D2.1 Methodology for Semantics Extraction from Multimedia Content

version 2.0-final

S. Petridis[1], N. Tsapatsoulis[1], D. Kosmopoulos[1], Y. Pratikakis[1], V. Gatos[1], S. Perantonis[1], G. Petasis[1], P. Fragou[1], V. Karkaletsis[1], K. Biatov[2], C. Seibert[2], S. Espinosa[3], S. Melzer[3], A. Kaya[3], R. Möller[3]

Distribution: Restricted

BOEMIE

Bootstrapping Ontology Evolution with Multimedia Information Extraction

[1]National Centre for Scientific Research "Demokritos" (NCSR)

[2]Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. (FHG/IMK)

[3]Hamburg University of Technology (TUHH)

[4]University of Milano (UniMi)

[5]Centre for Research and Technology Hellas (CERTH)
Tele Atlas (TA)

FP6-027538 **D2.1**

December 21, 2006

Project ref.no.	*FP6-027538*
Project acronym	*BOEMIE*
Project full title	*Bootstrapping Ontology Evolution with Multimedia Information Extraction*

Security (distribution level)	*Restricted*
Contractual date of delivery	*M08*
Actual date of delivery	*December 21, 2006*
Deliverable number	*D2.1*
Deliverable name	*Methodology for Semantics Extraction from Multimedia Content*
Document type	*Report*
Status & version	*version 2.0-final*
Number of pages	*107*
WP contributing to the document	*WP2*
WP / Task responsible	[1] *NCSR*
Other contributors	[2] *FHG/IMK,* [3] *TUHH*
Author(s)	*S. Petridis*[1], *N. Tsapatsoulis*[1], *D. Kosmopoulos*[1], *Y. Pratikakis*[1], *V. Gatos*[1], *S. Perantonis*[1], *G. Petasis*[1], *P. Fragou*[1], *V. Karkaletsis*[1], *K. Biatov*[2], *C. Seibert*[2], *S. Espinosa*[3], *S. Melzer*[3], *A. Kaya*[3], *R. Möller*[3]
Quality Assurance	*A. Ferrara*[4], *V.Tzouvaras*[5]
EC Project Officer	*Johan Hagman*
Keywords	*pattern recognition, semantics extraction, ontology, machine learning, image analysis, video analysis, video OCR, multimedia interpretation*
Abstract (for dissemination)	*This document outlines the "Methodology for Semantics Extraction from Multimedia Content" that will be followed in the framework of the BOEMIE project. The basic aim is to describe the architecture and methods that will be used in each particular modality (text, image, video-audio) for extracting semantics from multimedia content using an ontology-driven framework. Furthermore it aims at answering the key research questions that need to be addressed in order for ontology-driven semantics extraction from multimedia content to be possible. These key issues are: (a) how ontology guides the extraction process, and (b) how the multimedia information extraction provides the means for ontology population and enrichment.*

Contents

Executive Summary **7**

1 Introduction **13**

2 Architecture for Semantics Extraction from Multimedia Content **14**
2.1 Design principles . . . 14
2.2 The multimedia semantic model of BOEMIE . . . 14
2.3 Description of the methodology . . . 16
2.4 Support for ontology evolution . . . 22
2.5 Contribution to the state of the art . . . 24

3 Still Images **27**
3.1 Aim of image-based information extraction . . . 27
3.2 Overview of the methodology . . . 27
3.3 Methodology . . . 30
3.4 Semantic model usage . . . 34
3.5 Support to ontology enrichment . . . 34
3.6 Confidence handling . . . 35
3.7 Evaluation framework . . . 37
3.8 Use case: pole vault . . . 40

4 Video **44**
4.1 Aim of video-based information extraction . . . 44
4.2 Overview of the methodology . . . 44
4.3 Description of methodology . . . 45
4.4 Semantic model usage . . . 47
4.5 Support to ontology enrichment . . . 47
4.6 Confidence handling . . . 47
4.7 Evaluation framework . . . 47
4.8 Use case: pole vault . . . 48

5 Video OCR **51**
5.1 Introduction . . . 51
5.2 Overview of VOCR methodology . . . 51
5.3 Text detection methodology . . . 53
5.4 Text tracking Methodology . . . 57

6 Audio/Speech **62**
6.1 Aim of Audio/Speech-based information extraction . . . 62
6.2 Overview of the Methodology . . . 62
6.3 Description of the Methodology . . . 63
6.4 Semantic Model Usage . . . 66
6.5 Support to Ontology Enrichment . . . 66
6.6 Confidence Handling . . . 66
6.7 Evaluation Framework . . . 66
6.8 Use Case: Pole Vault . . . 67

7 Text **69**
7.1 Aim of Text-based information extraction . . . 69
7.2 Overview of the methodology . . . 69
7.3 Description of Methodology . . . 69
7.4 Semantic Model Usage . . . 74
7.5 Support to Ontology Evolution . . . 75
7.6 Confidence handling . . . 75

7.7 Evaluation Framework . . . 75
7.8 Use Case: High Jump . . . 76

8 Reasoning **79**
8.1 Concept Rewriting using Concrete Domains . . . 79
8.2 High-level interpretation . . . 81
8.3 Concluding Remark . . . 82

9 Multi-modal Data Fusion **83**
9.1 Aims of multi-modal data fusion . . . 83
9.2 Interdependence with other workpackages . . . 83
9.3 Measurable objectives . . . 84

10 Risk analysis **85**
10.1 Image . . . 85
10.2 Video . . . 85
10.3 VOCR . . . 87
10.4 Audio . . . 87
10.5 Text . . . 88
10.6 Reasoning . . . 88

11 Epilogue **89**

A Description Logics: The $\mathcal{SH}$ Family **101**
A.1 The Foundation: $\mathcal{ALC}$ . . . 101
A.2 Concrete Domains . . . 102
A.3 Transitive Roles . . . 103
A.4 Role Hierarchies and Functional Restrictions . . . 103
A.5 Number Restrictions and Inverse Roles . . . 104
A.6 Number Restrictions, ABoxes and Concrete Domains . . . 105
A.7 Nominals . . . 105
A.8 The research Frontier of SH Family . . . 106

List of Figures

1 The BOEMIE multimedia semantic model . . . 15
2 Modality–specific vs. modality–independent concepts . . . 16
3 Analysis: processing and interpreting a multimedia document . . . 17
4 A typical `xml` output of a single-media semantics extraction process . . . 19
5 Training – improving multimedia analysis . . . 20
6 Discovery – expanding multimedia analysis . . . 21
7 The bootstrapping process of BOEMIE . . . 23
8 A goal analysis example of the BOEMIE project . . . 24
9 Schematic Diagram of the proposed methodology for semantics extraction from still images 27
10 Meaning of shapes shown in Figure 3.2 . . . 28
11 An example of an xml document corresponding to the annotation of a still image . . . 29
12 Primitive (line) detection in a running event. . . . 32
13 Colour histogram of areas between parallel lines . . . 32
14 A set of images that may trigger a new concept creation (floodlight). . . . 35
15 Example of extracting the face MLC through segmentation-based or holistic approach . . . 36
16 Annotated image to be used for training . . . 41
17 Annotation example for MLC `athlete_face` . . . 41
18 `Pole` and `bar` detection using the Hough transform. . . . 43
19 Overview of video-based information extraction . . . 44
20 Annotated video . . . 49
21 Video annotation example in `xml` . . . 50
22 An example of artificial text in a video frame . . . 51
23 Flowchart of the proposed algorithm. . . . 52
24 Edge map of the frame in figure 22 . . . 53
25 The edge map after the dilation process. . . . 54
26 The edge map after the opening operation. . . . 54
27 Initial bounding boxes. . . . 55
28 Example of box splitting through horizontal edge projection thresholding . . . 55
29 Example of box splitting through vertical edge projection thresholding. . . . 56
30 Refined result . . . 56
31 A screenshot from the detection-evaluation application. . . . 58
32 Text area detection example . . . 58
33 (a) The result of text area detection, (b) the result of text line detection. . . . 59
34 (a) edge map of a text area, (b) horizontal projection of the area. . . . 59
35 (a) edge map of a text line, (b) vertical projection of the text line. . . . 59
36 Enhancing image quality after text block tracking. . . . 61
37 Methodology for audio segmentation, classification and recognition . . . 62
38 Methodology for audio event detection . . . 65
39 Information extraction from the textual part of a multimedia document. . . . 70

List of Tables

1 Measurable objectives for technologies dealing with semantics extraction from still images. 38
2 Ground truth content and evaluation parameters 39
3 Mid Level Concepts (MLCs) for the text modality 72
4 Types of relations for the text modality 73
5 Measurable objectives for technologies dealing with text-based information extraction . . 76
8 An abstract Tbox for the athletics domain 80

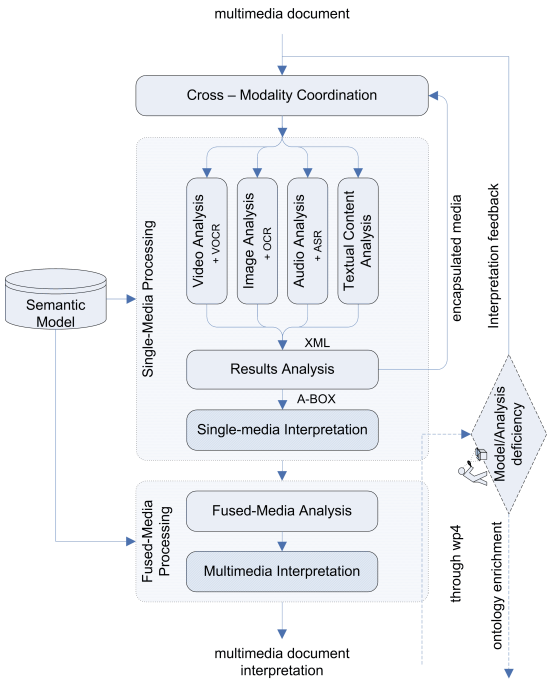
Executive Summary

Complex structured multimedia documents possess a rich variety of information appearing in different forms and combined under diverse schemata. Their analysis is a demanding operation calling for specific per-medium processing techniques to be developed, assembled and fused. This will enable multimedia document interpretation and adaptation in the context of an evolving domain application.

In BOEMIE, the objective of a Methodology for Semantics Extraction from Multimedia Content –as forseen in the DoW, p.44 – is to specify how information from the multimedia semantic model can be used to achieve semantic extraction from various modalities (text, image, video and audio) and to come up with an open architecture, which will communicate with the ontology evolution modules in WP4, accessing existing knowledge and providing back newly extracted information. This document describes the architectural and methodological choices that we believe will lead us to the fulfilment of the above objective.

Architecture

The design choices of the semantics extraction methodology have been guided by the core ontology-oriented architecture of the BOEMIE project. Ontology is a useful milieu for systematically fusing and interpreting multimedia analysis results. Nevertheless, an important issue is how to enable its interfacing with ontology-unaware media processing and machine learning techniques in an effective and transparent manner. The architecture for semantics extraction from multimedia content has been designed to advance the state of the art by (a) facilitating independent development of processing and learning techniques per medium (b) allowing transparent coordination of per-medium semantics extraction modules and (c) enabling reasoning-based feedback on semantics extraction results. Moreover, the architecture supports the evolution of the system, by requiring processing algorithms for each medium to be adaptable to the evolving domain utilising both supervised and unsupervised machine learning techniques.



Design principles The semantics extraction architecture allows us to deal with multimodal information and to bridge the gap between extraction techniques and ontology-based reasoning services.

- *Multimodal information fusion*: To deal with multimedia documents, semantics extraction is decomposed into two steps. First an analysis of each medium-specific subdocument is performed. Then, extraction results are fused, in order to take into account complementarity, redundancy and coherence of the extracted information. These steps may be repeated in a loop to account for (a) analysing embedded documents (such as OCR text in images) and (b) refining the analysis of one medium-specific document using information extracted from another.

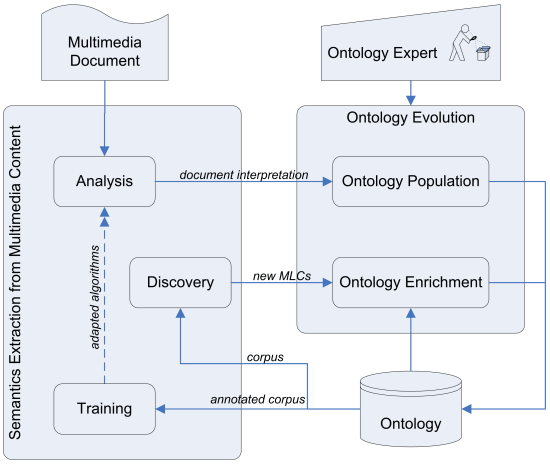
  Multimodal data fusion is explicitly supported by the ontology. Namely, for each modality, a set of modality-specific concepts is defined. Concepts across modalities are then associated with modality-independent concepts (Example: the *visual representation of an athlete*, such as its photo, and the *textual representation of an athlete*, such as its name, are concepts specific to the image and text modality respectively, related to the the modality–independent *athlete* concept).

- *Bridging the semantic gap*: Semantics extraction for a particular medium is further decomposed into two steps. First, segments inside a document are detected and classified using medium-specific

processing techniques. The development of specific methodologies to process each medium is a significant part of WP2, separately described in the following section. To allow linkage with the ontology, the set of possible classes per medium are mapped to an evolving subset of concepts of the ontology, referred to as as *mid-level concepts*. (Example: the *visual representation of a pole* is a mid level concept, under the condition that it is possible to detect directly a pole in an image, using the image analysis tools).

Once mid-level concepts instances in a document have been found, reasoning services, such as deduction and abduction, complement the document analysis by inferring the existence of instances of aggregate concepts, referred to as *high-level* concepts (Example: a *pole-vault event* is a high level concept, if its existence is deduced by the existence of a *pole* and a *athlete*, together with a suitable rule within the ontology). The findings of reasoning may then be used as a feedback, to refine the detection and classification of mid level concept instances.

Semantics extraction adaptability The semantics extraction methodology comprises three distinct modes of operation, namely *Analysis*, *Training* and *Discovery*. These three modes are related respectively to fundamental BOEMIE system functions: ontology population, adaptability of the analysis with respect to new content and ontology enrichment. Although the first mode of operation implements the main semantics extraction task, the second and third modes are essential to the applicability of the BOEMIE system to an evolving domain.



In summary:

- The *Analysis* mode of operation applies each time a new multimedia document becomes available to the BOEMIE system. Its task is to analyse and interpret the document using single-medium specific techniques followed by fusion of multimedia information.

- The *Training* mode of operation applies when new manually annotated content, or content inaccurately analysed so far, is available. Its task is to enhance the analysis modules given the available content, through the usage of supervised machine learning algorithms.

- The *Discovery* mode of operation applies when a significant amount of content is available in the BOEMIE system, that can lead to expansion of the semantic model, through augmenting the set of mid-level concepts. The new concepts correspond either to a refinement of existing mid-level concepts or to a specific cluster of instances so far classified as "unknown" that exhibit similarities among them. Clustering of instances is based both on instance similarity and discrimination with respect to a high-level concept which they are associated to.

Methodologies

Semantics extraction from a multimedia document breaks down to the semantics extraction from each modality-specific sub-document followed by the fusion of the extracted information. To this end, specific methodologies dealing with the image, video, audio and text modalities are developed. The task, for each modality is: (a) to detect and classify segments inside the given documents, corresponding to one among a given set of mid-level concepts (b) to identify clusters of similar segments in order to propose new mid-level concepts, and (c) to adapt to both an evolving set of mid-level concepts and to newly annotated content. In the sequel, the methodologies for processing still images, video, audio and text are summarised.

Still Images The aim of semantics extraction from still images is to provide information about the existence of image-specific mid-level concepts, their maps (unique region numbers that identify the image area that is covered by a particular mid-level concept), their low-level descriptors (e.g., scalable colour descriptor, etc) and complementary information about unknown image regions (i.e., MPEG-7 colour, texture and shape descriptors) which will allow for new mid-level concept identification.

We propose to combine two complementary approaches: region-based analysis (segmentation) and holistic image analysis (primitive detection). In the region-based approach each input image is partitioned into segments with the aid of the Watershed transform or similar image segmentation techniques. Then, for each segment a set of features, such as the MPEG-7 visual descriptors, are evaluated. Finally, based on these features, a trained classifier assigns a concept to each region. In the holistic approach, mid-level concepts are modeled through primitives (simple geometric objects such as lines, ellipses, etc., or composite objects whose detection is feasible through dedicated algorithms, e.g. face). It is foreseen that findings of the region-based and holistic approaches will be combined to produce the final mid-level concept instances which will populate the ontology. At the same time a confidence value indicating the belief that the detection and identification of a instance is correct will also be computed.

The use of unsupervised clustering and Resource Allocation Networks (RAN) will be examined for new concept detection. The assumption here is the following: If 'unknown' image segments with similar properties (low-level descriptor values) appear systematically in the given set of images then this is an indication of the existence of a new concept. Unsupervised clustering is a possible solution for grouping together unclassified image segments across a large set of images, so as to provide a suggestion to the ontology evolution module.

Video The aim here is to provide information about the existence of mid-level concept instances in video data, which are by nature distributed over time (Example: phases of an athletics event like the approach of a pole vaulter to the jump-off point). The methodology will follow a three step approach: pre-processing, feature extraction and classification.

During pre-processing, the incoming video is segmented into shots and frames from each shot are extracted to obtain information about objects present in the shot that can be tracked in time across the shot. To conduct feature extraction, analysis of the shot in terms of both global and local camera motion will be considered. In particular, we will examine using optical flow analysis to discover global camera motion, such as booming, dollying or tilting, and then detect local camera motion, by undoing the effect of global motion on consecutive frames, find blobs in the difference image and evolve a trajectory for these blobs along the succeeding frames.

The classification part of the methodology connects the extracted features, like object trajectories, to the mid-level concepts in the semantic model. Shot classification assigns mid-level concepts to each shot. Classification will be based on two complementary approaches: statistical analysis of motion patterns (global and local) and matching of object trajectories. The former approach tries to determine characteristic patterns of global and local motion in semantically similar shots based on statistical machine learning. The suitability of the Viola-Jones approach to object detection will be considered, combining simple filters, which extract relative motion of image regions over time, through the use of the boosting technique. The latter approach tries to match extracted object trajectories to modeled object trajectories. We consider modeling 3D constellations of objects like an athlete's head, hands and feet during a pole vault over time, matching projections of these model constellations to extracted 2D constellations.

Video OCR The procedure of retrieving text from video consists of 3 basic stages: text detection, text segmentation and recognition. We aim at providing an effective and computationally efficient algorithm for the spatial and temporal detection of artificial text in still video frames with a high recall rate. Incorporation of scene text detection algorithms will also be considered, examining whether these may work efficiently with many kinds of text without leading to deterioration of artificial text detection rates.

The proposed methodology for detecting artificial text in video frames is based on edge information: an edge map is created using Canny edge detector and morphological operations are used in order to connect the vertical edges and discard false alarms. A connected component analysis is then performed in order to determine potential bounding boxes for text areas, followed by horizontal and vertical projections, thus refining and splitting text areas in text lines. A fast tracking algorithm will also be developed

to track linear motion of artificial text, discarding the wrongly tracked text using confidence values. The tracking stage will result in quality enhancement through multi-frame integration and speed-up of the system since the slow stage of detection will be done periodically. Adaptive thresholding binarization techniques as well as resolution enhancement will be applied before the segmented and binarized images are fed to a commercial Optical Character Recognition (OCR) machine.

Audio The aim here is to analyse the audio stream, in order to extract a sequence of audio events, including speech and non-speech events, with their boundaries and their temporal relations for further semantic interpretation. The audio signal will be pre-processed and low-level features, such as Mel-frequency cepstral coefficients will be extracted. Then the audio signal will be separated into segments using the Bayesian Information Criterion (BIC). In parallel, a classifier will classify audio as speech or non-speech, based on an unsupervised calculation of the silence ratio on homogeneous sub-segments. The non-speech segments will then be clustered together, based on an extended BIC approach, in order to simplify audio event recognition.

To recognise the audio events, two approaches will be investigated: a first approach performs background sound normalisation, by adapting a background model using the Expectation-Maximisation algorithm. Foreground audio events, modeled through statistical GMM are then obtained by adapting the appropriate background model to foreground data using Maximum a-Posteriori Adaptation. A second approach to be examined in order to separate background and foreground sounds will be based on the Blind Source Separation technique, followed by a separate recognition of background and foreground sounds using a MAP classifier. To amend for often high-noise audio, we will examine taking advantage of potential dependence of events, organising them as a dependency network and developing a respective grammar.

Spoken names in speech segments will be recognised using a syllable-based speech recogniser, based on inverse dictionary search technique. The first-n-best syllable sequences of mixed syllables and phone will be identified using n-gram models followed by statistical string matching. We will investigate using these sequences to select the best candidates from the name list. Finally, methods of emotion recognition based on prosodic characteristics (pitch contour, energy contour) will also be investigated, to detect audio event interpretation.

Text The aim of extraction from the textual part of documents is to provide information about the existence of concepts such as names of persons, names of events, dates, ages, performance, etc., the relations that may occur between them (e.g. that a person with a name N1 has a performance P1), as well as the occurrence of terms for the various sporting events. The processing of textual content consist of five steps: Preprocessing, Named-Entity Recognition, Co-Reference Resolution, Normalisation and Relation Extraction.

Preprocessing involves tokenization, demarcation, sentence splitting, part-of-speech tagging, stemming, gazetteer lookup and shallow parsing. Algorithms implementing most of the above techniques are offered by respective modules of the Ellogon language engineering platform, which are, by design, domain independent and/or adaptable.

A Named-Entity Recogniser (NER) will be trained for each text-specific mid-level concept. We will examine both token-based NER, such as the Brill tagger, the Trigrams 'n' Tags (TNT) tagger and document-level NER, such as T-Rex. We will also investigate the combination of multiple systems through voting or stacking. Moreover, a rule-based co-reference resolution system will be developed in order to resolve anaphoric references to recognised entities corresponding to mid-level concepts within the same sentence or adjacent sentences or within the same paragraph, as well as to perform synonym detection. The system will handle standard co-reference problems such as name matching, pronoun-antecedent co-reference, and definite description co-reference. Finally, we will develop an adaptable normalisation system to transformation of certain types of concept instances to a predefined form. Normalisation may involve time expressions (e.g. dates), numeric expressions (e.g. performance), names (e.g. person names, event names).

Reasoning In the context of semantics extraction, reasoning services will aim at supporting low-level multimedia analysis and relate multimedia content to high-level knowledge in the domain ontology using

the results of low-level multimedia analysis. To that end, we will develop two non-standard reasoning services: concept rewriting using concrete domains and high-level multimedia interpretation respectively. To deal with the uncertainty of the perceptions and the vagueness of the concepts in the ontologies, a number of approaches to the reasoning problem under uncertainty and vagueness such as belief networks, Bayesian reasoning, fuzzy logics and rough mereology will also be investigated. The provided services will support medium-specific analysis by taking into account additional high-level knowledge about the domain such as HLCs, thus enabling a top-down view on documents which may lead to fine-tuning available extraction parameters to achieve better results.

Concept rewriting using concrete domains is an approach aiming to deal with the typically ambiguous extraction results of each single-medium methodology. This problem has been formally investigated in the literature and a unique solution can often not be found. Ambiguous results are considered those containing instances with constrained, but not exactly known, concrete domain attribute values. These results, in form of a query, are given as input to the reasoning service. Then, by taking into account the background knowledge, the reasoning service solves a logical entailment problem to provide the most specific concepts conforming to the extraction results. As a result, the ambiguity of the extraction results is reduced.

On the other hand, high-level multimedia interpretation of a multimedia object is achievable through the assertions delivered by multimedia analysis and reasoning, which can deduce further information using background knowledge and abduction. A detailed discussion of the abduction process can be found in the BOEMIE project deliverable D4.1. In a nutshell, the reasoning service is provided with the knowledge about high-level concepts, as well as with the assertions generated by low-level multimedia analysis as a consequence of the observations made. As a result, the abduction process may postulate the existence of further mid-level concepts instances or such that, at the end, observation fits to the background knowledge according to one or more possible explanations.

Multi-modal Data Fusion Multi-modal data fusion in BOEMIE aims to combine information stemming from the specific analysis of these modalities, in order to enable ontology evolution up to a degree that cannot be achieved using part of the modalities provided. To support open-architecture, fusion analysis will have no direct dependence with a modality-specific technique. Rather, the methodology is driven by the semantic model, relying on associations across modality-specific concepts.

Fusion will be investigated at several levels of data representation: numerical features, symbolic features as well as mid-level concepts and relations between them. We aim at improving the semantics extraction by allowing assembling information related to distinct properties of the event and raising the accuracy and confidence level about the extracted information. Fusion may also lead to information coherence across modalities, by guiding single-modality analysis, based on the extraction results of another modality. We will also investigate ways of balancing computational complexity by prioritising low-cost modal-specific analysis, until reaching an acceptable confidence level of extraction. Finally, to motivate concept discovery, we will examine ways to propagate the detection of a concept in one modality to its detection in others, through juxtaposition of concept instances associated to the same high-level concept.

1 Introduction

The aim of semantics extraction from multimedia content is the ontology-driven analysis of complex structured documents with multimedia content (video, image, audio and text). The results of extraction are used to evolve an ontology, which in turn is used to adapt extraction, thus enabling the bootstrapping process of the BOEMIE project.

In BOEMIE, the purpose of defining a Methodology for Semantics Extraction from Multimedia Content is described in the DoW, p.44:

> Through the proposed methodology we will *specify how information from the multimedia semantic model can be used to achieve semantic extraction from various modalities (text, image, video and audio).* The outcome of the proposed methodology will be *an open architecture, which will communicate with the ontology evolution modules in WP4, accessing existing semantic information and providing back newly extracted information* ... The architecture will also ... specify the interface for the extraction and fusion tools defined in WP2. Thus, it will be completely open to the replacement of the tools with new ones in the future. Additionally, the methodology will *cover the evaluation of the whole extraction process based on the separate evaluations of the approaches for single-media analysis.*

The BOEMIE Semantics Extraction from Multimedia Content Methodology is a complex process designed to facilitate independent development of processing and learning techniques per medium, to allow transparent coordination of per-medium analysis modules and to enable feedback of reasoning on extraction results to the analysis modules. In particular, the following activities are involved:

- *semantics extraction from still images*, concerning the detection and classification of image areas with domain-pertinent information, using both region-based and holistic approaches and based on the extraction of low-level image descriptors (e.g., scalable colour descriptor);
- *semantics extraction from video sequences*, concerning the detection and classification of spatio-temporal segments through analysis of global and local motion patterns or through model-based analysis of object trajectories;
- *video OCR*, concerning the detection, segmentation and recognition of text found in video sequences;
- *semantics extraction from audio/speech*, concerning the extraction of information about the existence of known audio events, events extracted using name recognition from speech data and non-speech audio events;
- *semantics extraction from text*, concerning the extraction from the textual part of documents information about the existence of names of persons, dates, etc., the relations that may occur between them as well as about the occurrence of terms for various domain-specific events;
- *coordination and fusion of multimedia content analysis*, concerning the combination of information stemming from the specific analysis of each modality, in order to enable semantics extraction and ontology evolution to a degree that cannot be achieved using the individual modalities ;
- *reasoning based multimedia interpretation*, concerning the extraction of high-level knowledge in the domain ontology based on multimedia content analysis (high-level multimedia interpretation) as well as supporting low-level multimedia analysis, based on concept rewriting using concrete domains.

The deliverable is organised as follows. In Section 2 the overall system architecture is introduced. In sections 3, 4, 5, 6 and 7, the modality-specific semantics extraction methodology from still images, video sequences, audio/speech and text is described. In Section 8, the reasoning services are explained, while in Section 9 the basic ideas about multimodal fusion are illustrated. Eventual risks pertaining to the proposed methodology and their handling are presented in Section 10. Finally, Section 11 concludes this document.

2 Architecture for Semantics Extraction from Multimedia Content

2.1 Design principles

Complex structured multimedia documents possess a rich variety of information appearing in all sorts of forms and combined under diverse ways. Their analysis is a demanding operation since a large number of specific per-media processing techniques need to be developed, assembled and fused in a way that enables their interpretation and adaptation to the context of a domain application.

