., Siegel, M.: Ontology-based Information Extraction with SOBA. In: Proceedings of the International Conference on Language Resources and Evaluation, pp. 2321–2324. ELRA (May 2006)
[22] Buitelaar, P., Cimiano, P., Paliouras, G., Spiliopoulou, M.: Proceedings of the ECAI 2008 Workshop on Ontology Learning and Population (OLP3) (2008)
[23] Castano, S., Peraldi, I.S.E., Ferrara, A., Karkaletsis, V., Kaya, A., Möller, R., Montanelli, S., Petasis, G., Wessel, M.: Multimedia Interpretation for Dynamic Ontology Evolution. Journal of Logic and Computation (September 2008)
[24] Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc., New York (2006)
[25] Ciravegna, F., Dingli, A., Petrelli, D.: Document Annotation via Adaptive Information Extraction. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, August 11-15 (2002)
[26] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Slattery, S.: Learning to construct knowledge bases from the World Wide Web. Artificial Intelligence 118, 69–113 (2000)
[27] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: Gate: an architecture for Development of Robust HLT Applications. In: Proceedings of ACL (2002)
[28] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, Phil. USA (2002)
[29] Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge (2005)
[30] Damerau, F.J.: Evaluating domain-oriented multiword terms from texts. Information Processing and Management 29(4), 433–447 (1993)
[31] Downey, O., Etzioni, D., Soderland, S., Weld, D.: Learning Text Patterns for Web Information Extraction and Assessment. In: Proceedings of the AAAI Workshop on Adaptive Text Extraction and Mining (2004)
[32] Drozdzynski, W., Krieger, H.-U., Piskorski, J., Schäfer, U., Xu, F.: Shallow processing with unification and typed feature structures – foundations and applications. Künstliche Intelligenz 1, 17–23 (2004)
[33] Ehrig, M., Haase, P., Stohanovic, N., Hefke, M.: Similarity for Ontologies – a Comprehensive Framework. In: Proceedings of the European Conference in Inf. Sys. (2005)
[34] Etzioni, O., Kok, S., Soderland, S., Cagarella, M., Popescu, A.M., Weld, D.S., Downey, D., Shaker, T., Yates, A.: Web-Scale Information Extraction in KnowItAll (Preliminary Results). In: Proceedings of the 13th International World Wide Web Conference (WWW 2004), New York, pp. 100–110 (2004)

[35] Etzioni, O., Kok, S., Soderland, S., Cagarella, M., Popescu, A.M., Weld, D.S., Downey, D., Shaker, T., Yates, A.: Unsupervised named-entity extraction from the Web: An experimental Study. Artificial Intelligence 165, 91–134 (2005)
[36] Euzenat, J., Pavel, S.: Ontology Matching. Springer, Heidelberg (2007)
[37] Faatz, A., Steinmetz, R.: Ontology Enrichment with texts from the WWW. In: Semantic Web Mining 2nd Workshop at ECML/PKDD-2002. Helsinki, Finland (2002)
[38] Fellbaum, C.: WordNet: An On-Line Lexical Database and Some of its Applications. MIT Press, Cambridge
[39] Faure, D., Nedellec, C., Rouveirol, C.: Acquisition of Semantic Knowledge using Machine Learning Methods: The System ASIUM, Technical Report number ICS-TR-88-16, Laboratoire de Recherche en Informatique, Inference and Learning Group, Universite Paris Sud (1998)
[40] Faure, D., Poibeau, T.: First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In: Proceedings of the ECAI 2000 Workshop on Ontology Learning (OL 2000) (2000)
[41] Fortuna, B., Mladevic, D., Grobelnik, M.: Visualization of Text Document Corpus. In: ACAI 2005 Summer School (2005)
[42] Fox, M.S., Barbuceanu, M., Gruninger, M., Lin, J.: An Organization Ontology for Enterprise Modelling. MIT Press, Cambridge (1998)
[43] Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: The c-value/nc-value method. International Journal on Digital Libraries 3(2), 115–130 (2000)
[44] Gruber, T.: Towards principles for the design of ontologies used for knowledge sharing. Int. J. of Human and Computer Studies 43, 907–928 (1994)
[45] Gómez-Pérez, A., Manzano-Macho, D.: A survey of ontology learning methods and techniques. Onto-web IST Project, Deliverable 1.5, http://www.ontoweb.aifb.uni-karlsruhe.de/Members/ruben/Deliverable%201.5
[46] Guarino, N., Welty, C.: Evaluating ontological decisions with ontoclean. Communications of the ACM 45(2), 61–65 (2002)
[47] Gupta, K.M., Aha, D., Marsh, E., Maney, T.: An Architecture for engineering sublanguage WordNets. In: Proceedings of the First International Conference On Global WordNet, pp. 207–215. Central Institute of Indian Languages, Mysore (2002)
[48] Gurevych, I., Malaka, R., Porzel, R., Zorn, H.: Semantic coherence scoring using an ontology. In: Proceedings of the HLT/NAACL (2003)
[49] Haase, P., Stojanovic, L.: Consistent Evolution of OWL Ontologies. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 182–197. Springer, Heidelberg (2005)
[50] Haase, P., van Harmelen, F., Huang, Z., Stuckenschmidt, H., Sure, Y.: A Framework for Handling Inconsistency in Changing Ontologies. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 353–367. Springer, Heidelberg (2005)
[51] Haase, P., Völker, J.: Ontology Learning and Reasoning — Dealing with Uncertainty and Inconsistency. In: da Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) URSW 2005 - 2007. LNCS (LNAI), vol. 5327, pp. 366–384. Springer, Heidelberg (2008)
[52] Hahn, U., Marko, K.G.: Ontology and Lexicon Evolution by Text Understanding. In: Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT 2002), Lyon, France (2002)

[53] Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France (1992)
[54] Harris, Z.: Mathematical Structures of Language. John Wiley & Sons, Chichester (1968); Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 268–275 (1990)
[55] Iria, J., Brewster, C., Ciravegna, F., Wilks, Y.: An Incremental Tri-Partite Approach To Ontology Learning. In: The 5th International Conference on Language Resources and Evaluation, May 24-25-26, pp. 24–25 (2006)
[56] Iwanska, L.M., Mata, N., Kruger, K.: Fully Automatic Acquisition of Taxonomic Knowledge from Large Corpora of Texts, pp. 335–345. MIT/AAAI Press (2000)
[57] Kietz, J.U., Maedche, A., Volz, R.: A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In: Proceedings of the ECAW 2000 Workshop Ontologies and Text, Juan-Les-Pins, France (2000)
[58] Kim, S.-S., Son, J.-W., Park, S.-B., Park, S.-Y., Lee, C., Wang, J.-H., Jang, M.-G., Park, H.-G.: OPTIMA: An Ontology Population System. In: 3rd Workshop on Ontology Learning and Population (July 2008)
[59] Kim, S., Alani, H., Hall, W., Lewis, P., Millard, D., Shadbolt, N., Weal, M.: Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web. In: Proceedings of Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002), the 15th European Conference on Artificial Intelligence (ECAI 2002), Lyon, France, pp. 1–6 (2002)
[60] Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., Cofino, T.: Glossary extraction and utilization in the information search and delivery system for IBM Technical SupportΣ. IBM System Journal 43(3) (2004)
[61] Krauthammer, M., Nenadic, G.: Term identification in the biomedical literature. Journal of Biomedical Informatics 37, 512–526 (2004)
[62] Lafferty, J.D., McCallum, A., Pereira, F.C.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML) (2001)
[63] Landauer, T.K., Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review 104, 211–240 (1997)
[64] Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: The Fifth International Conference on Systems Documentation, ACM SIGDOC (1986)
[65] Levenshtein, I.V.: Binary codes capable of correcting deletions, insertions and reversals. Cybernetics and Control Theory 10(8), 707–710 (1966)
[66] Li, W., McCallum, A.: Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. In: Proceedings of the 23rd Internationcal Conference on Machine Learning, pp. 577–584 (2006)
[67] Li, W., Blei, D., McCallum, A.: Nonparametric Bayes Pachinko Allocation. In: Uncertainty in Artificial Intelligence (2007)
[68] Lin, D., Pantel, P.: Induction of semantic classes from natural language text. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 317–322 (2001)

[69] Lin, D., Pantel, P.: Dirt - Discovery of Inference Rules from Text. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 323–328 (2001)
[70] Lin, D., Pantel, P.: Concept discovery from text. In: Proceedings of the International Conference on Computational Linguistics (COLING), pp. 577–583 (2002)
[71] Lisi, F.A.: Principles of Inductive Reasoning on the Semantic Web: A Framework for Learning in AL-Log. In: Fages, F., Soliman, S. (eds.) PPSWR 2005. LNCS, vol. 3703, pp. 118–132. Springer, Heidelberg (2005)
[72] Lozano-Tello, A., Gomez-Perez, A., Sosa, E.: Selection of Ontologies for the Semantic Web, pp. 413–416. Springer, Heidelberg (2003)
[73] Lozano-Tello, A., Gomez-Perez, A.: Ontometric: A method to choose the appropriate ontology. Journal of Database Management. Special Issue on Ontological Analysis, Evaluation, and Engineering of Business Systems Analysis Methods 15(2), 1–18 (2004)
[74] Maedche, A., Staab, S.: Semi-Automatic Engineering of Ontologies from Text. In: Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering (2000)
[75] Maedche, A., Staab, S.: Discovering Conceptual Relations from Text. In: Proceedings of ECAI 2000. IOS Press, Amsterdam (2000)
[76] Maedche, A., Staab, S., Nédellec, C., Hove, E. (eds.): IJCAI 2001 Workshop on Ontology Learning (2001), http://CEUR-WS.org/Vol-38/CEUR
[77] Maedche, A., Staab, S.: Ontology learning for the Semantic Web. IEEE Journal on Intelligent Systems 16(2), 72–79 (2001)
[78] Maedche, A., Staab, S.: Measuring Similarity between Ontologies. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 251–263. Springer, Heidelberg (2002)
[79] Maedche, A., Staab, S.: Ontology Learning. In: Handbook on Ontologies (2004)
[80] Maynard, D., Peters, W., Li, Y.: Metrics for evaluation of ontology based information extraction. In: Proceedings of the EON 2006 Workshop (2006)
[81] Meilicke, C., Völker, J., Stuckenschmidt, H.: Learning Disjointness for Debugging Mappings between Lightweight Ontologies. In: Gangemi, A., Euzenat, J. (eds.) EKAW 2008. LNCS (LNAI), vol. 5268, pp. 93–108. Springer, Heidelberg (2008)
[82] Mima, H., Ananiadou, S., Nenadic, G.: The atract workbench: Automatic term recognition and clustering for terms. In: Matoušek, V., Mautner, P., Mouček, R., Tauser, K. (eds.) TSD 2001. LNCS (LNAI), vol. 2166, p. 126. Springer, Heidelberg (2001)
[83] Mimno, D., Li, W., McCallum, A.: Mixtures of Hierarchical Topics with Pachinko Allocation. In: Proceedings of the 24th International Conference on Machine Learning, pp. 633–640 (2007)
[84] Morin, E.: Automatic Acquisition of Semantic Relations Between Terms from Technical Corpora. In: Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering - TKE 1999 (1999)
[85] Navigli, R., Velardi, P.: Enriching a Formal Ontology with a Thesaurus: an Application in the Cultural Heritage Domain. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, Sydney, Australia, pp. 1–9 (July 2006)
[86] Navigli, R., Velardi, P.: Ontology Enrichment Through Automatic Semantic Annotation of On-Line Glossaries. In: Staab, S., Svátek, V. (eds.) EKAW 2006. LNCS (LNAI), vol. 4248, pp. 126–140. Springer, Heidelberg (2006)
[87] Porzel, R., Malaka, R.: A task-based approach for ontology evaluation. In: ECAI 2004 Workshop on Ontology Learning and Population (2004)

[88] Quinlan, J.R.: Learning logical definitions from relations. Machine Learning 5, 239–266 (1990)
[89] Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 17–30 (1989)
[90] Sabou, M., Wroe, C., Goble, C., Stuckenschmidt, H.: Learning domain ontologies for semantic web service descriptions. Journal of Web Semantics 3(4) (2005)
[91] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
[92] Sangun, P., Juyoung, K., Wooju, K.: A Framework for Ontology Based Rule Acquisition from Web Documents in Web Reasoning and Rule Systems (2007)
[93] Sclano, F., Velardi, P.: TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. In: 9th Conf. on Terminology and Artificial Intelligence TIA 2007, Sophia Antinopolis (October 2007)
[94] Schütze, H.: Word space. Advances in Neural Information Processing Systems 5 (1993)
[95] Schutz, A., Buitelaar, P.: RelExt: A Tool for Relation Extraction from Text in Ontology Extension. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 593–606. Springer, Heidelberg (2005)
[96] Shadbolt, N., Berners-Lee, T., Hall, W.: The Semantic Web Revisited. IEEE Intelligent Systems 21(3), 96–101 (2006)
[97] Shamsfar, M., Barforoush, A.A.: An Introduction to HASTI: An Ontology Learning System. In: Proceedings of 6th Conference on Artificial Intelligence and Soft Computing (ASC 2002), Banff, Canada (June 2002)
[98] Shamsfard, M.: Designing the ontology learning Model, Prototyping in a Persian Text Understanding System, Ph.D. Dissertation, Computer Engineering Dept., AmirKabir University of Technology, Tehran, Iran (January 2003)
[99] Shamsfard, M., Barforoush, A.A.: The state of the art in ontology learning: a framework for comparison. Knowl. Eng. Rev. 18(4), 293–316 (2003), DOI: http://dx.doi.org/10.1017/S0269888903000687
[100] Shamsfar, M., Barforoush, A.A.: Learning Ontologies from Natural Language Texts. International Journal of Human-Computer Studies (60), 17–63 (2004)
[101] Schulte im Walde, S.: Clustering Verbs Semantically According to their Alternation Behaviour. In: Proceedings of the 18th International Conference on Computational Linguistics (COLINGS), pp. 747–753 (2000)
[102] Snow, R., Jurafsky, D., Ng, A.Y.: Semantic Taxonomy Induction from Heterogeneous Evidence. In: ACLY 2006 (2006)
[103] Snow, R., Jurafsky, D., Ng, A.Y.: Learning Syntactic Patterns for Automatic Hypernym Discovery. In: Proceedings of Advances in Neural Information Processing Systems (2004)
[104] Specia, L., Motta, E.: A hybrid approach for extracting semantic relations from texts. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, Sydney, Australia, pp. 57–64 (July 2006)
[105] Staab, S., Maedche, A., Nedellec, C., Wiemer- Hastings, P. (eds.): Proceedings of the Workshop on Ontology Learning (2000), http://CEUR-WS.org/Vol-31/CEUR
[106] Suchanek, F.M., Ifrim, G., Weikum, G.: LEILA: Learning to Extract Information by Linguistic Analysis. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, Sydney, Australia, pp. 18–25 (July 2006)

[107] Turney, P.D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
[108] Yang, H., Callan, J.: A Metric-based Framework for Automatic Taxonomy Induction. In: ACL 2009 (2009)
[109] Yarowsky, D.: Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In: COLING 1992, Nantes (1992)
[110] Velardi, P., Navigli, R., Cuchiarelli, A., Neri, F.: Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies. IOS Press, Amsterdam (2005)
[111] Velardi, P., Cucchiarelli, A., Petit, M.: A Taxomony learning Method and its Application to Characterize a Scientific Web Community. IEEE Transaction on Data and Knowledge Engineering (TDKE) 19(2), 180–191 (2007)
[112] Völker, J., Vrandečić, D., Sure, Y.: Automatic Evaluation of Ontologies (AEON). In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 716–731. Springer, Heidelberg (2005)
[113] Weber, N., Buitelaar, P.: Web-based Ontology Learning with ISOLDE. In: Proceedings of the Workshop on Web Content Mining with Human Language at the International Semantic Web Conference, USA (2006)
[114] Wei, W., Barnaghi, P.: Probabilistic Topic Models for Learning Terminological Ontologies. Transaction on Knowledge and Data Engineering 22(7), 1028–1040 (2010)
[115] Zavitsanos, E., Paliouras, G., Vouros, G.: Ontology Learning and Evaluation: A survey. Technical report, DEMO-(2006-3), NCSR Demokritos, Athens, Greece (2006)
[116] Zavitsanos, E., Paliouras, G., Vouros, G.: A Distributional Approach to Evaluating Ontology Learning Methods Using A Gold Standard. In: 3rd Ontology Learning and Population Workshop, ECAI 2008 (2008)
[117] Zavitsanos, E., Paliouras, G., Vouros, G.A., Petridis, S.: Learning Subsumption Hierarchies of Ontology Concepts from Texts. Web Intelligence and Agent Systems: An International Journal 8(1), 37–51 (2010)
[118] Zavitsanos, E., Paliouras, G., Vouros, G.A.: Gold Standard Evaluation of Ontology Learning Methods Through Ontology Transformation and Alignment. Transactions on Knowledge and Data Engineering (2010) (to appear)

Ontology and Instance Matching

Silvana Castano, Alfio Ferrara, Stefano Montanelli, and Gaia Varese

Università degli Studi di Milano
DICo, Via Comelico, 39, 20135 Milano, Italy
{castano,ferrara,montanelli,varese}@dico.unimi.it

Abstract. The growing need of sharing data and digital resources within and across organizations has produced a novel attention on issues related to ontology and instance matching. After an introductory classification of the main techniques and tools for ontology matching, the chapter focuses on instance matching by providing an accurate classification of the matching techniques proposed in the literature, and a comparison of the recent instance matching tools according to the results achieved in the OAEI 2009 contest. Ontology and instance matching solutions developed in the BOEMIE project for multimedia resource management and ontology evolution are finally presented.

1 Introduction

The growing need of sharing data and digital resources within and across organizations has produced a novel attention on issues related to ontology and instance matching. In this field, the existing solutions have reached a certain degree of maturity and they address a number of general requirements, such as the applicability to different ontology specification languages, the capability to deal with different levels of detail in describing the knowledge of interest, and the necessity to automate as much as possible the matching execution. For this reason, the recent research on ontology matching is more focused on investigating how the existing solutions can work together to enforce a dynamic and custom configuration of the matching process, rather than on designing new approaches/techniques [1,2]. Moreover, ontology matching approaches and tools are getting more and more important in the framework of Semantic Web applications, where not only conventional matching at the schema level, but also and especially matching at the instance level is becoming essential to support discovery and management of different individual descriptions referring to the same real-world entity [3,4,5,6].

This chapter is devoted to survey ontology and instance matching. In particular, after an introductory overview of ontology matching (Section 2), most of the chapter will be focused on instance matching techniques and tools (Section 3). Such a choice has a twofold motivation. The first motivation is related to the different maturity of the two research areas. Ontology matching is nowadays considered as a consolidated research area and a number of surveys already exist on this topic [7,8,9,10,11]. For this reason, we decided to recall an

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 167–195, 2011.

essential classification of ontology matching techniques and to focus the contribution on providing an up-to-date picture of ontology matching tools, including also the more recent ones participating to the 2008 and 2009 editions of the OAEI (Ontology Alignment Evaluation Initiative) contest on ontology matching. With respect to ontology matching, instance matching is a younger research area and, as such, the contribution of this chapter is to provide an accurate classification of the instance matching techniques proposed in the literature, and to present a comparison of the recent tools. Also in this case, we refer to the results of the OAEI 2009 contest, where a specific track on instance matching was organized for the first time. The second motivation for focusing more deeply on instance matching is related to the kind of problems that were faced in the BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction) EU FP6 project. In BOEMIE, the emphasis was on multimedia resource classification and management, for which the capability to match data descriptions at the instance level was actually demanded. The chapter will then describe ontology and instance matching solutions developed in the BOEMIE project, by highlighting their role and contributions to enforce multimedia resource management and ontology evolution (Section 4). Envisaged research trends for ontology and instance matching will conclude the chapter (Section 5).

We remark that in the chapter we use the expression "ontology matching" to denote matching at the schema-level (i.e., matching of concepts and properties represented by the so-called "TBoxes" of DL ontologies). The expression "instance matching" is used to denote matching at the data-level (i.e., matching of assertions represented by the so-called "ABoxes" of DL ontologies). However, it is important to mention that, especially in the matching community, the expression "ontology matching" is usually adopted to denote the activity of matching ontological knowledge in general (both TBoxes and ABoxes). In this case, the expression "concept matching" is used to specifically denote schema-level matching.

2 Ontology Matching

The problem of schema matching, and the more recent problem of ontology matching, have been widely investigated in the literature and a number of approaches and tools have been proposed both in the area of data and knowledge management. A reference survey on schema matching is given in [12], while ontology matching approaches and tools have been surveyed in [7,8,9,10]. Moreover, a book on this topic has been recently published [11]. On the basis of this wide literature, for the purpose of this chapter, we adopt a very general definition of ontology matching and we consider ontology matching as the process (automatically or semi-automatically performed) which takes two ontologies O_1 and O_2 as input and returns a set of mappings between them as output. Each resulting mapping specifies that a certain element of the ontology O_1 corresponds to (i.e., matches) a certain element of O_2.

2.1 Matching Techniques

An introductory classification of techniques for ontology matching distinguishes between two main families of techniques, namely *similarity-based techniques* and *reasoning-based techniques* as shown in Figure 1.

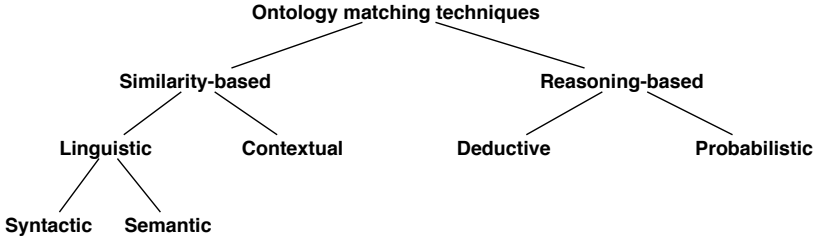


Fig. 1. Main families of ontology matching techniques

Similarity-based techniques. These techniques measure the degree of similarity of two ontology concepts, according to linguistic and contextual criteria and metrics.

Linguistic matching techniques group all the techniques evaluating similarity between ontology concepts on the basis of their names and the names of their properties. These techniques can work according to a syntactic or a semantic approach. By "syntactic" approach, we refer to the fact that only the string structure of the names that are matched is considered. Techniques for string matching [13] like those based on edit-distance [14,15], automata [16], bit-parallelism [17], or filtering [18] algorithms are examples of linguistic matching techniques implementing a syntactic approach. By "semantic" approach, we refer to the fact that linguistic techniques consider also the "meaning" of the names that are matched. Techniques relying on terminological relationships between terms like synonymy and hypernimy/hyponimy and on external dictionaries/thesauri like WordNet [19,20] are examples of linguistic matching techniques implementing a semantic approach.

Contextual matching techniques group all the techniques evaluating similarity between concepts on the basis of their contexts. The context of a concept c is seen as the set of properties, semantic relations, and other concepts that are involved in the ontological definition of c. Contextual matching techniques are typically implemented using graph matching algorithms that represent the context of c as a graph where nodes denote concepts and edges denote properties and semantic relations in the context of c, respectively. Graph matching algorithms evaluate the similarity between two concepts by measuring the topological similarity between their respective context graphs [21,22,23].

Reasoning-based techniques. Reasoning-based techniques consider the ontology matching problem as an inference problem involving two ontologies and an initial set of mappings, manually or automatically defined, between them.

The main goal of these techniques is to infer new mappings between the considered ontologies by applying reasoning techniques. In particular, the initial set of mappings is interpreted as a set of semantic relations holding between the concepts of the two ontologies and automatic reasoning techniques are exploited in order to derive the implications of mappings over the considered ontologies [24].

The main examples of reasoning-based matching techniques are based on deductive reasoning and, in particular, on propositional satisfiability (SAT) and Description Logics (DL). In the case of SAT-based techniques [25], the idea is to derive from the initial set of mappings new candidate mappings. A candidate mapping between two concepts is seen as a hypothetical semantic relation between them that is expressed as a propositional formula (i.e., an implication). The unsatisfiability of the propositional formula is checked by using SAT solvers. If the implication is satisfied, we can conclude that the candidate mapping is correct and can be added to the initial set of mappings. In techniques based on Description Logics, the idea is to see the two ontologies involved in the initial set of mappings as a new distributed TBox where the concepts of the two initial ontologies are correlated by bridge rules derived from the initial mappings [26]. Then, by exploiting subsumptions involving the ontology concepts of the distributed TBox new mappings can be inferred to enrich the initial set of mappings [27,28].