The design choices of the overall semantics extraction methodology (WP2) have been guided by the core ontology-oriented architecture of the BOEMIE project. The domain of application is modeled through an ontology (developed in WP3) whereas the output of semantics extraction is used to trigger both ontology population and enrichment (WP4). Ontology can be a useful milieu for systematically organising, fusing and interpreting multimedia analysis results. However, it prompts for devising a particular approach to enable its interfacing with ontology-unaware media analysis and machine learning techniques. Overall, the architecture for semantics extraction from multimedia content has been designed to meet the following criteria:

1. Facilitate independent development of processing and learning techniques per medium.

   Processing each one of the studied media (video, image, audio and text) is a complex task by itself. Reaseach for methodologies to deal with each one of them will be independely conducted and will result in and reusable components. Moreover, particular ontology-unaware processing and machine learning techniques should not be involved with ontology operations, but should rather be facilitated to take their input and give their output in the usual manner.

2. Transparent coordination of per-medium analysis modules

   Coordinating the analysis of a document across all media and fusing the analysis results should not be dependant on a particular medium processing choice. Rather, a transparent coordination should be ensured by agreeing on the types of the processing results. Coordination should also provide for routing encapsulated media (such as OCR extracted text) to the appropriate analysis module.

3. Enable reasoning-based feedback on analysis results

   Analysis results should be allowed to be reasoned on according to the application domain, modeled by the ontology. Moreover, media analysis modules should be able to take into account the results of reasoning, in order to adjust and improve their behaviour in the light of particular reasoning findings.

2.2 The multimedia semantic model of BOEMIE

Figure 1 depicts the BOEMIE multimedia semantic model implemented as an ontology consisting of several ontological modules. It is out of the scope of this document to cover in depth issues regarding ontology design[1]. However, methodology of Semantics Extraction from Multimedia Content relies on some key properties of the BOEMIE ontology, and therefore we informally introduce some ontology-specific terms, which may be unfamiliar to the reader. Namely, we will refer to the "T-BOX" of the ontology as a set of interrelated concepts and relations within the ontology and to the "A-BOX" of the ontology as instances of these concepts and relations. For a reader familiar with relational database systems, there is a rough correspondence between the T-BOX and the schema of a database, on one hand, and the A-BOX and the actual data stored on the other. The significant advantage of ontologies, as represented in description logics (DL) is that concepts and relations are defined in a way to allow specific formal reasoning to be applied (see section 8 of this document as well as D4.1).

Regarding the BOEMIE ontology, we highlight here its most important aspects affecting the functionality the methodology presented in this document:

[1] The reader is referred to Deliverables 3.1 and 3.2 for exact and detailed descriptions of the BOEMIE multimedia semantic model

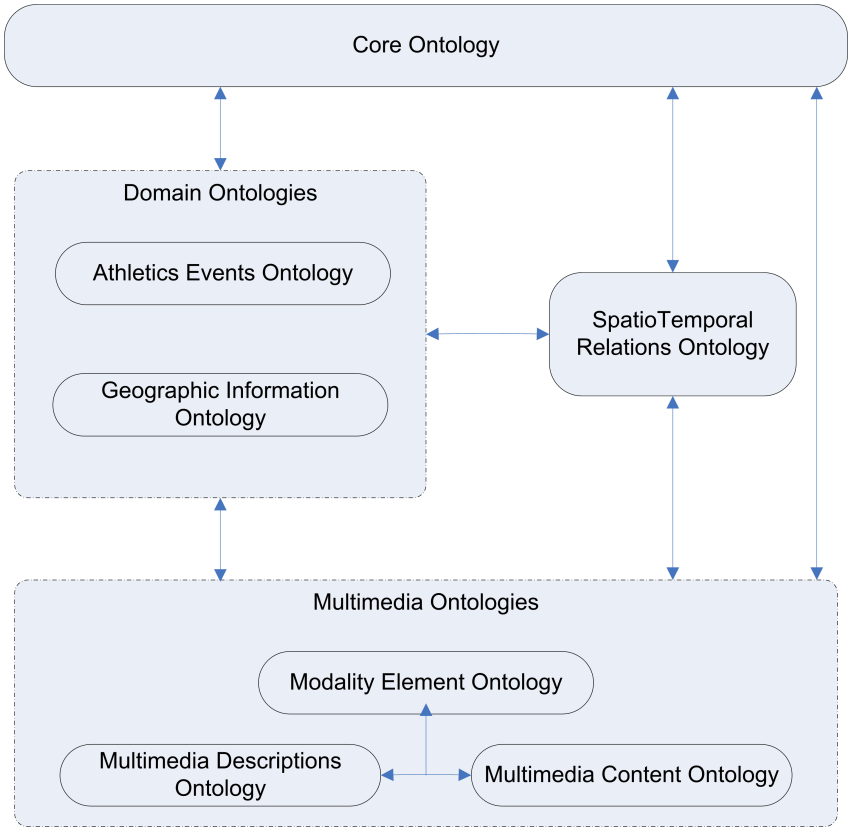


Figure 1: The BOEMIE multimedia semantic model

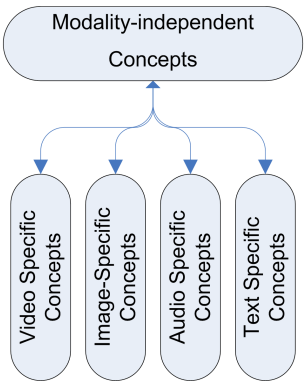


Figure 2: Modality–specific vs. modality–independent concepts

1. The existence of various modalities is explicitly taken into account. Namely, within the "modality elements" ontology, a set of modality-specific concepts are defined for each modality. These, may be associated via modality-independent concepts, defined in the "domain ontology" to enable cross-modality information fusion.

   As an example, the *image of an athlete* (`athlete_visual_element`) and *sound of an athlete* (`athlete_audio_element`) are concepts specific to the image and audio modality respectively. However, they both relate to the modality–independent *athlete* (`athlete`) concept.

2. According to a vocabulary agreed within the BOEMIE project, the modality-specific concepts are said to be mid-level, high-level or both.

   mid-level concepts (MLC) Those that can be directly instantiated by the relevant analysis module, using some modality-specific analysis technique. Notice that an MLC for one modality is not necessarily an MLC for some other.

   high-level concepts (HLC) Those that are instantiated by the reasoning services, by means of instantiated MLCs and rules within the ontology. These rules may also be modality-specific.

   As an example, assume that a hand and a face can both be directly identified by image analysis. Then, the `hand_visual_element` and the `face_visual_element` are both MLCs for the image modality. Also assume that a `man_visual_element` is associated within the ontology with a `hand_visual_element` and a `face_visual_element`. Then `man_visual_element` is an HLC for the image modality and can be instantiated through reasoning, once the relevant images of a `hand_visual_element` and a `face_visual_element` are found. Nevertheless, if `man_visual_element` can be further recognised by image analysis in a direct way, say by its silhouette, then this concept is characterised also as an MLC.

3. Mid-level concepts possess properties containing information pertinent to effectively applying the methodology proposed in section 2.3. These include the way instantiation has taken place (manual, confirmed by a human operator or automatically), specific parameters of the analysis module that allowed the instantiation as well as the confidence level of the instantiation.

2.3 Description of the methodology

The Semantics Extraction Methodology comprises three distinct modes of operation[2]. Each mode of operation implements a part of functionality required for BOEMIE to account for ontology population, adaptability of the analysis in respect to new content and direct involvement in ontology enrichment respectively. These functionalities will be available through respective application programming interfaces (APIs). In summary:

[2]These modes of operation should not be confused with patterns concerning ontology (see deliverable D4.1)

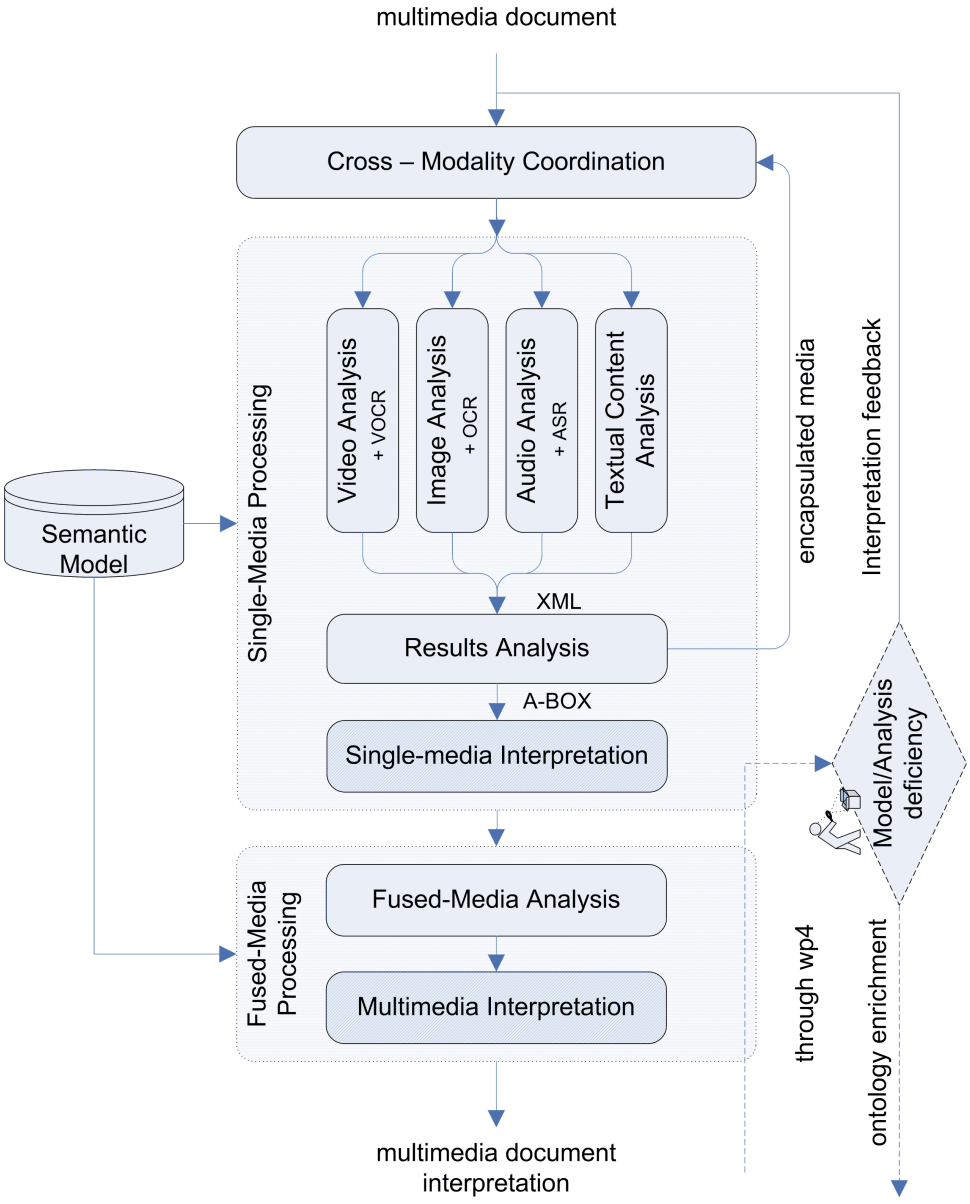


Figure 3: Analysis: processing and interpreting a multimedia document

Analysis. The first mode of operation applies each time a new multimedia document becomes available. Its task is to analyse and interpret the document using single-medium techniques followed by fusion of multimedia information. The semantic model is used by (a) defining the target classes as a subset of ontology concepts, referred to as mid-level concepts (see section 2.2) and (b) allowing reasoning on extraction results to be used as feedback in order to adapt the analysis process. The output of the semantics extraction processes is used for population and/or enrichment.

Training. The second mode of operation applies when new manually annotated content, or content inaccurately analysed so far by the single and/or fused analysis modules becomes available. Its task is to improve the analysis modules using the available annotated content and machine learning algorithms. This mode of operation does not lead to a modification of the T-BOX of the ontology.

Discovery. The third mode of operation applies when a significant amount of content is available, such that it can lead to the evolution of the semantic model, through extending the set of (mid-level) modality-specific concepts. The new concepts are discovered by unsupervised techniques, to clustering parts of the content that has so far been assigned to the same concept, while also enhancing the discrimination between aggregate concepts (referred to as high-level concepts). This mode of operation is particularly useful for discovering new concepts, by grouping parts of content so far labelled as "unknown".

2.3.1 Analysis: Processing a multimedia document

Figure 3 summarises the first mode of operation for semantics extraction from multimedia content, highlighting the processing modules as well as their interconnection. The input is a multimedia document, potentially containing video, image, audio and text. The current status of the semantic model (T-BOX) also needs to be available. On the other hand, the system has two types od output. The first is the document's specific A-BOX, containing instances that have been found using media-specific analysis and reasoning. The second is the interpretation of the multimedia document, which also accounts for redundant or missing media elements, thus signalling model and/or analysis deficiencies.

The initially given multimedia document (eg. a web page), is processed by a cross-modality coordinator module that separates it into subdocuments corresponding to different media formats (i.e. video, audio, image and text). Each one of these subdocuments is then given for analysis to the corresponding analysis module. The analysis modules, based on the semantic model, analyse the subdocuments so as to identify and classify elements within, such as 2-dimensional regions in images, word sequences in text, temporal segments in audio and spatio-temporal segments in video. The set of allowable classification labels, for each modality, corresponds to a subset of the T-BOX concepts of the ontology which are directly identifiable in the media and referred to as mid-level concepts[3] (MLCs). The analysis also identifies relations between the extracted elements, such as `adjacent to` in image, `before` in audio or `subject of` in text.

The output of the analysis modules is a set of `xml` files containing the list of the extracted elements and element relations together with sufficient properties to describe the related MLCs, the position of the elements in the subdocument, the extracted features used to conduct the analysis, as well as the confidence estimation of the classification. A typical `xml` file is given in figure 4. Subsequently, a *result analysis* module, parses the `xml` in order to (a) re-route potentially identified encapsulated media (such as OCR extracted text) to the cross-modality coordinator for further analysis and (b) update a document-specific A-BOX with the extracted information.

Once the single-modality results are in the form of an A-BOX, they are subjected to reasoning, in the context of the given ontology. Interpreting the results has a double impact. First, the A-BOX gets populated with instances of further concepts (high-level concepts or HLCs), using the implicit knowledge in the semantic model. Second, inadequacies of the ontology to fit the analysis (such as missing or redundant instances, according to the current ontology) are identified.

When all media have been processed, fused-media processing takes place. Using the information provided through the document A-BOX, in the form of concept instances and their properties, complementary information is used to identify instances of the fused-media concepts. Finally, reasoning on

[3]For an detailed definition of MLCs and HLCs, the reader is referred to section 2.2.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<DOCANALYSIS MODALITY="image">
  <DOCREF TYPE=FILE>
    <FILE>image-0234.jpg</FILE>
  </DOCREF>
  <SEGMENT_LIST>
    <SEGMENT TYPE>
      <ID>seg01</ID>
      <POS TYPE=MASKPOS>
         <MASKPOS>
            <MASKFILE>mask-A.png</MASKFILE>
            <MASKNUM>250</MASKNUM>
         </MASKPOS>
      </POS>
      <INFO TYPE=INFOFILE>
            <INFOFILE>features-A-01.xml</INFOFILE>
      </INFO>
    </SEGMENT>
    <SEGMENT>
       <ID>seg02</ID>
       <POS TYPE=BBOX>
         <BBOX><TOP>10</TOP><BOT>34</BOT>
          <WID>100</WID><HEI>30</HEI></BBOX>
       </POS>
    </SEGMENT>
  </SEGMENT_LIST>
  <MLCI_LIST>
    <MLCI>
      <ID>mlc01</ID>
      <ELEMENTREF>seg02</ELEMENTREF>
      <LABEL>textbox_visual_element<LABEL>
      <CONFIDENCE>0.95</CONFIDENCE>
      <INFO TYPE=INFOTEXT>
        <INFOTEXT>Jonn Smith in action</INFOTEX>
       </INFO>
   </MLCI>
   <MLCI>
     <ID>mlc02</ID>
     <ELEMENTREF>seg01</ELEMENTREF>
     <LABEL>pole_visual_element<LABEL>
     <CONFIDENCE>0.76</CONFIDENCE>
     <INFO TYPE=INFOFILE>
       <INFOFILE>classification-A-01.xml</INFOFILE>
     </INFO>
   </MLCI>
  </MLCI_LIST>
  <MLCRELI_LIST>
    <MLCRELI>
    <ID>01</ID>
    <LABEL>down-left</LABEL>
    <MLCIREF_LIST>
      <MLCIREF>mlc01</MLCIREF>
      <MLCIREF>mlc02</MLCIREF>
    </MLCIREF_LIST>
  </MLCRELI>
 </MLCRELI_LIST>
</DOCANALYSIS>
```

Figure 4: A typical `xml` output of a single-media semantics extraction process. The `xml` consists of three parts corresponding to (a) the segmentation of the document, (b) the classification of segments as modality-specific mid-level concepts and (c) their relations. The specification of segments may be done in several ways, depending on the document type and the segments considered. In the particular example, a reference to the pixel value of a reference mask of an image and a bounding box are used. Each mid-level concept instance is associated to a particular segment and additional information about the confidence of estimation is provided. Finally, a list of the identified relations between the extracted mid level concept instances are provided. Notice that the format allows to describe information about encapsulated media. In the particular example, a text-detection and OCR algorithm has detected text in a specified segment and the result of the OCR has been inlined in the `xml` file.

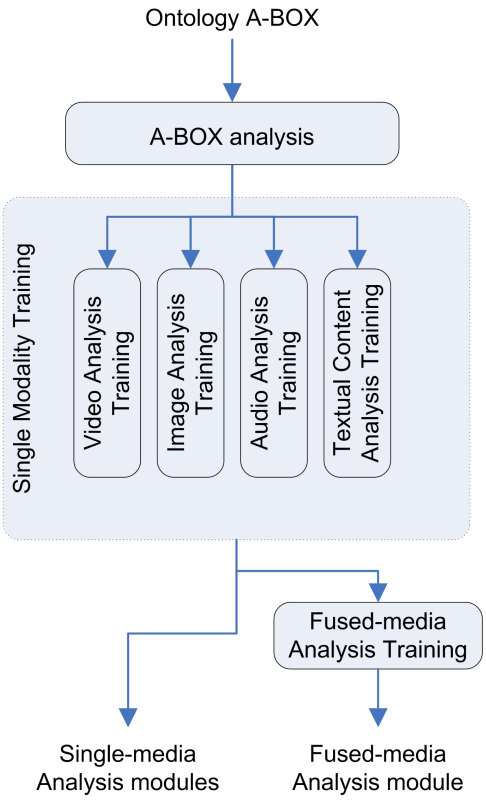


Figure 5: Training – improving multimedia analysis

the fused media instances is used to identify fused high-level concept instances as well as missing or redundant fused mid-level concept instances, by making use of the implicit knowledge in the semantic model.

The output of the system, i.e. the document's A-BOX, is then forwarded to the ontology evolute module describe in deliverable D4.1. Through operations defined within that system (dashed lines in Figure 3), a re-analysis of the same document may be requested, based on evidence that the analysis results did not satisfy the semantic model. In that case, hints based on reasoning results (such as missing or redundant instances or dominating high-level concepts), will be provided to guide and adapt the analysis of the given document. It is stressed, however, that a complete match between the semantic model and the analysis results may not be achieved, even after the re-analysis. In that case, a request for improving the analysis modules may be independently issued by the BOEMIE system. Improving the analysis modules is done through the second mode of operation, described in the following section.

2.3.2 Training: Improving multimedia analysis

Figure 5 summarises the methodology for improving multimedia analysis. The input is the ontology's A-BOX, comprising all the instances of multimedia elements so far analysed by the BOEMIE system. The outputs consist of improved versions of the single-media and fused-media analysis modules, expected to lead to improved analysis accuracy in the future.

By means of A-BOX analysis, each single-media analysis module is given a training set consisting of the corresponding MLCs instances, together with an assessment of whether the detection of these instances is true or not. The confidence of this assessment may vary according to if the population has been confirmed by a human operator and/or reasoning or if it has been judged as non-conformant to the semantic model. Each single-media analysis learning module applies machine learning algorithms on this data to improve the analysis, such that recall and precision rates for the given mid-level concepts

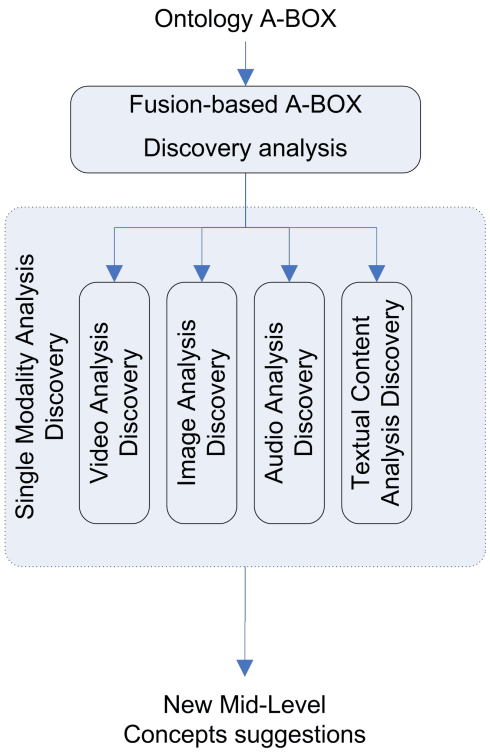


Figure 6: Discovery – expanding multimedia analysis

increases.

Once single-media analysis modules are enhanced, fused-media analysis learning takes place. The learning takes into account the new results to adjust parameters pertinent to fusion of modalities, such as the confidence levels attributed to each single-media analysis module. Both single and fused media analysis modules are then stored to be used for subsequent analysis requests (*Analysis*).

Notice that, since machine learning algorithms have finite capabilities, there is no guarantee that the analysis modules fully comply to the given A-BOX after learning has taken place. Actually, some analysis sub-modules may not be able to deal with the specific training procedure. Furthermore, it is to be expected that, since the analysis modules are potentially changed after applying this mode of operation, inconsistency may be introduced between the analysis modules and the ontology's A-BOX that was given as input. To ensure compatibility, using the *Analysis* mode of operation for all documents that have been used to populate the A-BOX is required.

This mode of operation also applies, as a specific case, during the initialisation phase of BOEMIE. An initial T-BOX modeling the application domain. together with an initial A-BOX formed by manually annotated corpus is required. Then, this methodology allows the analysis modules to be trained by means of the original A-BOX to build up a system able to identify the manually annotated MLC instances.

2.3.3 Discovery: Expanding multimedia analysis

Figure 6 summarises the methodology for expanding multimedia analysis. Similarly to the Schema 2, the input is the ontology's A-BOX, comprising all the instances of multimedia elements so far analysed by the BOEMIE system. The output is a suggestion for new modality specific mid-level concepts, which can ultimately lead to improved analysis of new multimedia content.

Based on analysis of the given A-BOX, a single-analysis module is fed with a set of instances, all of which are so far attributed to the same MLC. Furthermore, using evidence provided by cross-media information fusion and/or association with high-level concepts, instances in this set (denoted by S) are initially divided into subsets (S_1 and S_2). Instances of these subsets are assumed to be *implicitly*

associated with different concepts. The discovery modules strive to discover ways to discriminate between instances of the given subsets, using machine learning algorithms. In case of success, the analysis results are provided to the evolution module in order to propose new mid-level concepts, to be confirmed by a human expert.

Notice that although a complete discrimination between the subsets is the ideal case, a partial discrimination is also desirable. Namely, a sufficient goal of the analysis expansion module is *to discriminate between partial subsets of the original set* with high accuracy. The output of the analysis expansion module is a number of subsets (denoted by $\{O_i\}_{i=1}^{M}$) of the original set S which meet the following conditions: (a) the instances of each output subset O_i is itself a subset of one of the two input subsets (either S_1 or S_2) and (b) the instances belonging to each output subset can be discriminated against instances not-belonging to the subset with high accuracy.

To illustrate the methodology, we present two particular cases, in which the outcome of the analysis expansion is successful:

MLC split. The analysis expansion module discriminates all instances of S_1 against instances of S_2 with high overall accuracy. The output subsets are then the same as the input subsets: $O_1 = S_1$ and $O_2 = S_2$. The original MLC concept is proposed to be split into two MLCs, the instances of which correspond to the input subsets.

MLC discovery. When analysing a document, some of its parts (such as image segments) may not be classified to one of the MLC concepts of the current ontology. Nevertheless, these are kept to the ontology as instances of the concept "`unknown`". The discovery module may manage to identify a subset of the so far "`unknown`" instances, all associated to the same input subset. This may lead to the discovery of a new modality-specific mid-level concept, useful to identify the HLC associated with the respective input subset.

In any case, if the identified subsets are meaningfull, ontology enrichment will add (a) one or more MLCs to the modality-specific part of the ontology, together with rules that associate them with the MLC that subsumes them and (b) rules to associate each new MLC with the specific high level concept that lead to the formation of the subsets S_1 and S_2. The detailed enrichment procedure, which may give rise to the definition of further high-level concepts, is described in the ontology evolution methodology (deliverable D4.1)

We note also that, once new mid-level concepts are accepted, applying the training operation mode is necessary, in order to improve the analysis so as to comply with the new mid-level concepts. The analysis modules are then required to be able to identify the mid-level concepts by means of criteria similar to those that led to their discovery.

2.4 Support for ontology evolution

In BOEMIE, the semantics extraction process alternates in cycles with the ontology evolution process. The synergy of the respective tookits is established through, on one hand, the common multimedia semantic model they share and, on the other, the guidance of an integration component[4]. In this section, we describe the overall BOEMIE architecture from the viewpoint of semantics extraction, explaining the contribution of the proposed methodology to ontology population and enrichment, thus clarifying the contribution to the BOEMIE bootsrapping process.

The Ontology Evolution and Semantics Extraction toolkits interact significantly. Considerable effort has been put, though, in simplifying their interconnectivity, thus reducing the interface requirements. Figure 7 depicts the interconnection among the semantics extraction modes of operation on one side, and the ontology evolution processes on the other. Namely, there are two interfaces communicating the output of the semantics extraction toolkit to the ontology evolution toolkit: one concerning the output of a multimedia document (Analysis) and the other the identification of new mid-level concepts (Discovery). Notice that the outputs of the ontology evolution toolkit are communicated back to the semantics extraction toolkit indirectly (Analysis and Training), through the "evolved" ontology. This

[4]The reader is referred to deliverables D4.1, D3.1, D3.2 and D5.5 for detailed description of the ontology evolution, the multimedia semantic model and the integration component.

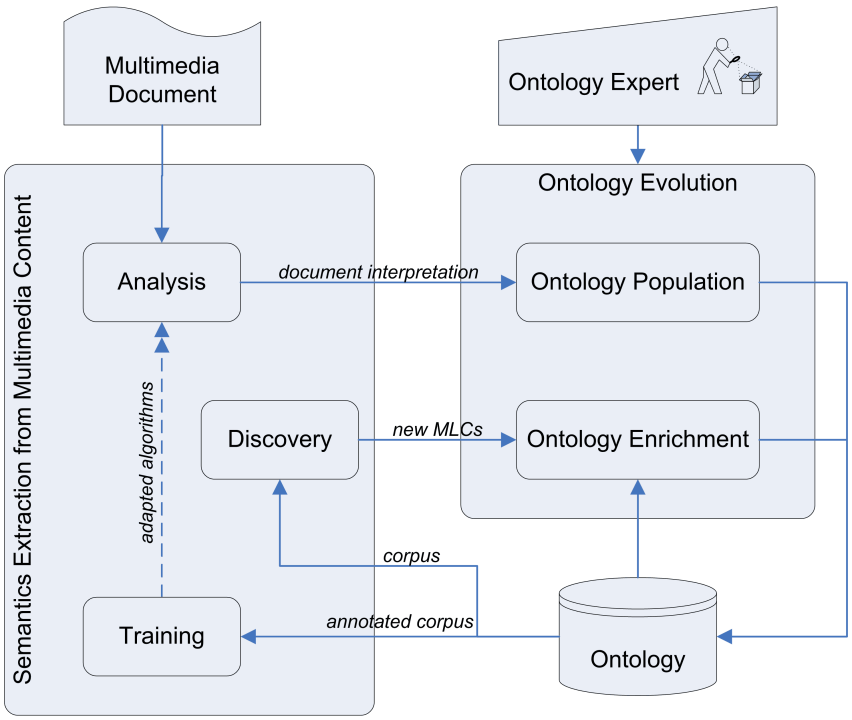


Figure 7: The bootstrapping process of BOEMIE, detailing the modes of operation. For simplicity, BOEMIE modules not directly relevant to semantics extraction have been omitted.

circular type of communication corresponds to the bootstrapping approach followed by the BOEMIE project.

The interoperabitlity between the toolkits may be described through three scenarios: (a) a new multimedia document enters the BOEMIE system in order to be analysed (b) the ontology's A-BOX contains statistically sufficient new information to allow potential expansion of the semantics extraction module and (c) the ontology's A-BOX contains statistically sufficient new information to allow potential enhancement of the semantics extraction module. In the following, we describe in detail each one of these scenarios.

Scenario 1. Let a new multimedia document enters the BOEMIE system. *Analysis* is applied in order to analyse the document. If the extraction results satisfy the ontology's T-BOX, they are passed on to the ontology evolution toolkit, which will eventually populate the ontology with new instances, corresponding to the knowledge extracted from the document. In case of mismatch between the extracted information and the ontology's T-BOX, the extraction results are judged by a human expert, who, using tools provided by the BOEMIE system, decides weather the mismatch is due to analysis deficiency and/or to a poor T-BOX. In the former case, a feedback is provided to the semantics extraction toolkit with the aid of the reasoning. Tha aim is to obtain better results. In the latter case, all relevant information, including non-conformant concept instances and/or unclassified (i.e., "`unknown`") instances, are kept in the ontology, in order to enable, at a later stage, ontology enrichment, as described in D4.1, and/or semantics extraction enhancement and expansion.

Scenario 2. Assume that a significant number of multimedia documents have already been processed by the BOEMIE system. Assume also that each single-modality semantics extraction module has potentially identified a significant number of important segments inside the processed documents which have not been classified into one of the known mid-level concepts. These have been kept in the ontology as instances of the "`unknown`" mid-level concept. Then, the ontology evolution tookit, queries the ontology to form the corpus relevant for mid-level concept discovery (see deliverable

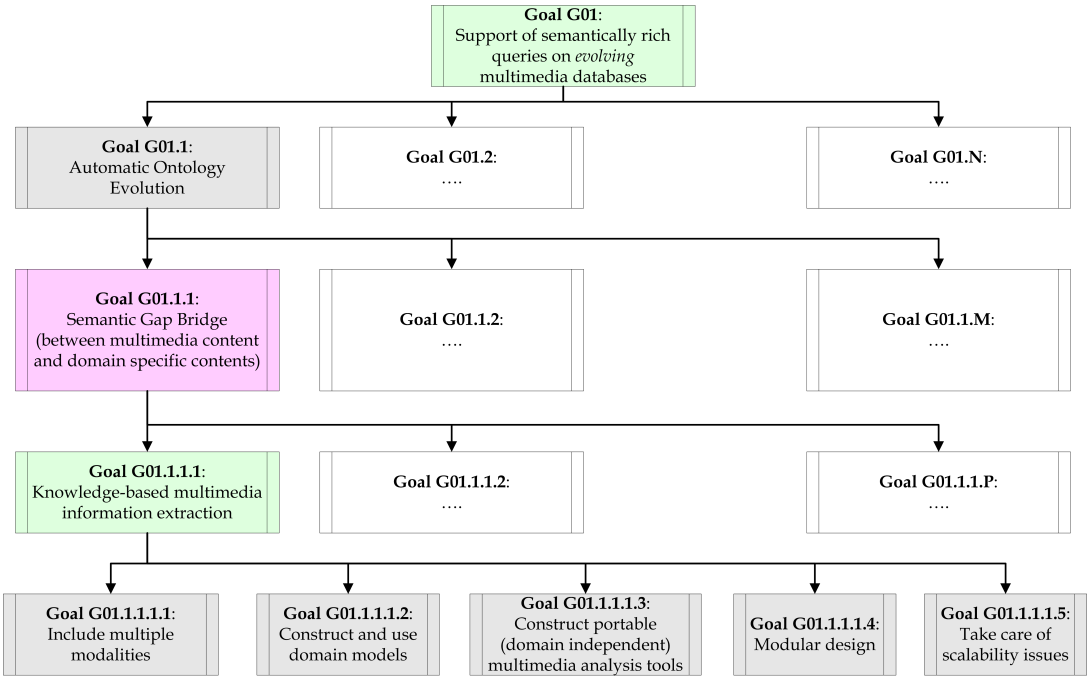


Figure 8: A goal analysis example of the BOEMIE project. *Note:* Other goals that may exist (e.g. G01.2...N) are not analysed because they do not affect the functionality of the semantics extraction methodology.

D4.1). Using the *Discovery* methodology, the respective single-modality modules strive to discover new concepts, either as refinement of the existing ones, or as new ones[5]. The results are given back to the ontology evolution tookit, for ontology enrichment.

Scenario 3. Let the ontology contain instances of MLC concepts which have not been detected automatically using Schema 1. There are three ways this may occur: (a) right after the ontology initialisation phase, where new manually annotated content has been provided, (b) when a significant number of occurrences of the first scenario has preceded, where the human expert has notified some deficient semantics extraction results and (c) when ontology enrichment has resulted in the addition of new mid-level concepts. In all these cases, there is reason for improvement of the semantics extraction modules, so as to take advantage the new annotated corpus, by means of machine learning supervised techniques. Hence, the interface particular to the *Training* mode of operation is applied, where the new annotated training instances are fed to their respective medium-specific module. Once learning is complete, the improved semantics extraction tools are stored, ready to be used when a new multimedia document enters the system (scenario 1).

2.5 Contribution to the state of the art

The architecture for semantics extraction from multimedia content implements one aspect of the BOEMIE methodology. In order to illustrate the innovative aspects of this architecture we will use the hierarchical goal analysis paradigm [Dix03]. One of the basic aims of the BOEMIE project is to support semantically rich queries on evolving multimedia databases. By analysing this goal down to lower level goals (see Figure 8) we face the need to construct knowledge-based multimedia information extraction methodologies. According to researchers working in this area (see Draper *et al* [Drap92a]) knowledge-based information extraction methodologies encounter the following problems:

[5]Note that refinement of an MLC concept and discovery of new MLC concept can both be viewed as refinement of an MLC concept, if one consider refinement of the concept "unknown" to a specific concept, and assumes the all MLC concepts subsume the "unknown" MLC concept

- Need to construct a brand new knowledge base,
- Domain dependency of the constructed tools, and
- Non-graceful scaling to large problems (difficulties in handling large datasets)

In the context of BOEMIE, one should add the need to:

- use information from multiple modalities and account for intermodality relations, and
- support modular and open design

The proposed architecture for semantics extraction from multimedia content has been designed for meeting all the above requirements in a unified manner. It presents a real advancement of the state of the art as there are no other such approaches in the corresponding literature (see *Deliverable D2.2-2.3: Semantics extraction from visual and non-visual-content*). Below we analyse in more detail the innovative aspects of the proposed architecture:

Integration of multiple modalities. To the best of the authors' knowledge, there are no other methodologies that utilise information from text, image, audio and video concurrently. Instead, knowledge-based approaches for information extraction in the corresponding literature include, at most, two modalities: either text and image [Alva04],[Dese05a],[Zaan04] or synchronised video and audio [Smit01], [Snoe05], [SP06]. The way the proposed architecture handles multimodal information per modality and as a whole (fused) is indicated in Figure 3.

Modular design. Modular design is not used explicitly in any of the multimodal approaches stated above. Multimodal information is handled in a woven way prohibiting graceful replacement of some parts of the system. In the proposed architecture, modular design is used both across modalities (see Figure 3), as well as within each modality (see the particular methodologies for semantic information extraction from images, text, audio and video in the following sections). Cooperation among the various components is achieved through well-defined input and output structures. The way each component operates, either as a whole (e.g., the image analysis component) or a modality-specific process (e.g., the segmentation process in the image modality), is transparent to the overall architecture.

Use of existing domain knowledge. Ontologies provide the means for constructing new knowledge-based systems [Nech91] by assembling reusable components. Several methodologies appeared in the literature that make use of ontologies. However, most of them are based on a single modality such as images [Hunt01] or video ([Meza04], [Dasi05]). Furthermore, in the aforementioned methodologies it is unclear in which way the use of ontologies enhances the analysis results. That is, ontologies are used, mainly, to specify and describe a particular domain but they do not affect the functionality of the modality specific tools. In the proposed architecture, the ontology provides to the various modalities a common semantic model enabling a formal fusion process based on description logic, but, more important, to adapt in a bootstrapped fashion the functionality of the modality-dependent information extraction tools. The *Analysis* mode of operation of the proposed architecture indicates the ontology-based feedback process during information extraction. Semantic model is also explicitly dealt with in each modality-specific analysis (see semantic model usages in sections 3, 4, 6, 7).

Portability to other domains. The embedding of learning techniques in multimedia analysis systems reduces or eliminates the domain dependency of the corresponding information extraction tools [Drap92b]. This view influences the work of several researchers, especially in the area of image and video analysis, resulting in methodologies ([Duyg02], [Laak04], [Lavr03], [Li03], [Lowe99], [Lowe04], [Metz04], [Viol01], [Yavl05]) that built modality specific concept models which are based on large numbers of primitive features. What distinguishes the proposed architecture from the aforementioned methodologies is the coordinated fashion through which portability to other domains can be achieved, as provided by the *Training* mode of operation. Learning is also supported within each modality as it is explained in the description of the corresponding methodologies. Therefore, the

multimedia information extraction tools which in the BOEMIE project will be constructed for the athletics' domain can be transferred to other domains without modifying the internal structures of any of the architectural components. The only requirement will be the usage of (a) a new domain ontology, and (b) training corpus through which the learnable components of the architecture will associate low-level features to domain concepts. Both learning techniques and low-level features, however, are internal in the architecture and transparent to the customisation process.

Scalability. Scalability handling is probably the most important innovation of the proposed architecture. Multimedia document databases are data volumes which constantly increase. In order to make sure that the evolving nature of the underlying multimedia database will not affect the performance of the semantic multimedia information extraction the proposed architecture allows dynamic expansion of the corresponding semantic model (ontology enrichment) as well as modification of modality dependent concept models. Schema 3 describes in detail this process while at the same time corresponding sections (enrichment handling) in the description of the various methodologies present modality specific details.

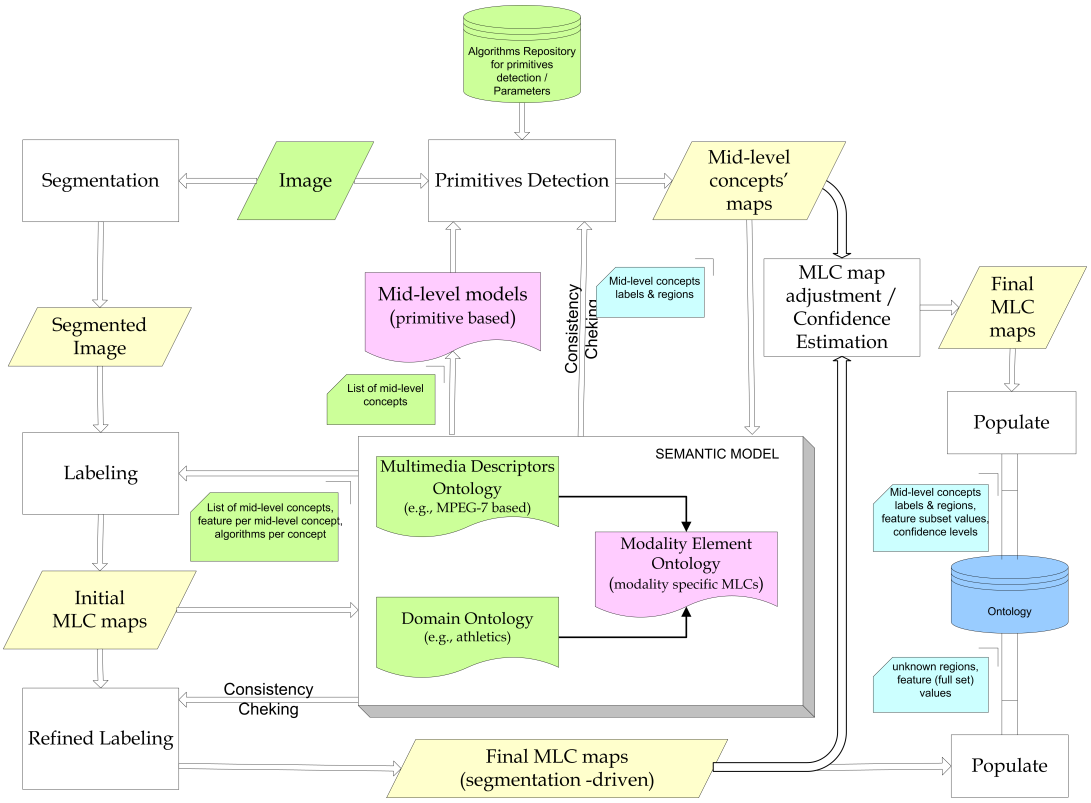


Figure 9: Schematic Diagram of the proposed methodology for semantics extraction from still images

3 Still Images

3.1 Aim of image-based information extraction

The aim of semantics extraction from still images is, for any input image, to provide information about the existence of image dependent MLCs (semantic labels such as `pole`, `podium`, `horizontal bar`, `sandpit`, `hurdle`, `lane`, etc.), their maps (unique region numbers that identify the image area that is covered by a particular mid-level concept), their low-level descriptors (e.g., scalable colour descriptor, etc) and complementary information about unknown image regions (i.e., MPEG-7 colour, texture and shape descriptors) which will allow for new mid-level concepts identification using the *Discovery* methodology described in section 2.3.

3.2 Overview of the methodology

The schematic diagram of the methodology for semantics extraction from still images is shown in Figure 3.2. It will be explained in Section 3.3 that the overall methodology combines two approaches for image analysis:

1. Region-based analysis (anticlockwise path in Figure 3.2) and,
2. Holistic image analysis (clockwise path in Figure 3.2)

The meaning of the various shapes shown in Figure 3.2 is explained in Figure 10.

3.2.1 Inputs to image analysis module

There are other BOEMIE modules that take advantage of the information extracted by the Image Analysis module. Information exchange with these modules is achieved with an ontology acting as an

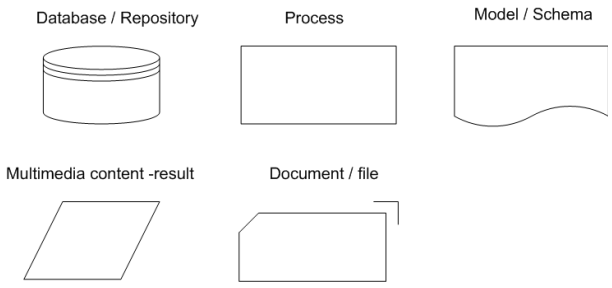


Figure 10: Meaning of shapes shown in Figure 3.2

intermediary. We will define the input expected from and output provided to those other modules. According to the DoW, p.48, WP2 modules receive information from:

Annotated content for still Images This information will be used to:

- train classifiers for labeling image regions with one of the available MLC labels (or with the label 'unknown' when no match with existing MLCs is detected),
- identify a proper set of algorithms for MLC detection using the holistic image analysis approach
- evaluate the performance of the above mentioned classifiers and algorithms. The annotated content for a still image is described through an `xml` file. The general structure and indicative contents of this file are shown in figure 11. Further details can be found in BOEMIE deliverable *D5.1: Content Collection and Annotation Strategy.*

Multimedia content & descriptor ontology (*see DoW, Task 3.1, p.49*). The multimedia ontology indicates the set of low-level descriptors whose computation is supported by the annotation tool. Based on this set, MLC labelling of an image region (segment) should be feasible. It is anticipated that:

- the minimum list of descriptors will include the MPEG-7 colour, shape and texture descriptors,
- the set of available descriptors will be extendable so as to include additional low level features that may be required, especially for the holistic approach.
- The list of available low-level descriptors is communicated to the image analysis module through an *xml* file.

Domain Ontology (*see DoW, Task 3.2, p.49*). The domain ontology provides the image analysis module with the list of Mid-Level Concepts that are present in a particular domain. These MLCs are structuring elements for high level concepts (HLCs), but recognition of HLCs is beyond the scope of semantics extraction from still images module.

The MLC list is required by both the region-based and the holistic approach. In the first case MLCs define the classes (with the addition of the 'unknown' class) to which an image segment should be classified to. This implies that proper classifiers should be trained to support the classification. In the holistic approach MLCs are needed in order for models (based on primitives) to be constructed (e.g., *"Lane "* = {*parallel lines, area with uniform colour between them*} and a set of algorithms for primitives' detection to be defined. Classifier training and modeling through primitives are both off-line procedures.

Multimedia Domain Model (*see DoW, Task 3.3, p. 50*) The multimedia domain model acts as a bridge between domain and multimedia ontologies. In the particular example of still images the multimedia domain model includes the MLC classifiers and the holistic detection algorithms along with their parameters (if any). The following information is, therefore, implicitly (through the MLC classifiers and holistic MLC detection algorithms) encoded in the image portion of the multimedia domain model:

```
<?xml version="1.0" ?>
<!--  Written on 25-Sep-2006 12:16:00 using the XML Toolbox  -->
- <root>
    <LocDir>E:\Projects\BOEMIE\09-content\IAAF\pictures\</LocDir>
    <htmlFile>20822.html</htmlFile>
    <OrigImage>20822_W400XH600.jpg</OrigImage>
    <SegImage>20822_W400XH600_Mask.png</SegImage>
    <NumOfMLCs>3</NumOfMLCs>
  - <MLClist>
      <ID>MLC01</ID>
      <MLClabel>Athletes_face</MLClabel>
      <Region>255</Region>
      <EstConfidence>0.85</EstConfidence>
      <InfoFile>20822_W400XH600_Athletes_face_255.xml</InfoFile>
      <ID>MLC02</ID>
      <MLClabel>Pole</MLClabel>
      <Region>250</Region>
      <EstConfidence>0.54</EstConfidence>
      <InfoFile>20822_W400XH600_Pole_250.xml</InfoFile>
      <ID>MLC03</ID>
      <MLClabel>OCRtext</MLClabel>
      <Region>245</Region>
      <EstConfidence>0.9</EstConfidence>
      <InfoFile>20822_W400XH600_OCRtext_245.xml</InfoFile>
    </MLClist>
    <NumOfRels>2</NumOfRels>
  - <MLCrelations>
      <ID>R01</ID>
      <Args>MLC01, MLC02</Args>
      <Relation>UP-RIGHT</Relation>
      <ID>R02</ID>
      <Args>MLC01, MLC03</Args>
      <Relation>OVER</Relation>
    </MLCrelations>
  </root>
```

Figure 11: A representative example of an xml document corresponding to the annotation results of a still image. The MLCs referred in this file are further described within the corresponding MLC xml files (in the above example for the MLC 'pole' further information is provided in the xml file named "20822_W400XH600_pole_250.xml".

- Low-level descriptor subsets that are required for the various MLCs to be recognised. Given that region labeling will be performed by classifiers, these subsets are useful for computing the corresponding descriptors, after a training process.
- Spatial relations between MLCs. This information can serve as a 'safeguard' during the image analysis process. For example, if the MLC `sandpit` appears in all images in a lower position than the MLCs `athletes_face` and `athletes_hands` (when the latter exist in the input image) then, once an MLC `athletes_face` has been identified and located, it is prohibited for regions that are above the `athletes_face` to be labeled as `sandpit`. This information will be used after the initial labeling.

Classifiers and holistic detection algorithms along with their parameters are communicated to the semantics extraction from still images module to allow MLC area detection (region labeling with MLCs).

3.2.2 Output of the image analysis module

There are various BOEMIE modules that use the image analysis results. According to the *DoW, p.48*, WP2 these modules are:

Multimedia and domain ontologies (*see DoW, Task 4.2, p.56*) The Image Analysis Module populates the domain ontology with Mid-Level Concept Instances (MLCIs) and the multimedia ontology with low-level descriptor instances (numeric and/or symbolic values). Furthermore, a link between an MLC instance and the corresponding low-level descriptor instances is preserved. Domain ontology population with MLCs includes a confidence value denoting the belief of the classifier that the label assigned to a particular image region is correct.

MCLIs and low-level descriptor instances are communicated to domain and multimedia ontologies through `xml` files.

Methodology for multimedia ontology enrichment (*see DoW, Task 4.1, p. 56*). In order to allow for ontology enrichment to take place (described in the *Discovery* methodology), computation of low-level descriptor values for all unclassified (labeled `unknown`) image segments should occur. In this case, the full set of descriptors is computed for each unknown image region. Unclassified MLCIs along with the corresponding low-level descriptor values are used to populate a temporary multimedia repository.

3.3 Methodology

3.3.1 Multimedia descriptors ontology

For the semantic information extraction from still images the definition and adoption of a multimedia ontology is of great importance. In the proposed methodology three architectures of such ontologies will be examined [Meza04], [Dasi05], [Hunt01] in order for the image analysis information to be semantically handled. For a detailed review on ontology architectures to be used for semantics extraction from still images see the BOEMIE deliverable *D2.2: Semantics extraction from visual content tools: State-of-the-art report.*

3.3.2 MLC detection module

As far as the image analysis is concerned two main approaches will be combined (see Figure 3.2):

- region-based analysis (segmentation) and,
- holistic image analysis (primitives detection)

In the region-based approach each input image is partitioned into segments [Prat06b] with the aid of the Watershed transform [Meye90] or other image segmentation techniques [Gonz02]. For each segment a set of properties (low-level descriptor values) is computed. In the case of the MPEG-7 visual descriptors

the MPEG-7 Experimentation Model [Ciep01] will be used for their computation. Based on the descriptor values a trained classifier assigns an MLC label to every region. Initial labeling is then checked for consistency by the semantic model (actually by the fused media analysis module – see Figure 3) and reasoning feedback is passed to the image analysis module. Interpretation feedback may, for example, lead neighboring segments initially assigned different MLC labels, to be merged into one region with a single MLC label. Conclusion of the consistency checking, therefore, will lead to a refined set of MLC maps.

In the holistic approach, MLCs are modeled through primitives (simple geometric objects such as lines, ellipses, etc., or composite objects whose detection is feasible through dedicated algorithms, e.g. face) [Belo02],[Viol01]. Accurate primitive detection accompanied with domain knowledge allows for mid-level concepts to be recognised. Mid-level concept modeling can be explicit (through the definition of rules and relations among primitives) or implicit (learned through classifier training). Once an initial set of MLC maps (mapping between MLCs and low-level descriptors) has been created consistency checking and guidance follows through the fused media analysis module. This may result in a second run of MLCs detection, triggered, for example, by the fact that an MLC was expected to be present in the image and was not found, or an MLC found was not expected to be present in the image.

MLC maps created by the segmentation (region-based) process and the holistic process are combined to produce the final MLC maps which will populate the ontology / repository. At the same time a confidence value indicating the belief that the detection and identification of an MLC is correct is also computed for each one of the MLCs.

3.3.3 MPEG-7 descriptors and other low-level features

In the region-based approach identification of MLCs will be based mainly on MPEG-7 low-level descriptors. In particular the MPEG-7 Colour, Shape and Texture descriptors [Chan01], [Eide03], [Eide04], [ISO], [Kosk02] will be computed for each image region. Additional low-level features that will be studied include:

- Shape Adaptive-DCT coefficients ([Siko95])
- Histograms of Oriented Gradients ([Dala05])
- Line orientation ([Prat06a])
- Moments of the RGB and /or LUV colour space ([Prat06a],[Vail01])
- Low order statistics of the energy wavelet coefficients ([Daub88],[Prat06b],[Unse95])
- MSAR texture features ([Mao92])

3.3.4 Mid-Level concepts identification

In the proposed methodology mid-level concepts are identified in two different ways:

Region-based For each image segment the MPEG-7 colour, shape, and texture descriptor values along with other low–level feature values are computed. These values are fed to a set of classifiers in order for an MLC label to be assigned to each image segment. Following that, connected regions with the same MLC label are merged producing the final map for each MLC.

Therefore the modelling of MLCs using the low-level descriptors is achieved through a set of classifiers. For each MLC a particular classifier is trained [Kosk02], [Laak04], [Li03]. Thus, given a subset of MPEG-7 descriptors and other low–level feature values the classifier decides whether the region, for which feature values have been computed, is a part of this particular MLC or not. The same process applies to the rest of the MLCs. In case none of the classifiers identifies the region the label `unknown` is assigned to it.

Holistic approach In the holistic approach there is no explicit modelling of MLCs based on MPEG-7 descriptors. Instead, MLCs are modelled via primitive geometric shapes (lines, arcs, ellipses etc.)

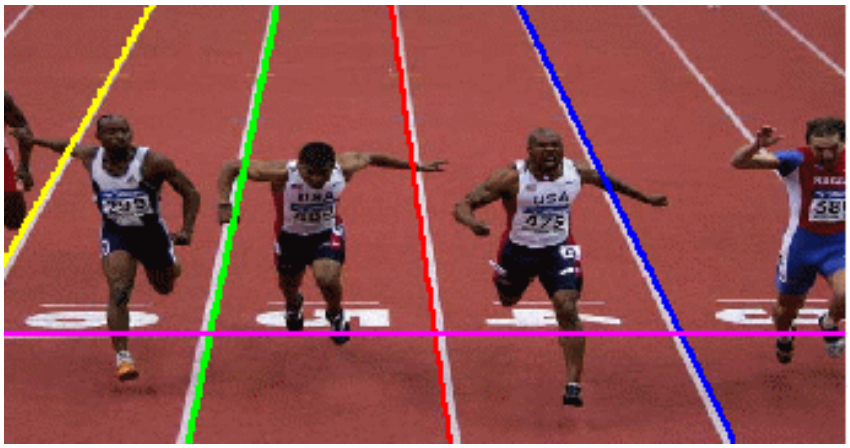


Figure 12: Primitive (line) detection in a running event. Lanes can be modeled as parallel lines with a large area of homogeneous colour between them

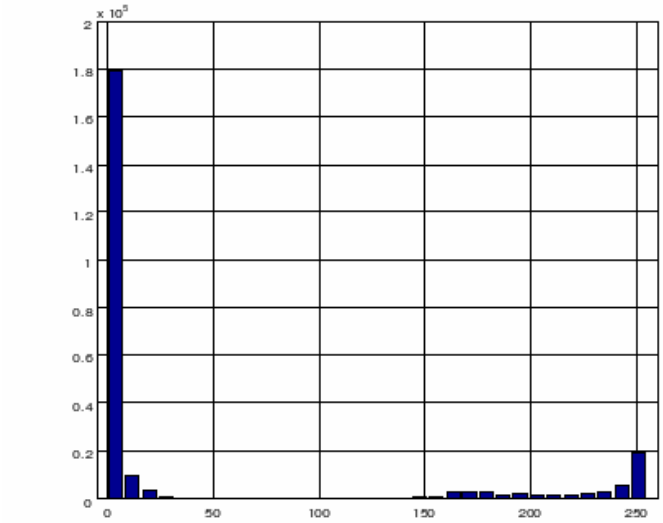


Figure 13: Colour histogram (colour model HSV, colour channel H, 32 bins) of areas between parallel lines of Figure 12

or extracted directly via dedicated algorithms (e.g., face detection) . However, in several cases dedicated object detection algorithms [Schn04],[Felz05], make use of MPEG-7 or similar descriptors [Lowe04].

As an example one may consider the MLC `lane` modelled as {`2 parallel lines`, `uniform colour area between them`}. The Hough transform [Gonz02] for line detection does not make explicit usage of any of the MPEG-7 descriptors. On the other hand there are several MPEG-7 colour descriptors (Scalable Colour Descriptor, Dominant Colour Descriptor, Colour Structure Descriptor) that can be used to indicate that an image area is of homogeneous colour. For example, Figure 12 shows the six most prominent lines of a photo taken from a running event, overlayed on the original image. The colour histogram (H channel of colour model HSV) of image areas lying between the parallel lines (see Figure 13) indicates clearly the homogeneity in colour of those areas. Therefore, the hypothesis that in the input image there are lanes is confirmed. The output of image analysis would be a map indicating the area that corresponds to a lane (an instance of the MLC `lane`). Obviously, in the case of the photo shown in Figure 12, several `lane` instances would be extracted.

3.3.5 Hierarchical search

The semantics extraction from still images module may use a hierarchical MLC search based on the:

- discriminative power of each MLC,
- ease of detection of MLCs

Let us consider the set of MLCs used to describe a particular domain: $\mathcal{M} = \{\text{MLC}_i\}_{i=1}^{N}$. We assume that high level concepts (HLCs) are modelled via subsets $\mathcal{H}_j$ of $\mathcal{M}$, along with relations among MLCs. The discriminative power of an MLC is inversely proportional to the number of subsets $\mathcal{H}_j$ that the MLC appears in. Once discriminative indices for the existing MLCs are available, it is natural to assume that MLCs with a higher discriminative power are searched first. This means that early decisions about HLC identification can be made, or, triggered search (see next section) can be activated.

Example Assume the HLCs `pole_vault`, `high_jump`, `javelin_throw,` and `hammer_throw` are modelled via the following sets of MLCS:

- `Pole_vault` →{ `pole`, `horizontal_bar`, `bed`, `pillars`, `athlete_face`, `athlete_leg`, `athlete_hand` }
- `High_jump` →{ `horizontal_bar`, `bed`, `pillars`, `athlete_face`, `athlete_leg`, `athlete_hand`}
- `Javelin_throw` →{ `javelin`,`sky`, `athlete_face`, `athlete_leg`, `athlete_hand`}
- `Hammer_throw` →{ `hammer`, `guard_net`, `sky`, `athlete_face`, `athlete_leg`, `athlete_hand`}

It is evident from the above sets that the MLCs with the highest discriminative power are `pole`, `javelin`, `hammer`, and `guard_net`, because they appear in only one set. Obviously, the least discriminative power have the MLCs `athlete_face`, `athlete_leg`, `athlete_hand` because they appear in all HLC sets.

It is important to note here, however, that relations among MLCs are also of particular importance for identifying MLCs. This means that though, for example, the set of MLCs used to model `high_jump` is a subset of the corresponding set for *pole_vault*, the particular spatial relations (in the case of images) among MLCs or motion trajectories of MLCs (in the case of a video sequence) make the two HLCs distinguishable.

An MLC with high discriminative power may be hard to detect in a particular medium (e.g., image). In this case hierarchical search for MLCs can be directed by the ease of MLC detection. In this way, successful detection of a particular MLC may facilitate, and render more robust, the detection of subsequent MLCs. As an example consider an image taken from a hammer throw event. Though the hammer has the most discriminative power, other MLCs such as `guard_net` and `sky` may be easier to detect. Therefore, the detection of these MLCs first may help in a robust detection of the hammer.

3.4 Semantic model usage

Both region-based and holistic approaches take advantage of information provided by the semantic model to perform the analysis in a 'guided' manner. For example:

- A multimedia descriptors ontology may indicate the available descriptors based on which an MLC should be identified.
- A modality element ontology indicates a set of MLCs that is likely to be present in the input image (e.g., if it is known that the input image comes from the athletics domain then MLCs such as pole, hurdles, javelin, sandpit may appear in it while others, like F1 racing cars, is very unlikely to be found)
- An image-based model may provide links between MLCs (e.g. pole), with multimedia ontology concepts (descriptors) and provide expected e.g., spatial relations among MLCs.

There are several ways in which the semantic model can support image analysis by taking into account high-level knowledge via reasoning. The way in which reasoning will be applied in BOEMIE to semantics extraction from images via two specific reasoning services is described in detail in sections 7.3 and 7.4.

3.5 Support to ontology enrichment

The aim of this section is to indicate the information that will be passed to the Ontology Evolution process in order for the ontology enrichment process to take place. This information is mentioned in DoW, p.56:

> More specifically, the extraction process will populate the ontologies with instances of the various concepts, together with their properties and will also provide unclassified entities extracted from the multimedia content which may lead to suggestions for the enrichment of the ontologies with new concepts and relations, through novelty detection. *This novelty detection is based on information from all different types of media being processed. ...* This notion of similarity will drive the evolution process.

It is clear that unknown MLCs cannot be directly recognised from the image analysis module because neither a class was assigned to them through the existing set of classifiers, nor a model through primitives for them exists. Image Analysis module labels all disjoint image regions that do not match with an existing MLC with the label 'unknown'. However, for each one of these regions the full set of low-level descriptor values and /or symbolic names is computed and stored in a temporary repository.

Upon notification by the repository that a large number of images containing regions labelled `unknown`, have been collected, the new concept detection process is activated.

Unsupervised clustering [Wall05] and Resource Allocation Networks (RAN) [Plat91],[Pert03] can be used for new concept detection. The assumption here is the following: If 'unknown' image segments with similar properties (low–level descriptor values) appear systematically in the given set of images then this is an indication of the existence of a new concept (which, in any case, should be confirmed by a human annotator). Unsupervised clustering is a possible solution for grouping together 'unknown' image segments across a large set of images, so as to provide a suggestion to the human annotator.

Example In Figure 14 the systematic appearance of floodlight in images may suggest a new concept creation:

It should be noted that support to ontology enrichment handling is an important reason for adopting the region-based approach in image analysis. In the holistic approach not known MLCs are ignored, that is, image analysis labels only those image areas that correspond to already known MLCs. The rest of the image area is considered as a single entity. Therefore, it does not allow for `unknown` segment grouping.

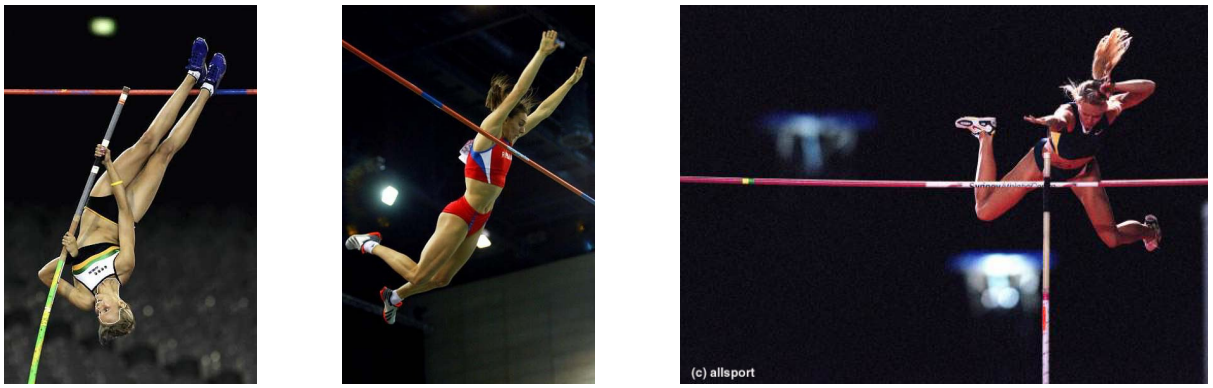


Figure 14: A set of images that may trigger a new concept creation (floodlight).

3.6 Confidence handling

Confidence estimation is required as stated in DoW, p.45:

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

For each mid-level concept a confidence measure should be computed. Given that there are two different channels (holistic and region -based) through which MLC regions are estimated a combination scheme should be defined so as a single confidence level is estimated.

We suppose that both the holistic and the region-based approach provide an individual confidence level for the estimation of the i-th mid-level concept, denoted as $c_{H,i}(I)$ and $c_{R,i}(I)$ respectively, where I is the input image. Let's also denote with $M_{H,i}(I)$ and $M_{R,i}(I)$ the masks of the MLC regions identified by the holistic and region-based approach, respectively. The term mask is used here to denote that every pixel of image I that belongs to the MLC region is denoted with one, while the remaining pixels are denoted with zero. For the estimation of a combined confidence measure a function $f\left(c_{H,i}(I), c_{R,i}(I), M_{H,i}(I), M_{R,i}(I)\right)$ needs to be defined.

A simple way for combined confidence level estimation is given below:

$$c_i(I) = \frac{\max\left(c_{H,i}(I), c_{R,i}(I)\right) \sum M_{H,i}(I) \cup M_{R,i}(I)}{\sum M_{H,i}(I) \cap M_{R,i}(I)} \tag{1}$$

where $\cup$ and $\cap$ denote the AND and OR operation, respectively and $c_i(I)$ is the estimated confidence level for i-th mid-level concept in image I.

As far as the individual confidence levels $c_{H,i}(I)$ and $c_{R,i}(I)$ are concerned their computation will be based either on pseudo-probabilistic density functions (holistic approach) or on the performance of the classifiers hat will be used to assign semantic labels to regions (region based approach). For, example a pseudo-probabilistic function can be defined to assign a confidence level based on the distance between the support vector (of the winning class) and the input vector, in the case of SVM classifiers.

In the example shown in Figure 15, we consider a hypothetical mid level concept (MLC) `face`. If the holistic and the region-based confidence levels for the identification of the MLC `face` are $c_{H,\texttt{face}}(I) = 0.5$ and $c_{R,\texttt{face}}(I) = 0.7$ respectively then the combined confidence (according to eq. 1), is:

$$c_{\texttt{face}}(I) = \max(0.7, 0.5)\frac{2294}{2760} = 0.5818$$

In the above equation the number of common pixels in the face maps $M_{H,\texttt{face}}$ and $M_{R,\texttt{face}}$ equals to 2294 while the total number of pixels covered by the union of $M_{H,\texttt{face}}$ and $M_{R,\texttt{face}}$ face maps equals

(a) Face Region Outline as a result of image segmentation

(b) Face map as a result of image segmentation

(c) Face Region Outline as a result of a dedicated face detection algorithm (holistic approach)

(d) Face map $M_{H,i}$ of a dedicated face detection algorithm (holistic approach)

Figure 15: Example of extracting the face MLC through segmentation-based or holistic approach

2760.

3.7 Evaluation framework

The evaluation of the technology for semantics extraction from still images is very important and aims to prove the appropriateness of the proposed methodology through quantitative measures.

3.7.1 Strategy

The evaluation of semantics extraction from still images involves the evaluation of the various technologies to be developed against measurable objectives.

The technologies that will be evaluated are:

MLC extraction This deals mainly with the accuracy of MLC area extraction.

Ontology population with new MLC instances The main check points are the correctness of (i) image region labelling, and (ii) estimated low-level descriptor values per MLC.

Ontology enrichment This refers to the performance evaluation of the algorithms (such as clustering [Wall05] and RAN [Plat91], etc), for new MLC detection.

Table 1, summarises the objectives, evaluation process and data requirements for the evaluation of semantics extraction from still images.

3.7.2 Content

Manually annotated images (created using the M-Ontomat tool) will be used as ground truth content. The annotation requirements have been stated in Section 3.2.1. Ground truth content will include a minimum of:

- 100 images per category for three categories of athletics events (jumping, throwing, running)
- 50 images for each of two events of a particular category
- 1-5 MLCs per image
- Images with varying content (e.g., background)
- Images of, continuous different stages of an athletic event (e.g. start, finish, jump upon hurdle, etc)

Table 2 summarises the properties of the ground truth content that will be used for the evaluation.

3.7.3 Quantitative measures

MLC area extraction In order to quantitatively evaluate the accuracy of MLC area detection the Mean Area Recall (MAR), and Mean Area Precision (MAP) measures are defined.

Let's consider that a set of N annotated images $I_j, j = 1, \ldots, N$, exist. We denote by AC_i^j the area (map) covered by the i-th MLC or unknown image region in the j-th annotated image. This area is denoted by pixels having the value 1, in contrast to non map pixels which have the value of 0. By EC_i^j we denote the estimated, by the MLC extraction process, area for the i-th MLC or unknown image region in j-th image. Area Recall (AR) and Area Precision (AP) are defined as follows:

$$AR_i^j = \frac{\sum AC_i^j \cup EC_i^j}{\sum AC_i^j} \tag{2}$$

$$AP_i^j = \frac{\sum AC_i^j \cup EC_i^j}{\sum EC_i^j} \tag{3}$$

Objective	Evaluation	Data requirements	Approach
MLC area extraction Detection of an image areas covered by MLC	**Scale** Hundreds of images. **Evaluation approach** Comparisons against areas annotated with the M-Ontomat tool. **Target** Area precision $>$ 60%, Area recall $>$ 60%	At least 300 dpi spatial resolution, 24 bit colour resolution	Region-based Holistic
Ontology population automated identification of new MLC instances and accurate computation of low-level descriptor values	**Scale** Hundreds of images. **Evaluation approach** Hide part of the annotated set of images and attempt to identify MLCs and corresponding low-level descriptor values. **Target** Correct labelling $>$ 70%, low-level descriptor values $\rightarrow$ average vector correlation with the ground truth $>$ 0.7	MLC instances to be identified $>$ 1000. MLCs from at least 5 different high level concepts A few tens of descriptors per MLC.	Region-based Holistic
Ontology enrichment Automated identification of new concepts and relations	**Scale** tens of images **Evaluation approach** hide part (up to 20%) of the MLCs and attempt to reconstruct them semi-automatically. **Target** to be able to identify a large part ($>$ 70%) of the hidden concepts.	New MLCs to be detected $>$ 5. MLCs from at least 3 different high level concepts	Clustering RAN Fuzzy Reasoning

Table 1: Measurable objectives for technologies dealing with semantics extraction from still images.

HLC	MLCs	Descriptors / primitives	Quantity
Jumping events: • pole_vault • high_jump	Pole, bed, pillars, horizontal bar, athlete_face, athlete_leg, athlete_hand	PEG-7 colour, texture and colour descriptors per MLC and per any unknown image region, Face, lines, arcs, homogeneous skin areas	100 images: 50 for pole vault, 50 for high jump. At least 3 MLCs per image
Throwing events: • hammer • javelin	Hammer, javelin, guard_net, athlete's_face, athletes_leg, athlete's_hand	PEG-7 colour, texture and colour descriptors per MLC and per any unknown image region, Face, lines, circles, homogeneous skin areas	100 images: 50 for hammer, 50 for javelin. At least 1 MLC per image
Running events • 100_meters • 100m_hurdles • marathon	Hurdle, lane athlete's_face, athletes_leg, athlete's_hand	PEG-7 colour, texture and colour descriptors per MLC and per any unknown image region, Face, lines, homogeneous skin areas	100 images: 50 each event At least 3 MLCs per image

Table 2: Ground truth content and evaluation parameters

where $\cup$ denotes the AND operation and $\sum$ denotes the sum of pixel values.

The Mean Area Recall (MAR), and Mean Area Precision (MAP) are the averages of the Area Recall and Area Precision over the whole set of annotated images:

$$MAR = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{ND_j}\sum_{i=1}^{ND_j} AR_i^j\right] \tag{4}$$

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{ND_j}\sum_{i=1}^{ND_j} AP_i^j\right] \tag{5}$$

where ND_j is the number of MLCs and disjoint unknown image regions in the j-th image. MAR and MAP values should both exceed 70% for the MLC area extraction process to be considered successful.

Computation of low-level descriptor values Accurate computation of low-level descriptor values will be evaluated with the help of the Average Vector Correlation Coefficient (AVCC) measure.

Let's denote by AC_i^j the concatenated vector of low-level descriptor values corresponding to the i-th MLC or unknown image region in the j-th annotated image. Let's also denote by EC_i^j the concatenated vector of low-level descriptor values computed by the MLC extraction process. The Vector Correlation Coefficient (VCC) is defined as follows:

$$VCC_i^j = \frac{{AC_i^j}^\top EC_i^j}{{AC_i^j}^\top AC_i^j + {EC_i^j}^\top EC_i^j}. \tag{6}$$

The Average Vector Correlation Coefficient is the average of the Vector Correlation Coefficient over the whole set of annotated images:

$$AVCC = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{ND_j}\sum_{i=1}^{ND_j} VCC_i^j\right] \tag{7}$$

where ND_j is the number of MLCs and disjoint unknown image regions in the j-th image. The AVCC value should exceed 0.7 for the low-level descriptor values computation process to be considered successful.

MLC labelling Correct MLC labelling will be evaluated with the help of the Mean Classification Error (MCE) measure.

Let's denote by AL_i^j the label of i-th MLC in the j-th annotated image. Let's also denote by EL_i^j the label assigned to the i-th MLC in the j-th image by the MLC extraction process. Labelling is correct when $AL_i^j = EL_i^j$ under the assumption that Area Recall (AR_i^j) and Area Precision (AP_i^j) both exceed 60%. The MCE is defined as follows:

$$MCE = \frac{1}{N}\sum_{j=1}^{N}\left[1 - \frac{1}{NM_j}\sum_{i=1}^{NM_j}(AL_i^j = EL_i^j)\right] \tag{8}$$

where NM_j is the number of MLCs in the j-th image for which and both exceed 60% and equals 1 when the labels and are the same while it equals 0 otherwise. MCE should be less than 0.3 for the labelling process to be considered successful.

Unknown Region Clustering Successful detection of new concepts will be evaluated through the Least Set Coherence Number (LSCN) measure.

Let's denote by R_i^j a region in the j-th annotated image assigned (manually) the i-th label. We denote by SR_i the set of all region instances having the i-th label. If we consider the k-th cluster of an unsupervised clustering process as a set denoted with CR_k then the Maximum Set Coherence (MSC) measure is defined as follows:

$$MSC_i = \max_k \frac{\|CR_k + SR_i|}{\|CR_k\| + \|SR_i\|} \tag{9}$$

where $\|\|$ denotes the number of elements of the corresponding set. The Least Set Coherence Number (LSCN) measure is defined as follows:

$$LSCN = 100\frac{1}{NL}\sum_{i=1} NL(MSC_i > 0.5) \tag{10}$$

where NL is the total number of available labels (known MLCs). MSC_i should be higher than 0.5 for at least 70% of the total number of known MLCs for the process of identifying new concepts to be considered successful. That is, LSCN should be greater than 70%.

3.8 Use case: pole vault

In this section, the Methodology for Semantics Extraction from Still Images is explained through the example of the High Level Concept `pole_vault`.

3.8.1 Inputs

Annotated content An example of an annotated image that can be used for training purposes is shown in Figure 3.8.1. For each region an `xml` file similar to that shown in Figure 17 is expected.

Domain Model The HLC `pole_vault` includes the following MLCs:

- `Pole`
- `Athlete's_face`
- `Athlete's_leg`
- `Athlete's_hand`

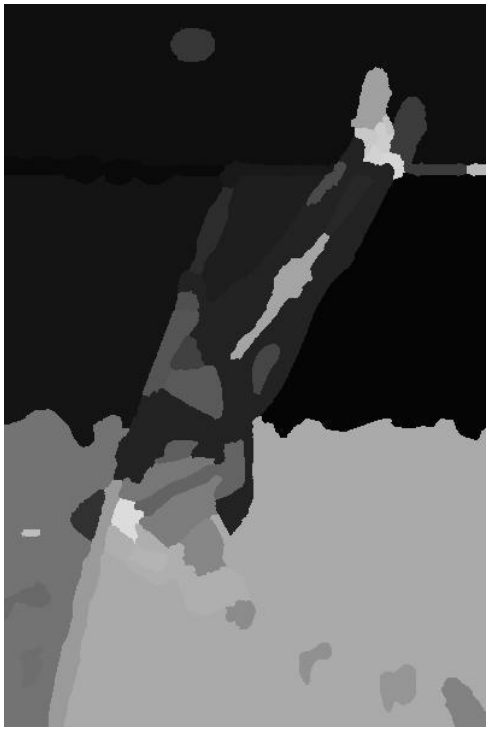


Figure 16: Annotated image to be used for training