Another approach for reasoning-based techniques is the probabilistic approach. The basic idea here is to calculate the probability that two concepts in two independent ontologies are similar or have the same instances. Many solutions have been proposed to address this problem using machine learning techniques [29,30] or Bayesian networks [31,32]. In both the approaches, given a set of initial mappings between two ontologies it is possible to infer new mappings between two concepts by considering the initial mappings as a projection of a concept of the first ontology on the second ontology. These projections express the joint probabilities holding in the first ontology in another probability space, represented by the second ontology.

2.2 Matching Tools

A number of tools for ontology matching are today available and they have been developed as concrete software applications of the above-mentioned matching techniques. These tools have been progressively enriched during time, starting from initial prototypes capable of performing a prefixed set of matching operations, up to modern matching engines capable of tailoring the matching execution according to the specific scenario at hand. To provide a basic classification, we define two dimensions where ontology matching tools are characterized according to i) the *composition of the dataset to match*, which determines the granularity and the cardinality of the ontology elements to consider for matching, and ii) the *configurability of the matching execution*, which determines the level of flexibility of the matching process. In this respect, we distinguish the following three generations of ontology matching tools.

- *First generation tools* (*meta-model generation*). These matching tools are mainly focused on the problem of schema matching. The dataset to match is constituted by the schema elements of the considered datasources, either structured (e.g., database) or semi-structured (e.g., XML, RDF(S), OWL). The matching execution is mostly *embedded* in the tool, in that the matching process follows a predefined workflow and personalization of the process is not allowed. Examples of tools belonging to this generation are ARTEMIS, Cupid, Glue, TSIMMIS, Clio, MAFRA [9].
- *Second generation tools* (*knowledge generation*). These matching tools focus on providing a wide suite of basic techniques with specific matching goals that can be combined in a flexible way. This kind of tools are mainly conceived to deal with the problem of ontology matching, with specific focus on the concept (TBox) level. The matching execution becomes *dynamic*, which means that the various techniques featuring a matching tool can be invoked alone or in combination to satisfy the specific need of the considered matching scenario. Examples of tools belonging to this generation are FOAM, OLA, PROMPT, COMA++, S-Match, HMatch [9].
- *Third generation tools* (*holistic generation*). These matching tools characterize the current state-of-the-art in the field. This kind of tools are characterized for being "all in one", in that they are capable of working on a dataset to match at both schema and instance level, with a matching execution *incremental/iterative*. This means that matching can require more than one execution to obtain the final result, and the results of the intermediate executions are used to support/refine the subsequent processing phases. Examples of tools belonging to this generation are ASMOV, DSSim, HMatch 2.0, RiMOM. A more detailed discussion of the instance matching capabilities provided by these tools is presented in Section 3.4.

A comparative overview of ontology matching tools is provided in Table 1. In particular, in our comparison, we focus on the most recent tools (second and third generation) that provided high-quality results in the 2008 and 2009 Ontology Alignment Evaluation Initiatives (OAEI) [33].

In Table 1, we observe that the capability to combine different kinds of matching techniques within a given matching execution characterizes most of the considered tools. In particular, the focus is on the support of similarity-based matching where both linguistic and contextual techniques are provided and available for combination (e.g., AFlood, ASMOV, Lily, RiMOM, SAMBO, SOBOM). In this case, the various techniques involved in a certain matching execution are separately invoked and the respective similarity results are finally combined in a weighted sum. In some tools, both similarity- and reasoning-based matching techniques are available for combination (e.g., AROMA, ASMOV, CIDER, TaxoMap). In this case, similarity-based techniques are invoked to calculate an initial set of corresponding elements, which is subsequently used as input for reasoning-based techniques to produce a final set of mappings. In some other tools, like AgrMaker, the focus is more on providing rules and operations to combine different sets of mappings than, on specifying the behavior of the matching execution

Table 1. Overview of the main ontology matching tools (second and third generation)

Tool	Kind of matching	Supported techniques	External resources
AFlood [34]	Similarity-based	Linguistic Contextual	WordNet
AgrMaker [35]	Similarity-based	Contextual	Initial mapping set
AROMA [36]	Similarity-based Reasoning-based	Linguistic	-
ASMOV [37]	Similarity-based Reasoning-based	Linguistic Contextual	WordNet
CIDER [38]	Similarity-based Reasoning-based	Linguistic	WordNet
DSSim [39]	Reasoning-based	Dempster-Shafer theory	Initial mapping set WordNet
GeRoMe [40]	Similarity-based	Contextual	-
KOSIMAP [41]	Reasoning-based	DL reasoning	-
Lily [42]	Similarity-based	Linguistic Contextual	-
MapPSO [43]	Reasoning-based	Discrete particle swarm optimization (DPSO)	Initial mapping set WordNet
RiMOM [44]	Similarity-based	Linguistic Contextual	-
SAMBO [45]	Similarity-based	Linguistic Contextual	Domain dictionaries WordNet
SOBOM [46]	Similarity-based	Linguistic Contextual	-
TaxoMap [47]	Similarity-based Reasoning-based	Linguistic	-

itself. Moreover, we note that linguistic matching techniques are provided by most of the considered tools, often with the support of external domain dictionaries and/or lexical systems (i.e., WordNet). This is due to the fact that linguistic-based techniques are widely recognized as a basic family of matching techniques, which can be invoked i) in a stand-alone way, when it is sufficient to match the linguistic features of the considered dataset, ii) in a combined way with other kinds of matching techniques (e.g., contextual), when the linguistic features are one of the aspect to consider during the matching process.

3 Instance Matching

In the recent years, the research work on ontology matching is gradually shifting from the level of concepts to the level of instances. This is mainly due to the increasing popularity of Web 2.0 and Semantic Web technologies, where data are usually provided with a poor (or totally missing) schema/metadata specification. The relevance of matching instances becomes even more important if

we consider that any real-world entity (e.g., a person, a place, an event) can appear on the web within a number of different documents with heterogeneous representations called *instances* or *individuals*. The problem of recognizing when different instances refer to the same real-world entity is a matter of instance matching as described in the following.

Instance matching. *Given two ontologies O_1 and O_2 as input, instance matching is defined as the process of comparing an instance (or individual) $i_1 \in O_1$ and an instance (or individual) $i_2 \in O_2$, in order to produce as output a similarity measure of i_1 and i_2 together with a mapping between their matching assertions.*

Instance matching is the process of evaluating the degree of similarity of pairs of instances across heterogeneous knowledge sources (e.g., OWL ABoxes, RDF documents, SCORM data) to determine whether they refer to the same real-world entity in a given domain. Usually, the higher is the similarity between two instances, the higher is the probability that they actually refer to the same real-world entity.

Approaches and techniques for instance matching are currently employed in a number of application fields. For example, in the Semantic Web, instance matching is exploited to address the so-called *identity recognition* problem [3]. In this field, instance matching has the goal to support discovery and reuse on the web of a unique identifier for the set of instance descriptions that is recognized as referring to the same real-world entity. In the field of semantic integration, instance matching can be used to determine the set of matching concepts to integrate in two considered knowledge sources. To this end, the similarity between two concepts is evaluated by measuring the "significance" in the overlap of their respective instance sets, and two instances are set as overlapping according to their level of similarity [4,5]. Moreover, instance matching is currently demanded in the field of ontology management where it is invoked to support domain experts in performing ontology changes through advanced, and possibly automated, techniques. For example, instance matching is used to correctly perform the insertion of new instances in a given ontology (i.e., *ontology population*) and to discover the possible similarity mappings between a new incoming instance and the set of instances already represented in the ontology. Mappings among instances can be exploited to enforce a query answering mechanism based on instance similarity. In Section 4, the BOEMIE project will be presented as an example of the possible role of instance matching to support ontology population.

3.1 Matching Techniques

Up to now, techniques for instance matching are mostly borrowed from those developed for *record linkage*. In the database community, record linkage is defined as "the task of quickly and accurately identifying records corresponding to the same real-world entity from one or more data sources" [48]. As this problem is very general, in the literature, it is known under different names (e.g., data

deduplication, duplicate detection), according to the specific requirements that need to be satisfied and to the goals that need to be pursued [4,49].

In the following, we will focus on the problem of instance matching by classifying existing approaches proposed for record linkage and by explaining how they can be/are being used for instance matching purposes.

Main approaches for record linkage were initially proposed for database applications, as a solution for deduplication. In particular, given a set of records $r_1, \ldots, r_n$ as input (i.e., the tuples belonging to one or more database relations), the deduplication process consists in firstly detecting different records referring to the same real-world entity (duplicates or matching records), and secondly in removing duplicates through appropriate record merge/unification operations. For the purpose of this survey, we will focus on record linkage techniques for duplicate detection, since they can be adapted to work on instance matching.

As shown in Figure 2, these techniques can be classified into two different categories, corresponding to two different levels of granularity: the *value-oriented techniques* and *record-oriented techniques*.

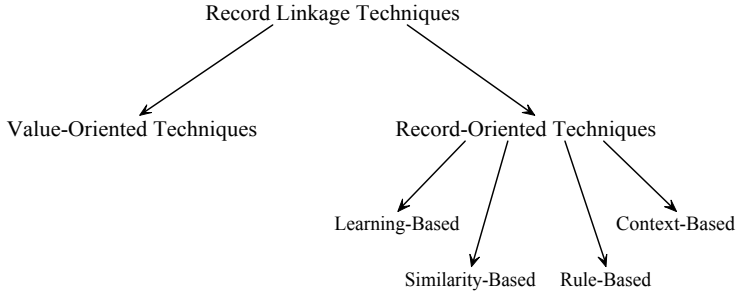


Fig. 2. A basic classification of existing record linkage techniques

For record linkage, a record r_i is represented as a vector $\underline{r_i} = [v_1, \ldots, v_m]$, where m is the number of its featuring attributes and v_j is the value of the j-th attribute. Given a pair of records r_1 and r_2, the goal of value-oriented techniques is to determine the similarity $sim(v_h, v_k)$ of values v_h and v_k, where $v_h \in \underline{r_1}$ and $v_k \in \underline{r_2}$, for each pair of corresponding attributes of r_1 and r_2. Record-oriented techniques aim at computing the overall similarity $sim(r_1, r_2)$ of r_1 and r_2, in order to determine whether r_1 and r_2 refer to the same real-world entity.

Value-oriented techniques. These techniques work at the value granularity under the assumption that the similarity level of two records r_1 and r_2 can be derived by matching the values of their comparable attributes. For each specific attribute datatype, appropriate matching techniques are provided to calculate the similarity of attribute values. As an example, approaches for matching numerical values use *conversion functions* to determine how to transform values of a source datatype (e.g., real values) into corresponding values of a target datatype (e.g., integer values). However, most of the work on value-oriented matching has

been focused on computing similarity of string attributes due to the fact that string data are the most frequently used datatypes in database and knowledge repositories for real-world entity descriptions. Different techniques have been developed in order to manage specific kinds of errors/differences within string values. *Character-based techniques*, like the Edit Distance, the Smith-Waterman Distance, and the Jaro Distance, are specifically suited for comparing string values and recognize typographical errors (e.g., "Computer Science", "Computer Science"). They basically compute the number of common characters of two strings. *Token-based techniques*, like the Cosine Similarity, TF-IDF, and the Q-Gram distance, are able to manage the use of different conventions for describing data (e.g., "John Smith", "Smith, John"). In this case, the similarity of two strings is calculated by analyzing their common patterns (tokens). Finally, *phonetic-based techniques*, like Soundex, NYSIIS, and Metaphone, try to measure the phonetic similarity of different strings, even if their textual representation is very different (e.g., "Kageonne", "Cajun"). These techniques analyze the position of consonants and vowels.

As we will see in Section 3.4, such techniques are still valid for string values in ontology instances and currently used by all the instance matching tools to perform string matching operations.

When the similarity value $sim(v_h, v_k)$ of each pair of corresponding attribute values of two considered records r_1 and r_2 has been calculated, it is possible to decide if, given a threshold, r_1 and r_2 can be classified as matching or non-matching records. The set of similarity values of single pairs of attribute values is then given as input to a decision engine, whose aim is to classify r_1 and r_2 as matching or non-matching records, by analyzing them as a whole. The decision engine works under the rules of a certain methodology. In the following, we present the main categories in which these methodologies can be classified, describing the record-oriented techniques used to compare and classify records.

Learning-based techniques. The idea behind learning-based techniques is to train a classifier in order to make it able to understand if two records refer to the same real-word entity or not. Thus, the classifier takes as input a set of instance pairs, together with the expected classification (i.e., matching or non-matching records). If the training set is adequate, the system will then be able to correctly classify new input data.

The main concern using these techniques is the need to find out a good training data set. In fact, the training input has to cover all the possible situations but, at the same time, it has to be general enough to make the system able to discover the correct classification functions. This is a non-trivial task and it usually requires a manual selection.

Different kind of learning-based techniques are available. The easiest method is the *supervised learning* technique. Using this approach, the system learns from a training input of already-classified record pairs. As proposed in [50], this method can also be exploited to improve the quality of record linkage results. In fact, the information obtained about the degree of similarity between two records can be propagated to each pair of their corresponding attribute values. In other words,

if two records are recognized as duplicates, all their corresponding attribute values can be considered equivalent as well.

Learning-based techniques require a great amount of high-quality and balanced training data. Therefore, those data have to be chosen carefully. In fact, the classifier need to receive as input not only examples in which the two compared records are clearly identical or examples in which records clearly refer to different real-world entities, but they also need record pair examples in which some kind of ambiguity is present. Only in that way the classifier can produce precise results. In order to automatically find out those kind of record pairs, it is possible to use the *active learning* technique. Specifically, those systems select, within non-classified data, instance pairs having an intermediate degree of similarity. Thus, a domain expert can manually classify selected record pairs and add them to the training set. As an example, ALIAS [51] automatically classify record pairs that clearly refer to the same real-world entity as well as record pairs that clearly denote different real-world entities, and automatically selects ambiguous record pairs, which instead have to be classified by humans.

As the quality of matching results produced by learning-based techniques depends on the quality of training data, record pairs within the training set have to be manually classified by a domain expert. An alternative approach is the *unsupervised learning* technique, which can be adopted to limit the manual effort required. This method uses clustering techniques in order to identify record pairs with similar features. The assumption behind unsupervised learning techniques is that record pairs with similar features belong to the same class (i.e., matching or non-matching records). In other words, all record pairs belonging to the same cluster, also belong to the same class. Thus, as pointed out in [52], using such techniques, it is possible to classify all input record pairs knowing the classification of just few record pairs belonging to each cluster. As proposed in [53,54], it is possible to train a classifier automatically selecting a set of classified record pairs which satisfy a specific criterion. For instance, record pairs selected as matching examples have to have a similarity degree that exceeds a given threshold value.

Another possible approach is to combine different learning-based techniques. The idea is to put already-classified data together with non-classified data, in order to reduce the amount of training information needed, still having good quality results. These methods are called *semi-supervised learning* techniques. An example of them is presented in [55].

Finally, we note that learning-based techniques are being recently proposed also in the field of ontology instance matching. For example, in [4], the authors propose to determine the set of matching instances stored in two considered ontologies by combining the results of different string matching functions (e.g., edit distance, cosine similarity) with a machine learning approach based on a SVM (Support Vector Machine) classifier. Different string matching functions are separately exploited to compare the values of the instance properties and to calculate their own set of mappings denoting the pairs of matching instances. The SVM classifier is then invoked to determine the final set of matching instances by

considering the various sets of (potentially different) mappings computed by the string matching functions. A set of matching instances calculated on a reference domain ontology is used as training set for the SVM classifier.

Similarity-based techniques. If no training data are available, the similarity degree of two records can be measured by considering the input records as long attribute values. In this case, it is possible to use the same methods used to compare attribute values, such as string matching functions. Another approach to measure the similarity degree between two records is to calculate the average similarity of each pair of their attribute values [56]. If some information about the relative importance of each attribute is available, the similarity of a record pair can be measured by calculating the weighted average of the similarity of each single pair of attribute values. The weight of each attribute can be manually specified by a domain expert [57] or it can be automatically determined through statistical analysis [58]. Finally, a further refinement of the instance matching process is to take into account the frequency each value occurs [59]. In particular, a pair of matching attribute values will receive a high weight if these values occur with a low frequency within the domain, while they will receive a low weight otherwise. For example, the surname "Smith" is very common, so the weight of two matching attributes sharing this value will be low. On the opposite, the surname "Zabrinsky" occurs very rarely, so the weight of two matching attributes sharing this value will be high. The idea is that records sharing a rare attribute value are more likely to refer to the same real-world entity.

The main drawback of similarity-based techniques is the identification of a right threshold, in a way that distinguishing matching from non-matching records is reasonable. For example, the problem is to decide if two records having a similarity measure of 0.5 have to be considered as matching or not.

Rule-based techniques. Rule-based techniques can be considered as a special case of similarity-based techniques. In fact, like similarity-based techniques, they assign a similarity value to each record pair but, differently from similarity-based techniques, they just produce a boolean output, namely 1 if the input records refer to the same real-world entity, and 0 otherwise. The idea behind these techniques is that, even if a key attribute is not available, it is still possible to identify a set of attributes that collectively are able to univocally distinguish each record [60]. This attribute set is usually determined by domain experts [61] and it can thus be exploited to identify heuristic rules which can help to find out records referring to the same real-world entity. For example, if two records denoting persons share the same value on attributes "Surname" and "Address", there is a very high probability that the considered records refer to the same person.

Rule-based techniques produce very precise matching results, but they have the drawback that they are domain-dependent and that it can be difficult to find good heuristic rules for the considered domain.

Context-based techniques. Context-based techniques are generally based on the idea of performing record matching by considering not only their attribute

values, but also their relationships with other records. In other words, records connected with the input records are considered to constitute their *context*. Thus, given two records r_1 and r_2, the similarity $sim(r_1, r_2)$ of r_1 and r_2 is computed by considering also the similarity value of each pair of records in the context of r_1 and r_2, respectively. An example of these techniques is the *collective model*, presented in [50]. Unlike classical methods based on the independent comparison of record pairs, this work proposes to analyze the records from one or more sources all together, by considering their shared attribute values. In particular, the process of finding duplicates is represented as an undirected graph where records sharing the same attribute values are linked together. A second example, namely the *iterative deduplication*, is presented in [62]. In this work, the records to analyze are first clustered, and then, all the records within the same cluster are matched, in order to find out duplicates. The deduplication process is iterative because matching records are linked together and, as new duplicates are discovered, the distance between clusters is updated, potentially leading to the discovery of new duplicates.

3.2 Optimizations

As the instance matching process often needs to take place in dynamic contexts and open networked scenarios, performance issues play a crucial role. Thus, one of the main concerns of instance matching is the time required to find out the correct mappings between individuals belonging to one or more ontologies. For this reasons, performance issues are even more important for instance matching. In fact, while records have a flat structure, where each property has an atomic value, instances can have a complex structure, where property values can be in turn instances.

The easiest way to identify instances which refer to the same real-world entity within two ontologies O_1 and O_2 is to compare each instance belonging to O_1 with each instance belonging to O_2. Thus, the total number of comparisons would be $n \cdot m$, where n and m denotes respectively the number of instances in O_1 and the number of instances in O_2. Another factor that influences the computational complexity of instance matching is the time needed to compare each single pair of instances.

Several optimization techniques have been developed. These were originally developed to improve the record linkage performances, but they can be applied in an analogous way to improve instance matching performances. Available optimization techniques can be divided in two categories: the ones aiming to reduce the number of comparisons between instances and the ones aiming to decrease the cost of each single comparison. Of course, those two classes of techniques can be combined to work at different levels of granularity.

Reduction of the number of comparisons. Many different solutions have been proposed in order to accurately select a subset of instances that are more likely to be similar to an input instance, avoiding to compare the input instance against all the instances within the ontology. Those techniques are based on

the idea of partitioning instances represented in an ontology by clustering together potentially matching instances. For example, *blocking* techniques divide instances belonging to a certain ontology in homogeneous and mutually exclusive subsets. These subsets are called *blocks*. Usually, instances are divided according to the value they assume on a strong identifying property, called *blocking key*. The assumption behind blocking techniques is that instances which refer to the same entity cannot be inserted into different subsets. Thus, each instance has to be compared only with instances belonging to the same block. As this method can increase the number of false negatives, it can be improved by repeating the blocking process using different blocking keys. Another well-known example of this kind of optimizations is the *sorted neighborhood* approach [61]. It works by sorting instances according to the value they assume on the property with the highest discriminating power and by only comparing instances within a shifting-window of a fixed dimension. In particular, each instance is matched against the other instances within the same window, while the window is progressively shifted to analyze the complete list of sorted instances. The assumption behind that method is that similar or matching instances cannot have different values on the property used for sorting. Again, in order to improve the quality results of the matching process, it is possible to repeat the execution using different sorting properties. That approach is called *multipass* technique. As an optimization of the sorted neighborhood technique, the *clustering* method is proposed. The idea is to build independent clusters of similar instances and apply the sorted neighborhood method on each resulting cluster in parallel. A further optimization of the sorted neighborhood technique is to dynamically change the size of the fixed window [63]. Obviously, the effectiveness of these approaches is based on the quality of values belonging to the property chosen for sorting. Null or inconsistent values can force potentially matching instances to be in different clusters. In that way, those instances cannot be compared. For this reason, the choice of the sorting property is done manually by a domain expert. A method for the automatic selection of the sorting property is proposed in [64]. According to the authors of that work, the choice of the sorting property can be done calculating the *identification power* of each property of the instances to compare, that is the evaluation of the level of accuracy, completeness and consistency of each property.

Reduction of the cost of each comparison. A different approach to improve instance matching performance is based on the idea that, each instance pair (i_1, i_2) can be classified as matching or non-matching analyzing only a subset of the corresponding property values of i_1 and i_2, instead of each one. In fact, that classification can usually be done comparing just the values that i_1 and i_2 respectively assume on their most identifying properties. Thus, the matching process can terminate when the knowledge about i_1 and i_2 is enough to classify them as matching or non-matching instances and further comparisons among their property values would be useless, as their results cannot change the classification choice [65]. As proposed in [52], it is possible to automatically choose the properties to compare using statistical heuristics.

3.3 Instance Matching vs. Record Linkage

Techniques for instance matching can rely on techniques for record linkage both at value and record level. In fact, the structure of an ontology instance, in terms of properties and values, is analogous to the structure of a record. However, the structure of instances presents additional features that require specific solutions in order to correctly perform the instance matching process. In addition, new peculiar requirements are originated from the ontology specification languages and associated data models. In the discussion of such peculiarities, we rely on the example of Figure 3 where a graphical representation of two sample ontology instances X and X' is provided.

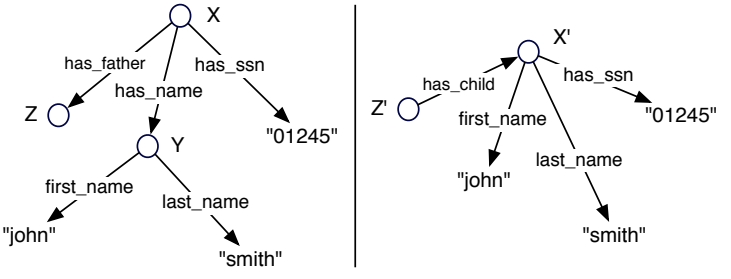


Fig. 3. An example of two ontology instances X and X', each describing the person John Smith

Structural heterogeneity. The structural heterogeneity can be seen at two levels: *language expressivity* and *design practices*. With language expressivity, we refer to the fact that the expressive power of ontology specification languages allows the definition of a number of structurally-different but semantically-equivalent instance representations. With design practices, we refer to the fact that many different methodologies for ontology design are currently available but consolidated and widely accepted ontology design patterns are still missing. Furthermore, advanced tools for supporting the ontology definition process are only partially available and the subjective choices of the ontology designer still have a key influence on the knowledge model of the resulting ontology and its quality. As an example, we can consider the Figure 3 where the instances X and X' denote the same individual John Smith even if two different representations are provided. This conclusion is supported by the fact that the same values (i.e., 01245, john, smith) are defined for three corresponding properties with identical labels (i.e., has_ssn, first_name, last_name).

Besides the capability to evaluate the level of similarity between property values, instance matching techniques have to go beyond heterogeneous representations by identifying the pairs of matching properties between two considered instances. To this end, it is important to stress the role of a dynamic (self) configuration of the instance matching process according to the specific features of the considered instances to be compared.

Implicit knowledge. We refer to the hierarchical organization of ontology elements. As argued in [5], the question is whether a concept is interpreted as a collection of instances annotated by itself alone, or whether the instances of its descendants in the hierarchy also belong to its extension. More generally, besides the explicitly defined set of instance assertions, additional implicit knowledge can be inferred and thus considered for instance matching purposes. Reasoning-based techniques can be invoked to this end. As an example, we can consider the instances X and X' of Figure 3. When only explicit knowledge is considered, the properties has_father and has_child are not taken into account for determining whether X and X' denote the same real-world entity. But when implicit knowledge is considered, if we assume that the property has_father is defined as inverseOf has_child, the assertion X' has_father Z' can be inferred and it can contribute to improve the results of matching for X and X'.