```
<?xml version="1.0" ?>
<!--  Written on 28-Sep-2006 01:49:47 using the XML Toolbox  -->
- <root>
   <LocalDir>20822_W400XH600_annotated</LocalDir>
   <OrigImage>20822_W400XH600.jpg</OrigImage>
   <SegImage>20822_W400XH600_Mask.png</SegImage>
   <MLClabel>Athletes_Face</MLClabel>
   <RegionNumber>255</RegionNumber>
   <DescrFolder>20822_W400XH600_Athletes_Face_255_descriptors</DescrFolder>
  </root>
```

Figure 17: Annotation example for MLC `athlete_face`

- `Horizontal Bar`
- `Pillars` (Vertical bars)

Multimedia Descriptors Ontology Available low-level descriptors and features include:

- Shape, Colour and Texture descriptors defined in MPEG-7
- Shape Adaptive DCT coefficients

Multimedia Domain Model This will be a set of:

- Classifiers trained to identify MLCs based on low-level descriptor and feature values
- Dedicated algorithms for the detection of primitives, through which MLCs can be identified
- Spatial relations among MLCs

3.8.2 Output

The output of Image Analysis module will be

- An automatically annotated image similar to the one shown in Figure 3.8.1.
- An xml file similar to that shown in Figure 8 that describes the image analysis results
- For each MLC or not recognised region an `xml` file similar to the one shown in Figure 17.

3.8.3 MLC Identification

Region-based approach For each MLC a list of properties consisting of low-level descriptors and features values will be created:

- `Pole` → {Shape descriptors, SA-DCT coefficients}
- `Athlete's_face` → {Colour descriptors, shape descriptors, texture descriptors, SA-DCT coefficients}
- `Athlete's_leg` → {Colour descriptors, shape descriptors, SA-DCT coefficients}
- `Athlete's_hand` → {Colour descriptors, shape descriptors, SA-DCT coefficients}
- `Horizontal Bar` → {Shape descriptors, SA-DCT coefficients}
- `Pillars` → {Shape descriptors, SA-DCT coefficients}
- `Mats`→ { Colour descriptors, SA-DCT coefficients}

Holistic approach For each MLC a list of primitives and /or algorithms will be created:

- `Pole` → {Line, arc, Generalised Hough Transform} (see Figure 18)
- `Athlete's_face` → {Face Detection}
- `Athlete's_leg` → {Skin Detection, Object recognition}
- `Athlete's_hand` → {Skin Detection, Object recognition}
- `Horizontal Bar` → {Line, Hough transform} (see Figure 18)
- `Pillars` → {Parallel vertical lines, Hough transform}

Figure 18: `Pole` and `bar` detection using the Hough transform.

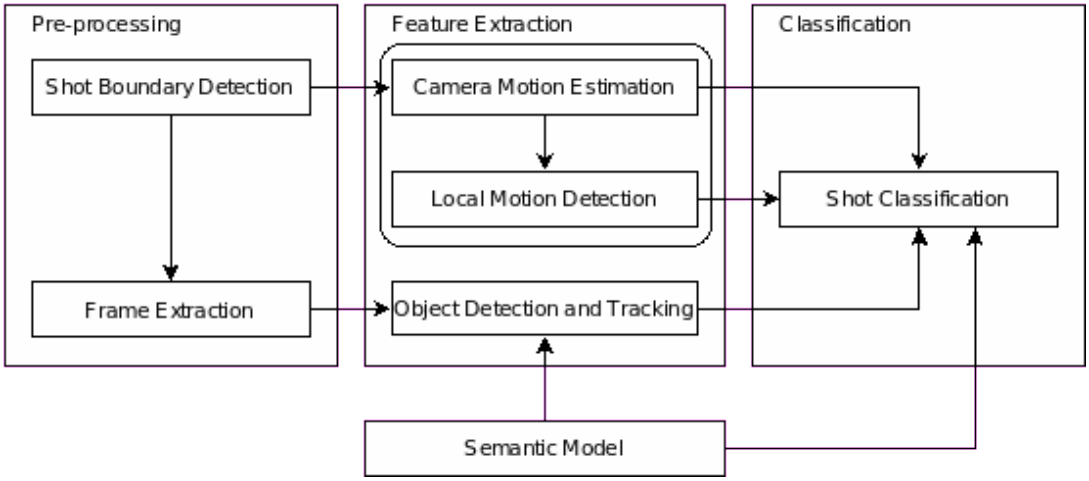


Figure 19: Overview of video-based information extraction. The dotted rectangle indicates that the statistical analysis uses relative motion patterns that may integrate camera and local motion.

4 Video

4.1 Aim of video-based information extraction

The aim of semantics extraction from video documents is to provide information about the existence of MLC instances (MLCIs) that are distributed over time, e.g. phases of an athletics event like the approach of a pole vaulter to the jump-off point. This scene information can be determined through statistical analysis of global and local motion patterns or through model-based analysis of object trajectories, i.e. extracted spatio-temporal relations for MLCs like javelins.

4.2 Overview of the methodology

The methodology for video-based information extraction follows the common approach used in pattern classification: pre-processing, feature extraction and classification. The components for each of these phases and their interactions are sketched in figure 19. In addition, the semantic model is used to retrieve the labels for MLC identification.

4.2.1 Inputs to the video-based information extraction

The required inputs to video-based information extraction concern three separate areas:

- The actual video data to be classified
- Annotations for training purposes
- Inputs from other BOEMIE components like the semantic model

Video Data The video data fed into the semantics extraction has to follow some basic requirements, at least in the current phase of the project, as defined in the "Content Collection and Annotation Strategy" document (D5.1). In the course of the project, some of these requirements should be relaxed because it is desirable to allow a wide variety of input formats and quality levels. Development will keep this in mind even when working with high quality material.

Semantics extraction accepts complete video documents, i.e. no pre-processing like segmentation or rescaling/recoding needs to be done externally.

Annotation Video-based semantics extractions involves training processes that need ground-truth information about the set of video documents from content collection. This information comes from the annotation effort and is also needed for evaluation of the methodology. Ideally, the following information should be available:

Shot boundaries for improvement or tuning to the sports domain of existing shot boundary detection algorithms

MLC annotations per shot (e.g. event phases) for the training of shot classifiers based on statistical analysis of motion patterns and for information about temporal relations between MLCs

HLC annotations per sequence of shots (e.g. complete events) that make the connection between MLCIs and their temporal relations on the one hand and the HLC on the other hand

Trajectories for specific objects for trajectory matching

At a minimum, in order to support an initial implementation of the methodology, the event phase annotations are needed for the statistical approach. The object trajectory annotations required for the trajectory matching approach are very time-consuming and should not be included in the first round of annotations.

Inputs from Other BOEMIE Components Video-based information extraction retrieves MLCs from the semantic model as labels for shot classification. It can also use temporal relations specified in the semantic model to check validity of detection results. For example, if an event phase is detected that does not follow the usual sequence of event phases, it may be a false detection.

4.2.2 Outputs of the video-based information extraction

Video-based information extraction provides extracted MLC instances and their spatio-temporal relations. The support to ontology enrichment is described below in section 4.5.

4.3 Description of methodology

4.3.1 Pre-processing

The pre-processing part of the methodology prepares the incoming video data for feature extraction and classification. This does not mean that this part is trivial; shot boundary detection and—to a lesser extent—frame extraction (when talking about key frames) are both subject to ongoing research.

Shot boundary detection The first task in video-based semantics extraction is the segmentation of the video input into shots. Following the terminology in [DelBi99], shots are "the set of frames between a camera turn-on and a camera turn-off or some other editing effect. Shots have perceptual continuity; they are the elementary segments of video (the syntactic atoms) and are both meaningful and humanly perceivable." ([DelBi99], p. 10).

In the sports domain, it may be necessary to split shots into even smaller segments: using the pole vault example, a broadcaster might show a pole vault attempt in one long shot, following the athlete with the camera. In that case, the different phases of the attempt cannot be distinguished on a shot-by-shot basis. Separating the long shot into phases would then involve some semantics, which the shot boundary detection will probably not be able to deal with. A different component might have to be added in a later step of the methodology.

Frame extraction Once the shot boundaries in the video have been detected, one way of continuing is to extract frames from each shot and get information about objects present in the shot that can be tracked in time across the shot to get spatio-temporal information. For this purpose, the first frame should be extracted. Another possibility would be to extract key frames and store them in the semantic model for future processing or for reference, e.g. when presenting video documents in a query result.

4.3.2 Feature extraction

The feature extraction part of the methodology reduces the pre-processed video data to a representation that subsequent shot classification can use. This means that feature extraction should here be seen in the context of that classification task; the individual components camera motion estimation, local motion detection and object detection and tracking involve their own classification tasks that deal with "lower-level" features.

Camera motion estimation The second way of continuing after shot boundary detection is to analyse the complete shot in terms of camera motion like booming, dollying or tilting. Some information can already be determined in the compressed domain as shown by Hesseler and Eickeler [Hess06] of particular interest here are the motion vectors generated by most video compression algorithms. Exploiting this information is desirable for efficiency reason, but it is quite possible that it is too unreliable for robust camera motion estimation. More reliable motion information may come from optical flow analysis, which however introduces a computational penalty.

Once the global camera motion has been established, it feeds on the one hand directly into shot classification; on the other hand, it forms the basis for subsequent local motion detection.

Local motion detection Local motion detection tries to find moving regions in a shot that do not follow the global camera motion. Again, a motion vector field—either gathered from the compressed domain or from optical flow analysis—can be the domain on which to work, this time taking into account, i.e. subtracting, the contribution of the global motion. Another possibility, as shown by Yi et al. [Yi03], is to undo the effect of global motion on two consecutive frames, find blobs in the difference image and evolve a trajectory for these blobs along the following frames. An approach like this may also lead to the forming of new MLCs, as described in section 4.5.

The results of local motion detection are made available to shot classification.

Object detection and tracking Object detection and tracking can track segments, regions or objects detected in one frame through subsequent frames.

The object trajectories found by this component are inputs to shot classification.

4.3.3 Classification

The classification part of the methodology connects the extracted features like object trajectories to the MLCs in the semantic model. It is therefore the actual semantics-forming step whose output is made available to multi-modality fusion.

Shot classification Shot classification assigns MLCs to each shot to form MLC instances. Of course, this is far more complex than it sounds: It is the most challenging aspect of the overall methodology. The current idea is to classify shots based on two complementary approaches:

- Statistical analysis of motion patterns (global and local)
- Matching of object trajectories

Statistical analysis of motion patterns The first approach tries to determine characteristic patterns of global and local motion in semantically similar shots based on statistical machine learning. It is useful because it requires little a priori knowledge, in the form of a set of annotated shots. The training algorithm will probably be inspired by the Viola-Jones approach to object detection [Viol01]: An overcomplete set of weak classifiers based on simple filters is searched and combined into a strong classifier through a boosting procedure. The intended filters work on motion vector fields and extract relative motion of image regions over time. This idea has recently been explored by Ke et al. [Ke05] , although they do not make use of boosting because of the large number of filters involved and the training time needed to explore the weak classifier set. However, it is possible to massively parallelise this selection procedure, and better filter selection can be expected through the use of boosting.

Matching of object trajectories The second approach tries to match extracted object trajectories to modelled object trajectories. An idea is to model 3D constellations of objects like an athlete's head, hands and feet during a pole vault over time and to match projections of these model constellations to extracted 2D constellations.

4.4 Semantic model usage

Video-based information extraction may use the semantic model to add robustness in terms of error resilience. If the semantic model indicates constraints on a particular MLC or a relation between MLCs, the extraction might determine that a candidate detection breaks the constraints. This may add a confidence penalty or lead to complete rejection of the candidate.

It is not yet clear if the video-based information extraction should take advantage of a reasoning process to combine MLCIs and their spatio-temporal relations to form HLCIs by itself or if this is left solely to the fusion process.

4.5 Support to ontology enrichment

The shot classification that video-based information extraction performs will no doubt lead to a number of unclassified shots. From time to time, an attempt can be made to determine commonalities between these shots through unsupervised learning. Hinting at the existence of a large enough, as yet unlabeled class of shots may lead to a new MLC in the ontology.

Another worthwhile attempt might be to determine recurring local motion patterns of as yet unidentified objects. These could either be concepts that are known, but presented in a new way in the video material, or new concepts that can be introduced in the ontology.

4.6 Confidence handling

Confidence values will be attached to the results from video-based semantics extraction to enable other BOEMIE components to judge the result quality. As has been described, several steps are involved in reaching the final result, each having some uncertainty in the respective in-between result. The final confidence value will therefore be some linear combination of individual confidence value. The corresponding weights and normalisation needed will be determined during development of the extraction system. It will be necessary to make sure that the confidence values match those of other modalities so that the results from video-based semantics extraction are not completely over- or underrated by the fusion or reasoning process.

4.7 Evaluation framework

This section describes the strategy for performance evaluation of video-based information extraction, calculated on a well-defined content set through established quantitative measures.

4.7.1 Strategy

The key result to evaluate is the final outcome from shot classification. All the steps in video-based information extraction contribute to this result, so there is a need to evaluate each individual step. The following list sketches evaluation methods for the components; the measures are described in the third subsection:

- Shot boundary detection can be evaluated through precision/recall or in the context of the boundary correspondence based error measure.
- Key frame extraction evaluation is an open research problem; so far, subjective evaluation has been employed.
- Camera motion estimation can be evaluated through precision/recall if it is classified into classes like pan, tilt or zoom.

- Local motion detection evaluation is unclear at the moment. Possibilities include defining classes like upward motion and using precision/recall, or extracting region trajectories and evaluating them through EDR.
- Object detection and tracking can—for the purpose of extracting object trajectories—be evaluated through similarity measures like EDR.
- Shot classification can be evaluated through the boundary correspondence based error measure or through precision/recall.

4.7.2 Content

For evaluation purposes, part of the manually annotated content will be retained as a test set instead of being used in training. The set sizes will depend on the amount of content that can be collected and annotated; it is expected that at least 50 shots per MLC can be reserved for testing.

4.7.3 Quantitative measures

Boundary correspondence based error measure The boundary correspondence based error measure was proposed by Eickeler and Rigoll [Eick00] in order to overcome several shortcomings of earlier error measures used in video indexing. First, shot boundary detection performance is evaluated in terms of missing and inserted boundaries as well as boundary displacement; this information is then used for evaluation of shot classification performance.

Precision/Recall These are the standard measures from information retrieval, measuring the ability of a system to detect all relevant information (recall), e.g. shots belonging to a specific class, and to detect only relevant information (precision). They can also be combined like in the F-measure.

Edit distance on real sequence (EDR) EDR is an adaption of edit distance as known from text analysis to object trajectories, introduced by Chen et al. [Chen05]. It measures the similarity of two trajectories to each other, in the face of noise, gaps and local time shifting. The minimum EDR can be used for retrieval of object trajectories.

4.8 Use case: pole vault

4.8.1 Inputs

Annotated content Figure 20 shows an example of an annotated pole vault attempt in a video. The individual event phases are annotated as well as the fact that this attempt failed.

Figure 21 shows the `xml` file that stores the video annotations. It records start and end timestamps for each phase.

Domain Model The HLC `Pole_Vault` includes the following MLCs:

- `Pole_Vaulter_Approach`
- `Pole_Vaulter_Plant_And_Take_Off`
- `Pole_Vaulter_Swing_And_Row`
- `Pole_Vaulter_Rockback`
- `Pole_Vaulter_Turn`
- `Pole_Vaulter_Fly_Away`
- `Pole_Vaulter_Failed_Landing`

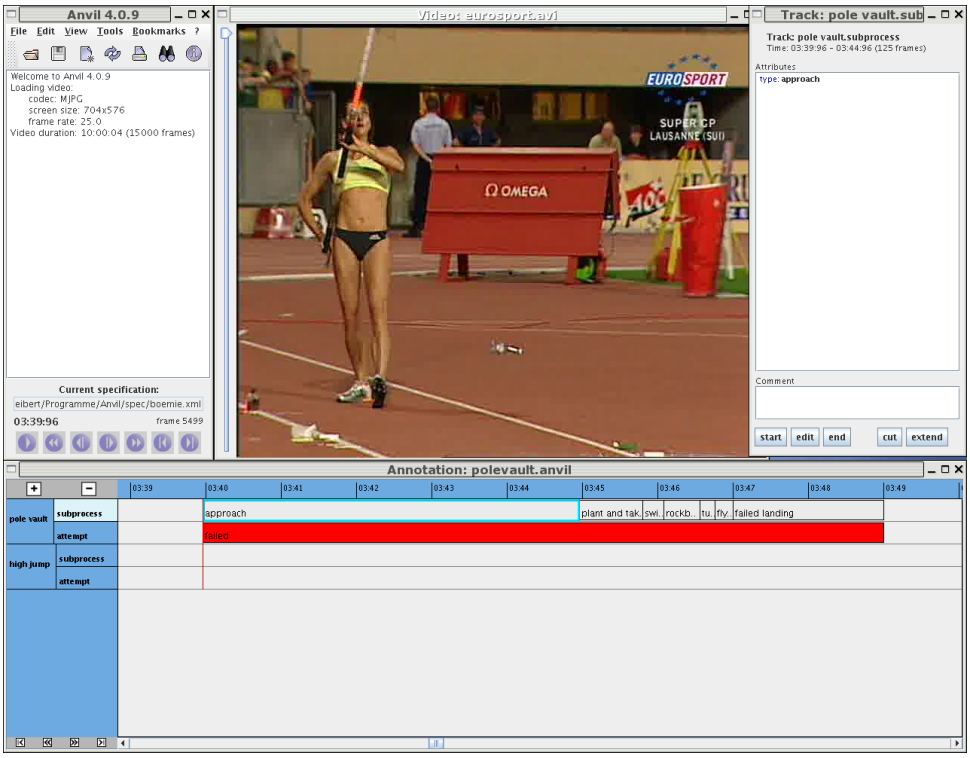


Figure 20: Annotated video

It also includes the temporal relations between the MLCs and between the HLC and the MLCs, of the types:

1. `Meets/met by`
2. `Precedes/preceded by`
3. `Starts/started by`
4. `Finishes/finished by`
5. `During/contains`

4.8.2 Outputs

For a given video, video-based semantics extraction will output the automatically detected MLC annotations in a format similar to that shown in figure 21. From this, it will calculate the temporal relations between MLCIs as an additional output.

4.8.3 MLC identification

Each MLC will be identified through shot classification according to

- characteristic camera motion
- characteristic local motion

  A combination of camera motion and local motion in the form of relative motion of regions against each other will also be considered.
- object trajectories (e.g. that of the pole)

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<annotation>
<head>
<specification src="/home/cseibert/Programme/Anvil/spec/boemie.xml" />
<video src="/tmp/cseibert/eurosport.avi" />
<info key="coder" type="String">
cseibert
</info>
</head>
<body>
<track name="pole vault.subprocess" type="primary">
<el index="0" start="219.96" end="224.96">
<attribute name="type">approach</attribute>
</el>
<el index="1" start="224.96" end="225.8">
<attribute name="type">plant and take off</attribute>
</el>
<el index="2" start="225.8" end="226.08">
<attribute name="type">swing and row</attribute>
</el>
<el index="3" start="226.08" end="226.56">
<attribute name="type">rockback</attribute>
</el>
<el index="4" start="226.56" end="226.76">
<attribute name="type">turn</attribute>
</el>
<el index="5" start="226.76" end="227">
<attribute name="type">fly away</attribute>
</el>
<el index="6" start="227" end="229">
<attribute name="type">failed landing</attribute>
</el>
</track>
<track name="pole vault.attempt" type="span" ref="pole vault.subprocess">
<el index="0" start="0" end="6">
<attribute name="type">failed</attribute>
</el>
</track>
</body>
</annotation>
```

Figure 21: Video annotation example in `xml`

Figure 22: An example of artificial text in a video frame

5 Video OCR

5.1 Introduction

Nowadays the number and the size of digital video libraries are increasing rapidly. This fact leads to an urgent need for fast and effective algorithms for information retrieval, from multimedia content, for applications in video indexing, editing or even video compression. Text in video and images proves to be a source of high-level semantics closely related to the concept of the video. Two types of text are found in images and video frames, artificial text and scene text. Artificial text is artificially superimposed on images or video frames at the time of editing. Scene text naturally occurs in the field of view of the camera during video capture. Scene text occurs on signs, banners etc. Artificial text can provide us with powerful information for television captured video indexing since this kind of text is added, in order to describe the content of the video or give additional information related to it. Artificial text presents some features and follows some characteristics in order to be readable from humans, like high intensity vertical edge strokes, colour homogeneity, contrast between text and background, horizontal alignment, various geometrical constraints etc. These are the features and constraints usually used by the text detection systems for distinguishing text areas from non-text areas. On the other hand, there are many challenges that have to be faced like, text embedded in complex backgrounds, with unknown colour, size, font or low resolution text. Scene text's structure is generally more random. It presents variability of size, font, colour, lighting conditions, orientation, transformations, style, alignment, even within words, occlusions, complex movement etc. Many of the features used for the detection of artificial text, like colour homogeneity, horizontal alignment or even high edge intensity, do not apply for artificial text. The procedure of retrieving text from video is usually called Video OCR and consists of 3 basic stages: text detection, text segmentation and recognition.

5.2 Overview of VOCR methodology

Text detection includes spatial and temporal (tracking) detection. The aim here is to give an effective and computationally efficient algorithm for the spatial and temporal detection of artificial text in still video frames with a high recall rate. The algorithm must produce one bounding box for every text line of the frame. Scene text recognition will also be examined, perhaps in individual algorithms since experience until now shows that no algorithm can work efficiently with all kinds of text. An attempt to incorporate scene text detection in an algorithm will probably cause artificial text detection rates fall.

The proposed methodology for detecting artificial text in video frames is based on edge information. First, an edge map is created using Canny edge detector. Then, morphological operations are used in order to connect the vertical edges and discard false alarms. A connected component analysis is performed to the edge map in order to determine a bounding box for every candidate text area. Finally, horizontal and vertical projections are calculated on the edge map of every box and a threshold is applied,

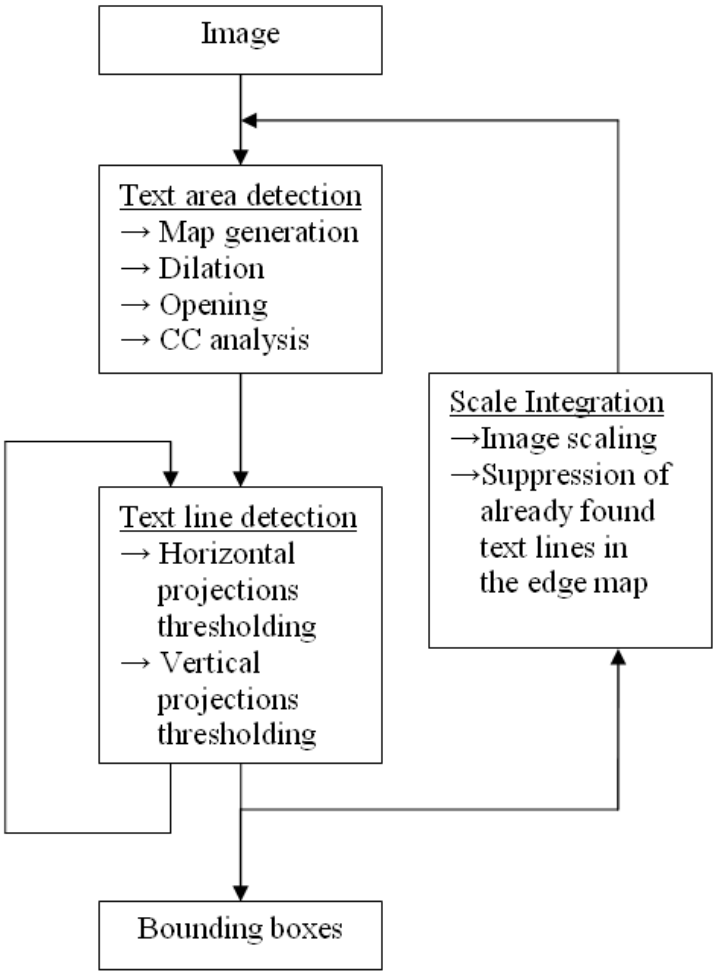


Figure 23: Flowchart of the proposed algorithm.

refining the result and splitting text areas in text lines. The whole algorithm is applied in different scales so fonts with different sizes are detected. Some experimental results show that the method is highly effective and efficient for artificial text detection. A detailed description of the algorithm can be found in section 5.3.

A fast tracking algorithm will also be developed. The algorithm will aim to track correctly and fast enough, simple linear motion of artificial text, since this will be the majority of text occurrences in the specific application, and discard the wrongly tracked text, that would worsen the results, using confidence values. The tracking stage will result in quality enhancement through multi-frame integration and speed-up of the system since the slow stage of detection will be done periodically. A detailed description of the algorithm can be found in section 5.4.

For the binarization stage, adaptive thresholding techniques will be used combined with connected component analysis. We will be based on binarization techniques for degraded documents [Gato06] as well as on text detection techniques for indoor/outdoor scene images [Gato05], both proposed by NCSR partner. Resolution enhancement will also be applied before the segmented and binarized images are fed to a commercial OCR machine.

For the evaluation of the whole procedure, 2 different methods will be used and compared: pixel-by-pixel comparing and evaluation through the OCR result.

5.3 Text detection methodology

The proposed text detection methodology exploits the fact that text lines produce strong vertical edges horizontally aligned and follow specific shape restrictions. Using edges as the basic feature of our system gives us the opportunity to detect characters with different fonts and colours since every character present strong edges, despite its font or colour, in order to be readable. Besides, artificial text is supposed to be readable. An example of artificial text is given in Figure 22. An example of artificial text in a video frame. The methodology consists of three basic stages: Text area detection, Text line detection and Scale Integration (Figure 23).

5.3.1 Text area detection

Step 1: Map generation First, the edge map of the image is produced. There are many masks and methods in the literature for computing the edge map of an image. For our algorithm we use Canny [Cann86] edge detector applied in greyscale images. Canny uses Sobel masks in order to find the edge magnitude of the image, in gray scale, and then uses non-Maxima suppression and hysteresis thresholding. With these two post-processing operations Canny edge detector manage to remove non-maxima pixels preserving the connectivity of the contours. Ideally the created edge map is a binarized image with the pixels of contours set to one (white) and the background equal to zero (black) (Figure 24).

Figure 24: Edge map of the frame in figure 22

Step 2: Dilation After computing the Canny edge map, a dilation by an element 5x21 is performed to connect the character contours of every text line (Figure 25). Experiments showed that a cross-shaped element has better results.

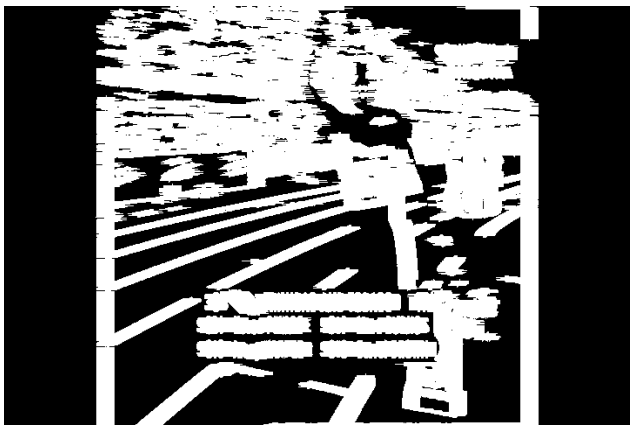


Figure 25: The edge map after the dilation process.

Step 3: Opening Then a morphological opening is used, removing the noise and smoothing the shape of the candidate text areas (Figure 26). The element used here is also cross-shaped with size 11x45. Every component created by the previous dilation with height less than 11 or width less than 45 is suppressed. This means that every edge which could not connect to a component larger than the element of the dilation will be lost. Unfortunately this operation may suppress the edges of text lines with height less than 12 pixels. However this is not so devastating since character of this size are either way not recognised in the final stage of the Video OCR system. Now every component represents a candidate text area.

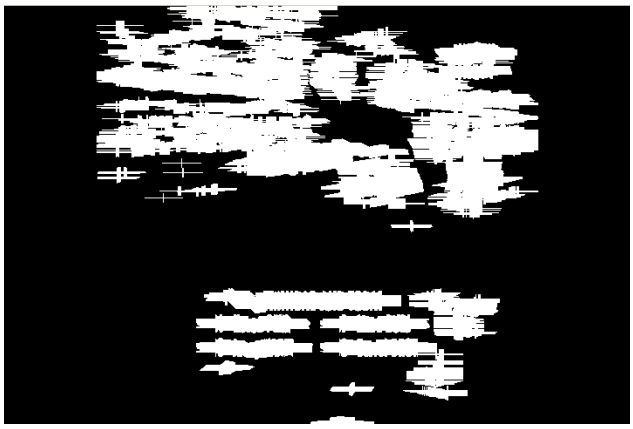


Figure 26: The edge map after the opening operation.

Step 4: Connected-component analysis Finally a connected component analysis gives the position of every component so the initial bounding boxes have been computed (Figure 27).

5.3.2 Text line detection using projections

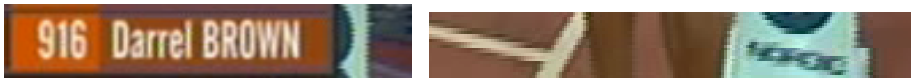
The previous stage has a high detection rate (recall) but relatively low precision. This means that most of the text lines are included in the initial text boxes. However some text boxes may include more than one text line as well as noise. This noise usually comes from objects with high intensity edges that connect to the text lines during the dilation process. This low precision also comes from detected

Figure 27: Initial bounding boxes.

bounding boxes which do not contain text but objects with high vertical edge density. To increase the precision and reject the false alarms we use a method based on horizontal and vertical projections.

Step 1 : Horizontal edge projection Firstly, the horizontal edge projection of every box is computed and lines with projection values below a threshold are discarded. In this way boxes with more than one text line are divided and some lines with noise are also discarded. Besides, boxes which do not contain text are usually split in a number of boxes with very small height and discarded by a next stage due to geometrical constraints.

(a) initial box - initial box's horizontal projection



(b) produced boxes

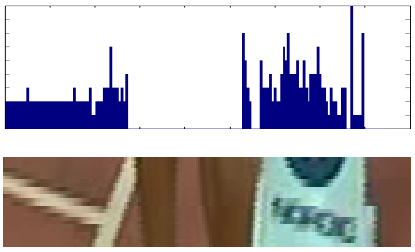
Figure 28: Example of box splitting through horizontal edge projection thresholding

Step 2 : Vertical edge projection Then, a similar procedure with vertical projection follows. This method would actually break every text line in words or even in characters. However this is not an intention of the algorithm so finally the vertically divided parts are reconnected, if the distance between them is less than a threshold which depends on the height of the candidate text line. In this way a bounding box will be split only if the distance between two words is very large which means that actually belong to different text lines or if a part of the candidate text line contain only noise.

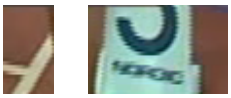
The whole procedure with horizontal and vertical projections is repeated three times in order to segment even the most complicated text areas (Figure 30).

5.3.3 Scale Integration

Using edge features in order to detect text gives to the method independence from text colour and different fonts. However this method clearly depends on the size of the characters. The size of the elements for the morphological operations and the geometrical constraints give to the algorithm the ability to detect text in a specific range of character sizes. With the values described above the algorithm is capable of



(a) initial box - initial box's vertical projection



(b) produced boxes

Figure 29: Example of box splitting through vertical edge projection thresholding.

Figure 30: Refined result

detecting accurately characters with height from 12 to 44 pixels. To overcome this problem we will adopt a multi-scale approach. The methodology described above is applied to the image in different scales and finally the results are fused to the initial scale. This fusion might be quite difficult if we consider that the same text might be detected in different scales so bounding boxes will overlap. To avoid that, the algorithm suppresses the edges of the already recognised characters in a scale before the edge map is passed to the next scale. For every scale, except for the initial scale a blurring filter is applied so the edges of the background become weaker compared to the edges of the text which still remain strong. This filter is not applied to the first scale because it would destroy the contrast of the small characters that already suffer the blurring caused by video compression. Taking into account that artificial text in videos usually does not contain very large characters and from the experience of related experiments we chose to use two scales for this approach: the initial and the one multiplied by a factor of 0.6. In this way the system can detect characters with height up to 80 pixels which was considered to be satisfying.

5.3.4 Evaluation method

Designing evaluation methods for text detection is an aspect that has not be studied extensively, yet. Very few related works have been published, moreover this works propose evaluation strategies with very complicated implementations or demand great effort for the generation of the ground truth. Many of the researchers use their own evaluation tool to test the success of their algorithm. This fact leads to the inability to compare the performance of the different algorithms which indubitably consists a barrier to the evolution of the area.

In our methodology, we will use as evaluation indicators the recall and precision rates on a pixel base. For the computation of the rates we need to calculate the number of the pixels for the ground truth bounding boxes, for the bounding boxes of the detection method and for their intersection. As final measure we use the F-measure, which is the harmonic mean of recall and precision. However this

method proved to have several drawbacks.

The first is that there is not an optimal way to draw the ground truth bounding boxes. This means that two boxes may be accurate enough for bounding a text line although they may not include exactly the same pixels. In other words, the result of the detection method may be correct although the evaluation method gives a percentage less than 100%. To overcome this problem one can segment the text pixels from the background pixels and then demand the presence of text pixels in the output bounding box. However this would make the detection evaluation depend on the performance of text segmentation which is something surely not desirable. In this work we follow a more simple strategy to solve this problem. The ground truth bounding boxes are drawn in a way that the margins between the text pixels and the edge of the box are equal for all text lines. Moreover, as last stage of the detection algorithm all bounding boxes grow by 8 pixels in width and height, providing a satisfying approximation of the ground truth.

Another important drawback is fact that this method actually measures the percentage of detected pixels. However the goal of the detection algorithm is not to detect maximal amount of pixels but the maximal number of characters. In other words, a text line must have influence to the final evaluation measure proportional to the number of containing characters and not to the number of its pixels. Unfortunately, the number of characters in a box cannot be defined by the algorithm but it can be approximated by the ratio width/height of the bounding box, if we assume that this ratio is invariable for every character and the spaces between different words in a text line is proportional to its height. In this way, every pixel counts for $\frac{1}{h^2}$ when calculating the recall and precision rates, where h is the height of the bounding box in which the pixel belongs.

$$Recall = \frac{\sum_{i=1}^{N} \frac{EGD_i}{hg_i^2}}{\sum_{i=1}^{N} \frac{EG_i}{hg_i^2}} \tag{11}$$

$$Precision = \frac{\sum_{i=1}^{M} \frac{EDG_i}{hd_i^2}}{\sum_{i=1}^{M} \frac{ED_i}{hd_i^2}} \tag{12}$$

where hg_i is the height of the i-th ground truth bounding box, EG_i is its number of pixels, EGD_i is the number of pixels of the intersection that belong to i-th ground truth bounding box, hd_i is the height of the i-th detection bounding box, ED_i is its number of pixels and EDG_i is the number of pixels of the intersection that belong to i-th detection bounding box.

A screenshot from a pilot evaluation application that implements the detection-evaluation algorithms is given in figure 31. Examples of the different stages can be found in figures 32-35.

5.4 Text tracking Methodology

A fast text tracking algorithm will also be developed. For the specific application of athletics video artificial text detection we can consider the text as either stationary or linearly moving. Actually, in most cases artificial text in this kind of video is stationary, however even in this kind of videos, text can slightly move by some pixels around its original position from frame to frame. Consequently, text tracking has to be done in order to enhance the text quality through multi-frame integration and speed-up the system since the slow stage of detection will be done periodically.

The text detection stage will ideally result in a set of bounding boxes for all the text occurrences in a frame. These boxes are going to be tracked from frame to frame till the next detection. A text occurrence in order to be readable has to stay in the video for at least 2 seconds. This fact results in several tens of frames for every text, so even with periodically text detection no text will be lost.

Motion is initially estimated by means of block-matching, since block-matching is suitable for rigid objects and characters are assumed to be rigid, changing neither their shape, orientation nor colour. As matching criterion the minimum mean absolute difference criterion is going to be used.