In this respect, instance matching techniques have to provide the capability to dynamically vary the number of assertions involved in the matching process according to the kind of knowledge that is actually considered (i.e., explicit vs. implicit). Moreover, the capability of invoking external support services, such as a reasoning service, is strongly required for providing a full suite of instance matching functionalities.

Id-oriented identification. We refer to the use of an URI-based mechanism for univocally identifying an ontology instance. Such an approach simplifies insertion and retrieval of ontology instances especially when distributed ontologies are considered. However, the use of a distinct URI for identifying each newly inserted instance can be considered as a "bad practice" since it hampers an incremental approach to knowledge definition, as discussed in [3]. The key problem is that the URI-based identification mechanism provides a "syntactic" instance identification which is useless for instance matching to determine when different instance descriptions refer to the same real-world entity. In various ontology specifications languages (e.g., OWL), it is possible to define functional properties to specify that a property has an identification role for an instance, thus providing a "value-oriented" identification mechanism. However, as a matter of fact, the specification of functional property constraints is not a widespread practice, and, in most cases, they are not explicitly defined in Semantic Web ontologies. Considering the example of Figure 3, it is intuitive for a human user to recognize that the instances X and X' refer to the same real-world entity since they have the same value for the has_ssn property that is an natural identification property for people. But in case that this property is not specified as functional, it is not trivial for an instance matcher to understand that, for the purpose of real-world entity identification, a matching value for the property has_ssn is more relevant than a matching value for other properties.

As a consequence, instance matching techniques have to provide the capability to capture and assess what we call the *identification power* of instance properties, apart from the availability of functional constraints specified in the ontology. More generally, the capability to distinguish between "featuring" and "non-featuring" instance properties is a basic functionality for an ontology

instance matcher and the support for their (semi) automatic detection is highly recommended in practical applications.

3.4 Matching Tools

A comparative overview of the main tools for instance matching is provided in Table 2 where we consider those tools that participated to the contest of OAEI 2009 [66]. The edition of 2009 was the first OAEI contest where the problem of instance matching has been explicitly considered with a focused evaluation track and ad-hoc datasets. The datasets of OAEI 2009 were conceived with the goal of evaluating tools with respect to their capability to deal with three different kinds of heterogeneity that can occur in the instance representation, namely *value heterogeneity*, *structural heterogeneity*, and *logical heterogeneity*.

For value heterogeneity, instances referring to the same real-world entity can appear in the datasets of OAEI 2009 with misspellings and other dissimilarities due to the use of different conventions/formats. For example, the person name "John Smith" of an instance Z can appear as "Jhn Smth" in instance Z' and as "Smith, John" in instance Z''. Moreover, the datasets of OAEI 2009 include different kinds of string data, ranging from short strings, such as person names, to longer texts, such as publication titles and abstracts. To deal with these kinds of value heterogeneity, all the tools in Table 2 provide string matching

Table 2. Overview of the main tools for instance matching

Instance Matching Tool	Value-Oriented Techniques	Record-Oriented Techniques	Supports Ontology Matching	Supported Languages
AFlood [34]	Jaro-Winkler string matching	Context-based	Yes	RDF OWL
ASMOV [37]	String matching based on [67]	Similarity-based Context-based	Yes	RDF OWL UMLS
DSSim [39]	Jaccard string matching	Context-based	Yes	RDF OWL SKOS
HMatch 2.0 [68]	QGram, Levenshtein, HMatch string matching	Similarity-based Context-based	Yes	RDF OWL
FBEM [69]	Levenshtein, TagLink string matching	Rule-based Context-based	No	RDF OWL
RiMOM [44]	RiMOM string matching	Context-based	Yes	RDF OWL

techniques for the comparison of property values. In this respect, we note that different results in terms of precision and recall are provided according to the specific technique adopted by the various tools. In particular, we observe that tools supporting general-purpose string matching techniques provide low performances when dealing with string transformations on long and/or complex texts (e.g., AFlood, DSSim, FBEM). This is due to the fact that general-purpose techniques like Jaccard or Levenshtein are basically conceived to detect data mistakes. High-quality results are provided in OAEI 2009 when ad-hoc string matching techniques are adopted (e.g., ASMOV, HMatch 2.0, RiMOM).

For structural heterogeneity, instances referring to the same real-world entity can appear in the datasets of OAEI 2009 with a different schema or with analogous schema but different property names. For example, the instances X and X' of Figure 3 can be used to provide two different representations of the person John Smith. To deal with structural heterogeneity, instance matching tools provide context-based and rule-based techniques for differently combining the results calculated by matching property values. In some tools, like AFlood and ASMOV, structural heterogeneity is managed by matching corresponding property names according to their level of depth in the instance representations and by iterating the comparison until a property value is reached. In some other tools, like HMatch 2.0 and FBEM, instances are internally represented through a flat structure where the level of property depth within an instance representation is not considered for matching. These two approaches to structural heterogeneity management provide similar results in therm of effectiveness. However, the use of a flat internal representation of instances is more promising in terms of computation time required for executing matching. As a final remark, we observe that all the considered tools, with the exception of FBEM, support the use of ontology matching techniques to discover mappings at the schema level with the aim at improving the effectiveness of instance matching by determining the pairs of corresponding properties values to compare when different instance structures are considered.

For logical heterogeneity, instances referring to the same real-world entity can appear in the datasets of OAEI 2009 with a different level of explicitly defined knowledge. For example, the fact that the instances Z and Z' belong to a class C can be explicitly defined for Z, but only implicitly defined for Z'. To deal with logical heterogeneity, instance matching tools rely on the use of reasoning. Some tools, like HMatch 2.0, invoke reasoning as an external service to make explicit the knowledge implicitly defined. Some other tools, like DSSim and RiMOM, support probabilistic reasoning and learning techniques that are used to refine an initial set of mappings computed with value-oriented techniques.

Finally, we note that the tools participating to OAEI 2009 provide better results in terms of precision, while recall values still need of improvements that can be achieved through the development of more flexible techniques for structural and logical heterogeneity management.

4 Ontology and Instance Matching in BOEMIE

In the BOEMIE project, ontology and instance matching techniques are employed to support semi-automated ontology evolution according to a bootstrapping approach, as will be described in the following.

In Table 3, we give an overview of some of the main research projects about ontology and instance matching that were active during the BOEMIE period. Some of these projects are specifically focused on the matching problem, and they directly contributed to the development of some existing matching tools. Some other projects are generically involved in the development of tools for ontology management, without a specific focus on ontology matching. However, these tools provide capabilities for comparing ontology elements and for defining mappings between them.

In the BOEMIE project, a novel methodology for ontology evolution is defined to evolve a domain ontology, called *BOEMIE ontology*, through continuous acquisition of semantic information from multimedia resources such as images, video, and audio. In this methodology, evolution is *pattern-driven* according to the results of a semantic interpretation process performed over the information extracted from the underlying multimedia sources. According to the selected evolution pattern, the BOEMIE ontology is semi-automatically evolved either through the insertion of new instances (ontology population) or through the addition of new concepts (ontology enrichment), by exploiting the results of the ontology and instance matching techniques of the HMatch 2.0 suite.

4.1 HMatch 2.0

In BOEMIE, the HMatch 2.0[1] system is exploited as a comprehensive matchmaking engine where different specialized components are invoked alone or in combination for performing ontology and instance matching according to the specific evolution scenario that need to be considered [70]. HMatch 2.0 is based on a modular architecture where each matching component addresses a specific task and interacts with the other components through appropriate interfaces [68]. In Figure 4, we highlight the HMatch 2.0 components that are mainly involved in the BOEMIE activities. In particular, the HMatchController is responsible for managing the HMatch 2.0 configuration by selecting the matching components to invoke and by supervising the execution of the overall matching process. The HMatch($\mathcal{L}$) and HMatch($\mathcal{C}$) components work at concept level and they provide linguistic and contextual matching functionalities, respectively. In particular, HMatch($\mathcal{L}$) provides a library of linguistic matching techniques for similarity-based ontology matching. In BOEMIE, HMatch($\mathcal{L}$) is invoked to discover similar concepts in external ontologies to provide possible reuse-suggestions during ontology enrichment (see Section 4.2). The HMatch($\mathcal{I}$) component has been specifically developed for BOEMIE to provide distance- and context-based matching techniques for ontology instance matching. In BOEMIE, HMatch($\mathcal{I}$) is invoked

[1] http://islab.dico.unimi.it/hmatch/

Table 3. Overview of the main research projects about ontology matching

Project	Main Contributions
CROSI http://www.aktors.org/crosi/	Development of a structural matching system capable of exploiting the rich semantics of the considered OWL ontologies. The CROSI Mapping System is defined to this end.
KnowledgeWeb http://knowledgeweb.semanticweb.org/	Network of Excellence focusing on Semantic Web technologies (including also ontology matching) in the areas of scientific research, education, and industry.
Linked Data http://linkeddata.org/	Research initiative providing techniques and tools for linking related data over the web.
NeOn http://www.neon-project.org/	Development of a service-oriented infrastructure, and associated methodology based on ontologies to enable intelligent access, integration, sharing and use of web data. Ontology matching is used to automatically find/combine knowledge provided by multiple online ontologies.
OKKAM http://www.okkam.org/	Construction of an Entity Name System (ENS), namely a service for matching any description of an instance against a repository of known instances and return the corresponding ENS-identifier. The FBEM matching tool has been developed to this end.
OpenKnowledge http://openk.org/	Support to knowledge sharing in a peer-to-peer network without any global agreement or a-priori knowledge. Matching is used to establish mappings among different network peers.
SEALS http://www.seals-project.eu/	Development of a platform providing an independent, open, scalable, extensible and sustainable infrastructure for the remote evaluation of semantic technologies by providing an integrated set of evaluation services and test suites, including ontology matching.
SEKT http://www.sekt-project.com/	Knowledge management through automated techniques for extracting meaning from the Web. The FOAM tool for ontology alignment and mapping has been developed to this end [9].
SEWASIE http://www.sewasie.org/	Tools and techniques for the development of semantically-enriched, virtual data stores. The MOMIS data integration tool is used to this end [9].
SWAP http://swap.semanticweb.org/	Knowledge discovery in Peer-to-Peer networks through schema-based matching techniques. The KAON platform for ontology management is used to this end [9].
TONES http://www.tonesproject.org/	Normalization and development of tools and techniques to establish mappings among different ontologies. An ontology modularization tool and various reasoning tools are used (e.g. CEL, RacerPro).

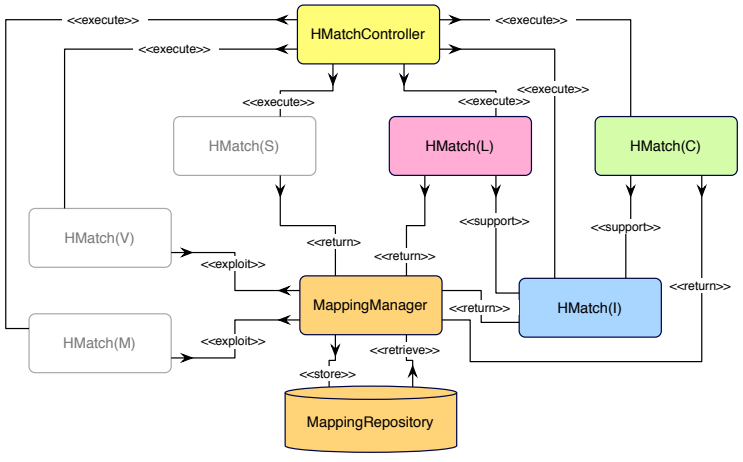


Fig. 4. HMatch 2.0: components and interactions

during ontology population to evaluate the similarity of multimedia ontology instances with the aim to determine when two descriptions refer to the same real-world entity (see Section 4.3). To this end, ontology matching components, namely HMatch($\mathcal{L}$) and HMatch($\mathcal{C}$), are exploited by HMatch($\mathcal{I}$) to deal with the problem of comparing ontology instances with structural heterogeneities as discussed in Section 3.3. Finally, the MappingManager is responsible for combining the results of the various HMatch 2.0 components and for storing the resulting mappings. As a further feature, we note that HMatch 2.0 is designed to interface an external reasoning service to support a sort of reasoning-based matching techniques where inference mechanisms are used to determine the complete set of concept/instance properties apart from their explicit definition in the considered ontology [2].

4.2 Ontology Matching

In BOEMIE, ontology matching techniques support the *enrichment* activity, which is defined as the activity of creating and framing new knowledge (e.g., new concepts, new properties) for the BOEMIE domain ontology. The need of introducing new concept definitions arises when the existing knowledge in the domain ontology is not sufficient to explain the new incoming instances extracted from the considered multimedia documents. This unexplained information represents a *concept proposal* $\bar{c}$ and it is expressed as an aggregation of axioms. Starting from $\bar{c}$, the domain expert can perform a set of refinements over it, for example by choosing a name for the new concept and/or by (re)defining its axioms. The final concept proposal is subsequently inserted in the BOEMIE ontology.

[2] Currently, the Racer reasoning system is configured as the reference reasoning service for HMatch 2.0.

Ontology matching techniques of HMatch 2.0 are exploited during enrichment to support the domain expert with a set of *suggestions* that can be exploited for partial/full reuse in the design of the concept proposal $\bar{c}$. Enrichment suggestions consist in a set of external concept definitions harvested from external knowledge sources (e.g., Semantic Web ontologies, Web directories, RDF repositories) and matching $\bar{c}$ using the linguistic matching component HMatch($\mathcal{L}$) of HMatch 2.0.

In BOEMIE, each concept c is characterized by a *terminological equipment* $TE(c) = \{t_1, \ldots, t_n\}$, namely a set of terms featuring the concept c specification. The terminological equipment of a concept includes its name, the name of its properties, and the name of all the concepts related to it (i.e., the ones occurring in its constituting axioms). To build the terminological equipment $TE(c)$, a normalization process is executed to determine the basic word-forms and to tokenize the composite terms that appear in the concept c specification. Moreover, by relying on the lexical dictionary WordNet, $TE(c)$ is enriched with other terms that are semantically-related to the terms featuring the concept c specification (e.g., synonyms and hyperonyms).

Given two concepts c_1 and c_2 and their respective terminological equipment $TE(c_1)$ and $TE(c_2)$, the *Linguistic Affinity* $LA(t_i, t_j)$ is calculated by HMatch($\mathcal{L}$) for each pair of terms (t_i, t_j), where $t_i \in TE(c_1)$ and $t_j \in TE(c_2)$. The linguistic affinity function returns a value in the range $[0, 1]$, and can be evaluated, by means of three different strategies.

- **Syntactic**: using a string matching algorithm (i.e., QGram, i_Sub).
- **Semantic**: using a thesaurus or a lexical system (i.e., WordNet).
- **Combined**: using a combination of syntactic and semantic strategies.

The similarity value $sim(c_1, c_2)$ of two concepts c_1 and c_2 is in the range $[0, 1]$ and it is calculated as follows.

$$sim(c_1, c_2) = \frac{2 \cdot |M|}{|TE(c_1)| + |TE(c_2)|}$$

where $M = \{(t_i, t_j) \mid t_i \in TE(c_1), t_j \in TE(c_2), LA(t_i, t_j) \geq th\}$ is the set of pairs of matching terms belonging to the terminological equipment of c_1 and c_2, th is a similarity threshold denoting the minimum level of matching required for considering two terms as matching terms, and $|M|$, $|TE(c_1)|$ and $|TE(c_2)|$ denote the cardinality of sets M, $TE(c_1)$, and $TE(c_2)$, respectively.

External suggestions are cataloged and indexed in a local repository to support efficient data retrieval during ontology enrichment. In particular, given a concept proposal $\bar{c}$, all the external concepts matching $\bar{c}$ are retrieved from the repository and presented in form of suggestions to the domain expert.

Example. As an example of ontology enrichment, we consider a concept proposal CP_1 defined as follows.

$$CP_1 \sqsubseteq \exists hasPart.PoleVaultAttempt$$
$$CP_1 \sqsubseteq \exists hasPart.HorizontalBar$$
$$CP_1 \sqsubseteq \exists hasPart.Pillar$$
$$CP_1 \sqsubseteq \exists hasPart.Pole$$

The concept CP_1 describes an entity which is associated with the objects "PoleVaultAttempt", "HorizontalBar", "Pillar", and "Pole".

During harvesting from external knowledge sources, the Athlete.owl ontology[3] is analyzed. This is a small Semantic Web ontology (about one hundred triples) modeling concepts in the athletics domain, like athletes and various Olympic sport competitions. To discover suggestions for possible reuse, the terminological equipment of CP_1 is matched against the terminological equipment of each concept belonging to the Athlete.owl ontology by using the HMatch($\mathcal{L}$) component of HMatch 2.0. In particular, we match CP_1 against the concept PoleVault in Athlete.owl. The concept PoleVault is defined as follows.

$$PoleVault \sqsubseteq SportCompetition$$
$$PoleVault \sqsubseteq JumpingEvent$$
$$PoleVault \sqsubseteq \exists hasPart.PoleVaultAttempt$$
$$PoleVault \sqsubseteq \exists hasPerformance.Performance$$

The terminological equipment of CP_1 and PoleVault are generated as follows.

$$TE(CP_1) = \{CP_1, have, part, pole, vault, attempt, horizontal, bar, pillar\}$$

$$TE(PoleVault) = \{pole, vault, sport, competition, jump, event,$$
$$have, part, attempt, performance\}$$

By using HMatch($\mathcal{L}$), $sim(CP_1, PoleVault)$ is calculated as follows.

$$sim(CP_1, PoleVault) =$$
$$= \frac{2 \cdot |M|}{|TE(c_1)| + |TE(c_2)|} =$$
$$= \frac{2 \cdot 5}{10 + 9} = 0.53$$

In BOEMIE, the similarity threshold $th = 0.5$ is used, then the concept PoleVault is considered as a matching concept of CP_1 and it is provided to the domain expert as a suggestion. The domain expert exploits the PoleVault suggestion to modify the concept proposal CP_1. In particular, the placeholder CP_1 is replaced with the name PoleVault. Thus, the final committed concept is defined as follows.

$$PoleVault \sqsubseteq \exists hasPart.HorizontalBar$$
$$PoleVault \sqsubseteq \exists hasPart.Pillar$$
$$PoleVault \sqsubseteq \exists hasPart.Pole$$
$$PoleVault \sqsubseteq \exists hasPerformance.Performance$$

After commitment, the new concept PoleVault is inserted in the BOEMIE ontology.

[3] http://www.mindswap.org/2004/athlete.owl

4.3 Instance Matching

In BOEMIE, instance matching techniques are employed to support the *population* activity, which is defined as the activity of correctly framing in the BOEMIE domain ontology the new incoming instance(s) extracted from multimedia resources. In this respect, the HMatch($\mathcal{I}$) component of HMatch 2.0 is invoked to automatically discover whether a new incoming instance matches one or more instances already stored in the domain ontology.

The matching process of HMatch($\mathcal{I}$) starts with the acquisition of the two ontology instances to compare in form of ABoxes. The subsequent matching stage is based on the comparison of instance properties and corresponding property values. To this end, each instance in HMatch($\mathcal{I}$) is represented as a tree (*instance tree construction*) where property values are nodes, and properties are labeled edges (see the example of Figure 3). Matching is then performed by traversing in postorder the instance trees and by collecting all the pairs of property values that are candidate for matching, namely all the pairs of leaf values that have a matching property at the first-level in their respective trees. The pairs of matching properties are detected by relying on the ontology matching functionalities of HMatch 2.0. In case of BOEMIE, the matching task was simplified by the fact all the instances were defined according to the same TBox. Thus, each pair of candidate matching values has identical first-level property in their respective trees. Given two instances i_1 and i_2 and the set C^{i_1,i_2} of their candidate matching values, the instance similarity $sim(i_1, i_2)$ is calculated as follows:

$$sim(i_1, i_2) = \frac{\mid \{(v_i, v_j) \mid (v_i, v_j) \in C^{i_1,i_2} \wedge LA(v_i, v_j) \geq th\} \mid}{\mid P^{i_1} \cup P^{i_2} \mid}$$

where $LA(v_i, v_j)$ is the linguistic affinity function introduced in Section 4.2, th is a similarity threshold in the range $[0, 1]$, and P^{i_1}, P^{i_2} are the sets of first-level properties of i_1 and i_2, respectively. Since in case of instance matching the function $LA(v_i, v_j)$ is used for comparing property values instead of concept/property names, we do not rely on WordNet for linguistic matching, but we exploit the Edit Distance function for linguistic affinity evaluation.

Given a new incoming instance i, the set $SIM(i)$ determines those instances stored in the BOEMIE ontology that match i as follows: $SIM(i) = \{i' | sim(i, i') \geq th_2\}$, where th_2 is a similarity threshold (in BOEMIE $th_2 = 0.5$). A new incoming instance i is inserted in the BOEMIE ontology as a new instance if $SIM(i) = \emptyset$, otherwise the instance i is stored in the BOEMIE ontology by defining an appropriate same_as relation with each matching instance $i' \in SIM(i)$.

Example. As an example of ontology population, we consider the following instance i_1 belonging to the BOEMIE ontology.

$$\begin{aligned} (i_1, \text{“Michal Bieniek”}) &: \text{hasName} \\ (i_1, \text{“Poland”}) &: \text{hasCountry} \\ (i_1, 188) &: \text{hasHeight} \\ (i_1, 2.36) &: \text{hasPerformance} \end{aligned}$$

Moreover, we consider the following new incoming instance i_2, extracted from the analysis of a web page in the athletics domain.

$$(i_2, \text{"Michal Bieniek"}) : \mathsf{hasName}$$
$$(i_2, \text{"Poland"}) : \mathsf{hasCountry}$$
$$(i_2, 71) : \mathsf{hasWeight}$$
$$(i_2, 2.32) : \mathsf{hasPerformance}$$

Both individuals i_1 and i_2 are instances of the concept Athlete, they share some properties (i.e., hasName, hasCountry, and hasPerformance), have the same name (i.e., Michal Bieniek), and come from the same country (i.e., Poland). However, we know the height of i_1, that is unknown in case of i_2, and we know the weight of i_2 that is unknown in case of i_1. As a first step in the matching procedure, we create the instance trees of the two instances and we define the set of candidate matching values as follows.

$$C^{i_1,i_2} = \{(\text{"Michal Bieniek"}, \text{"Michal Bieniek"}), (\text{"Poland"}, \text{"Poland"}), (\text{"2.36"}, \text{"2.32"})\}$$

Then, we execute the linguistic affinity function of the pairs of values of C^{i_1,i_2} by setting a threshold $th = 0.8$. By relying on the edit distance metric, we obtain that the name and the nationality of the two instances are matching, since they have the same values, while $LA(\text{"2.36"}, \text{"2.32"}) = 0.88$. The set $P^{i_1} \cup P^{i_2}$ is the following.

$$P^{i_1} \cup P^{i_2} = \{\mathsf{hasName}, \mathsf{hasCountry}, \mathsf{hasHeight}, \mathsf{hasPerformance}, \mathsf{hasWeight}\}$$

As a result, the instance similarity $sim_i(i_1, i_2)$ is calculated as follows.

$$sim_i(i_1, i_2) = \frac{3}{5} = 0.6$$

According to the default similarity threshold of BOEMIE $th_2 = 0.5$, the instances i_1 and i_2 are considered as matching instances, that is instances referring to the same real-world entity (i.e., the athlete Michal Bieniek). Thus, i_2 is inserted in the BOEMIE ontology together with a same_as relation between i_1 and i_2.

5 Concluding Remarks

In this chapter, ontology and instance matching approaches have been surveyed with special focus on instance matching and on ontology matching tools belonging to the so-called third generation. For the development of next-generation matching tools and for the design of effective applications in the framework of knowledge management and Semantic Web, we envisage the following trends for ontology and instance matching.

Matching for lightweight integration. In recent years, the growing need of sharing data and digital resources within and across organizations has produced a novel attention on data integration topics with the aim to investigate

novel applications for information discovery and sharing in open systems. In this direction, integration approaches need to move towards more "lightweight" techniques, typically suited for Web-scale, Semantic Web, and P2P environments. Conventional mediator-based architectures leave the floor to emergent peer-oriented architectures, where flexible schema/instance matching techniques are required by each peer for mapping discovery with other node schemas. In this respect, additional matching requirements will have to be satisfied, such as the capability to calibrate the accuracy of the matching execution according to a given set of time/space constraints.

Matching for semantic coordination. The emerging popularity of social networks and community-oriented collaboration platforms requires appropriate techniques and tools to effectively manage (i.e., coordinate) the (potentially large) bulk of data that are received from other external users during interactions. The development of integrated coordination platforms is then required, capable of addressing all the aspects concerned with data and knowledge acquisition, storage, and evolution in such a dynamic scenario. The role of linguistic and instance matching is prominent in this respect for the nature and the variability of the datasets to compare, which require more articulated approaches characterized by high scalability and the capability to correctly match poorly-structured and/or fully-unstructured data like, plain texts and simple annotations.

References

1. Shvaiko, P., Euzenat, J.: Ten Challenges for Ontology Matching. In: Chung, S. (ed.) OTM 2008, Part II. LNCS, vol. 5332. Springer, Heidelberg (2008)
2. Castano, S., Ferrara, A., Montanelli, S.: Dealing with Matching Variability of Semantic Web Data Using Contexts. In: Pernici, B. (ed.) CAiSE 2010. LNCS, vol. 6051, pp. 194–208. Springer, Heidelberg (2010)
3. Bouquet, P., Stoermer, H., Mancioppi, M., Giacomuzzi, D.: OkkaM: Towards a Solution to the Identity Crisis on the Semantic Web. In: Proc. of the 3rd Italian Semantic Web Workshop, Pisa, Italy (2006)
4. Wang, C., Lu, J., Zhang, G.: Integration of Ontology Data through Learning Instance Matching. In: Proc. of the 2006 IEEE/WIC/ACM Int. Conference on Web Intelligence (WI 2006), Washington, DC, USA, pp. 536–539 (2006)
5. Isaac, A., van der Meij, L., Schlobach, S., Wang, S.: An Empirical Study of Instance-Based Ontology Matching. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 253–266. Springer, Heidelberg (2007)
6. Engmann, D., Maßmann, S.: Instance Matching with COMA++. In: Proc. of the Workshop on Datenbanksysteme in Business, Technologie und Web (BTW 2007), Aachen, Germany (2007)
7. Kalfoglou, Y., Schorlemmer, M.: Ontology Mapping: the State of the Art. The Knowledge Engineering Review 18(1) (2003)
8. Noy, N.: Semantic Integration: a Survey of Ontology-based Approaches. SIGMOD Record, Special Issue on Semantic Integration 33(4) (2004)

9. INTEROP: State of the Art and State of the Practice Including Initial Possible Research Orientations. Deliverable D8.1, NoE INTEROP - IST Project n. 508011 - 6th EU Framework Programme (2004)
10. Shvaiko, P., Euzenat, J.: A Survey of Schema-based Matching Approaches. Journal on Data Semantics IV (2005)
11. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
12. Rahm, E., Bernstein, P.: A Survey of Approaches to Automatic Schema Matching. The VLDB Journal 10(4) (2001)
13. Navarro, G.: A Guided Tour to Approximate String Matching. ACM Computing Surveys 33(1), 31–88 (2001)
14. Levenshtein, V.: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady 10(8) (1966)
15. Cormode, G., Muthukrishnan, S.: The String Edit Distance Matching Problem with Moves. In: Proc. of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2002), San Francisco, CA, USA, pp. 667–676 (2002)
16. Ukkonen, E., Wood, D.: Approximate String Matching with Suffix Automata. Algorithmica 10(5), 353–364 (1993)
17. Baeza-Yates, R.A.: Text-Retrieval: Theory and Practice. In: Proc. of the IFIP 12th World Computer Congress on Algorithms, Software, Architecture - Information Processing, Amsterdam, The Netherlands, pp. 465–476 (1992)
18. Navarro, G., Baeza-Yates, R.A., Sutinen, E., Tarhio, J.: Indexing Methods for Approximate String Matching. IEEE Data Engineering Bulletin 24(4), 19–27 (2001)
19. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic Schema Matching with Cupid. In: Proc. of the Int. Conference on Very Large Data Bases (VLDB 2002), Hong Kong, China, pp. 49–58 (2002)
20. Castano, S., De Antonellis, V., De Capitani Di Vimercati, S.: Global viewing of heterogeneous data sources. IEEE Transactions on Knowledge and Data Engineering 13(2), 277–297 (2001)
21. Jeh, G., Widom, J.: SimRank: a Measure of Structural-Context Similarity. In: Proc. of the 8th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining (KDD 2002), Edmonton, Alberta, Canada, pp. 538–543 (2002)
22. Shasha, D., Wang, J.T.L., Giugno, R.: Algorithmics and Applications of Tree and Graph Searching. In: Proc. of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2002), Madison, Wisconsin, USA, pp. 39–52 (2002)
23. Melnik, S., Garcia-Molina, H., Rahm, E.: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. In: Proc. of the 18th Int. Conference on Data Engineering (ICDE 2002), San Jose, CA, USA (2002)
24. Meilicke, C., Stuckenschmidt, H., Tamilin, A.: Repairing Ontology Mappings. In: Proc. of the 22nd Conference on Artificial Intelligence (AAAI 2007), Vancouver, BC, Canada, pp. 1408–1413 (2007)
25. Giunchiglia, F., Shvaiko, P.: Semantic Matching. Knowledge Engineering Review 18(3) (2003)
26. Borgida, A., Serafini, L.: Distributed Description Logics: Assimilating Information from Peer Sources. Journal on Data Semantics I, 153–184 (2003)
27. Castano, S., Ferrara, A., Lorusso, D., Näth, T.H., Möller, R.: Mapping Validation by Probabilistic Reasoning. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 170–184. Springer, Heidelberg (2008)
28. Meilicke, C., Stuckenschmidt, H., Tamilin, A.: Reasoning Support for Mapping Revision. Journal of Logic and Computation 19(5), 807–829 (2009)

29. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Learning to Map between Ontologies on the Semantic Web. In: Proc. of the 11th Int. Conference on World Wide Web (WWW 2002), Honolulu, Hawaii, USA, pp. 662–673 (2002)
30. Lacher, M., Groh, G.: Facilitating the Exchange of Explicit Knowledge Through Ontology Mappings. In: Proc. of the 14th Int. FLAIRS Conference, Key West, FL, USA, pp. 305–309 (2001)
31. Peng, Y., Ding, Z., Pan, R.: Uncertainty in Ontology Mapping: A Bayesian Perspective. In: Proc. of the Information Interpretation and Integration Conference (I3CON), Gaithersburg, MD, USA (2004)
32. Smith, A., Elkan, C.: A Bayesian Network Framework for Reject Inference. In: Proc. of the 10th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining (KDD 2004), New York, NY, USA, pp. 286–295 (2004)
33. Euzenat, J.: The Ontology Alignment Evaluation Initiative
34. Seddiqui, M.H., Aono, M.: An Efficient and Scalable Algorithm for Segmented Alignment of Ontologies of Arbitrary Size. Web Semantics: Science, Services and Agents on the World Wide Web 7(4), 344–356 (2009)
35. Cruz, I.F., Palandri Antonelli, F., Stroe, C.: AgreementMaker Efficient Matching for Large Real-World Schemas and Ontologies. In: Proc. of the 35th Int. Conference on Very Large Databases (VLDB 2009), Lyon, France, pp. 1586–1589 (2009)
36. David, J., Guillet, F., Briand, H.: Association Rule Ontology Matching Approach. Int. Journal on Semantic Web and Information Systems 3(2), 27–49 (2007)
37. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 235–251 (2009)
38. Gracia, J., Lopez, V., D'Aquin, M., Sabou, M., Motta, E., Mena, E.: Solving Semantic Ambiguity to Improve Semantic Web based Ontology Matching. In: Proc. of the 2nd Int. Workshop on Ontology Matching, Busan, Korea (2007)
39. Nagy, M., Vargas-Vera, M., Motta, E.: DSSim - Managing Uncertainty on the Semantic Web. In: Proc. of the 2nd Int. Workshop on Ontology Matching, Busan, Korea (2007)
40. Kensche, D., Quix, C., Chatti, M., Jarke, M.: GeRoMe: A Generic Role Based Metamodel for Model Management. Journal on Data Semantics VIII, 82–117 (2007)
41. Reul, Q., Pan, J.Z.: KOSIMap: Ontology Alignments Results for OAEI 2009. In: Proc. of the 4th Int. Workshop on Ontology Matching, Chantilly, VA, USA (2009)
42. Wang, P., Xu, B.: Lily: Ontology Alignment Results for OAEI 2009. In: Proc. of the 4th Int. Workshop on Ontology Matching, Chantilly, VA, USA (2009)
43. Bock, J., Liu, P., Hettenhausen, J.: MapPSO Results for OAEI 2009. In: Proc. of the 4th Int. Workshop on Ontology Matching, Chantilly, VA, USA (2009)
44. Li, J., Tang, J., Li, Y., Luo, Q.: RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. IEEE Transactions on Knowledge and Data Engineering (TKDE) 21(8), 1218–1232 (2008)
45. Lambrix, P., Tan, H.: SAMBO-A System for Aligning and Merging Biomedical Ontologies. Web Semantics: Science, Services and Agents on the World Wide Web 4(3), 196–206 (2006)
46. Xu, P., Tao, H., Zang, T., Wang, Y.: Alignment Results of SOBOM for OAEI 2009. In: Proc. of the 4th Int. Workshop on Ontology Matching, Chantilly, VA, USA (2009)
47. Hamdi, F., Safar, B., Niraula, N.B., Reynaud, C.: TaxoMap in the OAEI 2009 Alignment Contest. In: Proc. of the 4th Int. Workshop on Ontology Matching, Chantilly, VA, USA (2009)

48. Gu, L., Baxter, R., Vickers, D., Rainsford, C.: Record Linkage: Current Practice and Future Directions. Technical report, CSIRO Mathematical and Information Sciences, Canberra, Australia (2003)
49. Zhou, R., Hansen, E.A.: Domain-Independent Structured Duplicate Detection. In: Proc. of the 21st National Conference on Artificial Intelligence (AAAI 2006), Boston, Massachusetts, USA (2006)
50. Singla, P., Domingos, P.: Multi-Relational Record Linkage. In: Proc. of the 3rd KDD Workshop on Multi-Relational Data Mining, Seattle, WA, USA (2004)
51. Sarawagi, S., Bhamidipaty, A.: Interactive Deduplication Using Active Learning. In: Proc. of the 8th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining (KDD 2002), Edmonton, Alberta, Canada, pp. 269–278 (2002)
52. Verykios, V., Elmagarmid, A., Houstis, E.: Automating the Approximate Record-Matching Process. Information Sciences - Informatics and Computer Science: An Int. Journal 126(1), 83–98 (2000)
53. Christen, P.: A Two-Step Classification Approach to Unsupervised Record Linkage. In: Proc. of the 6th Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia, pp. 111–119 (2007)
54. Christen, P.: Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification. In: Proc. of the 14th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, pp. 151–159 (2008)
55. Pasula, H., Marthi, B., Milch, B., Russell, S., Shpitser, I.: Identity Uncertainty and Citation Matching. In: Proc. of the Conference on Advances in Neural Information Processing Systems (NIPS 2002), Vancouver, BC, Canada, pp. 1401–1408 (2002)
56. Dey, D., Sarkar, S., De, P.: Entity Matching in Heterogeneous Databases: A Distance Based Decision Model. In: Proc. of the 31th Annual Hawaii Int. Conference on System Sciences (HICSS 1998), Kohala Coast, Hawaii, USA, pp. 305–315 (1998)
57. Dey, D., Sarkar, S., De, P.: A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases. IEEE Transactions on Knowledge and Data Engineering 14(3), 567–582 (2002)
58. Guha, S., Koudas, N., Marathe, A., Srivastava, D.: Merging the Results of Approximate Match Operations. In: Proc. 30th Int. Conference on Very Large Databases (VLDB 2004), Toronto, Canada, pp. 636–647 (2004)
59. Winkler, W.: Frequency-Based Matching in Fellegi-Sunter Model of Record Linkage. Statistical research report series rr/2000/06, US Bureau of the Census, Washington, DC, USA (2000)
60. Wang, Y., Madnick, S.: The Inter-Database Instance Identification Problem in Integrating Autonomous Systems. In: Proc. of the 5th Int. Conference on Data Engineering (ICDE 1989), Washington, DC, USA, pp. 46–55 (1989)
61. Hernández, M., Stolfo, S.: Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data Mining and Knowledge Discovery 2(1), 9–37 (1998)
62. Bhattacharya, I., Getoor, L.: Iterative Record Linkage for Cleaning and Integration. In: Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2004), New York, NY, USA (2004)
63. Yan, S., Lee, D., Kan, M., Giles, L.: Adaptive Sorted Neighborhood Methods for Efficient Record Linkage. In: Proc. of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2007), Vancouver, BC, Canada, pp. 185–194 (2007)
64. Bertolazzi, P., De Santis, L., Scannapieco, M.: Automatic Record Matching in Cooperative Information Systems. In: Proc. of the ICDT Int. Workshop on Data Quality in Cooperative Information Systems (DQCIS 2003), Siena, Italy (2003)

65. Newcombe, H.: Handbook of Record Linkage. Oxford University Press, Inc., Oxford (1988)
66. Euzenat, J., Ferrara, A., Hollink, L., Isaac, A., Joslyn, C., Malaisé, V., Meilicke, C., Nikolov, A., Pane, J., Sabou, M., Scharffe, F., Shvaiko, P., Spiliopoulos, V., Stuckenschmidt, H., Sváb-Zamazal, O., Svátek, V., Trojahn dos Santos, C., Vouros, G.A., Wang, S.: Results of the Ontology Alignment Evaluation Initiative 2009. In: Proc. of the 4th Int. Workshop on Ontology Matching, Chantilly, VA, USA (2009)
67. Lin, D.: An Information-Theoretic Definition of Similarity. In: Proc. of the 15th Int. Conference on Machine Learning, San Francisco, CA, USA, pp. 296–304 (1998)
68. Castano, S., Ferrara, A., Lorusso, D., Montanelli, S.: The HMatch 2.0 Suite for Ontology Matchmaking. In: Proc. of the 4th Workshop on Semantic Web Applications and Perspectives (SWAP 2007), Bari, Italy (2007)
69. Stoermer, H., Rassadko, N.: Results of OKKAM Feature Based Entity Matching Algorithm for Instance Matching Contest of OAEI 2009. In: Proc. of the 4th Int. Workshop on Ontology Matching, Chantilly, VA, USA (2009)
70. Castano, S., Ferrara, A., Montanelli, S.: Matching Ontologies in Open Networked Systems: Techniques and Applications. Journal on Data Semantics V (2006)

A Survey of Semantic Image and Video Annotation Tools

Stamatia Dasiopoulou, Eirini Giannakidou, Georgios Litos,
Polyxeni Malasioti, and Yiannis Kompatsiaris

Multimedia Knowledge Laboratory, Informatics and Telematics Institute,
Centre for Research and Technology Hellas
{dasiop,igiannak,litos,xenia,ikom}@iti.gr

Abstract. The availability of semantically annotated image and video assets constitutes a critical prerequisite for the realisation of intelligent knowledge management services pertaining to realistic user needs. Given the extend of the challenges involved in the automatic extraction of such descriptions, manually created metadata play a significant role, further strengthened by their deployment in training and evaluation tasks related to the automatic extraction of content descriptions. The different views taken by the two main approaches towards semantic content description, namely the Semantic Web and MPEG-7, as well as the traits particular to multimedia content due to the multiplicity of information levels involved, have resulted in a variety of image and video annotation tools, adopting varying description aspects. Aiming to provide a common framework of reference and furthermore to highlight open issues, especially with respect to the coverage and the interoperability of the produced metadata, in this chapter we present an overview of the state of the art in image and video annotation tools.

1 Introduction

Accessing multimedia content in correspondence with the meaning pertained to a user, constitutes the core challenge in multimedia research, commonly referred to as the *semantic gap* [1]. The current state of the art in automatic content analysis and understanding supports in many cases the successful detection of semantic concepts, such as persons, buildings, natural scenes vs manmade scenes, etc. at a satisfactory level of accuracy; however, the attained performance remains highly variable when considering general domains, or when increasing, even slightly, the number of supported concepts [2,3,4]. As a consequence, the manual generation of content descriptions holds an important role towards the realisation of intelligent content management services. This significance is further strengthened by the need for manually constructed descriptions in automatic content analysis both for evaluation as well as for training purposes, when learning based on pre-annotated examples is used.

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 196–239, 2011.

The availability of semantic descriptions though is not adequate per se for the effective management of multimedia content. Fundamental to information sharing, exchange and reuse, is the interoperability of the descriptions at both syntactic and semantic levels, i.e. regarding the valid structuring of the descriptions and the endowed meaning respectively. Besides the general prerequisite for interoperability, additional requirements arise from the multiple levels at which multimedia content can be represented including structural and low-level features information. Further description levels induce from more generic aspects such as authoring & access control, navigation, and user history & preferences. The strong relation of structural and low-level feature information to the tasks involved in the automatic analysis of visual content, as well as to retrieval services, such as transcoding, content-based search, etc., brings these two dimensions to the foreground, along with the subject matter descriptions.

Two initiatives prevail the efforts towards machine processable semantic content metadata, the Semantic Web activity[1] of the W3C and ISO's Multimedia Content Description Interface[2] (MPEG-7) [5,6], delineating corresponding approaches with respect to multimedia semantic annotation [7,8]. Through a layered architecture of successively increased expressivity, the Semantic Web (SW) advocates formal semantics and reasoning through logically grounded meaning. The respective rule and ontology languages embody the general mechanisms for capturing, representing and reasoning with semantics. They do not capture application specific knowledge. In contrast, MPEG-7 addresses specifically the description of audiovisual content and comprises not only the representation language, in the form of the Description Definition Language (DDL), but also specific, media and domain, definitions; thus from a SW perspective, MPEG-7 serves the twofold role of a representation language and a domain specific ontology.

Overcoming the syntactic and semantic interoperability issues between MPEG-7 and the SW has been the subject of very active research in the current decade, highly motivated by the complementary aspects characterising the two aforementioned metadata initiatives: media specific, yet not formal, semantics on one hand, and general mechanisms for logically grounded semantics on the other hand. A number of so called *multimedia ontologies* [9,10,11,12,13] issued in an attempt to add formal semantics to MPEG-7 descriptions and thereby enable linking with existing ontologies and the semantic management of existing MPEG-7 metadata repositories. Furthermore, initiatives such the W3C Multimedia Annotation on the Semantic Web Taskforce[3], the W3C Multimedia Semantics Incubator Group[4] and the Common Multimedia Ontology Framework[5], have been established to address the technologies, advantages and open issues related to the creation, storage, manipulation and processing of multimedia semantic metadata.

[1] http://www.w3.org/2001/sw/
[2] http://www.chiariglione.org/mpeg/
[3] http://www.w3.org/2001/sw/BestPractices/MM/
[4] http://www.w3.org/2005/Incubator/mmsem/
[5] http://www.acemedia.org/aceMedia/reference/multimedia_ontology/index.html

In this chapter, bearing in mind the significance of manual image and video annotation in combination with the different possibilities afforded by the SW and MPEG-7 initiatives, we present a detailed overview of the most well known manual annotation tools, addressing both functionality aspects, such as coverage & granularity of annotations, as well as interoperability concerns with respect to the supported annotation vocabularies and representation languages. Interoperability though does not address solely the harmonisation between the SW and MPEG-7 initiatives; a significant number of tools, specially regarding video annotation, follow customised approaches, aggravating the challenges. As such, this survey serves a twofold role; it provides a common framework for reference and comparison purposes, while highlighting issues pertaining to the communication, sharing and reuse of the produced metadata.

The rest of the chapter is organised as follows. Section 2 describes the criteria along which the assessment and comparison of the examined annotation tools is performed. Sections 3 and 4 discuss the individual image and video tools respectively, while Section 5 concludes the paper, summarising the resulting observations and open issues.

2 Semantic Image and Video Annotation

Image and video assets constitute extremely rich information sources, ubiquitous in a wide variety of diverse applications and tasks related to information management, both for personal and professional purposes. Inevitably, the value of the endowed information amounts to the effectiveness and efficiency at which it can be accessed and managed. This is where semantic annotation comes in, as it designates the schemes for capturing the information related to the content.

As already indicated, two crucial requirements featuring content annotation are the interoperability of the created metadata and the ability to automatically process them. The former encompasses the capacity to share and reuse annotations, and by consequence determines the level of seamless content utilisation and the benefits issued from the annotations made available; the latter is vital to the realisation of intelligent content management services. Towards their accomplishment, the existence of commonly agreed vocabularies and syntax, and respectively of commonly agreed semantics and interpretation mechanisms, are essential elements.

Within the context of visual content, these general prerequisites incur more specific conditions issuing from the particular traits of image and video assets. Visual content semantics, as multimedia semantics in general, comes into a multilayered, intertwined fashion [14,15]. It encompasses, amongst others, thematic descriptions addressing the subject matter depicted (scene categorisation, objects, events, etc.), media descriptions referring to low-level features and related information such as the algorithms used for their extraction, respective parameters, etc., as well as structural descriptions addressing the decomposition of content into constituent segments and the spatiotemporal configuration of these segments. As in this chapter semantic annotation is investigated mostly with

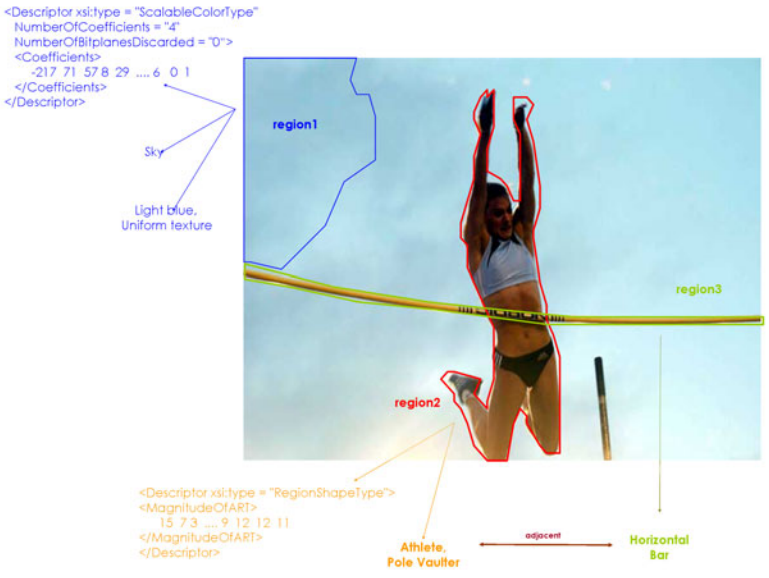


Fig. 1. Multi-layer image semantics

respect to content retrieval and analysis tasks, aspects addressing concerns related to authoring, access and privacy, and so forth, are only shallowly treated.

Figure 1 shows such an example, illustrating subject matter descriptions such as "Sky" and "Pole Vaulter, Athlete", structural descriptions such as the three identified regions, the spatial configuration between two of them (i.e. region2 above region3), and the ScalableColour and RegionsShape descriptor values extracted for two regions. The different layers correspond to different annotation dimensions and serve different purposes, further differentiated by the individual application context. For example, for a search and retrieval service regarding a device of limited resources (e.g. PDA, mobile phone), content management becomes more effective if specific temporal parts of video can be returned to a query rather than the whole video asset, leaving the user with the cumbersome task of browsing through it, till reaching the relative parts and assessing if they satisfy her query.

The aforementioned considerations intertwine, establishing a number of dimensions and corresponding criteria along which image and video annotation can be characterised. As such, interoperability, explicit semantics in terms of liability to automated processing, and reuse, apply both to all types of description dimensions and to their interlinking, and not only to subject matter descriptions, as is the common case for textual content resources.

In the following, we describe the criteria along which we overview the different annotation tools in order to assess them with respect to the aforementioned considerations. Criteria addressing concerns of similar nature have been grouped together, resulting in three categories.

2.1 Input and Output

This category includes criteria regarding the way the tool interacts in terms of requested / supported input and the output produced.

- Annotation Vocabulary. Refers to whether the annotation is performed according to a predefined set of terms (e.g. lexicon / thesaurus, taxonomy, ontology) or if it is provided by the user in the form of keywords and free text. In the case of controlled vocabulary, we differentiate the case where the user has to explicitly provide it (e.g. as when uploading a specific ontology) or whether it is provided by the tool as a built-in; the formalisms supported for the representation of the vocabulary constitute a further attribute. We note that annotation vocabularies may refer not only to subject matter descriptions, but as well to media and structural descriptions. Naturally, the more formal and well-defined the semantics of the annotation vocabulary, the more opportunities for achieving interoperable and machine understandable annotations.
- Metadata Format. Considers the representation format in which the produced annotations are expressed. Naturally, the output format is strongly related to the supported annotation vocabularies. As will be shown in the sequel though, where the individual tools are described, there is not necessarily a strict correspondence (e.g. a tool may use an RDFS[6] or OWL[7] ontology as the subject matter vocabulary, and yet output annotations in RDF[8]). The format is equally significant to the annotation vocabulary as with respect to the annotations interoperability and sharing.
- Content Type. Refers to the supported image/video formats, e.g. jpg, png, mpeg, etc.