The mean absolute difference (MAD) is defined as:

$$MAD(d_1, d_2) = \frac{1}{|R|} \sum_{(x,y)\in\Re} |g(x.y) - g(x, +d_1, y + d_2)| \tag{13}$$

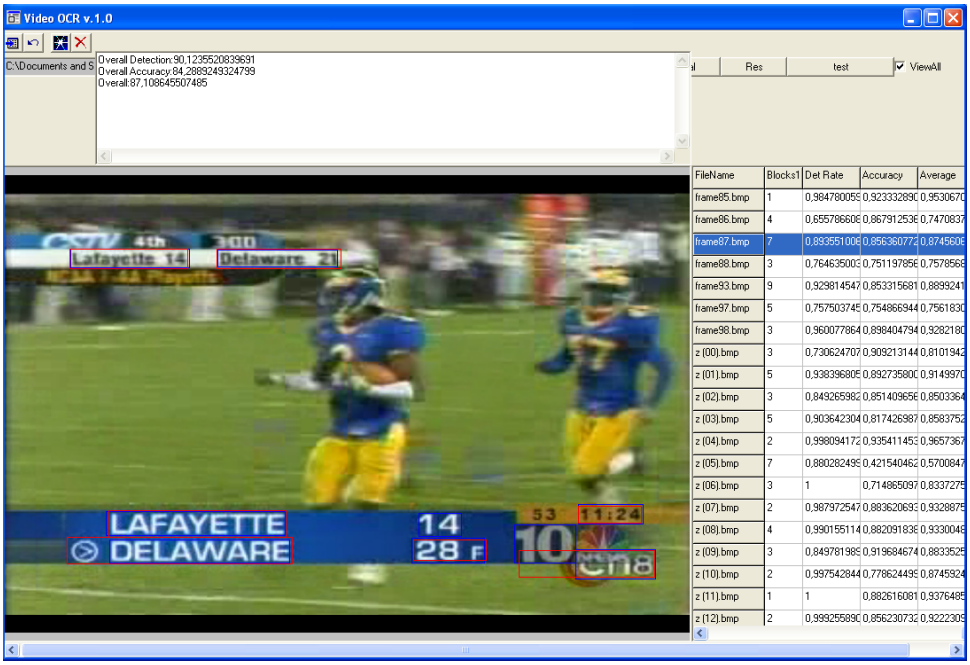
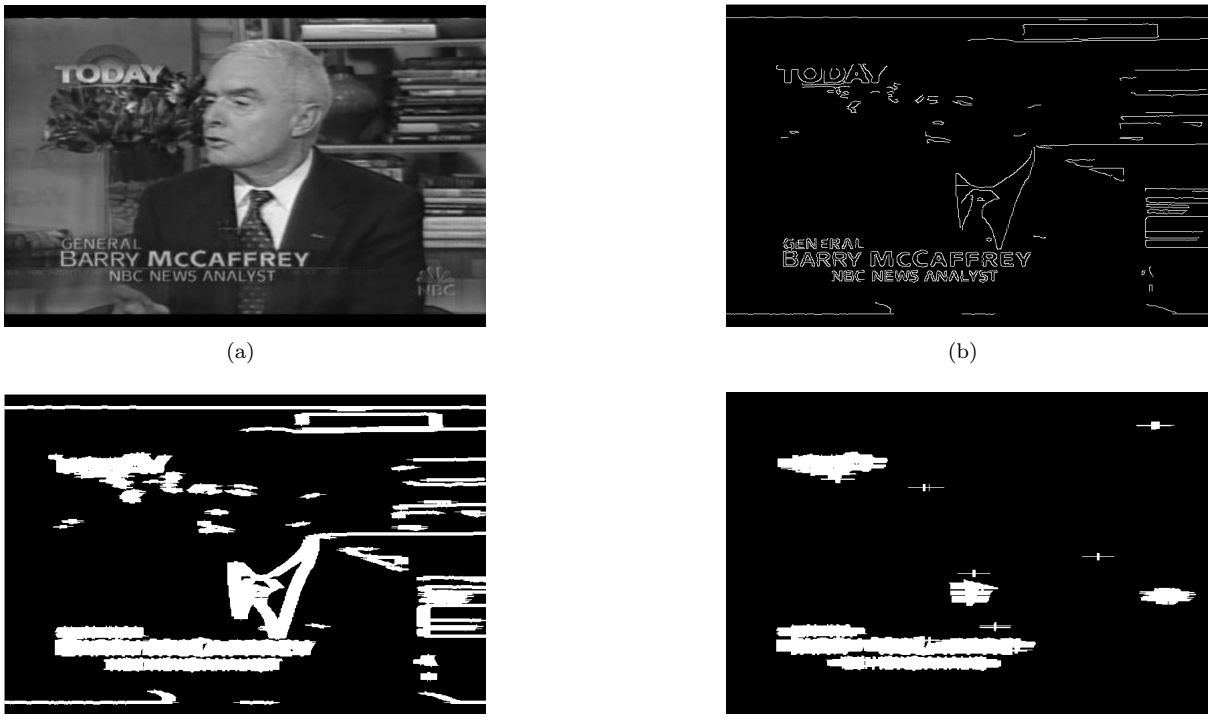


Figure 31: A screenshot from the detection-evaluation application.



(a) (b)

(c) (d)

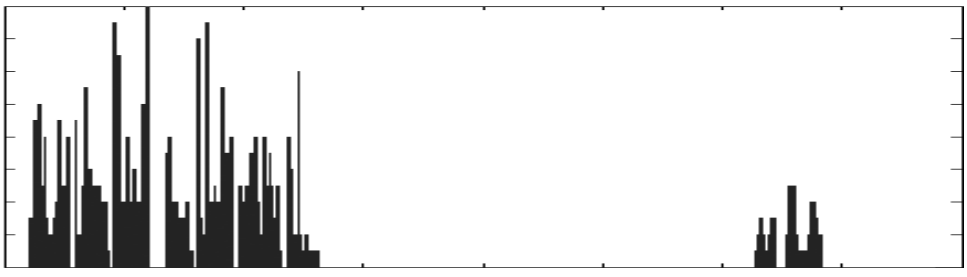
Figure 32: Text area detection example

(a) (b)

Figure 33: (a) The result of text area detection, (b) the result of text line detection.

(a) (b)

Figure 34: (a) edge map of a text area, (b) horizontal projection of the area.



(a)

(b)

Figure 35: (a) edge map of a text line, (b) vertical projection of the text line.

where R specifies the block for which the translation vector has to be calculated, $g(x, y)$ is the gray scale image included in every bounding box derived from the detection stage. The use of colour information in this stage will be examined. The displacement estimate $(d1, d2)$ for block R is given as the displacement where the MAD value is minimal. The search area is restricted to $|d1|, |d2| \leq search\_range$ and derived from the speed of fast-scrolling credit titles. For high scrolling speed, search_range has to be large so the procedure will be more time consuming.

This procedure will lead us in a series of images for every box initially tracked by the text detection stage. Many false alarms initially detected, will be discarded since they will be present in only a few frames. However some objects with high edge intensity may still remain detected as text and tracked from frame to frame. To eliminate these false alarms we can examine the motion trail of every tracked object. Assuming that all text objects will have static or linear movement we can discard all objects that move in non linear trails.

The motion trail of text in video is defined as the temporal ordering of the centre points of the tracked text blocks: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ where (X_k, Y_k) is the centre point coordinates for the k-th text block. To quantitatively measure the confidence of the motion trail, we use a straight line $y = ax + b$ to approximate the motion trail. For N points in motion trail, the parameter vector $p = (a, b)'$ can be estimated in the Least Mean Square sense as: $p = (X'X)^{-1}X'Y$ where

$$X = \begin{pmatrix} X_1 & 1 \\ X_2 & 1 \\ \cdots & \cdots \\ X_N & 1 \end{pmatrix} \text{ and } Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_N \end{pmatrix}$$

The Least Mean Square error will be used as a measure for approximation:

$$err = \frac{1}{N} \sum_{1}^{N} (Y_i - a \cdot X_i - b) \tag{14}$$

A large value of the above measure shows complicated non-linear movement so the tracked object is discarded. Moreover having the information of the movement direction of text in a video we can demand specific values for the parameter vector $p = (a, b)$. In this way we can discard even linear movements with gradient greater than a threshold, knowing that text moves mainly horizontally or vertically.

After tracking the text blocks, an averaging will be applied for every text object image series in order to enhance the image quality (see Fig. 36). This procedure will result in a more smoothed background so the text image will be much more successfully binarized and recognised.

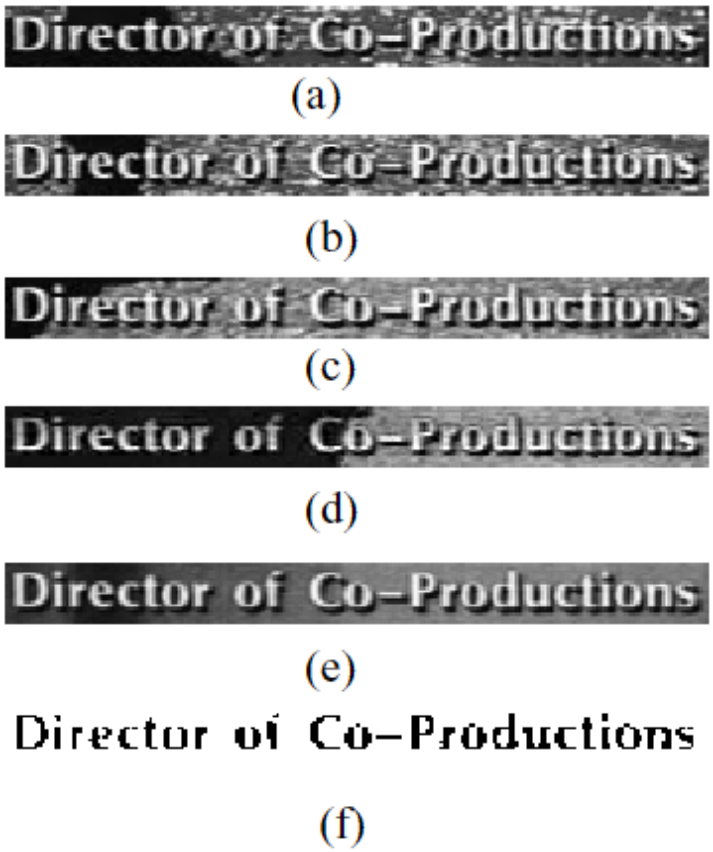


Figure 36: Enhancing image quality after text block tracking.

6 Audio/Speech

6.1 Aim of Audio/Speech-based information extraction

The aim of semantics extraction from audio/speech is, for any input audio, to provide the information about the existence of known audio events, events extracted using name recognition from speech data and non-speech audio events recognised from speech and non-speech data, their position with respect to other events, their intensity.

6.2 Overview of the Methodology

The schematic diagram of the methodology for semantic extraction from audio/speech is shown in Figure 37.

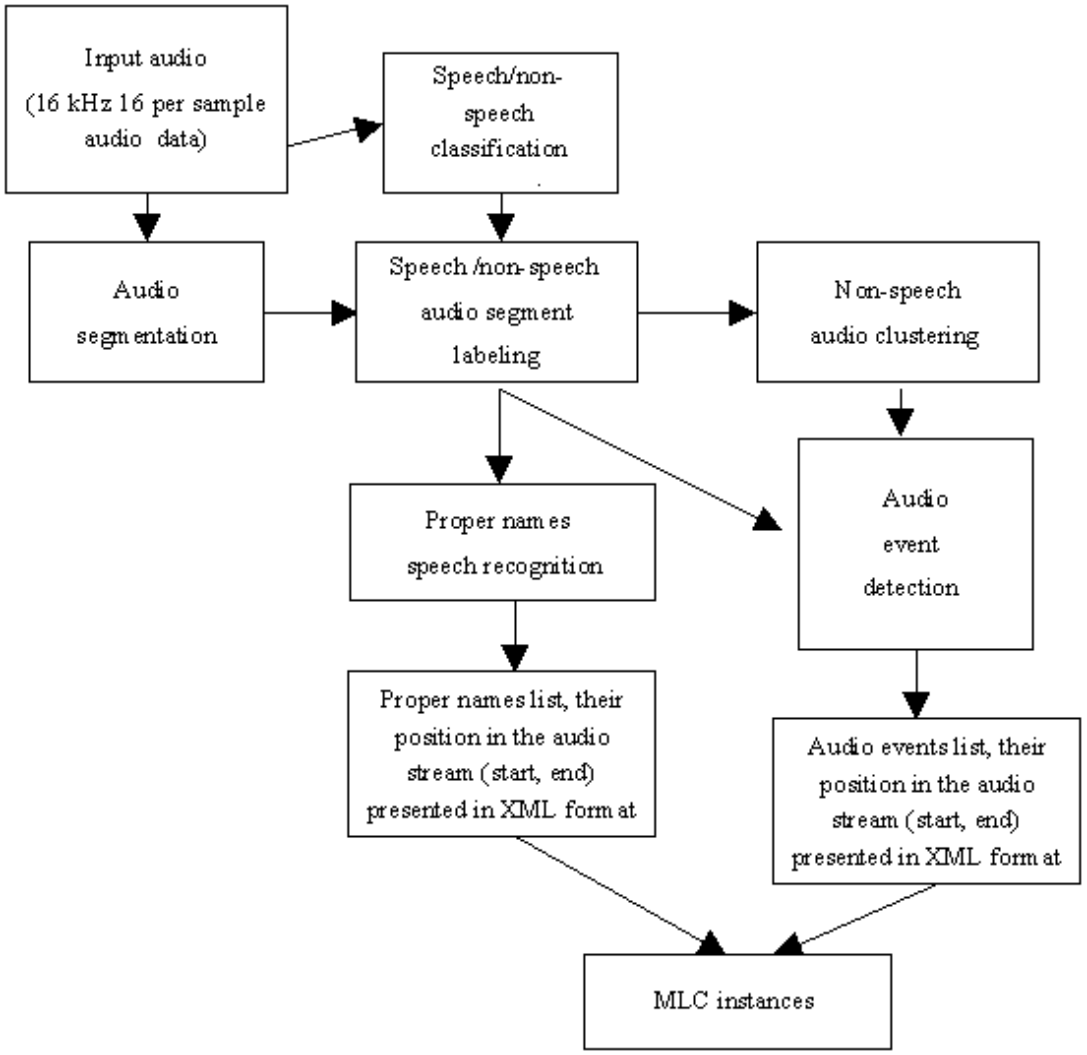


Figure 37: Schematic diagram of the proposed methodology for audio segmentation, classification and recognition

6.2.1 Inputs to the Audio/Speech-based Information Extraction

Input for audio/speech-based information extraction will be audio stream, 16 kHz audio signal with 16 bit for one audio sample.

6.2.2 Outputs of the Audio/Speech-based Information Extraction

Audio/speech based IE will include MLC' related with audio events.

6.3 Description of the Methodology

Audio processing will be used to extract audio events from both spoken and non-speech segments. These events will be used to find their corresponding context in the information fusion step where they will help to improve the performance of the information extraction.

6.3.1 Audio events segmentation

At the beginning the audio signal will be pre-processed and low-level features will be extracted. For audio segmentation usually mel-frequency cepstral coefficients, their delta and delta energy are used. Then the audio signal will be separated into segments using Bayesian Information Criterion (BIC). BIC can separate audio stream into statistically homogeneous regions. The reliable segmentation is possible when minimal size of segment is not less then 1 second. BIC is model selection algorithm. In general way BIC is defined as

$$BIC(M) = \log L(X, M) - \log(N) \tag{15}$$

where $L(X, M)$ denotes segment X likelihood given by the model M, N is the number of data points, $\#(M)$ is the number of free parameters in the model and λ is a tuning parameter. Only $\lambda = 1$ corresponds to the strict definition of BIC. In practice, the value of λ giving the best segmentation performance is different than 1 and depends on the features used. In order to estimate turn point between two segments and that have and frames respectively the Δ BIC value is computed as:

$$\Delta BIC = \frac{1}{2} n_i \log |\Sigma_i| + \frac{1}{2} n_j \log |\Sigma_j| - \frac{1}{2}(n_i + n_j) \log |\Sigma_{ij}| + \lambda P \tag{16}$$

where d is the dimension of the feature vector, Σ_{ij} is the covariance matrix of the data points from two segments c_i and c_j, Σ_i is the covariance matrix of the data points from the segment c_i, Σ_j is the covariance matrix of the data points from the segment c_j and penalty P is

$$P = \frac{1}{2}(d + \frac{1}{2} d(d+1)) \log(n_i + n_j) \tag{17}$$

ΔBIC is a distance between two Gaussian models which describe the same audio data with a hypothetic segment's turn. A negative value of ΔBIC indicates that the model that describes the data as a two Gaussian process fits better than the model that describes the data as a one-Gaussian process. In the segmentation process we follow in the algorithm described in [Chen98].

6.3.2 Speech/non-speech classification

Speech/non-speech classification is independent from audio data segmentation and can be applied in parallel. The speech/non-speech classifier classifies each one second audio segment as speech or non-speech. For speech/non-speech classification the approach based on unsupervised calculation silence ratio is used. This method includes unsupervised optimal self-segmentation of the audio segment into small, homogeneous sub-segments with the size corresponding to the syllabic rates. The homogeneity is defined on a base of the average amplitude and a zero-crossing calculated for each frame. A measure of the homogeneity is entropy. In described method a relative ratio between the average amplitudes of the neighboring homogeneous segments is calculated. For a speech signal this ratio has specific value. This value is defined as a threshold and is evaluated on a short pure speech signal. As a discriminative feature a percent of sub-segments having high relative amplitude ratio within 1 sec interval is used. In the process of the classification for each 1 sec assigns a label speech or non-speech [Biatov04].

6.3.3 Non-speech audio segments clustering

After audio segmentation, speech/non-speech separation and labelling resulting segment as a speech or non-speech, the non-speech segments are processed using clustering algorithm. In standard clustering approach using BIC, ΔBIC is used to make a pair-wise comparison between audio segments. If ΔBIC is positive and maximal the segments are merged, if ΔBIC is negative, they are not. This comparison continues until there are no more pairs of the segments with a positive ΔBIC (stop criterion).

Our algorithms [Biatov05], [Biatov06] extend the use of the BIC for audio clustering by constraining which pairs of segments can be considered for merging. We only permit two segments to be merged if they have the same patterns of similarity (global similarity) with all the other segments. The global similarity depends on how a segment is similar to all other segments taking part in the clustering. The global similarity features of each segment are presented as a vector. Each component of this vector is the local distance between this segment and each other segment. For local inter-segment similarity ΔBIC was used. Other local metric based on adapted cross likelihood ration can be also used. In the process of the clustering we are looking for a pair of segments having distance between global similarity vectors less than predefined threshold. Then the best pair with the minimal distance is selected to be merged. The process of the clustering stops when no more pairs satisfying these global similarity constraints and no more segments with a positive ΔBIC exist.

Audio event clustering unites similar audio events in one cluster. Using audio event clustering simplifies audio event recognition. Recognition for one member of the cluster gives result for all other members. Audio events clustering gives possibility to have more data for cluster model adaptation that yields more precise statistical model and leads to more reliable audio event recognition.

6.3.4 Audio events classification

Non-speech segments can provide the information that describes highlights for sport events and could be used with other modalities for sport events interpretation. Non-speech information includes applause, rhythmic applause (supporting sportsman before the start), silence, noise of stadium, signal of the start, sound of downed pole, bell, musical break, rhythmic music, that accompanying some kind of sport events and other groups of events. Non-speech audio events could be divided into two parts – background audio events and foreground audio events. Usually background audio events have much more longer duration. Examples of audio background events are noise of stadium, background music. Human speech can be also considered as background sound. Usually foreground events are shorter than background events. Examples of foreground events are applause, start signal, bell, sound of downed pole. Foreground audio events could appear on different audio backgrounds including the speech background and for recognition of foreground audio events requires normalisation of background sounds or separation of foreground sounds from of background sounds. Two techniques will be tested for separation background and foreground sounds.

First approach for audio events recognition is based on background sounds normalisation. We consider that background model is Gaussian Mixture Model (GMM). For each background data training data from available training data are extracted. Each background model is trained using Expectation-Maximisation (EM) algorithm. Foreground audio events are also described by statistical GMM and these models for each foreground event are obtained by adapting appropriate background model to foreground data using Maximum A Posteriori Adaptation (MAP). The result of recognition of foreground event is the ratio between likelihood of foreground and background models calculated for the same audio data.

The process of audio events classification includes two phases. Fist we recognise which background model fits the best for the region of interest in non-speech segments. The background model giving the highest likelihood on the tested non-speech segment is selected. Then the most probable foreground event is selected. Let X_i denote a tested segment, M_{bg}- background model that has the best match with the tested segment, $M_{fg}(M_{bg})$- foreground model that is obtained by adaptation of background model M_{ng}, $P(X_i|M_{fg})$ - average probability per fame of segment X_i under condition of foreground model $M_{fg}(M_{bg})$, $P(X_i|M_{bg})$ - average probability per frame of segment X_i under condition of background model M_{bg}. LR is likelihood ratio of two conditional probabilities:

$$LR = \frac{P(X_i|M_{fg})}{P(X_i|M_{bg})} \tag{18}$$

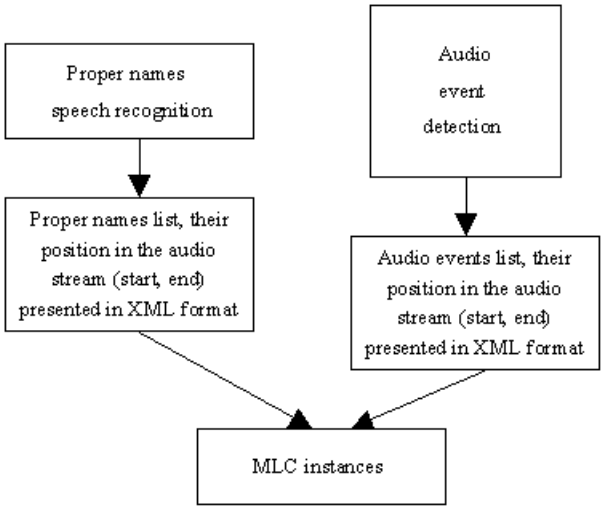


Figure 38: Schematic diagram of the proposed methodology for audio event detection for tested audio segments using adapted GMM

If likelihood ratio LR is more or equal than predefined threshold the hypothesis that tested foreground audio event is appear in non-speech segment X_i is accepted, otherwise not. This process is repeated thought all the segments.

The second approach is based on separation background and foreground sounds. For background and foreground source separation Blind Source Separation (BSS) can be used. In our task we will apply BSS to separate foreground and background sound from stereo audio signal. BSS is based Independent Component Analysis (ICA). Applying ICA to the feature set extracted from the mixture of audio sources does decomposition of the feature set in the components which are independent from each other. The independent components that correspond to the background and foreground sound can be identified and separated. Separated background and foreground sounds finally will be recognised separately using MAP classifier.

The result of audio events recognition will be a sequence of audio events with their boundaries and their temporal relations for further semantic interpretation.

Unfortunately very often the sport audio is noised by different kind of noise. Sometimes intensity of the noise is very high. In noisy condition the performance of audio events recognition could be low. For improvement the recognition rate of audio events the dependency of these events could be used. Some dependent events could be organised in the dependency network that can be considered as the grammar of low level events. The dependency network will be used for more reliable low-level events recognition.

Each low level audio event is described by their own GMM model trained on development data. The dependency between audio events is not very complex and for simplicity the transition probability can considered as equal. The united statistical model of low level events presents structural statistical model and can be used for events decoding using Viterbi decoder. Let suppose that X is observation, a sequence of feature vectors extracted from non-speech segments, M is statistical structural model and S is the sequence of possible low level audio events. The Viterbi algorithm will be used to choose the optimal sequence of low level audio events:

$$S_{opt} = \max P(S|M, O) \tag{19}$$

6.3.5 Spoken names recognition in speechdata

The segments labelled as speech will we used also as an input to the speech recogniser. Spoken names in speech segments will be recognised using a syllable-based speech recogniser. We will be based on a currently developed for the German language (see [Eickeler03]) which we wll extend and adapt for the English language where appropriate. The syllable-based approach is based on using inverse dictionary search technique to recognise a proper name. First N best syllables sequences of mixed syllables and

phone based on n-gram (syllables or hybrid phone & syllable language model) are identified and then using statistical string matching the best candidates from the name list using dictionary can be find. Knowledge of frequent syllable (phone) insertion, deletion and substitution can be incorporated in the statistical search stage to make it more accurate. It is also possible to create finite state grammar from N-best list to reduce complexity of the search.

The speech of commentator is noisy. This leads to mismatch between training audio condition and test audio condition in which speech is recognised. We would like to use some techniques for noise separation and normalisation. For noise separation we will investigate BSS some and other techniques of speech enhancement.

Proper name list is dependent from the type of sport event. To reduce perplexity of speech recognition task we suggest for each type of audio event (pole vault, high jump and etc) use audio event dependent vocabulary list. Unlikely is that the name of marathon runner will appear in context of high jump. Context dependent vocabulary reduces perplexity of speech recognition task and improves performance of speech recognition in highly noisy environment.

6.3.6 Emotion recognition in speech data

The speech of commentator in sport recording is often emotional. In audio event detection task is interested to recognise the emotional state of commentator. The emotional highlight will be useful for audio sport event interpretation, for example for high result highlight. The traditional way to indicate the emotional state of talker (commentator) is using prosodic features. The pitch contour and energy contour can indicate emotional stress in the speech. The number of pitch period indicates the changes in speed of speech. The significant acceleration in the speed of speech indicates the emotional stress of the talker. Method of emotion recognition based on prosodic characteristics of talker in sport recording will be investigated.

6.4 Semantic Model Usage

Audio/speech based information extraction will mainly use the domain ontology as an indexing structure: the instances which already exist in the ontology can be used for labelling audio material.

The domain ontology indicates the MLCs that are likely will be presented in the input audio data.

The domain ontology can also provide temporal relations among MLC, for example greeting is followed by start signal and start signal is followed by bell, after downed pole newer happened applauses.

6.5 Support to Ontology Enrichment

The audio/speech information extraction component is able to extract MLC which after further processing during the fusion stage, are forwarded to ontology evolution. The ontology enrichment can be performed when in the region of interest after audio events recognition appear events that do not match to any existing MLC. The temporal boundary of unknown events can be extracted during unsupervised non-speech data clustering. These unlabeled events will be considered as new events and together with other modality results will drive the ontology evolution process.

6.6 Confidence Handling

For each identified instance a confidence measure will be computed during the information extraction phase. Since the information extraction phase is implemented through a combination of recognisers/classifiers based on machine learning techniques, the confidence figure is also a combination of the confidence figure of each algorithm.

6.7 Evaluation Framework

6.7.1 Strategy

The evaluation strategy will evaluate the output of the following extraction steps:

1. Non-speech audio events extraction.
2. Spoken proper names extraction.

For non-speech audio events extraction and proper names evaluation will be used detection rate and false alarm rate measures.

The evaluation is provided using labelled by hand reference audio data.

We define that the event is correctly recognised when the distances between left and right boundaries of recognised event and left and right boundaries of the same referenced event are less then tolerance threshold. Otherwise the event is considered as incorrectly recognised.

In the first step of evaluation for each event according to the above criterion we decide this event is correctly recognised or not. Then the detection rate and false alarm rate is calculated:

$$\text{detection rate} = \frac{\#\text{ correctly recognized events}}{\#\text{ really present events}}$$

$$\text{false alarm rate} = \frac{\#\text{ falsely recognized events}}{\#\text{ really present events} + \#\text{ falsely recognized events}}$$

We define that the word is correctly recognised when the distances between left and right boundaries of recognised word and left and right boundaries of the same referenced word are less then the tolerance threshold. Otherwise the word is considered as incorrectly recognised.

In the first step of evaluation for each word (proper name) according to the above criterion we decide this word is correctly recognised or not. Then the detection rate and false alarm rate is calculated:

$$\text{detection rate} = \frac{\#\text{ correctly recognized words}}{\#\text{ really present words}}$$

$$\text{false alarm rate} = \frac{\#\text{ falsely recognized words}}{\#\text{ really present words} + \#\text{ falsely recognized words}}$$

6.7.2 Content

Manually annotated audio data with the different background will be used to do evaluation of audio events, spoken proper names and their boundaries.

6.8 Use Case: Pole Vault

In this section the methodology for semantics extraction from textual documents is explained through an example for the high level concept `pole_vault`.

Modality Element Ontology The HLC `pole_vault` should include the following MLCs:

1. `support`
2. `applause`
3. `exult`

Multimedia Descriptors Ontology Available low-level descriptors and features should include:

- Mel-Cepstral Coefficients (MFCC)
- delta MFCC
- delta delta MFCC
- Zero Crossing Rate (ZCR)
- Short Term Energy (STE)
- Fundamental Frequency (F0)

Multimedia Domain Model This will be a set of classifiers trained to identify MLCs using low-level features.

6.8.1 Output

The output of audio event detection module

- for each recognised MLC or for not recognised region will be an `xml` file (transcriber format) that contains the boundaries of region (event) and labels for recogniser regions. This output gives automatically the temporal relations (one event is before of after then the other) between recognised events.