2.2 Annotation Level

This category addresses attributes of the annotations per se. Naturally, the types of information addressed by the descriptions issue from the intended context of usage. Subject matter annotations, i.e. thematic descriptions with respect to the depicted objects and events, are indispensable for any application scenario addressing content-based retrieval at the level of meaning conveyed. Such retrieval may address concept-based queries or queries involving relations between concepts, entailing respective annotation specifications. Structural information is crucial for services where it is important to know the exact content parts associated with specific thematic descriptions, as for example in the case of semantic transcoding or enhanced retrieval and presentation, where the parts of interest can be indicated in an elaborated manner. Analogously, annotations intended for training purposes need to include low-level features descriptions and moreover to provide support for their linking with domain notions. Similarly, administrative

[6] http://www.w3.org/TR/rdf-schema/
[7] http://www.w3.org/TR/owl-features/
[8] http://www.w3.org/RDF/

descriptions may or may not be of significance. To capture the aforementioned considerations, the following criteria have been used.

- Metadata Type. Refers to the annotation dimension. For the purposes of this overview, we identify the following types:
  - content descriptive metadata addressing subject matter information,
  - structural metadata describing spatial, temporal and spatioteporal decomposition aspects
  - media metadata referring to low-level features, and
  - administrative, covering descriptions regarding the creation date of the annotation, the annotation creator, etc.
- Granularity. Specifies whether the annotation describes the content assets as a whole or whether it refers to specific parts of it.
  - For image assets, annotation may refer to the whole image, usually termed as scene or global level annotation, or it may refer to specific spatial segments, for which case the terms region-based, local and segment-based annotation are commonly used
  - For video assets, annotation may refer to the entire video, temporal segments (shots), frames (temporal segments with zero duration), regions within frames, or even to moving regions, i.e. a region followed for a sequence of frames. It worths noting that due to the more complex structural patterns applicable for video, many tools besides the annotation functionality provide corresponding visualisation functionalities through the use of timelines. Thereby, the associations of subject matter annotations with respect to the video structure can be easily inspected.
- Localisation. This criterion relates to the supported granularity, and refers to the way in which a part of interest is localised within a content asset. We discriminate two cases with respect to whether localisation is performed automatically (through some segmentation or shot detection algorithm embedded in the tool) or whether manual drawing services are provided.
- Annotation expressivity. Refers to the level of expressivity supported with respect to the annotation vocabulary. For example, in the case an ontology is used for subject matter descriptions, some tools may support only concept based annotation, while others enable to create annotations representing relations among concepts as well.

2.3 Miscellaneous

This category summarises additional criteria that do not fall under the previous dimensions. The considered aspects relate mostly to attributes of the tool itself rather than of the annotation process. As such, and given the scope of this chapter, in the description of the individual tools that follows in the two subsequent Sections, these criteria are treated very briefly.

- Application Type: Specifies whether the tool constitutes a web-based or a stand-alone application.

– Licence: Specifies the kind of licence condition under which the tool operates, e.g. open source, etc.
– Collaboration: Specifies whether the tool supports concurrent annotations (referring to the same media object) by multiple users or not.

3 Tools for Semantic Image Annotation

In this Section we describe prominent semantic image annotation tools with respect to the dimensions and criteria outlined in Section 2. As will be illustrated in the following, Semantic Web technologies have permeated to a considerable degree the representation of metadata, with the majority of tools supporting ontology-based subject matter descriptions, while a considerable share of them adopts ontological representation for structural annotations as well. In order to provide a relative ranking with respect to SW compatibility, we order the tools according to the extend to which the produced annotations bear formal semantics.

3.1 KAT

The K-Space Annotation Tool[9] (KAT), developed within the K-Space[10] project, implements an ontology-based framework for the semantic annotation of images. Figure 2 depicts a screenshot using the KAT 0.2.1 release to annotate the pole vaulter and pole regions in an image depicting a pole vault attempt.

KAT's annotation framework [16] is based on the Core Ontology of Multi-Media (COMM) [13]. COMM extends the *Descriptions & Situations (D&S)* and *Ontology of Information Objects (OIO)* design patterns of DOLCE [17,18], while incorporating re-engineered definitions of MPEG-7 description tools[19,20]. As such, COMM models the various annotation levels and their linking (e.g. of descriptive and structural annotations), while providing MPEG-7 based structural and media descriptions of formal semantics.

KAT currently supports descriptive and structural annotations. A user loaded ontology provides the vocabulary and semantics for the subject matter descriptions. The latter are strictly concept based, i.e. considering the aforementioned annotation example it is not possible to annotate the pole as being next to the pole vaulter, and may refer to the entire image or to specific regions of it. The localisation of image regions is performed manually, using either of the rectangle and polygon drawing tools. COMM provides the definitions for the structural and localisation semantics, leaving them hidden to the user. The supported input ontology languages include RDFS and OWL, and the produced annotations are in OWL.

It should be noted that the COMM based annotation framework implemented by KAT is media independent, i.e. additional content types can be supported as long as respective media management functionalities (e.g. video player) are

[9] htpps://launchpad.net/kat
[10] http://www.k-space.eu/

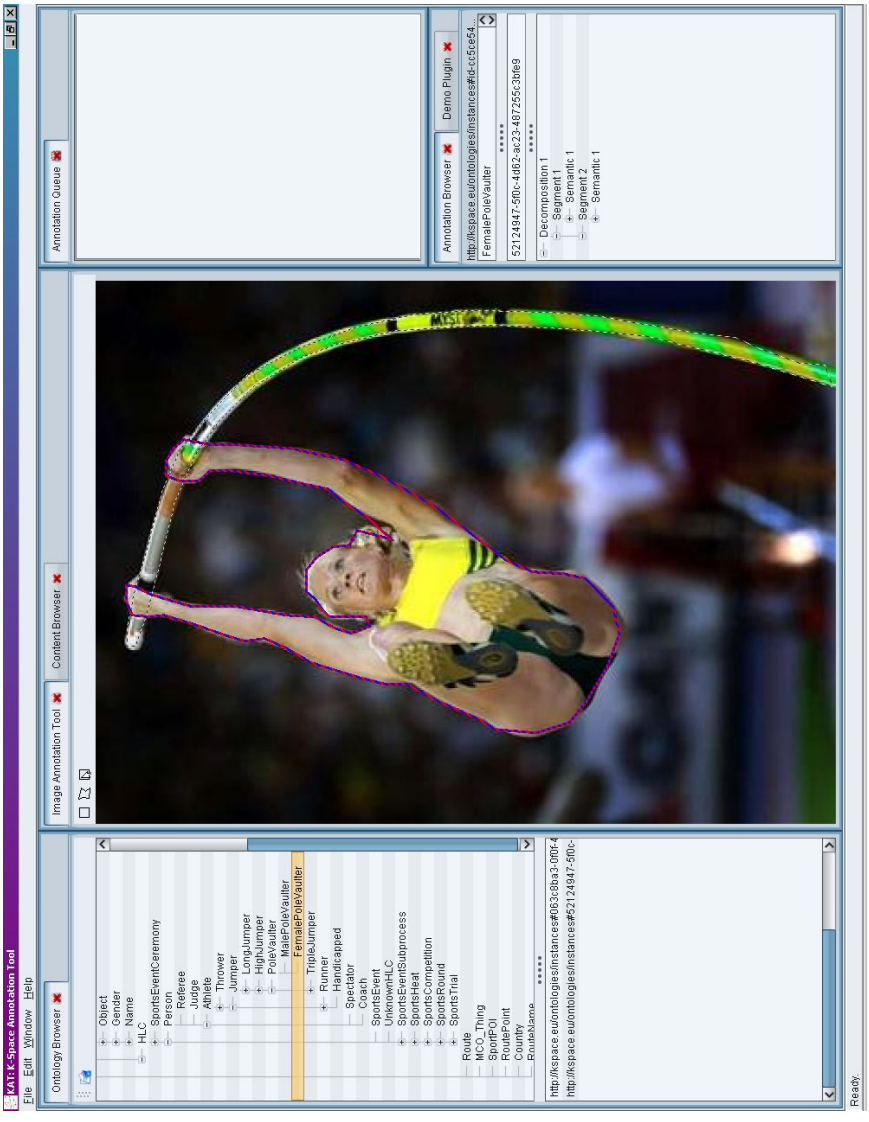


Fig. 2. Example image annotation using KAT

included. Furthermore, the COMM based annotation scheme renders quite straightforward the extension of the annotation dimensions supported by KAT. For example, COMM provides means to represent low-level features and additionally to associate them with the corresponding extraction algorithm and its parameters. Thus, assuming the availability of descriptor extraction capability, KAT could support media annotations as well.

3.2 PhotoStuff

PhotoStuff[11], developed by the Mindswap group[12], is an ontology-based image annotation tool that supports the generation of semantic image descriptions with respect to the employed ontologies. Figure 3 illustrates a screenshot of PhotoStuff 3.33 Beta, used during this overview; following the previous example, two regions have been annotated: the one depicting the female pole vaulter localised using a rectangle and the one depicting the pole, for whose localisation a polygon has been used.

PhotoStuff [21] addresses primarily two types of metadata, namely descriptive and structural. Regarding descriptive annotations, the user may load one or multiple domain-specific ontologies from the web or from the local hard drive, while with respect to structural annotations, two internal, hidden to the user, ontologies are used: the Digital-Media[13] ontology and the Technical[14] one. The two ontologies model the different multimedia content and multimedia segments types in accordance with the MPEG-7 specifications. Furthermore, they provide a simple schema for linking content instances (or parts of it) with the depicted domain-specific instances and its respective low-level descriptors. Specifically, the *depicts* property of FOAF[15] and its inverse, i.e. *depiction*, are used to link a media instance to the depicted content and vice versa, while the properties *descriptor* and *visualDescriptor* provide connection with low-level descriptors. However, nor the representation neither the extraction of such descriptors is addressed.

It is worth noticing that the modeling of content structure reminds a simplified version of well known multimedia ontologies, including Hunter's [9], the acemedia Multimedia Content[16] ontology and the Rhizomik ontology [11]. Specifically, only part of the content and segment class hierarchy has been retained, in combination with a minimal set of decomposition and localisation properties, such as the properties *regionOf*, *startFrame* and *coords*.

As aforementioned, additional types of metadata can be addressed as long as an appropriate ontology is loaded. For example, authoring metadata can be generated if the Dublin Core[17] element set is used in addition to the domain-specific

[11] http://www.mindswap.org/2003/PhotoStuff/
[12] http://www.mindswap.org/
[13] http://www.w3.org/2004/02/image-regions♯
[14] http://www.mindswap.org/ glapizco/technical.owl♯
[15] http://xmlns.com/foaf/0.1/
[16] http://www.acemedia.org/aceMedia/results/ontologies.html
[17] http://dublincore.org/documents/dces/

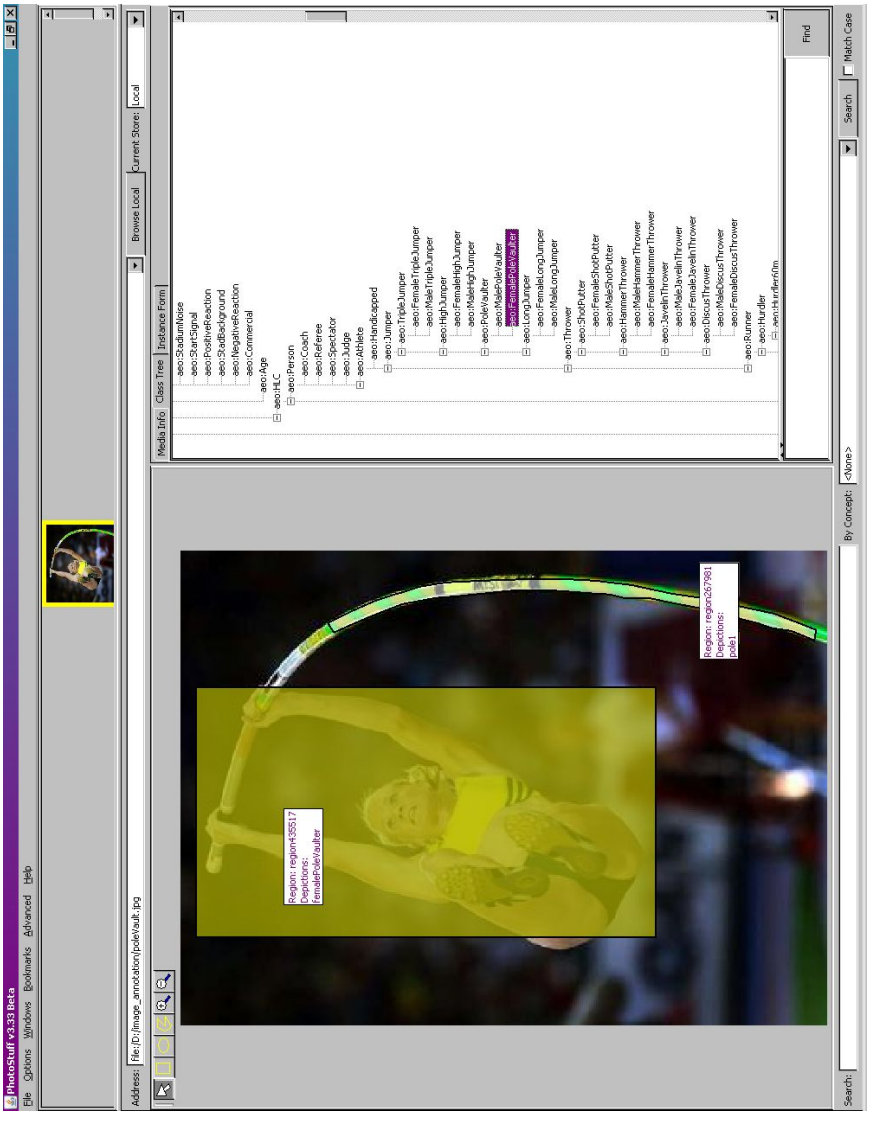


Fig. 3. Example image annotation using PhotoStuff

ontologies. The supported ontology languages are OWL and RDF/RDFS, while the generated annotations are expressed in RDF. Annotations can be attached to either the entire image or to specific regions, using one of the available drawing tools, that is circle, rectangle, and polygon (as an approximation to free hand drawing). Notably, annotations may refer not only to concept instantiations, but also to relations between concept instances already identified in an image. As additional functionalities, PhotoStuff allows keyword-based search through the generated semantic annotations, editing of previously created annotations, as well as parsing and translation of embedded media metadata such as EXIF[18] and IPTC[19].

3.3 AktiveMedia

AktiveMedia[20], developed within AKT[21] and X-Media[22] projects, is an ontology-based cross-media annotation system addressing text and image assets. Figure 4 illustrates a screenshot of the image annotation mode for the AktiveMedia 1.9 release, for the previously considered pole vault annotation example.

In image annotation mode, AktiveMedia supports descriptive metadata with respect to user selected ontologies, stored in the local hard drive [22]. Multiple ontologies can be employed in the annotation of a single image; unlike PhotoStuff though, a single ontology is displayed each time in the ontology browser. AktiveMedia provides also localisation metadata through a simple built-in schema that defines corresponding properties for the representation of coordinates, as well as the linking of media-specific to domain-specific instances through a *hasAnnotation* property.

Annotations can refer to image or region level. To describe an entire image, AktiveMedia provides three free text fields, namely title, content and comment. Utilising the text mode, the respective user entered descriptions can be subsequently annotated with respect to an ontology. Region based annotations are associated to either rectangular or circular regions of the image, and are directly associated with a domain-specific concept.

The supported ontology languages include RDFS and OWL, as well as older semantic web languages such as DAML and DAML-ONT; RDF is used for the representation of the generated annotations. Contrary to Photostuff which uses URIs to identify the class to which an instance belongs, AktiveMedia explicitly models the ontology to which the descriptive annotations refer through a *usesOntology* property, and nests correspondingly the values of *hasConcept* and *hasAnnotationText*, i.e. the class and corresponding instance names. As such, the semantics of generated RDF metadata, i.e. the annotation semantics as it entails from the respective ontology definitions, are not direct but require additional processing to retrieve and to reason over.

[18] http://www.digicamsoft.com/exif22/exif22/
[19] http://www.iptc.org/
[20] http://www.dcs.shef.ac.uk/ ajay/html/cresearch.html
[21] http://www.aktors.org/akt/
[22] http://www.x-media-project.org/

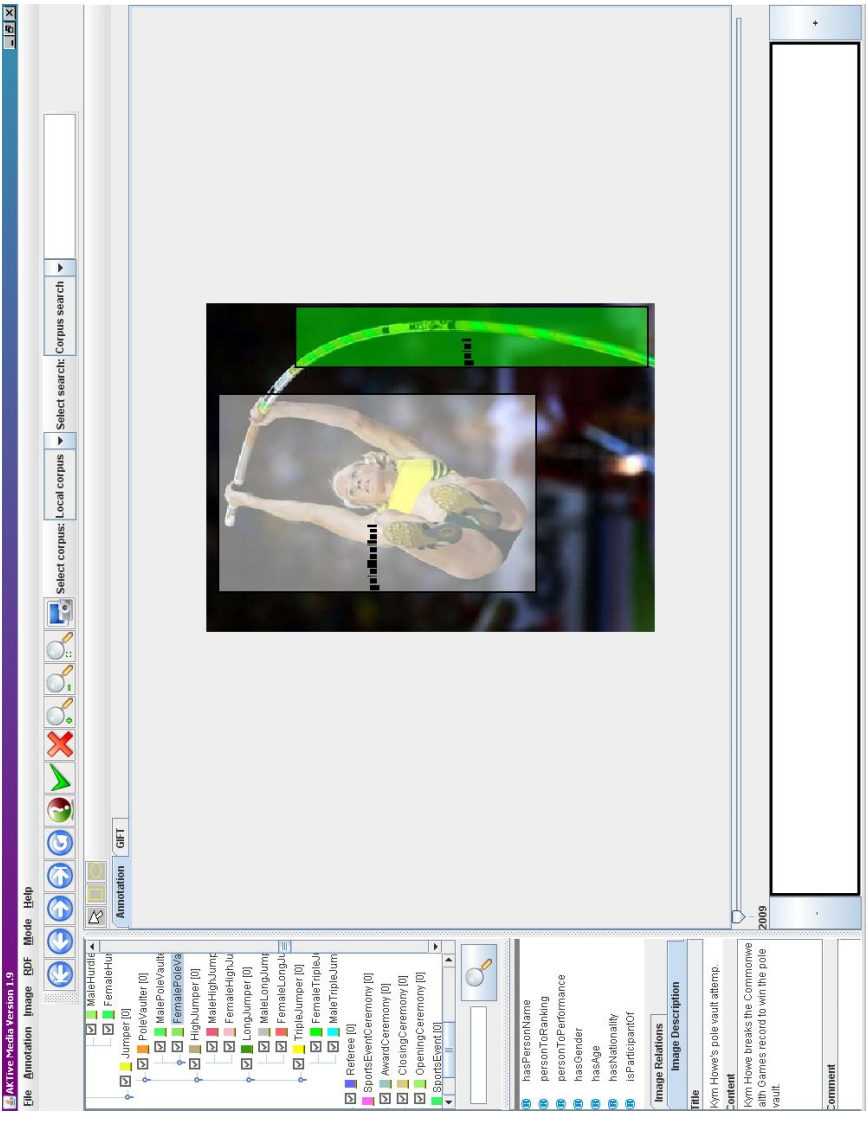


Fig. 4. Example image annotation using AktiveMedia

An interesting feature of AktiveMedia, though not directly related to the task of image annotation, is its ability to learn during textual annotation mode, so that suggestions can be subsequently made to the user, thus realising semi-automatic text annotation. Such facility can prove beneficial when considering the free text and keyword annotations that a user may enter when annotating an image as a whole.

3.4 M-OntoMat-Annotizer

M-Ontomat-Annotizer[23], developed within the aceMedia[24] project, enables the ontology-based representation of associations between domain specific concepts and their respective low-level visual descriptors. Figure 5 illustrates a screenshot of the latest release, namely v0.60, where in the context of the pole vault annotation example, selected descriptors have been extracted and associated to the female pole vaulter and pole instances.

In order to formalise the linking of domain concepts with visual descriptors, M-Ontomat-Annotizer [23] employs the Visual Annotation Ontology (VAO) and the Visual Descriptor Ontology (VDO) [24], both hidden to the user. The VAO serves as a meta-ontology allowing to model domain specific instances as prototype instances and to link them to respective descriptor instances through the *hasDescriptor* property. The VDO[25] models in RDFS the core MPEG-7 visual descriptors (i.e. colour, texture, shape, motion, and localisation)[20]. As in the previous cases, the domain specific instances are in accordance with the domain ontology loaded by the user.

The domain specific instances, and by analogy the extracted descriptor instances, may refer to a specific region or to the entire image. For the identification of a specific region the user may either make use of the automatic segmentation functionality provided by the M-Ontomat-Annotizer or use one of the manually drawing tools, namely the predefined shapes (rectangle and ellipse), free hand and magic wand. To further facilitate the identification of the intended image parts, region merging is also supported. Thereby under-segmentation phenomena can be alleviated, while the annotation of compound objects becomes significantly faster (e.g. merging a face and body region to create a person annotation).

The supported input ontology languages are RDFS and DAML, while the generated annotations are in RDFS. It should be noted that compared to PhotoStuff which provides a corresponding *hasDescriptor* property, M-Ontomat-Annotizer provides in addition both the means to extract descriptors and an ontology to formally represent them. However, it lacks structural descriptions, i.e. explicit representation of spatial decomposition instances and direct descriptive annotations. In a following release within the K-Space project, M-Ontomat 2.0[26]

[23] http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html
[24] http://www.acemedia.org/aceMedia
[25] http://www.acemedia.org/aceMedia/files/software/m-ontomat/acemedia-visual-descriptor-ontology-v09.rdfs
[26] http://mklab.iti.gr/m-onto2

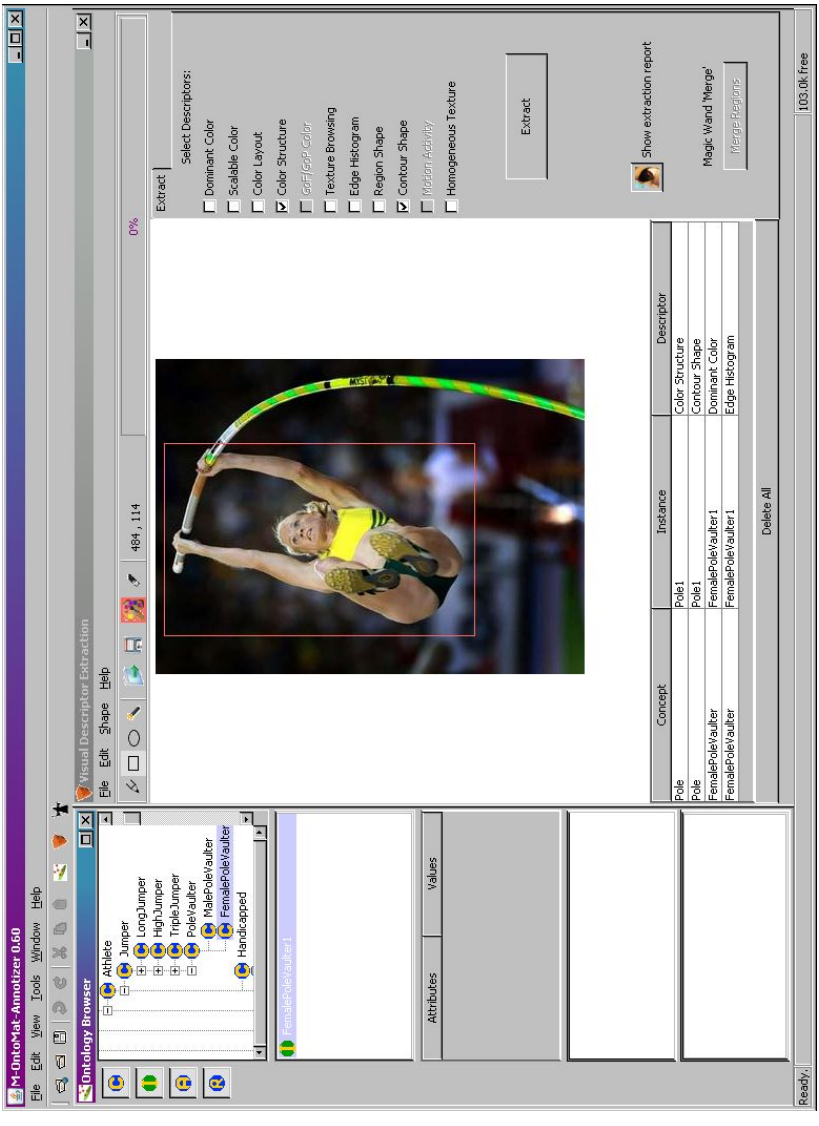


Fig. 5. Example image annotation using M-Ontomat-Annotizer

provides support for descriptive and structural annotations in the typical semantic search and retrieval sense.

3.5 Caliph

Caliph[27] is an MPEG-7 based image annotation tool that supports all types of MPEG-7 metadata among which descriptive, structural, authoring and low-level visual descriptor annotations. In combination with Emir, they support content-based retrieval of images using MPEG-7 descriptions. Figure 6 illustrates two screenshots corresponding to the generic image information and the semantic (descriptive) annotation tabs.

Contrary to the aforementioned tools, Caliph allows descriptive annotations only at image level [25]. The descriptions may be either in the form of free text or structured, in accordance to the SemanticBase description tools provided by MPEG-7 (i.e. Agents, Events, Time, Place and Object annotations [26]). The so called semantic tab (illustrated at the right part of Figure 6) allows for the latter, offering a graph based interface. A subset of the relations specified in MPEG-7 are available; it is not clear though how to extend them, while additional issues emerge to users unfamiliar with MPEG-7 tools with respect to which relations and how should be used.

3.6 SWAD

SWAD[28] is an RDF-based image annotation tool that was developed within the SWAD-Europe project[29]. The latter ran from May 2002 to October 2004 and aimed to support the Semantic Web initiative in Europe through targeted research, demonstrations and outreach activities. Although the SWAD tool [27] has not been maintained since, we chose to provide a very brief description here for the purpose of illustrating image annotation in the Semantic Web as envisaged and realised by that time, as a reference and comparison point for the various image annotation tools that have been developed afterwards.

Figure 7 illustrates a screenshot of SWAD's web-based interface. Different tabs allow to insert descriptions regarding who or what is depicted in the image (person, object, event), when and where it was taken, and additional creator and licensing information as described in the respective SWAD deliverable[30]. When entering a keyword description, the respective Wordnet[31] hierarchy is shown to the user, assisting her in determining the appropriateness of the keyword and in selecting descriptions of further accuracy. The number of RDF vocabularies the tool utilises is quite impressive, including FOAF, the Dublin Core element set, RDFiCalendar[32] as well as an experimental by the time namespace for WordNet, the latter in an attempt towards explicit subject matter semantics.

[27] http://www.semanticmetadata.net/features/
[28] http://swordfish.rdfweb.org/discovery/2004/03/w3photo/annotate.html♯
[29] http://www.w3.org/2001/sw/Europe/
[30] http://www.w3.org/2001/sw/Europe/reports/report_semweb_access_tools/♯WN
[31] http://wordnet.princeton.edu/
[32] http://www.w3.org/2002/12/cal/

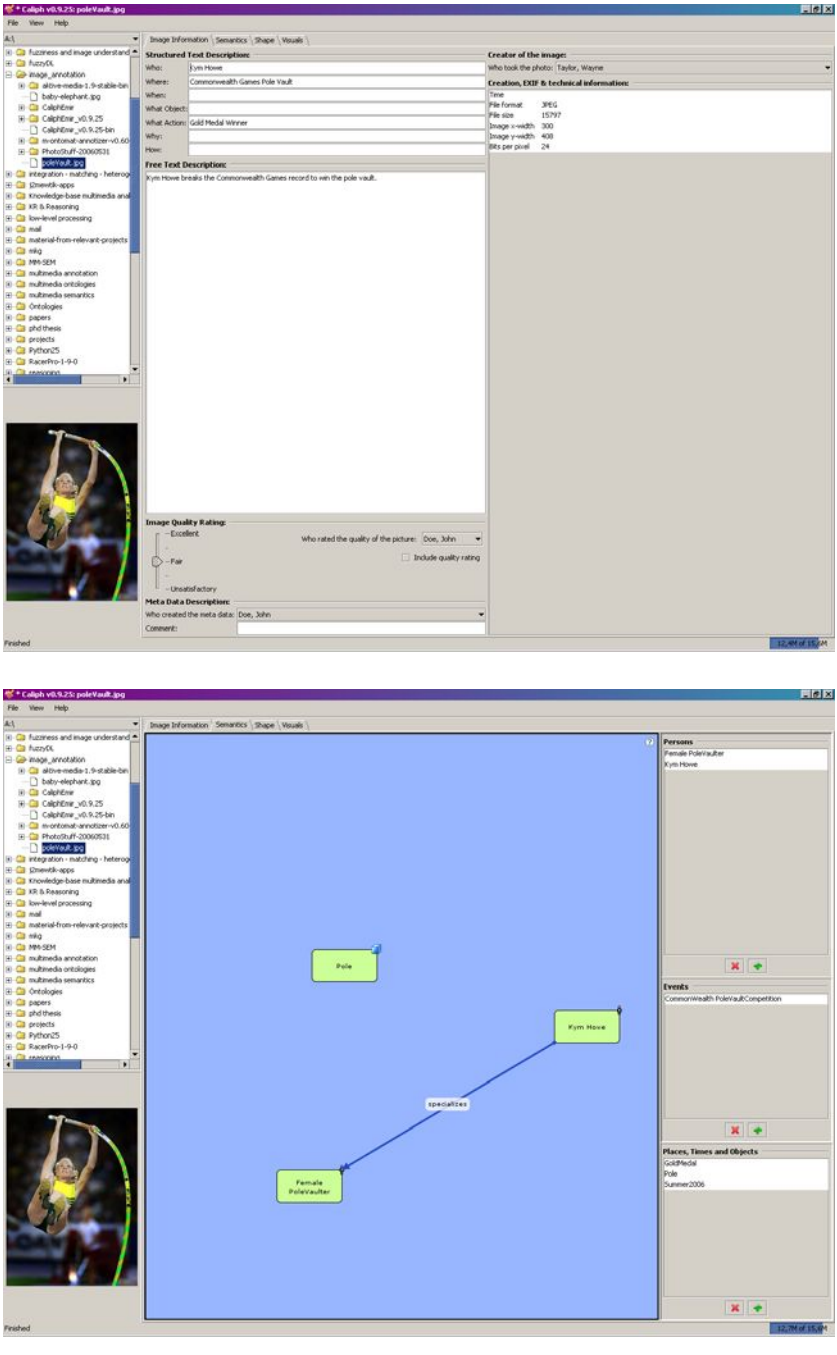


Fig. 6. Example image annotation using Caliph; generic (image information) and (semantic) descriptive annotation tabs

Fig. 7. Example image annotation using the SWAD annotation tool

3.7 LabelMe

LabelMe[33] is a database and web-based image annotation tool, aiming to contribute in the creation of large annotated image databases for evaluation and training purposes [28]. It contains all images from the MIT CSAIL[34] database, in addition to a large number of user uploaded images. Figure 8 depicts a screenshot using LabelMe to annotate the pole vaulter and pole objects of the example image.

LabelMe [28] supports descriptive metadata addressing in principle region-based annotation. For each image, randomly selected from the database or user uploaded, the user my annotate as many objects as desired in order to further enrich already annotated images or provide new ones. There is no functionality for adopting a controlled vocabulary; instead each user may enter as many words as she considers appropriately in order to precisely describe the annotated object. For the localisation of regions, a manual drawing facility is provided. Specifically, the user defines a polygon enclosing the annotated object through a set of control points. Defining a polygon that equals the entire image allows for scene level annotations; we note though, that such behaviour rather diverges from the intended goal, i.e. the construction of a large, rich and open data set of annotated objects.

The resulting annotations are stored in XML format, with the choice of XML based on portability and extensibility concerns. A proprietary schema is followed, including attributes such as *filename*, *folder*, and *object* that allow to represent information regarding the image and its location, and the annotation itself. Additional elements under the *object* attribute, allow to represent the various words ascribed to the annotated object, the coordinates of the polygon, the date the annotation was created, and whether it has been verified by the use or not.

Summing up, LabelMe addresses image annotation from a rather different perspective than the rest of the tools. Its focus on requirements related to object recognition research, rather than image search and retrieval, entails different notions regarding the utilisation, sharing and purpose of annotation. In a way, it is closer to M-Ontomat-Annotizer, but lacking formal domain specific as well as low-level descriptors representation; in addition the extraction of descriptors and their linking with domain concepts is left up to the algorithms using the annotations to implement object recognition.

3.8 Application-Specific Image Annotation Tools

Apart from the afore described semantic image annotation tools, a variety of application-specific tools are available. Some of them relate to Web 2.0 applications addressing tagging and sharing of content among social groups, while others focus on particular application domains, such as medical imaging, that impose additional specifications pertaining to the individual application context.

[33] http://labelme.csail.mit.edu/
[34] http://web.mit.edu/torralba/www/database.html

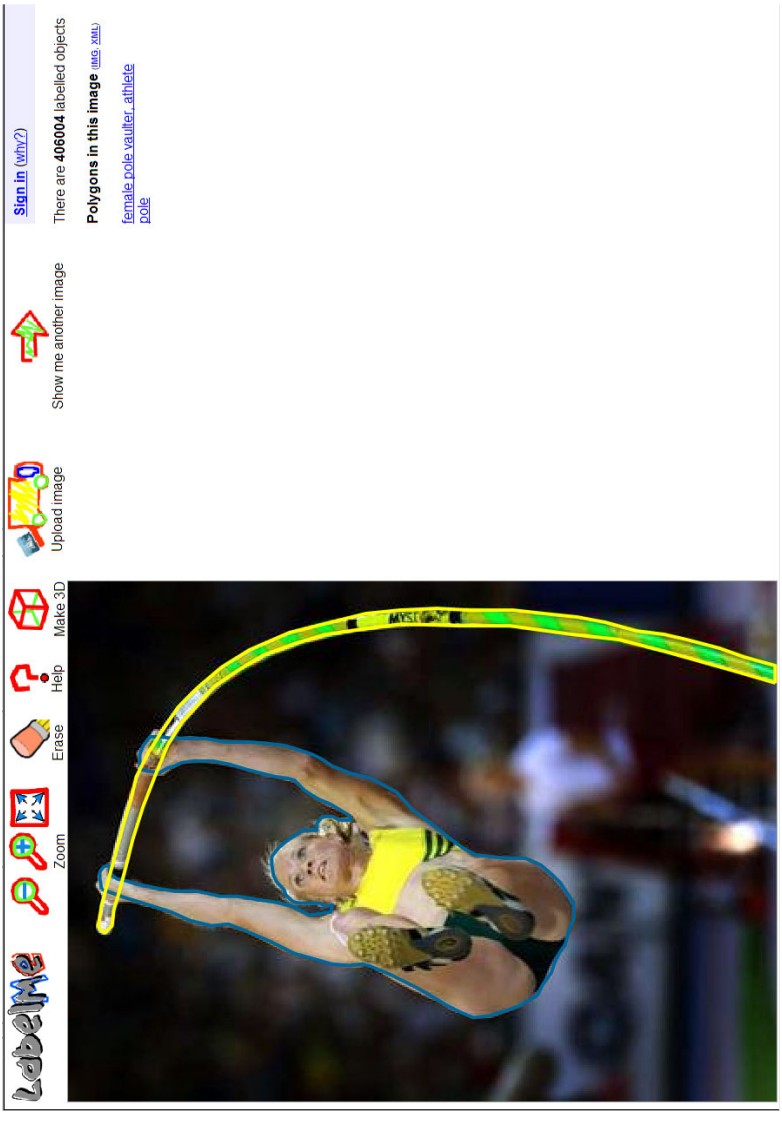


Fig. 8. Example image annotation using LabelMe

Aspiring to specific usages, these tools induce different perspectives and specifications on the annotation process. In the following, we briefly go through some representative examples.

iPad (image Physician Annotation Device) supports clinicians in the semantic annotation of radiological images [29]. Using the provided drawing facilities the user selects the regions of interest and attaches to them descriptions referring to anatomical, pathological and imaging observations. Utilising radiology specific ontologies, iPad enhances the annotation procedure by suggesting more specific terms and by identifying incomplete descriptions and subsequently prompting for missing parts in the description (e.g. "enlarged" is flagged as incomplete while "enlarged liver" is acceptable). The created annotations are stored in XML based on a proprietary schema, which can be subsequently transformed into different standard formats such as DICOM[35] and HL7[36] in order to support seamless and effective interchange of medical data across heterogenous systems. Furthermore, aspiring to enhance interoperability with Semantic Web technologies, translation to OWL is also provided.

FotoTagger[37] builds on the paradigm of the popular Web 2.0 application of Flickr[38]. It comes both as a Web-based and a standalone application, allowing users to attach tags to specific image regions with the purpose of enhancing content management in terms of accessing and sharing it. It supports descriptive and structural metadata, where region localisation is performed through a rectangle drawing facility. The produced descriptions are in RDF/XML following a proprietary schema[39] that models the label constituting the tag, its position (the label constitutes a rectangle region in itself), and the position of the rectangle that encloses the annotated region in the form of the top left point coordinates and width and height information. Furthermore, general information about the image is included such as image size, number of regions annotated, etc. Oriented towards Web 2.0, FotoTagger places significant focus on social aspects pertaining to content management, allowing among others to publish tagged images to blogs and to upload/download tagged images to/from Flickr, while maintaining both FotoTagger's and Flickr's descriptions.

Given the general purpose scope of the current survey, elaborating into the various application specific tools and the particular annotation aspects they introduce falls beyond the intended scope. It is worth noting though that as the corresponding literature shows, interoperability, even when not necessarily in conformance with the SW notion, constitutes a major concern.

3.9 Discussion

The aforementioned overview reveals that the utilisation of Semantic Web languages for the representation, interchange and processing of image metadata has

[35] http://www.rsna.org/Technology/DICOM/
[36] http://www.hl7.org/
[37] http://www.fototagger.com/
[38] http://www.flickr.com/
[39] http://www.cogitum.com/fototagger/

permeated semantic image annotation. This is particularly evident for subject matter descriptions, where from the examined tools only Caliph and LabelMe follow a different approach. Caliph though is more oriented towards content-based annotation and retrieval in the "traditional" multimedia community sense, and thus adopts the MPEG-7 perspective. The choice of a standard representation shows the importance placed on creating content descriptions that can be easily exchanged and reused across heterogenous applications, and works like [10,11,30] provide bridges between MPEG-7 metadata and the Semantic Web and existing ontologies. The case is different with LabelMe, where the tool serves a very specific purpose that of creating a large object annotated database, and does not address retrieval tasks. Even in this case though, one can speculate that adopting a more formal vocabulary the descriptions added by users could be better exploited.

The representation of structural and localisation information appears to be also wide established, illustrating that there is a considerable need to attach descriptions to specific content parts. It is interesting that in all tools supporting such kind of description, an ontology has been used (Caliph is the exception following the MPEG-7 decomposition schemes), which is hidden from the user. Thus unlike subject matter descriptions, where a user can choose which vocabulary to use (in the form of a domain ontology, a lexicon or user provided keywords), structural descriptions are tool specific. The different ontologies used by the tools reflect the undergoing efforts towards making structural semantics explicit and the variations witnessed due to the loose semantics of the corresponding MPEG-7 definitions on which these ontologies are based on [31,12]. Media related information on the other hand in terms of low-level descriptors can be represented in a rather straightforward manner, practically eliminating interoperability issues. The choice of whether or not to include support for media related annotations depends on whether the tool aims to contribute to analysis tasks as well.

Summing up, the choice of a tool depends primarily on the intended context of usage, which provides the specifications regarding the annotation dimensions supported, and subsequently on the desired formality of annotations, again related to a large extend to the application context. Thus for semantic retrieval purposes, where semantic refers to the SW perspective, KAT, PhotoStuff, SWAD and AkiveMedia would be the more appropriate choices. In cases that domain semantics need to be associated with low-level representations a tool like M-Ontomat-Annotizer or KAT should be selected. Finally, when adopting a strict MPEG-7 perspective is required, then a tool like Caliph should be preferred. We note the difference between MPEG-7 metadata, i.e. XML descriptions according to the respective Description Schemes, and MPEG-7 compliant metadata that can be as well in RDFS or OWL. Table 1, summarises the comparative study of the examined image annotation tools with respect to the *Input & Output* and *Annotation Level* criteria described in Section 2. Regarding the miscellaneous criteria (see Section 2.3), as illustrated in the individual tools descriptions, none provides supports for collaborative annotation. Web-based and stand-alone are

Table 1. Image annotation tools summarisation. In the *A*nnotation Vocabulary field, "U" denotes user-entered vocabularies, while "T" refers to vocabularies embedded within the tool, and thus hidden to the user.

Tool	**Input & Output**		**Annotation level**			
	Metadata Format	*Annotation Vocabulary*	*Metadata Type*	*Granularity*	*Localisation*	*Expressivity*
KAT	OWL	U: domain ontology (RDFS/OWL) T: COMM	descriptive, structural	image, region-based	rectangle, polygon	concepts
PhotoStuff	OWL	U: domain ontology (OWL) T: Digital Media, Technical ontologies	descriptive, structural, administrative	image, region-based	rectangle, circle, polygon	concepts, relations
AktiveMedia	RDF	U: domain ontology (RDFS/DAML/OWL), free text T: customised structural schema	descriptive	image, region-based	rectangle, circle	concepts
M-Ontomat Annotizer	RDF	U: domain ontology (RDFS/DAML) T: VAO, VDO	descriptive, media	image, region-based	rectangle, eclipse, polygon, free hand	concepts
Caliph	MPEG-7/XML custom XML	U: free text, keywords T: MPEG-7	descriptive, structural, media, administrative	image	N/A	concepts relations
SWAD	RDF	U: free text, keywords T: Dublin Core, FOAF, WordNet	descriptive, administrative	image	N/A	concepts relations
LabelMe	custom XML	U: free text, keywords	descriptive	image region-based	polygon	concepts

equally popular choices, and all tools are freely available for non-commercial use[40].

4 Tools for Semantic Video Annotation

The increase in the amount of video data deployed and used in today's applications not only caused video to draw increased attention as a content type, but also introduced new challenges in terms of effective content management. Image annotation approaches as described in the previous section can be employed for the description of static scenes found in a video stream; however, in order to capture and describe the information issuing from the temporal dimension featuring a video object, additional requirements emerge.

In the following we survey typical video annotation tools, highlighting their features with respect to the criteria delineated in Section 2. In addition to tools that constitute active research activities, we also examine representative video annotation systems that despite no longer maintained, are still accessible and functional; however, tools that are neither maintained nor accessible have not been considered. In the latter category fall tools such as VIDETO[41], Ricoh Movie Tool[42], or LogCreator[43]. It is interesting to note that the majority of these tools followed MPEG-7 for the representation of annotations. As described in the sequel, this favourable disposition is still evident, differentiating video annotation tools from image ones, where the Semantic Web technologies have been more pervasive.

4.1 VIA

The Video and Image Annotation[44] (VIA) tool has been developed by the MK-Lab[45] within the BOEMIE[46] project. A snapshot of the interface of the tool, during a shot annotation of a video file is shown in Figure 9. The shot records a pole vaulter holding a pole and sprinting at the jump point.

VIA supports descriptive, structural and media metadata of image and video assets. Descriptive annotation is performed with respect to a user loaded OWL ontology, while free text descriptions can also be added. Administrative metadata follow a customised schema internal to the tool, including information about the creator of the annotations, the date of the annotation creation, etc. A customised XML schema is also used for the representation of structural information, allowing for example to nest a video segment as part of a video and to define its start and end frame / time interval. The produced metadata can be exported either in XML or as in a more human readable format in textual format.

[40] In many cases, the source code is available for research purposes
[41] http://www.zgdv.de/zgdv/zgdv/departments/zr4/Produkte/videto/
[42] http://www.ricoh.co.jp/src/multimedia/MovieTool/
[43] http://project.eia-fr.ch/coala/demos/demosFrameset.html
[44] http://mklab.iti.gr/project/via
[45] http://mklab.iti.gr
[46] http://www.boemie.org

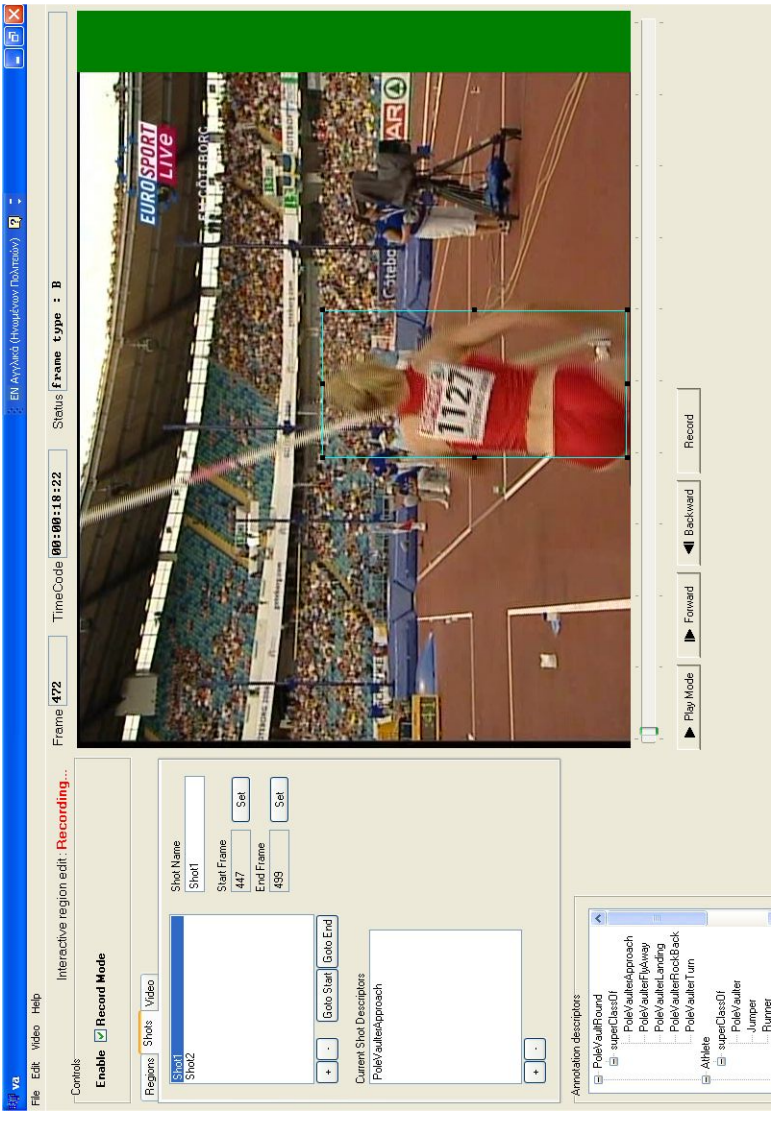


Fig. 9. Example video annotation using VIA

Regarding image (and by consequence frame) annotation, the granularity levels supported include the entire image and specific still regions. The localisation of regions is performed either semi-automatically, providing the user a segmented image and allowing her to correct it by region merging, or manually, using one of the drawing functionalities provided, i.e. free hand, polygon, circle, rectangle. In the case of image annotation, the tool supports additionally the extraction of MPEG-7 visual descriptors per each annotated region, based on MPEG-7 XM [32], so the annotation outcome can be used as a training set for semantics extraction algorithms.

Regarding video annotation, the supported annotation granularity may refer respectively either to the entire video, video segments, moving regions, frames or even still regions within a frame. The annotation can be performed in real time, on MPEG-1 and MPEG-2 videos, using an interface consisting of three panels. The first one is concerned with region annotation, in which the user selects rectangular areas of the video content and subsequently adds corresponding annotations. The other two panels are used for annotation at shot and video level respectively. Shot boundaries are defined manually, by selecting its start and end frames. An important feature about region annotation is that the user can drag the selected region whereas at the same time the video is playing, so as to follow the movement of the desired region.

The annotations performed with VIA can be saved as annotation projects, so that the original video, the imported ontologies, and the annotations can be retrieved and updated at a later time. VIA is publicly available.