6.8.2 MLC Identification

Likelihood ratio based approach For each MLC a list of properties consisting of low-level features will be created:

1. `support` → {MFCC,delta MFCC, delta delta MFCC, ZCR, STE, F0}
2. `Applause` → {MFCC, delta MFCC, delta delta MFCC, ZCR, STE, F0}
3. `Exult` → {MFCC,delta MFCC, delta delta MFCC, ZCR, STE, F0}

7 Text

7.1 Aim of Text-based information extraction

The aim of extraction from the textual part of BOEMIE documents is to provide information about the existence of Mid-Level Concepts – MLCs (e.g. names of persons, names of events, dates, ages, performance, etc.), the relations that may occur between them (e.g. that a person with a name N1 has a performance P1), as well as about the occurrence of terms for the various sporting events.

The output of the text-based information extraction is then provided to the single media interpretation module (see Figure 3) which associates some of the retrieved MLCs in order to form High-Level Concept (HLC) instances (e.g. an athlete's instance with information on the athlete's name, age, nationality, etc.) exploiting the relations retrieved between these MLCs.

As it is depicted in Figure 3, the result of this reasoning process is given to the "Fused-Media Analysis" module which combines the results of the text-based interpretation with the interpretation results of the other media that may occur in the same multimedia document that is being processed (e.g. a web page that has a textual part and one or more still images), in order to produce the document interpretation that will then feed the ontology evolution mechanism.

7.2 Overview of the methodology

Text-based information extraction involves three distinct processing phases which correspond to the three modes of operation specified in Section 2 of this deliverable:

- The *Analysis* mode of operation applies each time a new multimedia document becomes available. The run-time system is activated which takes as input the textual part of the multimedia document aiming to identify MLC instances and their relations. The run-time system is composed of modules trained to process text for the events under examination and is depicted in Figure 39.

- The *Training* mode of operation applies when manually annotated content is available. The domain dependent modules of text-based information extraction (e.g. the named entity recognition module, the relation extraction module depicted in Figure 39), will have to be trained or re-trained in case new annotated content becomes available.

- The *Discovery* mode of operation applies when a significant amount of content is available such that it can lead to expansion of the semantic model through augmenting the modality-specific MLCs. In the case of text-based information extraction, this part of methodology refers to the use of techniques in order to locate in the textual part of the collected content interesting clusters of strings that may correspond to new MLCs. For example, in case the current ontology does not cover "performance", clusters of numeric expressions, that is numbers with measurement units, may be located that may finally lead to the introduction of a new MLC for "performance". This mode of operation refers also to the application of term extraction techniques, for the identification of new terms for already known HLCs (e.g. the term pole in pole-vault events) or terms related to currently unknown HLCs (e.g. the term "high jump" in documents relevant to the event "high-jump" that is currently not covered in the ontology).

7.3 Description of Methodology

As it was previously mentioned, information extraction from text involves three phases. All phases share the same preprocessing stage (see Figure 39), which is responsible for extracting lexical and syntactic information from the text to be processed. Different parts of this information may then be exploited during each of the three different processing phases.

This section presents first the preprocessing stage and continues describing the three phases in the following order: training phase (Schema 2), run-time phase (Schema 1) and clustering-term extraction phase (Schema 3).

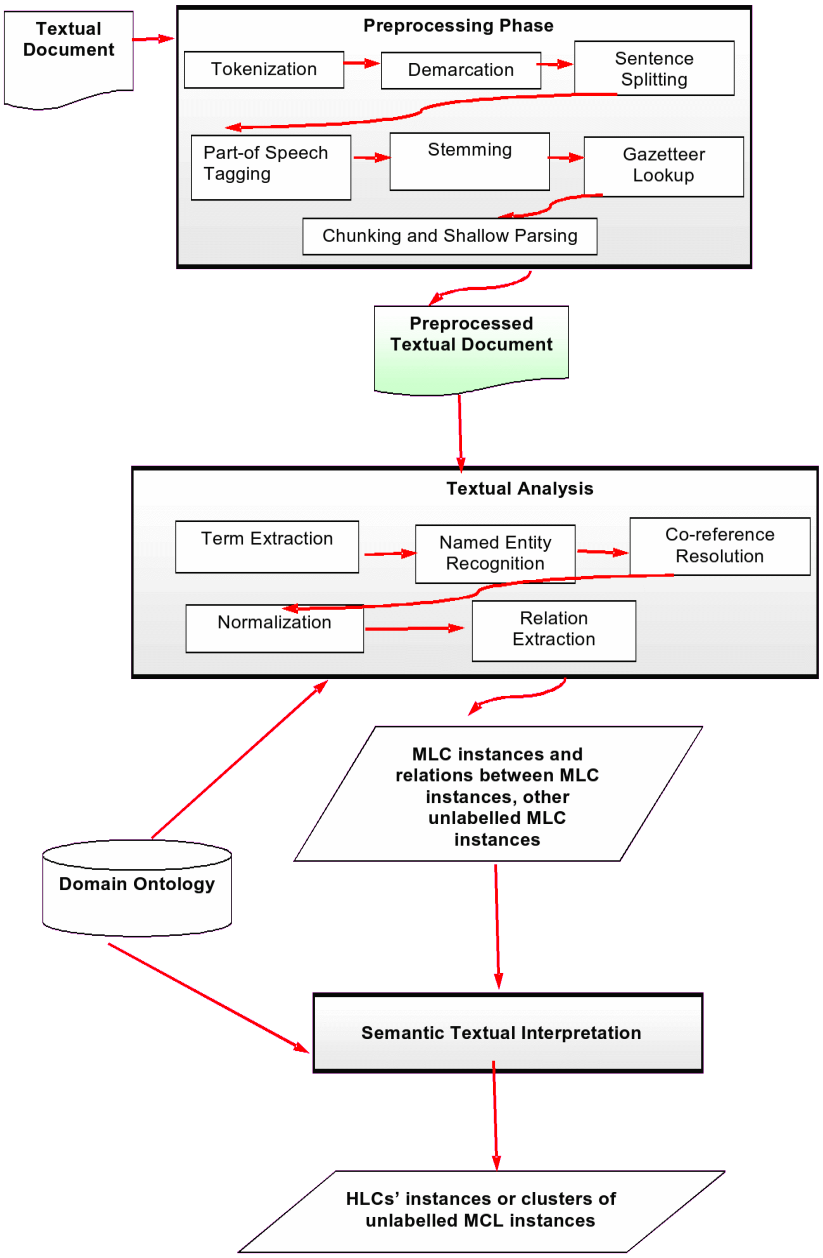


Figure 39: Information extraction from the textual part of a multimedia document.

7.3.1 Preprocessing

Preprocessing involves modules for the extraction of lexical and syntactic information from the textual part of the multimedia documents available in BOEMIE.

More specifically, preprocessing involves modules for: tokenization, demarcation, sentence splitting, part-of-speech tagging, stemming, gazetteer lookup and shallow parsing. Each of these modules is described below.

Tokenisation The tokenizer is responsible for separating text into word units (tokens), where a token can be roughly defined as a sequence of non-space characters. For tokenization, a tokeniser offered by the Ellogon language engineering platform will be used (HTokeniser), which is able to tokenize texts in the English language. HTokeniser not only identifies tokens, but also classifies them into categories encoding information like the character capitalisation/type of the characters contained in the token or whether the token belongs to a set of specific types (like ABBREVIATION, TIME, DATE, etc.). As HTokeniser is HTML/`xml` aware, it recognises HTML/`xml` markup tags and classifies them into proper categories (i.e. HTML_TAG, HTML_COMMENT, HTML_SCRIPT, etc.). HTokeniser follows a rule-based approach and is not domain dependent.

Demarcation The demarcation is responsible for identifying portions of text that refer to a single piece of information. Its primary use is to segment HTML documents that contain multiple items on a single document (e.g. a results table, a news item, an image's caption, etc.). Following a rule-based approach, the demarcator is domain independent.

Sentence Splitting The sentence splitter is responsible for identifying sentence boundaries, based on information obtained by the tokenizer and demarcation components. The sentence splitter can process both plain and HTML documents. In case of HTML documents it additionally examines the HTML tags identified by the tokenizer to produce additional sentence types, such as table rows or list elements. The sentence splitter follows also a rule-based approach and does not depend on thematic domains.

Part-of-speech tagging The part of speech (POS) tagger is an adaptable component, as it is based on a machine learning algorithm (transformation-based error-driven learning (TBED), a machine learning technique that has been successfully used for this task in a wide range of languages). The POS tagger has been trained for the English language on a large annotated corpus.

Stemming Stemming is the procedure of transforming a word into it's stemmed or root form. For example the stemmed form of the words "computation" and "computer" is "comput". A rule-based approach will be applied in order to transform each word into its equivalent stemmed form.

Gazetteer Lookup This module identifies in the processed text, strings that correspond to already known named entities, terms, etc., that are found in the ontology (domain, geographic) or other knowledge sources to be used in BOEMIE (e.g. Wordnet).

Shallow Parsing Shallow parsing tries to identify phrases (noun phrases, verb phrases, etc.) appearing in a text and associate some of them with syntactic roles (for example, that a specific noun phrase has a subject role in a sentence with respect to a specific verb). A rule-based approach will be applied.

7.3.2 Training

The task of this phase is to train the domain-dependent modules of the textual analysis sub-system (see Figure 39) according to the requirements of the domain and the application.

The input to this phase is the ontology along with the ontology-based manually annotated corpus. The output is the set of trained modules. The domain-dependent modules that need to be trained (or re-trained) when manually annotated textual content becomes available are the following: Named-Entity Recognition, Co-Reference Resolution, Normalisation and Relation Extraction.

name	round_name	sport_name	event_name
age	date	sport_gender	city
gender			country
nationality			
performance			
ranking			

Table 3: Mid Level Concepts (MLCs) for the text modality

Named Entity Recognition A Named Entity Recogniser (NER) should be trained on the annotated corpus for each text-specific MLC defined in the ontology (see Table 3) in order to recognise and classify in the processed text those named entities which correspond to these MLCs. In the case that entities appearing in the text cannot be categorised into one of the known MLCs, those entities are given as input to Phase 3.

NER systems nowadays are built entirely with the help of machine learning algorithms, as research on the application of machine learning on the task has been extensive during the last decade, with almost any known classification algorithm applied to it. Machine learning based NER systems can be roughly classified into the following categories:

Token-based NER This category involves systems that examine each token (word) in order to classify it as a part of a named entity of a specific category, or as not being part of a named entity. Named entities are considered as series of successive tokens classified as belonging to a named entity of a semantic category.

Phrase-based NER Systems in this category segment texts into phrases, that can be possible named entities (i.e. noun phrases), thus separating the problem of entity recognition from this of classification. Once recognised, phrases are classified into the various semantic categories, or as not being named entities.

Document-level NER Being the most recent addition to machine learning approaches for NERC, systems belonging to this category combine the results of the aforementioned subsystems and perform some basic filtering over their results. They examine documents as a whole in order to locate and classify named entities. An interesting property of systems in this category comes from the fact that they perform a search that is not located to processing a few tokens, but all tokens are processed simultaneously, offering the ability to perform co-reference resolution along with NERC. As a result, some systems of this category operate as complete information extraction systems.

Of course, there are also systems that utilise a combination of classifiers that can belong to any of the above categories, in order to maximise the accuracy of the results. Almost all of the approaches (independent of their category) utilise information from a preprocessing phase, such as morphological cues (i.e. capitalisation), contextual information (i.e. presence of specifiers in the vicinity of inspected entity), or presence in predefined lists (gazetteers) in order to decide.

For the purposes of BOEMIE, we will examine approaches from at least two categories (token-based and document-level NER). Among the techniques we are going to examine are various token-based NER classifiers such as the Brill tagger [Bril95], the Trigrams 'n' Tags (TNT) tagger (http://www.coli.uni-saarland.kde/~thorsten/tnt/), the CRF++ toolkit (http://chasen.org/ ~taku/software/CRF++/) or the YamCha toolkit (http://chasen.org/~taku/software/yamcha). Regarding document-level NER systems, the T-Rex (http://tyne.shef.ac.uk/t-rex/index.html) (the successor of Amilcare (http://nlp.shef.ac.uk/amilcare/)). Finally, we will try to investigate the combination of multiple systems through voting or stacking.

Co-reference Resolution Co-reference resolution is the process of identifying whether two or more text portions refer to the same named entity. A co-reference module for an extraction system should handle the following co-reference problems:

Name matching This involves the retrieval of names referring to the same entity (e.g. "Tatiana Lebedeva", "T. Lebedeva", "Lebedeva").

hasAge(name$_i$, age$_i$)
hasGender(name$_i$, gender$_i$)
hasNationality(name$_i$, nationality$_i$)
hasPerformane(name$_i$, performance$_i$)
hasRanking(name$_i$, ranking$_i$)
hasDate(round_name$_i$, date$_i$)
hasStadium(sport_name$_i$, stadium_name$_i$)
hasSportStartDate(sport_name$_i$, date$_i$)
hasSportEndDate(sport_name$_i$, date$_i$)
hasSportCity(sport_name$_i$, city$_i$)
hasEventCity(event_name$_i$, city$_i$)
hasCountry(event_name$_i$, country$_i$)
hasEventStartDate(event_name$_i$, date$_i$)
hasEventEndDate(event_name$_i$, date$_i$)

Table 4: Types of relations for the text modality

Pronoun-antecedent co-reference Pronouns like 'he', 'she', 'they', 'them', etc. must be associated with their antecedents, resolving them to a domain relevant named entity if possible.

Definite description co-reference This type of co-reference would hold between 'Tatiana Lebedeva' and 'the athlete', or 'Tatiana Lebedeva' and 'The Olympic Gold Medal Champion'.

A co-reference resolution system will be trained in order to resolve anaphoric references to recognised entities corresponding to MLCs within the same sentence or adjacent sentences or within the same paragraph, as well as to perform synonym detection.

Nowadays, both machine learning based approaches (such as [Ng02]) as well as rule-based ones ([Bont02, Dimi02]) are used for co-reference resolution. For the purposes of BOEMIE, a rule-based approach will be examined. Especially for name matching various token matching algorithms will be examined [Vala04, Koti06, Cast05].

Normalisation Normalisation is responsible for the transformation of certain types of MLC instances to a predefined form. Normalisation may involve time expressions (e.g. dates), numeric expressions (e.g. performance), names (e.g. person names, event names). For example, the normalised form of a date appearing in a text like "13 December 06" will be "13/12/2006", if the template for dates is defined as "DD/MM/YEAR".

For the purposes of BOEMIE a normalisation component will be trained for each known MLC in order to transform MLC instances into a predefined format. The MLCs that require such a trainable module are the following: expressions denoting dates, durations; numeric expressions denoting performance, age, ranking of an athlete; names of sports, events. In the case of new MLCs requiring normalisation, the ontology expert needs to define the normalisation format for them.

Extraction of relations The relation extraction module locates in the text relations between the identified MLC instances. These relations correspond to the properties of the text-specific HLCs defined in the ontology (the predicate names in Table 4).

Initially, a pattern language will be used in order to create manually patterns for each relation exploiting the results of the previous processing stages (i.e. pre-processing, named entity recognition, co-reference resolution). At a second step, we plan to examine the use of machine learning techniques for learning the patterns as well as combination of rule-based and machine learning techniques.

7.3.3 Analysis

In this phase, the text-based information extraction system, in which the domain-dependent modules have already been trained (in the Training Phase presented above), receives as input the textual part of a multimedia document (note that is not annotated). This textual part is first being processed by the pre-processing sub-system and the results are then fed to the trained textual analysis sub-system which may produce the following:

1. instances of MLCs
2. named entities, numeric expressions, time expressions, that have been recognised by the NER module but which could not be categorised into one of the known MLCs; these feed the modules involved in Phase 3 (see below)
3. relations between instances of MLCs;

The output of the text-based information extraction is then provided to the single media reasoning module which associates some of the retrieved MLCs in order to form High-Level Concept (HLC) instances exploiting the extracted relations. The result of this reasoning process is given to the "Fused-Media Analysis" module which combines the results of the text-based reasoning with the reasoning results of the other media that may occur in the same multimedia document that is being processed (e.g. a web page that has a textual part and one or more still images), in order to produce the document interpretation that will then feed the ontology evolution mechanism.

7.3.4 Discovery

When a significant amount of textual content is available (what "significant" means in terms of content size is still to be defined), this processing phase is activated in order to detect interesting types of information that could not be classified during the analysis phase. Such processing can be triggered by the ontology enrichment activity of ontology evolution in order to support the ontology expert to define new MLCs (see section 6.2 of Deliverable D4.1).

Types of interesting information may include for example clusters of time expressions which present common properties (e.g. use of "second" as a measurement unit and being in a specific range of values). This cluster may lead the ontology expert, during ontology evolution (see deliverable D4.1) to form a new type of MLC for athlete's performance. Another type of interesting information may be terms related either to known HLCs or to unknown ones. Again this may be helpful input for ontology evolution.

Several methods have been proposed in the literature for term identification and extraction. Among the most successful ones are statistical methods, which usually try to determine the significance of each word with respect to other words in a corpus, based on word occurrence frequencies. TF/IDF [Salt75] is usually employed for this task ([Ahma94, Dam93]) possibly combined with other methods such as latent semantic indexing [Fort05] or taking into account co-occurrence information among phrases [Fran00]. Clustering techniques also play an important role in term identification: recognisable entities can be clustered into groups based on various similarity measures, with each cluster being a possible term (comprised of synonyms). Approaches like ([Kiet00, Agir00, Faat02]) employ clustering techniques and other resources like WordNet 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 quite often ([Gupt02, Haas05a]). Finally, morphological clues (such as prefixes and suffixes) can be extremely useful for some domains: suffixes like "-fil" and "-it is" quite often mark terms in medical domains [Haas05b, Haas05c].

For the purposes of BOEMIE, we will examine simple statistical methods and clustering techniques exploiting also the results of previous processing stages (pre-processing, textual analysis).

7.4 Semantic Model Usage

Text-based IE exploits the semantic model in the following ways:

- During pre-processing, the gazetteer lookup module locates inside the processed texts existing instances of MLCs, names of ontology concepts and properties which can be exploited during the subsequent processing phases.
- During the training phase, the ontology-based annotated content is used for the training of the domain-dependent modules of textual analysis.
- During the analysis phase, the named entity recognition (NER) module locates new instances of MLCs, and the relation extraction module identifies relations between instances of MLCs which correspond to properties of HLCs. Furthermore, single-media semantic reasoning exploits ontology axioms in order to associate some of the MLC instances found in a document under a specific HLC instance.
- During the discovery phase, the aim is to identify interesting clusters of text expressions or terms to support ontology evolution.

7.5 Support to Ontology Evolution

Text-based information extraction can provide to ontology evolution the following:

- instances of text-specific MLCs
- instances of HLCs resulted from text-based reasoning which form part of the multimedia document interpretation after fusion
- clusters of textual expressions which may correspond to new MLCs or terms which may guide the naming of new concepts or properties.

7.6 Confidence handling

For each identified MLC instance and each relation a confidence measure will be computed during the information extraction phase. Since text-based information extraction involves various modules, some of which may be implemented through a combination of recognisers/classifiers, the final confidence figure is also a combination of intermediate confidence figures. The confidence mechanism will be able to decide the degree of confidence for the recognition of an MLC instance or relation and feed this figure to the text-based reasoning module.

7.7 Evaluation Framework

7.7.1 Strategy

The evaluation strategy will assess the output of the following extraction steps:

1. Named entity recognition and their categorisation into existing MLCs
2. Extraction of relations
3. Identification of instances of hidden MLCs
4. Identification of instances of hidden relations between MLCs

The evaluation strategy will follow two scenarios. The first scenario will evaluate the ability of text-based information extraction to recognise and categorise correctly the instances of existing MLCs (step 1 above) and the relations between these instances (step 2). The second scenario will evaluate the ability of text-based information extraction to “discover” instances of hidden MLCs (step 3) and hidden relations (step 4). Table 5 summarises the objectives, evaluation process and data requirements for text-based information extraction.

Objective	Evaluation	Data requirements
Extraction of instances of existing MLCs and relations from large collections of textual content	*Scale:* Hundreds of texts. *Evaluation approach*: Comparisons against corpus annotated manually (gold-standard) with the Ellogon text annotation tool. *Target*: Precision > 80%, Recall > 80%	300 annotated documents per sporting event for all the events in the three categories
Identification of instances of hidden MLCs and relations	*Scale:* Hundreds of texts *Evaluation approach*: Hide the instances of some MLCs and relations and attempt to identify them *Target*: be able to identify > 70% of the hidden instances	MLC instances to be identified > 1000. Instances of relations to be identified > 100.

Table 5: Measurable objectives for technologies dealing with text-based information extraction

7.7.2 Content

Manually annotated textual content will be used as ground truth. Annotated content will conform to the following specifications:

- 300 annotated documents per sporting event for all the events in the three categories (jumping, running, throwing).
- Each annotated document must contain at least 5 MLC/HLC instances.

7.7.3 Quantitative measures

Evaluation will follow standard practice in the information extraction community of comparing system output against a hand-annotated "gold-standard". Measures like *Recall*, *Precision* and *F-measure* will be used. *Recall* is a measure of how many entities from the gold-standard were annotated in the system output and *precision* is a measure of how many of the entities in the system output actually occur in the gold-standard. It is possible for a system to score well for recall (i.e. finding a high number of the names annotated in the gold-standard) while scoring badly for precision (i.e. characterising as names a large number of strings that are not annotated as such in the gold-standard). The standard way to score a system's performance in general is to compute *F-measure* which averages across recall and precision. More precisely,

$$F = 2\frac{recall \cdot precision}{recall + precision}$$

.

7.8 Use Case: High Jump

In this section the methodology for semantics extraction from textual documents is explained through an example for the high level concept `high_jump`.

Suppose that the following text was given:

> Blanka Vlasic, the winner in Oslo, took her second Golden League victory in the women's High Jump. The Croatian saw off many of her main rivals for next month's European title including Sweden's Kajsa Bergqvist (1.97m), the World champion, and Russian Olympic champion Yelena Slesarenko (1.91m 7th, nursing an ankle injury by the look of it). Belgium's Tia Hellebaut equalled her two meters national record set last week in Paris but lost on count-back to Vlasic.

Initially, the pre-processing modules will be applied in order to perform tokenization, demarcation, sentence splitting, part of speech tagging and shallow parsing. For example the output of the shallow parsing module for the first sentence of the example may have the following form:

```
[NP Blanka Vlasic], [NP the winner] [PP in Oslo] [VP took]
[NP her second Golden League victory] [PP in the women's High Jump]
```

where NP stands for noun phrase, PP for prepositional phrase and VP for verb phrase.

During named entity recognition (NER) in textual analysis, the named entities identified are annotated:

> <NE TYPE = 'NAME>Blanka Vlasic </NE>, the winner in <NE TYPE = CITY> Oslo </NE>, took her second <NE TYPE = 'EVENT_NAME'> Golden League </NE> victory in the ¡NE TYPE = 'GENDER''> women</NE>'s <NE TYPE = 'SPORT'> High Jump</NE>. The <NE TYPE = 'NATIONALITY> Croatian <NE>saw off many of her main rivals for next month's European title including ¡NE TYPE = 'NATIONALITY>Sweden</NE>'s <NE TYPE = 'NAME> Kajsa Bergqvist (¡NE TYPE = PERFORMANCE> 1.97</NE>m), the World champion, and ¡NE TYPE = 'NATIONALITY> Russian </NE> Olympic champion <NE TYPE = 'NAME> Yelena Slesarenko (¡NE TYPE = PERFORMANCE> 1.91</NE>m <NE TYPE = RANKING>7th </NE>, nursing an ankle injury by the look of it). <NE TYPE = 'NATIONALITY> Belgium</NE>'s <NE TYPE = 'NAME> Tia Hellebaut </NE> equalled her ¡NE TYPE = PERFORMANCE>two meters</NE> national record set last week in <NE TYPE = 'CITY'> Paris </NE> but lost on count-back to <NE TYPE = 'NAME> Vlasic <NE>.

The named entities categorised as MLC instances are listed below:

$$
\begin{aligned}
\text{name}_1 &= \text{“Blanka Vlasic”} \\
\text{name}_2 &= \text{“Kajsa Bergqvist”} \\
\text{name}_3 &= \text{“Yelena Slesarenko”} \\
\text{name}_4 &= \text{“Tia Hellebaut”} \\
\text{nationality}_1 &= \text{“Croatian”} \\
\text{nationality}_2 &= \text{“Sweden”} \\
\text{nationality}_3 &= \text{“Russian”} \\
\text{nationality}_4 &= \text{“Belgium”} \\
\text{gender}_1 &= \text{“women”} \\
\text{sport\_name}_1 &= \text{“High Jump”} \\
\text{ranking}_1 &= \text{“1st”} \\
\text{ranking}_2 &= \text{“7th”} \\
\text{performance}_1 &= \text{’“1.97m”} \\
\text{performance}_2 &= \text{“1.91m”} \\
\text{performance}_3 &= \text{“2.00m”}
\end{aligned}
$$

The co-reference resolution module initially performs synonym detection which recognises, for example, that “Blanka Vlasic” and “Vlasic” are two alternatives of the same person's name.

The relation extraction module identifies a number of relations between the MLC instances retrieved:

$$
\begin{aligned}
(\texttt{name}_1, \texttt{nationality}_1) &: \texttt{hasNationality} \\
(\texttt{name}_2, \texttt{nationality}_2) &: \texttt{hasNationality} \\
(\texttt{name}_3, \texttt{nationality}_3) &: \texttt{hasNationality} \\
(\texttt{name}_4, \texttt{nationality}_4) &: \texttt{hasNationality} \\
\texttt{sport\_name}_1 &: \texttt{Sport\_Name} \\
\texttt{gender}_1 &: \texttt{Sport\_Gender} \\
(\texttt{name}_1, \texttt{sport\_name}_1) &: \texttt{participatesAtSport} \\
(\texttt{name}_2, \texttt{sport\_name}_1) &: \texttt{participatesAtSport} \\
(\texttt{name}_3, \texttt{sport\_name}_1) &: \texttt{participatesAtSport} \\
(\texttt{name}_4, \texttt{sport\_name}_1) &: \texttt{participatesAtSport} \\
(\texttt{name}_1, \texttt{ranking}_1) &: \texttt{hasRankingPosition} \\
(\texttt{name}_3, \texttt{ranking}_2) &: \texttt{hasRankingPosition}
\end{aligned}
$$

$(\texttt{name}_2, \texttt{performance}_1)$: hasPerformance
$(\texttt{name}_3, \texttt{performance}_2)$: hasPerformance
$(\texttt{name}_4, \texttt{performance}_3)$: hasPerformance

8 Reasoning

The main goal of this chapter is to present reasoning approaches for multimedia analysis. The reader is referred to the appendix A for an introduction of description logics (DLs) that form the logical foundations of the ontology languages and enable the definition of reasoning services is provided. Here, we will present two non-standard reasoning services that we propose for high-level multimedia interpretation in the BOEMIE project. Finally, this chapter will conclude with some remarks on the open research problems confronted.

One of the major goals of the BOEMIE project is to automate knowledge acquisition from multimedia content by combining multimedia extraction and ontology evolution in a bootstrapping process. In particular, workpackage two deals with semantics extraction from multimedia content. Reasoning services have to perform following tasks to support workpackage two:

- Support low-level multimedia analysis, which aims to extract semantics from multimedia content (concept rewriting using concrete domains)
- Relate multimedia content to high-level knowledge in the domain ontology using the results of low-level multimedia analysis (high-level multimedia interpretation)

Low-level multimedia analysis has the task of extracting semantics from multimedia content in order to deliver an appropriate set of concept and role assertions (w.r.t. a domain ontology) as input for the higher-level multimedia interpretation. In what follows, we analyze our approach focusing in image analysis.

In the BOEMIE project, a collection of images from the athletics domain will serve as the training data for image analysis. In order to gain the necessary training data, certain regions of these images will be manually annotated with so called MLCs (mid-level-concepts). MLCs are concepts from the domain ontology that can be associated with multimedia features and thus extracted using image analysis techniques. Given a new image, image analysis first segments it into regions and then analyses each region using certain low-level features such as colour, shape or texture. Finally, image analysis has the goal to identify and label each region (or some set of regions) as instances of MLCs by using the knowledge gained through the training data.