4.2 VideoAnnEx

The IBM VideoAnnEx[47] annotation tool addresses video annotation with MPEG-7 metadata. Although the project within which VideoAnnEx was developed has finished and the tool is no longer maintained, VideoAnnEx is accessible and provides an illustrative case of content annotation in accordance to the MPEG-7 initiative. A screenshot of the annotation interface of the tool is shown in Figure 10.

VideoAnnex supports descriptive, structural and administrative annotations according to the respective MPEG-7 Description Schemes. Descriptive metadata may refer at the entire video, at specific video segments (shots), or even at still regions within keyframes. The tool supports default subject matter lexicons in XML format, and additionally allows the user to create and load her own XML lexicon, design a concept hierarchy through the interface menu commands, or insert free text descriptions.

As illustrated in Figure 10, the VideoAnnEx annotation interface consists of four components. On the upper right-hand corner of the tool is the Video Playback window with shot information. It allows standard VCR operations (such as play, pause, etc.) and loads video files in MPEG-1 or MPEG-2 format. On the upper left-hand corner of the interface is the Shot Annotation panel with

[47] http://www.research.ibm.com/VideoAnnEx/index.html

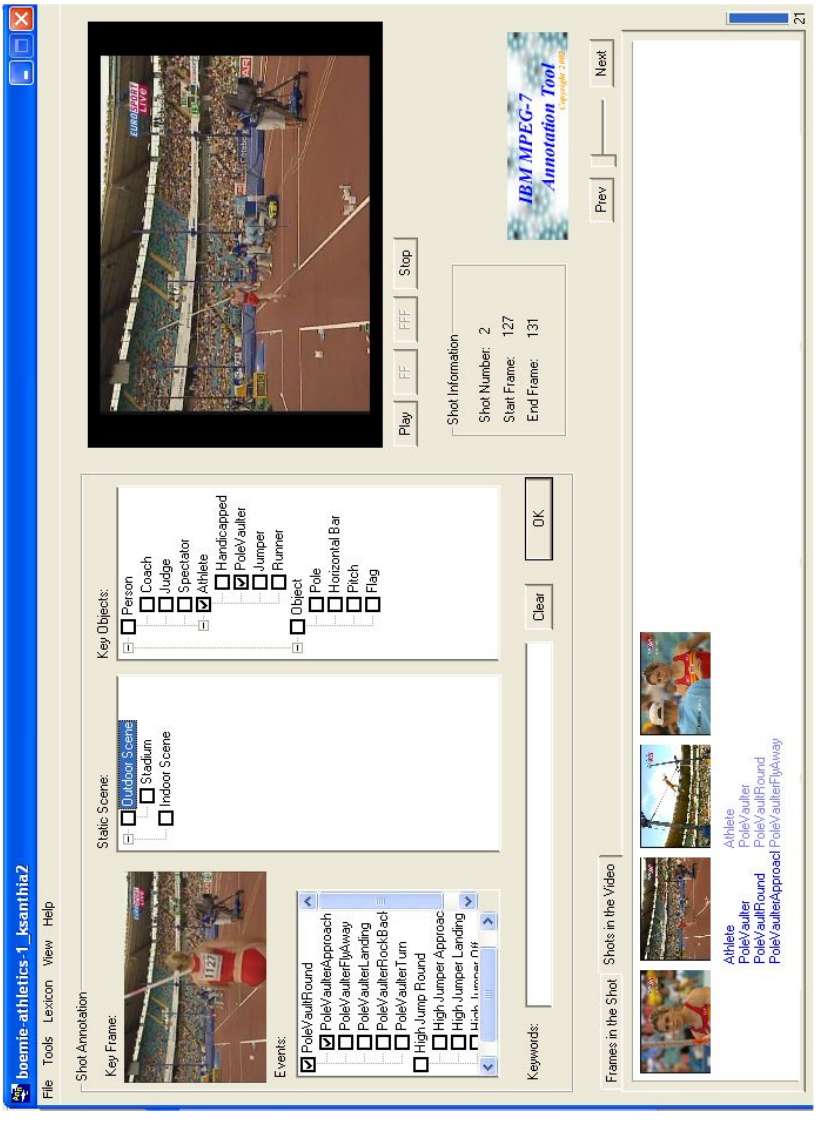


Fig. 10. Example video annotation using VideoAnnEx

a key frame image display. The tool supports either automatic shot detection or loading of customised video segmentation lists. In the space between the two display windows, the concept hierarchy of the loaded XML lexicon is displayed.

On the bottom part of the tool, two views are available of the annotation preview: one contains the I-frames of a shot and the keyframes of each shot in the video, respectively. The user may see under the keyframe of each shot, the annotation this shot has received, up to this point. A fourth component, not shown in Figure 10, is the region annotation pop-up window for specifying annotated regions using a rectangle. After the text annotations are identified on the shot annotation window, each description can be associated with a corresponding rectangular region on the selected key frame of that shot.

It worths noticing an extra feature this tool offers, which is annotation learning. This utility assist the annotator in finding similar shots and labeling them with the same descriptions. VideoAnnEx runs on Windows platforms and can be used under the IBM terms of use[48].

4.3 Ontolog

Ontolog[49] is a tool for annotating video and audio sources using structured sets of terms/concepts. It is a java application, designed and developed as part of a Ph.D. thesis in the Norwegian University of Science and Technology. Though not maintained the past four years, the source code is available upon request. A screenshot of a video annotation process is shown in Figure 11.

Ontolog addresses various types of metadata, including descriptive, structural and administrative. Descriptive annotations are inserted according to one or more RDFS ontologies, imported or created by the user. The user can further enrich the subject matter descriptions by introducing additional properties. For the representation of administrative metadata, Ontolog provides by default two ontologies, namely the Dublin Core Element Set and the Dublin Core Qualified Element Set. Structural descriptions referring to video segments are created in correspondence with user-defined intervals, following the simplified structure representation defined in the Ontolog Schema[50] ontology. The produced annotations are in RDF.

Ontolog's interface consists of four components: a Media Panel, an Ontology Editor, a Logging Panel and a Property Editor. The media panel handles the video assets that are contained in an annotation project. For media loading either Quicktime (for Java) or the JMF framework can be used (and the corresponding media formats). The Ontology Editor provides mechanisms for the definition of concept hierarchies; properties defining relations between concepts can be specified in the Property Editor. Each property may optionally specify what kind of concept it may be applied to (domain) and what kind of values it may take (range).

[48] http://www.ibm.com/legal/
[49] http://www.idi.ntnu.no/ heggland/ontolog/
[50] http://www.idi.ntnu.no/ heggland/ontolog/ontolog-schema#

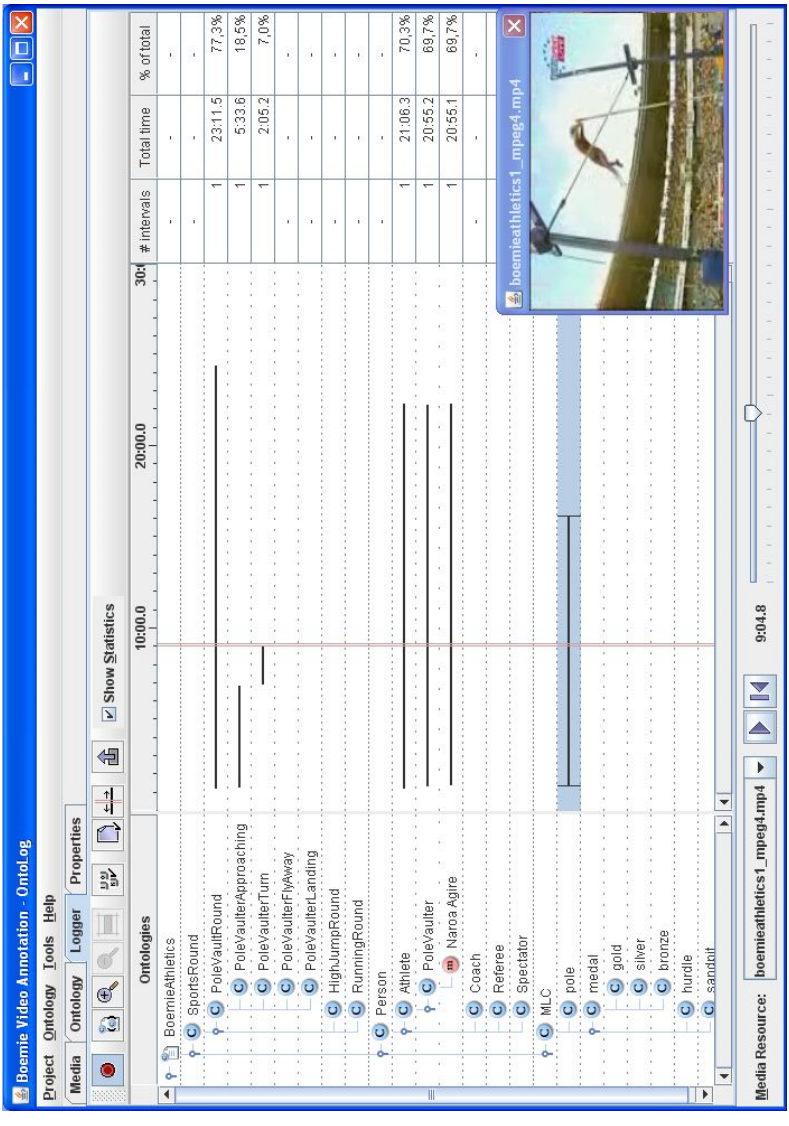


Fig. 11. Example video annotation using Ontolog

OntoLog's logging interface is shown in Figure 11. The left panel contains the ontologies the user is working with. The right panel displays a horizontal timeline with the annotation intervals corresponding to each concept in the ontology (referred to as "annotation strata", in the context of this tool). Each stratum consists of a series of interval lines along the time axis, indicating the positions of the media resource where the concept is present. The strata corresponding to collapsed concepts (concepts with subconcepts that are not currently displayed in the tree) are shown as of lines of varying thickness. This is because they represent an aggregation of the strata beneath them in the hierarchy. The time intervals are specified manually, i.e. automatic or semiautomatic temporal segmentation is not supported.

An extra feature the tool offers involves the extraction of simple statistics, such as the length of the intervals per concept/instance, the percentage of this length with regard to the total length of the media resource, etc. In addition, the resulting set of annotation intervals (i.e. strata) serves as a visual index to the media file, with dynamic level of detail due to the tree-based, aggregating visualisation technique. Moreover, the logging panel provides a SMIL export function. This produces a SMIL file [33], specifying a "virtual edit" of the selected media resource, namely a concatenation of the intervals related to the currently selected concept. For instance, a user may create a SMIL version of the "Olympics 2008" video with just the parts with running events. Concluding, Ontolog is accompanied with the Ontolog Crawler software[51], which implements many search queries and facilitates the task of retrieval.

4.4 Advene

Advene[52] (Annotate Digital Video, Exchange on the NEt) is an ongoing project in the LIRIS[53] laboratory at University Claude Bernard Lyon. Advene addresses a twofold goal, namely to provide an annotation model for sharing descriptions about digital video documents, and to serve as an authoring tool for visualising and accessing hypervideos, i.e. videos augmented with annotations. A screenshot of the interface of the tool during a video annotation is shown in Figure 12.

Annotation in Advene is performed according to user-created schemas which group together descriptions of related annotation dimensions (i.e. subject matter, administrative, etc.). Schemas including concept level descriptions are referred as annotation types, while schemas defining relations between concepts, comprise the relation types. Each annotation type defines in addition a content type for its annotations, in the form of a MIME type (text/plain,text/XML, image/jpeg, audio/wav, etc.). If the type is text/XML, it can be further constrained by a structured description (e.g. using DTD). Analogously, a relation type defines a content type for its instances. In addition, it specifies the number of participating annotations and their respective types. The generated annotations may contain

[51] http://folk.ntnu.no/heggland/ontolog-crawler/login.php
[52] http://liris.cnrs.fr/advene/
[53] http://liris.cnrs.fr/

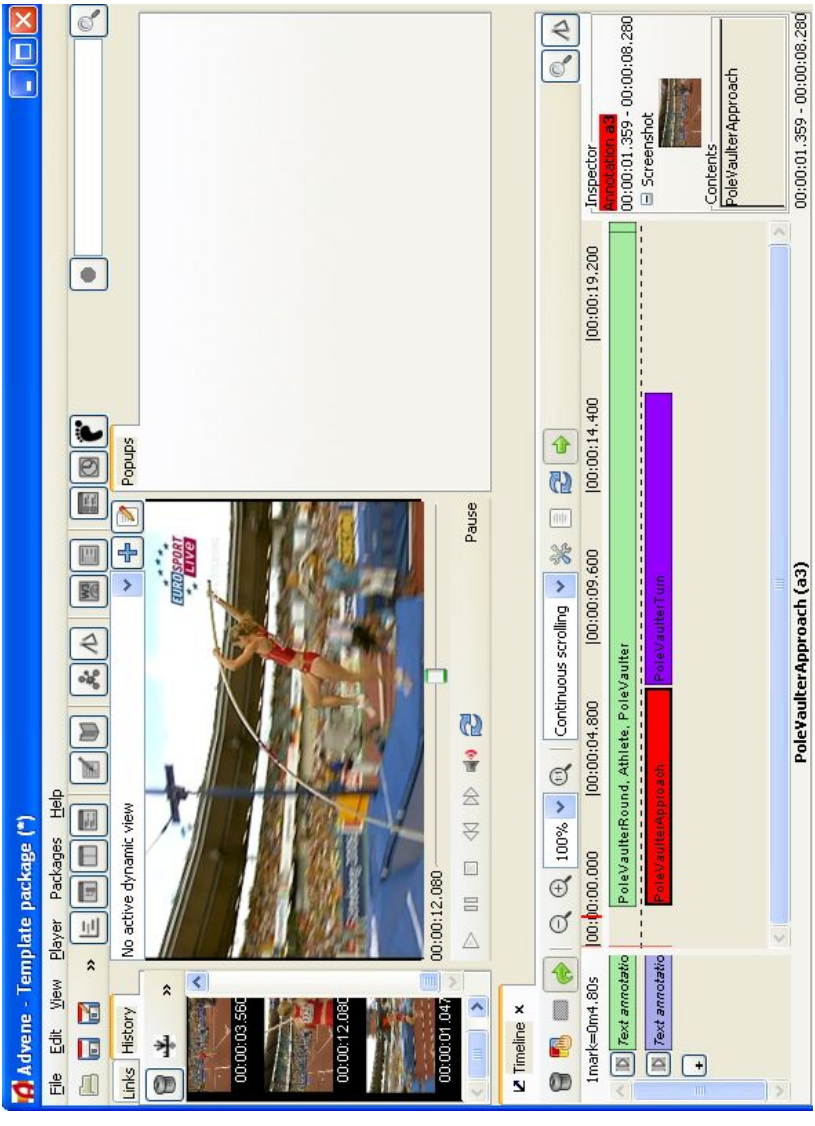


Fig. 12. Example video annotation using Advene

descriptive, administrative and structural information and may pertain to the entire video or to temporal segments of it. The output is stored in XML format.

Advene uses the VLC video player[54] that supports various audio and video formats, such as MPEG-1, MPEG-2, MPEG-4, DivX, mp3, ogg, and so on, as well as DVDs, VCDs, and various streaming protocols. The tool offers the ability to dynamically control the video player based on the annotations, as well as to define dynamic visualisation means (views). Moreover, it allows multiple ad-hoc views of annotations (e.g. timeline, tree-view, transcription, etc) and the annotations' content may be displayed as SVG caption on the video. The annotations along with the views may be shared in packages independently from the audiovisual material, through an embedded web server which dynamically generates XHTML[55] documents, using data taken from the annotations.

The main focus of Advene is not so much to support the annotation task itself, but rather to offer visualisation means and the functionalities afore described, so as to facilitate the management of readily available annotation metadata. This accounts for the variety of annotation formats that the tool supports, among which TXT files where each line contains the start time, the end time and the contents of the annotation separated by tabs, SRT[56] subtitle files, XI[57] XML files, EAF[58] files produced with ELAN, PRAAT[59] files, CMML[60] files, Anvil files, MPEG-7 files containing only free text annotations, AnnotationGraph[61], Shotdetect and IRI files[62]. Advene is distributed under the GPL conditions and runs on Linux, Windows and MacOS platforms.

4.5 Elan

Elan[63], developed at the Max Planck Institute for Psycholinguistics[64], is an annotation tool designated primarily for linguistic purposes, involving issues related to analysis of language, sign language and gestures in audio and video resources. A screenshot showing a video annotation along with the user interface of Elan is shown in Figure 13.

The tool addresses exclusively descriptive annotations, where an annotation may be a sentence, word or gloss, and in general any description of a feature observed in the media file. The user may also create and use her own vocabularies, containing frequently used terms, so that she avoids repetitive typing of the same

[54] http://www.videolan.org/vlc/
[55] http://www.w3.org/TR/xhtml1/
[56] http://www.matroska.org/technical/specs/subtitles/srt.html
[57] http://www.ananas.org/xi/index.html
[58] http://www.let.kun.nl/sign-lang/echo/ELAN/ELAN_intro.html
[59] http://www.fon.hum.uva.nl/praat/
[60] http://www.anodex.net/
[61] http://sourceforge.net/projects/agtk
[62] http://www.iri.centrepompidou.fr/
[63] http://www.lat-mpi.eu/tools/elan/
[64] http://www.mpi.nl

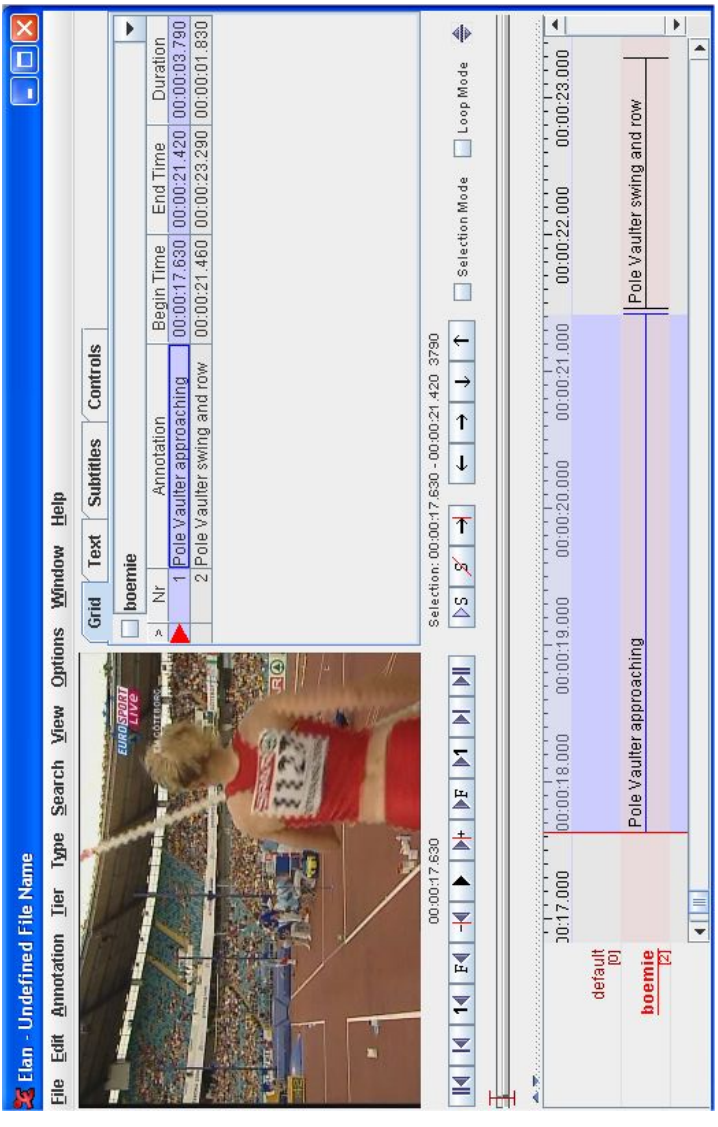


Fig. 13. Example video annotation using Elan

term. The produced metadata is in XML format and refer either to the entire video or to temporal segments of it.

Annotations, in Elan, can be created on multiple layers, called tiers which can be hierarchically interconnected, so that annotations in a referring tier are linked to annotations on a referred tier. This feature pertains to the linguistic design and multi-language support of the tool, so that different tiers correspond to different translations. However, it can also be used so as to simulate a structural description of the content (parent tiers describe video objects and children tiers describe segments of the former) or, in general, produce annotations containing meta information about other annotations.

In the upper left part of the interface of Elan is the media player. The kind and number of supported video formats depend upon the media framework the user has installed. There are three supported media players, that is Windows Media Player, QuickTime and JMF. Below the player window, there are the media control buttons. Apart from the standard VCR operations, the tool supports browsing based on frames and on user-assigned annotations. The lower part of the interface includes the timeline viewer. There are multiple timelines, one for each particular tier. The timeline viewer displays the tiers and their annotations, whereby each annotation corresponds to a specific time interval. With regard to the localisation of the video content, the user has to manually select the intervals, she wants to annotate.

Further, the tool offers keyword-based and regular expression based search functionalities that facilitate the task of retrieval, as well as it supports a variety of import/export functions with formats, such as Shoebox/Toolbox[65], CHAT[66], Transcriber[67], Praat[68], SMIL[33], etc. Elan is distributed under the GPL conditions and runs on Windows, MacOS and Linux platforms.

4.6 Anvil

Anvil[69] is a tool that supports audiovisual content annotation, but which was primarily designed for linguistic purposes, in the same vein as the previously described tool. It was developed as part of a Ph.D. thesis at the Graduate College for Cognitive Science[70] and the German Research Center for Artificial Intelligence (DFKI[71]). A screenshot showing a video annotation along with the user interface of Anvil v4.7.7 is shown in Figure 14.

Anvil [34] supports descriptive, structural and administrative annotations of video or audio objects that refer to the entire assets or to temporal segments of them. User-defined XML schema specification files provide the definition of the vocabulary used in the annotation procedure. The output is an XML file

[65] http://www.sil.org/computing/catalog/show_software.asp
[66] http://childes.psy.cmu.edu/
[67] http://trans.sourceforge.net/en/history.php
[68] http://www.fon.hum.uva.nl/praat/
[69] http://www.anvil-software.de/
[70] http://www.ps.uni-sb.de/gk/kog/cognition.html
[71] http://www.dfki.de/web

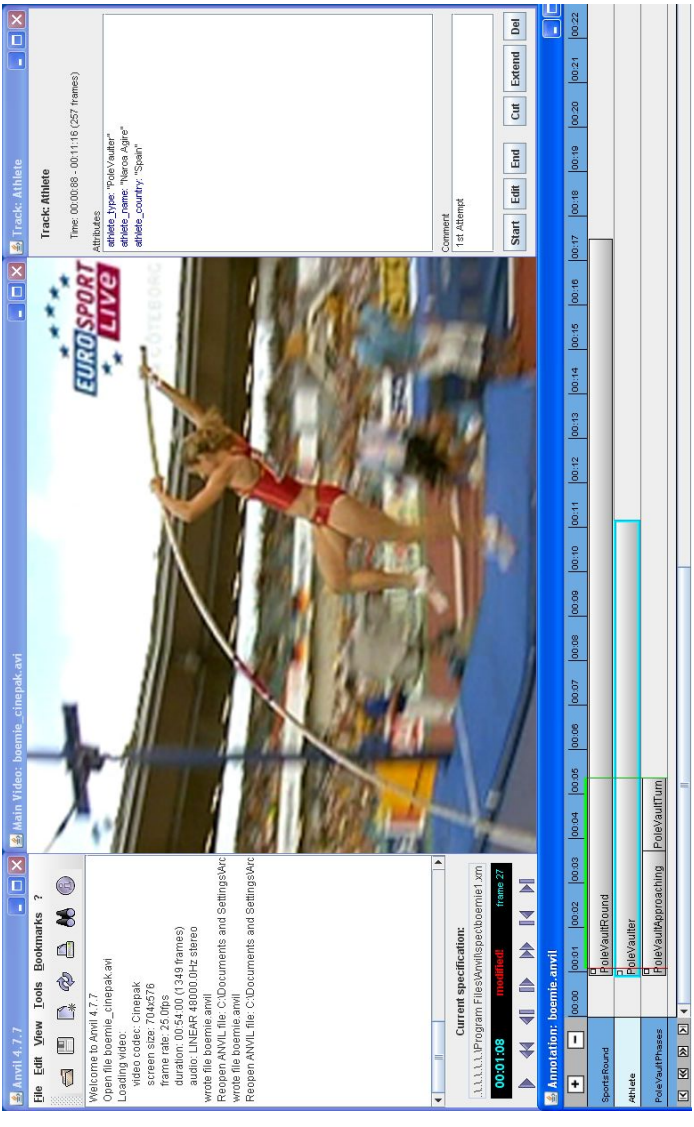


Fig. 14. Example video annotation using Anvil

containing administrative information in its head segment, while its body includes the descriptive metadata along with structural information regarding the temporal localisation of the possible video segments. Recently, Anvil has been extended to support spatiotemporal annotation as well by allowing annotations to be attached to specific points [35]; interpolation functionalities and arbitrary shapes constitute future extensions.

The tool uses hierarchical user-defined layers, in exactly the same way as described in the previous tool. Its interface consists of the media player window, the annotation board and the metadata window. The player loads files in AVI and MOV format and supports standard video controls, including frame-by-frame stepping. The annotation board contains except for the standard timeline, a waveform timeline, a pitch/intensity timeline and timelines for each described concept. The latter timelines follow the hierarchy of the concept definition in the XML file and may be collapsed or not for better viewing. As in most described tools, also in Anvil, the user has to manually define the temporal segments that wants to annotate.

Anvil can import data from the phonetic tools PRAAT[72] and XWaves which perform speech transcriptions. Moreover, it can export data to SPSS[73] and Statistica[74] for statistical analysis of the annotated data. As in more tools described in this Section, Anvil offers functionalities that allow search in the annotations, facilitating, thus, the retrieval task. It also allows the creation of bookmarks that correspond to the favorite annotations of each user. Anvil is written in Java, runs on Windows, Macintosh and Unix (Solaris/Linux) platforms and it is publicly available upon request.

4.7 Semantic Video Annotation Suite

The Semantic Video Annotation Suite[75] (SVAS), developed by Joanneum research Institute of Information Systems & Information Management[76], targets the creation of MPEG-7 video annotations. Figure 15 illustrates a screenshot of the 1.5 release.

SVAS [36] encompasses two tools: the Media Analyzer, which extracts automatically structural information regarding shots and key-frames, and the Semantic Video Annotation Tool (SVAT), which allows to edit the structural metadata obtained through the Media Analyzer and to add administrative and descriptive metadata, in accordance with MPEG-7. The administrative metadata include information about the creator, the production date, the video title, shooting and camera details, and so forth.

The descriptive annotations correspond to the MPEG-7 semantic description tools deriving from the SemanticBase DS allowing to capture subject matter

[72] http://www.fon.hum.uva.nl/praat/
[73] http://www.spss.com/statistics/
[74] http://www.statsoft.com/products/products.htm
[75] http://www.joanneum.at/en/fb2/iis/products-solutions-services/semantic-video-annotation.html
[76] http://www.joanneum.at/en/jr.html

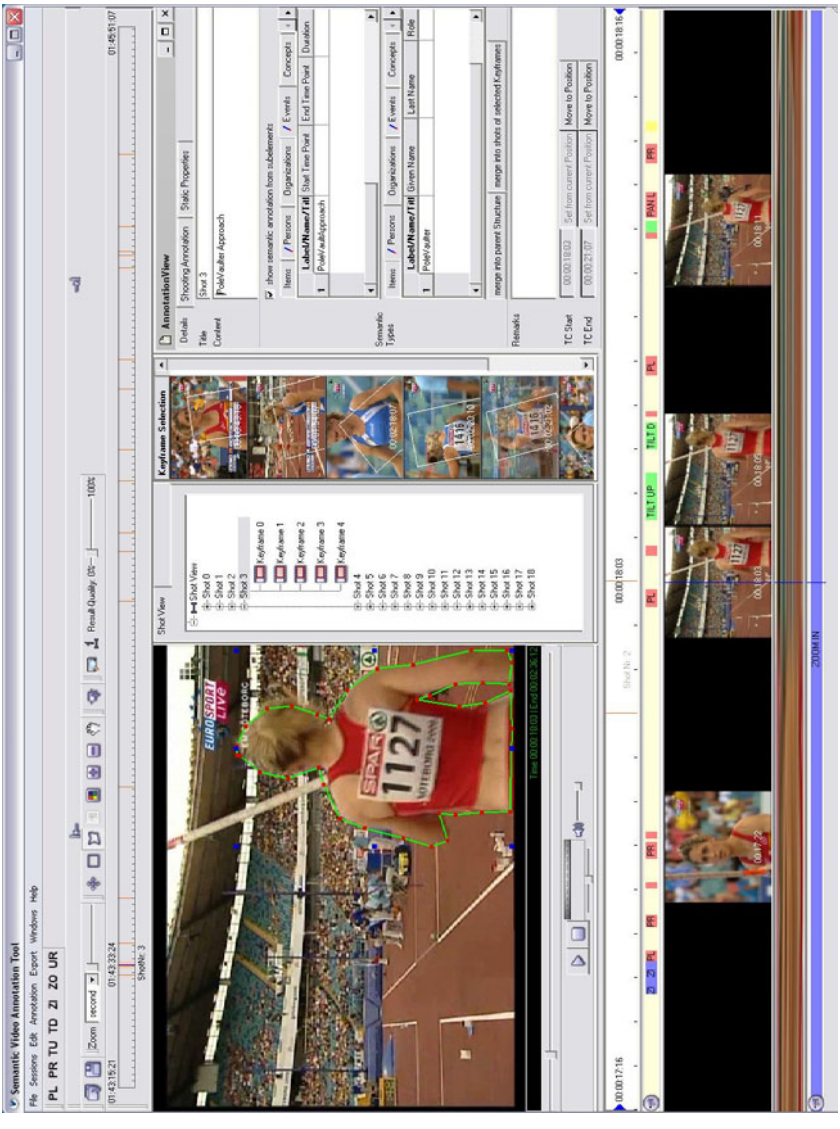


Fig. 15. Example video annotation using SVAT

descriptions regarding persons, places, events, objects, and so forth, and may refer either to shot (video segment) or region level. Regarding the latter, the localisation of specific regions in a key frame (or any other frame) can be performed either manually using the provided bounding box and polygon drawing facilities, or by deploying automatic image segmentation. Once the location of an object of interest is determined, SVAT provides an automatic matching service in order to detect similar objects throughout the entire video. The detection results are displayed in a separate key-frame view, where for each of the computed key frames the detected object is highlighted. The user can partially enhance the results of this matching service by removing irrelevant key-frames; however more elaborate enhancement such as editing of the detected region's boundaries or of its location is not supported. The annotations entered for a specific region can be copied by one mouse click to all matching objects within the video, thus reducing massively the manual annotation time required. All views, including the shot view tree structure, can be exported to a CSV file and the metadata is saved in an MPEG-7 XML file. SVAS is publicly available.

4.8 Application-Specific Video Annotation Tools

Apart from the afore described semantic video annotation tools, a number of additional annotation systems have been proposed that aspiring to specific application contexts induce different perspectives on the annotation process. To keep the survey comprehensive, in the following we examine briefly some representative examples.

Vannotea[77] is a tool for collaborative indexing, browsing, annotation and discussion of video content [37], developed by the University of Queensland. Contrary to the afore described annotation tools, Vannotea's primary focus consists in providing support for collaborative, real-time, synchronous video conferencing services. Interoperability concerns, in conjunction with the requirements for simple and flexible annotations, led to the adoption of an XML-based description schemes. Building on a simplified translation of the respective MPEG-7 and Dublin Core descriptions, Vannotea metadata can be easily transformed into the corresponding standardised representations through the use of XSLT. It is worth noticing that Vannotea builds on the Annotea initiative, a W3C activity aiming to advance the sharing of metadata on the Web. Advocating W3C standards, Annotea adopts RDF based annotation schemes and XPointer[78] for locating the annotations within the annotated resource.

ProjectPad[79] is a web-based system for collaborative media annotation and management tailored to distributed teaching and learning applications. Similarly to Vannotea, ProjectPad focused on providing synchronous interaction in terms of creation and editing of digital media collections and learning object metadata, for the purpose of supporting thematic content organisation, search

[77] http://www.itee.uq.edu.au/ eresearch/projects/vannotea/index.html
[78] http://www.w3.org/XML/Linking
[79] http://dewey.at.northwestern.edu/ppad2/

Table 2. Video annotation tools summarisation. In the *A*nnotation Vocabulary field, "U" denotes user-entered vocabularies, while "T" refers to vocabularies embedded within the tool, and thus hidden to the user.

Tool	**Input & Output**		**Annotation level**			
	Metadata Format	*Annotation Vocabulary*	*Metadata Type*	*Granularity*	*Localisation*	*Expressivity*
VIA	XML	U: domain ontology (OWL), free text T: customised structural XML schema	descriptive, structural administrative	video, video segment, frame, moving region, image, still region	time interval, free hand, polygon, rectangle	concepts
Ontolog	RDF	U: domain ontology (RDFS) T: Dublin Core ES T: Ontolog Schema ontology	descriptive, structural administrative, structural	video, video segment	time interval,	concepts, relations
VideoAnnex	MPEG-7/XML	U: XML, free text T: MPEG-7	descriptive, structural, administrative	video, video segment, frame, still region	time interval, rectangle	concepts relations
Advene	custom XML	U: free text (specific format)	descriptive, structural administrative	video, video segment	time interval,	concepts relations
Elan	custom XML	U: free text, keywords	descriptive	video video segment	time interval	concepts
Anvil	custom XML	U: XML Schema T: customised structural XML schema	descriptive, structural administrative	video, points video segment	time interval	concepts
SVAT	MPEG-7/XML	U: free text, keywords T: MPEG-7	descriptive, structural administrative	video, video segment frame, still region	time interval	concepts

and retrieval services. Annotations can be attached to the entire video (audio) asset or to specific temporal segments (spatial segments correspondingly in the case of images). Content is identified via Uniform Resource Identifiers[80] (URIs), while for the representation and storage of metadata both XML and RDF are supported.

The Video Performance Evaluation Resource Kits Ground Truth[81] (ViPER-GT) tool has been developed by the Language And Media Processing (LAMP) lab, at the University of Maryland, with the aim to assist in the evaluation of approaches addressing automatic semantic video analysis. ViPER-GT enables the creation and editing of frame-by-frame annotations at scene and object level, providing a number of predefined shape drawing facilities for the localisation of objects. To speed up the process of annotation, the automatic propagation of descriptions is supported. Specifically, by choosing to copy a description from one frame to another, the description is assigned to all frames in between as well. In case of object level descriptions, subsequent editing allows to adjust the exact position at each frame. Object level descriptions can be also propagated through dragging while the video is playing. ViPER-GT uses a simple proprietary XML-based format, which for the case of descriptive annotations can be edited by the user so as to include additional attributes.

For a more detailed list and pointers to additional tools, the reader is referred to the Tools&Resources[82] report of the W3C Multimedia Semantics Incubator Group.

4.9 Discussion

As illustrated in the aforementioned descriptions, video annotation tools make a rather poor utilisation of Semantic Web technologies and formal meaning, XML being the most common choice for the capturing and representation of the produced annotations. The use of MPEG-7 based descriptions, may constitute a solution towards standardised video descriptions, yet raises serious issues with respect to the automatic processing of annotations, especially the descriptive ones, at a semantic level. The localisation of temporal segments is performed mostly manually, indicating the issues involved in automatically identifying the time interval corresponding to the semantic notion addressed by the annotation; only Advene, SVAT and VideoAnnex perform automatic shot detection. Furthermore, VideoAnnex, VIA and SVAT are the only ones that offer selection and annotation of spatial regions on frames of the video, as well. Anvil has recently presented a new annotation mechanisms called spatiotemporal coding aiming to support point and region annotation, yet currently only points are supported.

A challenging issue in video annotation concerns the representation of structural and by consequence temporal information in an effective manner so as to avoid overwhelming volumes of metadata. This issue has been already pointed

[80] http://www.isi.edu/in-notes/rfc2396.txt
[81] http://viper-toolkit.sourceforge.net/
[82] http://www.w3.org/2005/Incubator/mmsem/wiki/Tools_and_Resources

out in relevant studies on multimedia ontologies and the resulting metadata complexity, while it should be noted that many of the MPEG-7 based video annotation tools follow simplified translations in order to avoid the cumbersome and complex MPEG-7 specifications. Finally, it is interesting to note, that although descriptors representation, if not extraction, constitutes a consideration for image annotation tools, this is not the case for video tools.

Table 2 summarises the comparative study of the examined video annotation tools with respect to the *Input & Output* and *Annotation Level* criteria described in Section 2. Regarding the miscellaneous criteria, as illustrated in the individual tools descriptions, no tool provides support for collaborative annotation and all tools are stand-alone applications, publicly available for non-commercial use[83].

It worths noticing that most annotation tools offer a variety of additional functionalities, in order to satisfy varying user needs. Facilitating the retrieval task seems to be a common demand, since almost all the tools have embedded mechanisms for allowing the user to efficiently search and/or navigate through the annotations. Moreover, the visualisation of annotations is enhanced by the annotated concepts' timeline views that most of the tools support. Concluding, we should add that the choice of a tool depends primarily on the intended context of usage, which provides the specifications regarding the annotation dimensions supported, and subsequently on the desired formality of annotations.

5 Conclusions

In the previous Sections, we reviewed representative examples of well known image and video annotation tools with respect to a number of criteria, such defined as to provide a common framework of reference for assessing the suitability and interoperability of annotations under different context of usages.

The afore presented overview suggests that semantic image annotation tools appear to follow up with relevant research advances. Domain specific ontologies are supported by the majority of tools for the representation of subject matter descriptions. Moreover, influenced by initiatives addressing multimedia ontologies, many tools utilise corresponding ontologies for the representation of structural, localisation and low-level descriptors information. With the exception of KAT though, the defined ontologies constitute simplified versions of corresponding state of the art initiatives. Consequently, given the detail of modelling provided by the state of the art ontologies, a reasonable expectation would be to investigate the use of those ontologies in manual annotation tools, especially with respect to practical scalability and complexity concerns [38,39].

Semantic video annotation tools on the contrary, present a rather gloomy scenery with respect to interoperability concerns both at semantic and syntactic level. Almost none of the examined tools supports the use of ontologies for

[83] In many cases, the source code is available for research purposes

descriptive annotations. The case is similar for structural and localisation information, where proprietary schemas are used in proprietary formats. VideoAnnEx and SVAT following the MPEG-7 specifications alleviate to an extend interoperability issues by promoting specific annotation vocabularies and schemes. Yet, apart from the XML-based issues regarding the lack of declarative semantics, the free text formats of MPEG-7 semantic descriptions perpetuate the limitations related to keyword-based search and retrieval. Consequently, a general subject of consideration relates to the low outreach and uptake of results in multimedia annotation research to practical video annotation systems [40].

However, the level of correspondence between research outcomes and implemented annotation tools is not the sole subject for further investigation. Research in multimedia annotation, and by consequence into multimedia ontologies, is not restricted to the representation of the different annotation dimensions involved. A critical issue is the delineation of multimedia specific annotation schemes, i.e. the conceptualisation and modelling of how the various annotations pertaining to multimedia assets can be interlinked in a scalable, yet effective manner. Apart from research activities conducted individually [9,41,42,30,13,12] collective initiatives have been pursued. The W3C Multimedia Semantics Incubator Group[84] (MMSEM), constitutes a prominent such activity that has produced a number of comprehensive reports including "Image annotation on the Semantic Web"[85], Multimedia Vocabularies[86] and Tools&Resources[87], as well as a proposal towards a "Multimedia Annotation Interoperability Framework"[88]. As a continuation of the efforts initiated within MMSEM, further manifesting the strong emphasis placed upon achieving cross community multimedia data integration, two new W3C Working Groups have been charted, the Media Annotation[89] and Media Fragments[90] WGs. The objective of the Media Annotation WG is to provide an ontology infrastructure to facilitate cross-community data integration of information related to multimedia objects in the Web, while the Media Fragments one addresses the identification of temporal and spatial media fragments in the Web using URIs.

Concluding, semantic image and video annotation constitute particularly active research fields, faced with intricate challenges. Such challenges issue not only from implications related to the sheer volume of content available, but also from the dynamically evolving context of intelligent content management services as delineated by the growth of Semantic Web technologies, as well as by new powerful and exciting concepts introduced by initiatives such as Web 2.0, Linked-Data[91] and Web Services.

[84] http://www.w3.org/2005/Incubator/mmsem/
[85] http://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/
[86] http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies/
[87] http://www.w3.org/2005/Incubator/mmsem/wiki/Tools_and_Resources
[88] http://www.w3.org/2005/Incubator/mmsem/XGR-interoperability/
[89] http://www.w3.org/2008/01/media-annotations-wg.html
[90] http://www.w3.org/2008/WebVideo/Fragments/
[91] http://linkeddata.org/

Acknowledgement

This work was partially supported by the European Commission under contract FP6-027538 BOEMIE. We would also like to express our gratitude to the various tool authors who provided us with useful feedback.

References

1. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22, 1349–1380 (2000)
2. Hauptmann, A., Yan, R., Lin, W.: How many high-level concepts will fill the semantic gap in news video retrieval? In: 6th ACM International Conference on Image and Video Retrieval (CIVR), Amsterdam, The Netherlands, pp. 627–634 (2007)
3. Snoek, C., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., Worring, M.: Adding semantics to detectors for video retrieval. IEEE Transactions on Multimedia 9, 975–986 (2007)
4. Hanjalic, A., Lienhart, R., Ma, W., Smith, J.: The holy grail of multimedia information retrieval: So close or yet so far away. IEEE Proceedings, Special Issue on Multimedia Information Retrieval 96, 541–547 (2008)
5. Nack, J.: Mpeg-7: Overview of description tools. IEEE MultiMedia 9, 83–93 (2002)
6. Salembier, P., Manjunath, B., Sikora, T.: Introduction to MPEG 7: Multimedia Content Description Language (2002)
7. van Ossenbruggen, J., Nack, F., Hardman, L.: That obscure object of desire: Multimedia metadata on the web, part 1. IEEE MultiMedia 11, 38–48 (2004)
8. Nack, F., van Ossenbruggen, J., Hardman, L.: That obscure object of desire: Multimedia metadata on the web, part 2. IEEE MultiMedia 12, 54–63 (2005)
9. Hunter, J.: Adding Multimedia to the Semantic Web: Building an MPEG-7 Ontology. In: Proc. The First Semantic Web Working Symposium (SWWS), California, USA (July 2001)
10. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Integration of OWL ontologies in MPEG-7 and TV-anytime compliant semantic indexing. In: Persson, A., Stirna, J. (eds.) CAiSE 2004. LNCS, vol. 3084, pp. 398–413. Springer, Heidelberg (2004)
11. Garcia, R., Semantic Integration, O.C.: Retrieval of Multimedia Metadata. In: Proc. International Semantic Web Conference (ISWC), Galway, Ireland (2005)
12. Dasiopoulou, S., Tzouvaras, V., Kompatsiaris, I., Strintzis, M.: Capturing mpeg-7 semantics. In: Proc. International Conference on Metadata and Semantics (MTSR), Corfu, Greece (2007)
13. Arndt, R., Troncy, R., Staab, S., Hardman, L., Vacura, M.: COMM: Designing a well-founded multimedia ontology for the web. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 30–43. Springer, Heidelberg (2007)
14. Jorgensen, C., Jaimes, A., Benitez, A., Chang, S.: A conceptual framework and empirical reserach for classifying visual descriptors. J. of the American Society for Information Science and Technology (JASIST) 52, 938–947 (2001)
15. Hollink, L., Schreiber, G., Wielinga, B., Worring, M.: Classification of user image descriptions. Int. J. Hum.-Comput. Stud. 61, 601–626 (2006)

16. Saathoff, C., Schenk, S., Scherp, A.: Kat: the k-space annotation tool. Poster Session, Int. Conf. on Semantic and Digital Media Technologies (SAMT), Koblenz, Germany (2008)
17. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002)
18. Gangemi, A.: Ontology design patterns for semantic web content. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 262–276. Springer, Heidelberg (2005)
19. MPEG-7 MDS: ISO/IEC 15938-5:2003 information technology. Multimedia Content Description Interface - Part 5: Multimedia Description Schemes, 1st Edition (2001)
20. MPEG-7 Visual: ISO/IEC 15938-3:2001 information technology. Multimedia Content Description Interface - Part 3: Visual, 1st Edition (2001)
21. Halaschek-Wiener, C., Golbeck, J., Schain, A., Grove, M., Parsia, B., Hendler, J.: Annotation and provenance tracking in semantic web photo libraries. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 82–89. Springer, Heidelberg (2006)
22. Chakravarthy, A., Ciravegna, F., Lanfranchi, V.: Aktivemedia: Cross-media document annotation and enrichment. In: Poster Proceedings of 5th International Semantic Web Conference (ISWC), Athens, GA, USA (2006)
23. Petridis, K., Anastasopoulos, D., Saathoff, C., Timmermann, N., Kompatsiaris, Y., Staab, S.: M-ontoMat-annotizer: Image annotation linking ontologies and multimedia low-level features. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4253, pp. 633–640. Springer, Heidelberg (2006)
24. Simou, N., Tzouvaras, V., Avrithis, Y., Stamou, G., Kollias, S.: A visual descriptor ontology for multimedia reasoning. In: Proc. of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux, Switzerland (2005)
25. Lux, M., Becker, J., Krottmaier, H.: Caliph & emir: Semantic annota-tion and retrieval in personal digital photo libraries. In: Eder, J., Missikoff, M. (eds.) CAiSE 2003. LNCS, vol. 2681. Springer, Heidelberg (2003)
26. MPEG-7: ISO/IEC 15938. Multimedia Content Descritpion Interface (2001)
27. Miller, M., McCathieNevile, C.: Semantic web tools to help authoring: A semantic web image annotation tool. In: SWAD-Europe Deliverable 9.3 (2001)
28. Russell, B., Torralba, A., Murphy, K., Freeman, W.: Labelme: A database and web-based tool for image annotation. International Journal of Computer Vision 77, 157–173 (2008)
29. Rubin, D., Rodriguez, C., Shah, P., Beaulieu, C.: ipad: Semantic annotation and markup of radiological images. In: Proc. of Annual American Medical Informatics Association (AMIA) Symposium, Washington, DC, pp. 626–630 (2008)
30. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support between mpeg-7/21 and owl in ds-mirf. IEEE Trans. Knowl. Data Eng. 19, 219–232 (2007)
31. Troncy, R., Celma, O., Little, S., Garcia, R., Tsinaraki, C.: Mpeg-7 based multimedia ontologies: Interoperability support or interoperability issue? In: Proc. Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies (MARESO), Genova, Italy, pp. 2–16 (2007)
32. MPEG-7 XM: MPEG-7 Visual eXperimentation Model (XM), Version 10.0, Doc. N4062. ISO/IEC/JTC1/SC29/WG11 (2001)
33. Rutledge, L.: Smil 2.0: Xml for web multimedia. Internet Computing 5, 78–84 (2001)

34. Kipp, M.: Anvil - a generic annotation tool for multimodal dialogue. In: Proc. 7th European Conf. on Speech Communication and Technology (Eurospeech), Aalborg, Denmark (2001)
35. Kipp, M.: Spatiotemporal coding in anvil. In: Proc. 6th International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco (2008)
36. Schallauer, P., Ober, S., Neuschmied, H.: Efficient semantic video annotation by object and shot re-detection. Posters and Demos Session, 2nd International Conference on Semantic and Digital Media Technologies (SAMT), Koblenz, Germany (2008)
37. Schroeter, R., Hunter, J., Kosovic, D.: Vannotea - a collaborative video indexing, annotation and discussion system for broadband networks. In: Proc. of Workshop on Knowledge Markup and Semantic Annotation (K-CAP), Florida, US (2003)
38. Hausenblas, M., Bailer, W., Bürger, T., Troncy, R.: Deploying multimedia metadata on the semantic web. Posters and Demos Session, 2nd International Conference on Semantic and Digital Media Technologies (SAMT), Genoa, Italy (2007)
39. Vacura, M., Svátek, V., Saathoff, C., Ranz, T., Troncy, R.: Describing low-level image features using the comm ontology. In: Proc. 15th International Conference on Image Processing (ICIP), San Diego, California, USA, pp. 49–52 (2008)
40. Bürger, T., Hausenblas, M.: Why real-world multimedia assets fail to enter the semantic web. In: Proc. of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM), Whistler, British Columbia, Canada (2007)
41. Lagoze, C., Hunter, J.: The abc ontology and model. Journal of Digital Information 2 (2001)
42. Troncy, R., Bailer, W., Hausenblas, M., Hofmair, P., Schlatte, R.: Enabling multimedia metadata interoperability by defining formal semantics of MPEG-7 profiles. In: Avrithis, Y., Kompatsiaris, Y., Staab, S., O'Connor, N.E. (eds.) SAMT 2006. LNCS, vol. 4306, pp. 41–55. Springer, Heidelberg (2006)

Subject Index

Author Index