This means that, once an image is segmented into regions, extraction tools used for image analysis investigate regions of the image in a bottom-up approach using low-level features. Obviously, both image segmentation and feature extraction are non-deterministic tasks due to the uncertainties involved in the analysis methods applied, and the imprecision and incompleteness of available knowledge about the domain.

Despite the uncertainty, image analysis has to recognise visual representations of MLCs in the image. At this point, non-standard reasoning services can support image analysis by taking into account additional high-level knowledge about the domain such as HLCs (high-level concepts). This will enable image analysis to gain a top-down view on an image and fine-tune available parameters to achieve better results. Non-standard reasoning services for high-level multimedia interpretation construct the interconnection between workpackage two and four in the BOEMIE project.

In workpackage two, the reasoning services offered for multimedia interpretation consider each multimedia object in isolation. Some of the multimedia objects will be analysed by considering more than one modality. Even if more than one modality is taken into account, the reasoning services still regard a single multimedia object. Contrary to this, workpackage four considers multimedia objects not separately but as a whole in order to apply its methodology for ontology evolution.

To sum it up, non-standard reasoning services for high-level multimedia interpretation, which we will discuss next in detail, work on single multimedia objects and the results cumulate to serve as input for ontology evolution.

8.1 Concept Rewriting using Concrete Domains

Typically, semantics extraction from multimedia is managed by an extraction algorithm. According to some strategy, this algorithm will apply low-level feature extractors on certain regions of an image and obtain some characteristic value(s) as a result for each region.

$$\begin{aligned}
\texttt{woman_high_Jump_event} &\equiv \texttt{high_jump_event} \sqcap \exists \texttt{hasPart.Female_Jumper} \\
\texttt{female_jumper} &\equiv \texttt{jumper} \\
\texttt{male_jumper} &\equiv \texttt{jumper} \\
\texttt{male_jumper} &\equiv \neg\ \texttt{female_jumper} \\
\texttt{female_jumper} &\equiv (\leq\ \texttt{jumps}\ 2.09)
\end{aligned}$$

Table 8: An abstract Tbox for the athletics domain

In an ideal case, comparison of these values with the values of MLCs, which are observed in the training data, will provide for an unambiguous labeling of each region with a MLC. However, in most cases a complete match of these values will not be possible. In such situations, the extraction algorithm could benefit from a reasoning service that can classify the measured values of a region with respect to the values of MLCs. This will support the extraction algorithm to decide on further actions such as fine-tuning low-level feature extractor parameters.

In this section we propose a reasoning approach for such situations by using a hypothetical example from the athletics domain. Suppose that the following MLC definitions are given in the multimedia ontology:

$$\texttt{Horizontal\_Bar} \equiv \texttt{Sport\_Equipment} \sqcap (\leq\ \texttt{S}\ 50) \cap (\geq\ \texttt{S}\ 30)$$

$$\texttt{Pole} \equiv \texttt{Sport\_Equipment} \sqcap (\leq\ \texttt{S}\ 20) \cap (\geq\ \texttt{S}\ 10)$$

where S is a concrete domain attribute, e.g., size. Now, suppose that for a region named o_1 in an image, the extraction algorithm applies the low-level feature extractor f_1 that yields the value for S, e.g., the interval between 10 and 50.

First, the extraction algorithm notifies the reasoning service about this information by adding following assertions:

$$(o_1, i) : \texttt{S}$$

$$10 \leq i \leq 50$$

Later it asks the reasoning service for the concept, of which o_1 can be an instance of. This query can be formalised as logical entailment problem:

$$\Sigma \cup \Gamma_1 \models \alpha$$

where Σ is the background knowledge, Γ_1 represents the two assertions just added and α is a set of concept names such that there is no α' with $\alpha' = \alpha$. The reasoning service is required to deliver an α, which contains as many as possible concept names from Σ. Moreover, for concrete domain value intervals, for which no concept name exists in Σ, the most specific concept should be created and added to α. This can be expressed as follows:

$$10 \leq i \leq 50 \equiv_\Sigma C_1 \sqcup C_2 \sqcup \ldots C_n \sqcup D$$

where $C_i \in CN$ (CN stands for concept names in Σ) and D is the most specific concept that can be defined. The problem of rewriting a concept given a terminology has been investigated in the literature [Baad00]. A unique solution to this problem can often not be found. In our specific example reasoning service may return the following answer:

$$C \equiv \texttt{Horizontal\_Bar} \sqcup \texttt{Pole} \sqcup (20 \leq \texttt{S} \leq 30)$$

In natural language this result means that the image region o_1 shows either a horizontal bar or a pole or an unknown MLC, which has a S value between 20 and 30. Given this information, the extraction algorithm can make a better decision. E.g., it may decide to adjust some parameters such as segmentation parameters and apply the low-level feature extractor f_1 again. Alternatively, it may exploit another feature extractor to narrow down the MLC candidates for this image region.

8.2 High-level interpretation

In the ideal case, high-level multimedia interpretation of a multimedia object is achievable solely through the assertions delivered by multimedia analysis and reasoning can deduce further information using background knowledge. Suppose that the textual analysis of an image, which is a screenshot of a TV broadcast and contains some textual information, extracts following words: women high jump event. The concept `Woman_High_Jump_Event` is defined as a high level-concept in the domain ontology. It is an example of concepts, which are extractable by one modality and not extractable by another one. In this case the concept `Woman_High_Jump_Event` is extractable by text analysis but not extractable by image analysis. We assume that the axioms shown in Table 8 are given in the domain ontology (Tbox). Using the background knowledge in the domain ontology the reasoning service can deduce that the women high jump event shown in this image includes a female jumper, who cannot jump more than 2.09 meters, which is the current world record in women high jump.

It is realistic to expect that multimedia analysis will in many cases not be able to deliver enough assertions to interpret a multimedia object at a high level. In such situations an abductive approach is required for high-level interpretation[6]. In a nutshell, given a set of facts and assertions, the abduction process tries to find an explanation so that all assertions hold. The abduction process is represented by the formula:

$$\Sigma \cup \Delta \cup \Gamma_1 \models \Gamma_2$$

where Σ is the background knowledge, Δ is the sought-after explanation and finally Γ_1 and Γ_2 represent different kind of assertions. Γ_1 contains bona fide assertions, which are believed to be true by default, and Γ_2 contains assertions, which are to be entailed by the abduction process. In description logics terminology, Σ forms the Tbox, whereas Γ_1 and Γ_2 constitute the Abox. The assertions Γ_1 and Γ_2 are generated by low-level multimedia analysis as a consequence of the observations made: Given some training data, low-level multimedia analysis has the task of recognising objects in new multimedia content and identifying constraints among them. Typically these constraints may be of spatial or temporal nature.

Consider the following hypothetical example: In a given image the extraction algorithm recognises a person, a horizontal bar, a mat and another object that may be a javelin or a pole. Due to the lack of high-level knowledge, in this case the `Pole_Vault_Event`, the extraction algorithm cannot classify the unknown object. The extraction algorithm notifies the reasoning service about all recognised regions in the image, which constitute Γ_1:

- $\texttt{Person}(i_1)$
- $\texttt{Horizontal\_Bar}(i_2)$
- $\texttt{Mat}(i_3)$
- $\texttt{Javelin} \sqcup \texttt{Pole}(i_4)$

Moreover it notifies the reasoning service about Γ_2. Γ_2 may contain some spatial constraints among the objects in Γ_1 and thus represent the belief that they are related to each other and are probably part of a high-level object:

- $\texttt{NextTo}(i_1, i_2)$
- $\texttt{OnTopOf}(i_2, i_3)$
- $\texttt{NextTo}(i_1, i_4)$

The reasoning service is provided with the knowledge about high-level concepts, which is represented by the symbol Σ in the abduction formula. For instance, the high-level concept `Pole_Vault_Event` may be described in the domain ontology as follows:

$$\begin{aligned}\texttt{pole\_vault\_event} \equiv & \texttt{jumping\_event} \sqcap \exists \texttt{hasEquipment.Pole} \sqcap \\ & \exists \texttt{hasFacility.Foam\_Mat} \sqcap \exists \texttt{hasFacility.HorizontalBar}\end{aligned}$$

As a result, the abduction process finds the explanation below and delivers it to the extraction algorithm:

[6] A more detailed discussion of the abduction process can be found in the BOEMIE project deliverable D4.1

- $\texttt{Pole}(i_4)$

As demonstrated by the example, abduction can guide the extraction algorithm in situations where the extraction algorithm can only assign a disjunction of MLCs to a region.

8.3 Concluding Remark

Multimedia analysis will most likely utilise a learning algorithm in order to exploit the knowledge contained in the training data. There are many known learning algorithms such as decision tree learning, statistical learning or inductive learning that can be applied in here. These algorithms are based on different techniques, which require different kind of guidance. Accordingly, reasoning services must be tailored to the needs of the specific learning algorithm employed in multimedia analysis.

Despite the simplicity of the examples presented in this chapter, it can be observed that reasoning services have to deal with the uncertainty of the perceptions and the vagueness of the concepts in the ontologies to deliver valuable results. Many approaches to the reasoning problem under uncertainty and vagueness are known in the literature e.g. belief networks, Bayesian based reasoning [Pear88], fuzzy logics [Dubo93] or rough mereology [Polk96]. Finding the most suitable approximate reasoning approach for semantics extraction from multimedia requires further investigation.

9 Multi-modal Data Fusion

In this section, we provide a preliminary outline of the methodology for combining extracted information from multiple media. Emphasis is given on the aims of multi-modal data fusion and the interdependence of the methodology with other workpackages. Some preliminary measurable objectives are also given.

9.1 Aims of multi-modal data fusion

Real-world events in BOEMIE are described through the visual (still images, video), the auditory and the textual modalities. Multi-modal data fusion aims to combine information stemming from the specific analysis of these modalities, in order to enable ontology evolution up to a degree that cannot be achieved using part of the modalities provided. In particular, the aims of multi-modal data fusion considered in BOEMIE are the following:

Improve semantics extraction Combine different data sources, to improve understanding of a real-world event. Semantics extraction can be improved through:

- assembling information related to distinct properties of the event (*complementary information*)
- improving the accuracy and confidence level about the extracted information, in case of sources agreement (*redundant information*)
- assist modal-analysis, in case of sources disagreement, so as to achieve sources analysis coherence (*coherent information*)

Motivate concept discovery Propagate a concept discovered by one modality to its identification from other modalities, through juxtaposition of the modalities instances.

In particular, since discovery of new concepts based on similarity may be conducted with more or less difficulty depending on the modality, discovery of a concept by one modality does not exclude that the same concept can also be identified at other media. The association of a concept to its instances in modalities, other than the discovering one, can be done by analysis of media instances concurring with the discovering modality instances.

Balance computational complexity Prioritise low-cost modal-specific analysis to avoid full-media analysis, until reaching acceptable confidence level of extraction.

Data analysis comes with a computational cost, which may differ in terms of the media to be analysed. In case of redundant sources of information, the overall analysis cost can be reduced by first applying low-cost analysis, which, when conducted with high confidence level, may narrow down the overall classification tasks and, hence, allow avoiding full analysis.

9.2 Interdependence with other workpackages

Methodology for multi-modal data fusion in BOEMIE respects both the ontology-driven approach followed in the BOEMIE project as well as the choices of information representation. In particular:

Ontology evolution driven process (WP 4) As any modal-specific analysis, fusion analysis is triggered by the ontology evolution process and its results are eventually populating and/or enriching the ontology.

- Inputs of the fusion analysis are provided by the ontology, either as direct instances of the ontology's modal concepts or indirectly, through URI to corresponding `xml` files. Fusion analysis has no direct interaction with any modality specific analysis.
- Outputs of the fusion analysis are given back to the ontology as instances of the ontology's corresponding concepts. When needed, information not possible to be contained in the ontology will be provided through `xml` files. Usage of temporary population, to take advantage of the inference engine for complementary consistency checking may be demanded.

Choice of information representation (WP 3) Information in BOEMIE is expected to be represented in many levels: as (a) (numerical) low-level features, (b) (symbolic) mid-level features (c) by (symbolic) mid-level concepts and (symbolic or numeric) relations between them and (e) (implicit) high-level knowledge within the ontology.

- Fusion of data stemming from different modalities will be considered for the above levels, both symmetrically and asymmetrically with specific aims and techniques each time.
- The output representation of information fusion will be dependant on its inputs, but will always be of the abovementioned level types. Especially, when inputs of fusion analysis contain confidence score levels, the fusion output will also have a confidence level score.

9.3 Measurable objectives

- Population: Augment the performance (recall, precision) of the integrated system up to a statistical significant level
- Enrichment: Enrich the ontology with concepts specific to two or more modalities, when concept discovery has been done only through one modality.
- Complexity: Enable balancing the complexity of the overall system vs. its performance, maintaining the performance higher than a unimodal approach.

10 Risk analysis

In this section we identify the main risk factors that may influence the project plan and cause delays or that may influence the quality of the semantics extraction from multimedia content. The handling strategy for each of them is briefly presented.

This section serves the purpose of identifying as early as possible the major risks, so that the development of the methodology will address these risks early in the development cycle. This approach may guarantee minimisation of delays through early prototyping and testing.

10.1 Image

The main risks that can be identified in the Methodology for Semantics Extraction from Still Images are related to *MLC detection failure* and *incorrect MLC labelling*. The first type of risks is common to every object recognition algorithm that appears in the literature. Avoiding such risks will lead to a real advance in the area of Computer Vision as well as in several other disciplines. Actually it is one of the aims of BOEMIE to use domain knowledge in order to improve the MLC detection process.

The second type of risks (incorrect MLC labelling) is common in the area of artificial intelligence. In addition, one should take into consideration that, in BOEMIE, MLC labelling will apply to image regions created by image segmentation techniques. The latter are known for their sensitivity on several factors such as low contrast, image noise, colour diffusion, etc. Handling of these risks will also lead to an advance in the corresponding state of the art. The following table describes the main risks that are anticipated in the Semantics Extraction for Image Methodology:

1. **Risk** MLC detection failure: *missing MLC*

   Handling (a) Implementation of several holistic detection methods
   (b) Combination of holistic and region based approaches for the final MLC area detection
   (c) Guidance from the fused analysis module (use of fusion information and reasoning services) through the interpretation feedback.

2. **Risk** MLC detection failure: *incorrect MLC area (false alarm)*

   Handling (a) Creation of alternative models for each one of the MLCs
   (b) Cross checking, within the image analysis module, (combination of holistic and region based approaches), of the assigned MLC labels.
   (c) Guidance from the fused analysis module (use fused information and reasoning services) through the interpretation feedback.

3. **Risk** Incorrect MLC labelling due to the classifier

   Handling (a) Use of multiple classifiers per MLC and a voting scheme
   (b) Use of rule-based classification schemes
   (c) Cross checking, within the image analysis method, of the assigned MLC labels.

4. **Risk** Incorrect MLC labelling due to segmentation failure.

   Handling (a) Implementation of several image segmentation techniques
   (b) Use of guidance from the fused analysis module for segmentation parameters tuning.
   (c) Combination of holistic and region based approaches for the final MLC area detection.

10.2 Video

The risk analysis for video-based semantics extraction was conducted based on the risk theme categories identified by [Bass06], as far as they seemed to be applicable to a research project as opposed to commercial software development. They are not used here directly, but served as a guideline to think about risks associated with the methodology.

The presentation of the individual risks consists of a short risk description, an assessment of the severity of the risk and a description how the risk will be handled, by trying to avoid it or addressing it if it does occur.

1. **Risk** Failure to provide relevant information to the overall system. High severity: signifies complete component failure if not handled.

   Handling *Avoidance:* The ongoing design of the overall system is closely followed and taken into account in the methodology. *Addressing:* Communicate with partners to identify exactly why the provided information is not relevant and what can be done to increase information quality.

2. **Risk** Late identification of missing or underspecified steps in the methodology that are costly to add in retrospect. Medium severity: signifies degraded component quality and schedule slip if not handled.

   Handling *Avoidance:* Areas in the methodology are flagged that show lack of progress because of potentially missing or underspecified steps. *Addressing:* Identify missing step, evaluate how important it is for the success of the methodology, propose schedule for implementation if it cannot be left out or worked around.

3. **Risk** Late identification of missing information to provide to the overall system. Medium severity: signifies degraded component quality and schedule slip if not handled.

   Handling *Avoidance:* Semantic model development is followed and new detectable concepts or relations are either implemented immediately or flagged for potential problems. *Addressing:* Identify missing information, evaluate its importance for overall system performance and cost for providing it, decide on inclusion.

4. **Risk** Failure of statistical analysis to train classifiers due to insufficient amount of annotated training data. High severity: signifies failure of major part of methodology if not handled

   Handling *Avoidance:* Methods for reducing the required amount of training data in the literature are investigated during tool development. *Addressing:* Estimate additional amount of training data required, decide if further annotation can be achieved in time or if above methods should be applied.

5. **Risk** Failure of statistical analysis to train classifiers due to uninformative features. High severity: signifies failure of major part of methodology if not handled

   Handling *Avoidance:* Existing research indicates that the features to be used are informative [Ke05]. Feature responses are monitored on collected content during tool development. Other possible feature types are kept in mind and tool development will proceed in a way that allows flexibility in feature choice. *Addressing:* Switch to or add a different, probably more complex feature type.

6. **Risk** Failure of statistical analysis to train classifiers for short event phases due to insufficient motion information. Medium severity: signifies degraded component quality if not handled.

   Handling *Avoidance:* Avoidance: Existing research has used a similar approach successfully on video events of only a few frames. [Ke05] *Addressing:* Combine subsequent short event phases into one longer phase and train a classifier for that phase.

7. **Risk** High classification error of statistical approach. Medium severity: signifies degraded component quality if not handled.

   Handling *Avoidance:* A desired positive rate and false alarm rate can be specified in this approach. Failure to reach these rates can be due to uninformative features (see above), or because the training has to be aborted prematurely, e.g. for time reasons. *Addressing:* Even a classifier with a relatively high error can be useful in the methodology as a step in reducing the amount of event phase candidates, but it will have to be supplemented by a different, possibly rule based approach.

8. **Risk** Insufficient accuracy of shot boundary detection. Medium severity: signifies degraded component quality if not handled.

Handling *Avoidance:* Several shot boundary detection algorithms can be applied and the one that performs best can be chosen. *Addressing:* The event phase detection does not have to rely on shot boundary information because it is able to locate the events in time itself. Therefore, it can be run on an unsegmented video if necessary.

9. **Risk** Missing or insufficient tool support for implementation of methodology. Medium severity: signifies schedule slip if not handled.

   Handling *Avoidance:* The tool basis has already been used and extended in previous projects, and is well known. Implementation of the methodology will incorporate the tool basis in a modular fashion. *Addressing:* Identify which support is missing or insufficient, improve or replace it.

10.3 VOCR

Regarding VOCR, the risks are given in the following:

1. **Risk** Grayscale edge map is not sufficient for text area detection.

   Handling Color information will be also involved.

2. **Risk** Noise removal and shape smoothing step is not effective.

   Handling In addition to morphological operations, masks and "shrink and swell" transformations will be studied.

3. **Risk** Text tracking fails.

   Handling Examine alternative matching criteria for motion estimation.

10.4 Audio

The main risks in audio events detection and proper names recognition are related to the events detection failure and incorrect proper names recognition. There are some reasons that lead to events detection failure and incorrect proper names recognition. One reason is a presence of other audio events in audio cocktail simultaneously with the detected events as background sounds. The same problem is for proper names recognition. The other reason is absence of the data in the training set related to the events that appear in real life audio recording. For example, absence of speech in audio corpus with emotions that appears in sport recording. Other reasons also related with specific speaker pronunciation or accent and absence of sufficient data for training and evaluation results.

1. **Risk** Failure in audio events recognition due simultaneous audio cocktail in the background.

   Handling Different audio events in cocktail sound are often not synchronized in temporal and in spectral areas. This gives possibility for applying specific method capable to work with the cocktail sound such as missing feature method, model selection and source separation.

2. **Risk** Failure in proper names recognition due to the absence in the training data pronunciation or accent of the tested speaker.

   Handling Applying pronunciation modeling that includes using different pronunciation for the same word.

3. **Risk** Failure in proper names recognition due to the absence in the training data the data with the emotions that had tested speaker.

   Handling Put in training corpus enough data to have enough coverage for all phonemes in different emotional states.

4. **Risk** Failure in proper names recognition due to the masking the speech by high intensity background noise.

   Handling Applying appropriate methods that can recognize reliable regions and handle masked feature by reconstruction using training data (missing feature method).

10.5 Text

For text-based information extraction, the following risks have been identified

1. **Risk** Lack of semantically rich textual content for certain MLCs or HLCs

   Handling Use of various sources of content to ensure the collection of adequate content

2. **Risk** Low performance in terms of recall and precision

   Handling Use of various techniques, rule-based, and machine learning based, as well as their combination in order to reach the target performances.

3. **Risk** Text-based reasoning cannot handle numeric data (e.g. performance s, ranking numbers)

   Handling One option may be for textual analysis to convert numeric data to symbolic ones (e.g. name performance ranges); another option is to examine extensions of the reasoning services in order to handle such data

4. **Risk** The performance of reasoning services is not sufficient to deal with the amount of instances coming from large documents

   Handling Reasoning services will be improved in order to exploit large structures

5. **Risk** Low performance in extraction from textual content related to or produced from other media (e.g. image captions, transcripts from speech

   Handling The source of textual content will be taken into account during textual analysis. Contextual information (either image meta-data or information extracted from the rest textual content) will be taken into account in order to improve system's performance

10.6 Reasoning

Regarding reasoning the risks are given in the following:

1. **Risk** Semantics extraction results are not properly translated to ABox assertions, which serve as input for inference services for high-level multimedia interpretation.

   Handling In a number of cases proper translation will require machine learning and data analysis techniques. The techniques of choice can be adjusted to deliver better results or different techniques can be applied.

2. **Risk** Domain model(s) are not appropriate which leads to the failure of inference services for high-level multimedia interpretation.

   Handling The domain model(s) must be modified or enhanced in close collaboration with the semantics extraction tools and the inference services in order to achieve better results.

3. **Risk** Despite suitable input (ABoxes and domain ontologies), results delivered by inference services for high-level multimedia interpretation not satisfactory.

   Handling The early evaluation of the results will help to enhance the reasoning approach.

11 Epilogue

This document has elaborated on a methodology for semantics extraction from multimedia content. We proposed a modular architecture to cope with complexity arising from processing of multiple media. Distinct modes of operation, namely (*Analysis*, *Training* and *Discovery*) have been defined to account for ontology population, adaptability to an evolving domain and active participation to its evolution. The *semantic gap* has been addressed through the definition of a set of concepts (the *mid-level* concepts) acting as an interface between modality-specific processing techniques and formal reasoning. We took account of the semantic model to guide the analysis process and to allow fusion of information stemming from multiple sources.

Furthermore, for each medium examined (image, video, audio and text) a particular methodology to allow semantics extraction on an evolving domain has been crafted. In all cases, the purpose has been defined to detect and classify particular segments of the media (such as image regions, video shots, audio segments and test phrases) to a class associated with a concept of the semantic model. To this end, particular techniques will be investigated per medium. To name a few: extraction of low-level image descriptors (such as MPEG7) followed by segmentation and holistic detection of primitives for image processing, statistical analysis of motion patterns and matching of object trajectories in video processing, blind source separation and inverse search dictionary for audio/speech recognition, and document-level named-entity recognition for text processing. For illustration purposes, an example from the athletics domain has been used to show the expected output of each of the methodologies in a non-trivial use-case. As a special task, investigation of algorithms for detecting text in video has also been foreseen.

Semantics Extraction from Multimedia Content in an evolving domain is an ambitious task. Many issues, that seem trivial to the unaware reader, are a real challenge given the current state of the art in specific medium processing techniques. In most parts of the methodology, alternative per medium algorithms have been described to decrease the risk of inadequate extraction results. Furthermore, we also expect to improve the overall information extraction results based on complementarity and redundancy of information found in each media. Nevertheless, we will consider updating and improving the methodology as needed, based on feedback from the early versions of the respective semantics extraction toolkits.

References

[Achi91] E. Achilles, B. Hollunder, A. Laux and J.-P. Mohren, "KRIS: Knowledge Representation and Inference System – User guide", *Technical Report D91-14*, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), 1991.

[Agir00] E. Agirre, O. Ansa, E. Hovy and D. Martinez, "Enriching Very Large Ontologies Using the WWW", *In Workshop on Ontology Construction of the European Conference of Artificial Intelligence* (ECAI-00), 2000.

[Ahma94] K. Ahmad, A. Davies, H. Fulford and M. Rogers. *What is a term? The Semi Automatic Extraction of Terms from Text*. Amsterdam: John Benjamins Publishing Company, 1994.

[Alva04] C. Alvarez, A. Oumohmed, M. Mignotte, and J.-Y. Nie, " Toward Cross- Language and Cross-Media Image Retrieval", *Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum. CLEF 2004*, Springer: Proceedings Multilingual Information Access for Text, Speech and Images, Vol. 3491 of LNCS, pp. 676687, Bath, UK, September 2004.

[Baad91a] F. Baader and P. Hanschke, "A schema for integrating concrete domains into concept languages", *In Proc. of the 12th Int. Joint Conf. on Artificial Intelligence* (IJCAI'91), pp. 452–457, 1991.

[Baad91b] F. Baader and B. Hollunder, "KRIS: Knowledge Representation and Inference System", *SIGART Bulletin*, vol. 2(3), pp. 8–14, 1991.

[Baad92a] F. Baader and P. Hanschke, "Extensions of concept languages for a mechanical engineering application", *In Proc. of the 16th German Workshop on Artificial Intelligence (GWAI'92), volume 671 of Lecture Notes in Computer Science*, pp. 132–143. Springer Verlag, 1992.

[Baad92b] F. Baader, B. Hollunder, B. Nebel, H.-J. Profitlich and E. Franconi, "An empirical analysis of optimization techniques for terminological representation systems", *In Proc. of the 3rd Int. Conf. on the Principles of Knowledge Representation and Reasoning* (KR'92), pp. 270–281. Morgan Kaufmann Publishers, 1992.

[Baad93] F. Baader and B. Hollunder, "Cardinality restrictions on concepts", *Technical Report RR-03-48*, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), 1993.

[Baad94] F. Baader, E. Franconi, B. Hollunder, B. Nebel, and H.-J. Profitlich, "An empirical analysis of optimization techniques for terminological representation systems or: Making KRIS get a move on", *Applied Artificial Intelligence. Special Issue on Knowledge Base Management*, vol. 4, pp. 109–132, 1994.

[Baad96] F. Baader, M. Buchheit and B. Hollunder, "Cardinality restrictions on concepts", *Artificial Intelligence*, vol. 88(1–2), pp. 195–213, 1996.

[Baad00] F. Baader, R. Küsters and R. Molitor, "Rewriting concepts using terminologies", *In Proceedings of the 7th International Conference on Knowledge Representation and Reasoning (KR2000)*, pp. 297–308, San Francisco, CA, 2000. Morgan Kaufmann Publishers.

[Baad03] F. Baader and W. Nutt, "Basic description logics", *In Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors, The Description Logic Handbook: Theory, Implementation, and Applications*, pp. 43–95. Cambridge University Press, 2003.

[Bass06] Bass, Len; Nord, Robert; Wood, William; Zubrow, David. "Risk Themes Discovered Through Architecture Evaluations (CMU/SEI-2006-TR-012)". Technical Report. Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 2006.

[Bech03] S. Bechhofer, V. Haarslev, R. Möller and P. Crowther, "The dig description logic interface", *In Proceedings of the International Workshop on Description Logics* (DL-2003), Rome, Italy, 2003.

[Belo02] M. S. Belongie and J. Puzicha, "Shape matching object recognition using shape contexts", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*,vol. 24(24), pp. 509 – 522, 2002.

[Bera01] D. Berardi, D. Calvanese and G. De Giacomo, "Reasoning on uml class diagram using description logic based system", *In Proc. of the KI-2001 Workshop W6 on Applications of Description Logics ADL-01*, vol. 44. http://ceur-ws.org/Vol-44/, 2001.

[Biatov04] K. Biatov, "An Extraction of Speech Data from Audio Stream Using Unsupervised Pre-Segmentation", in Proc. UkrObraz-04, Kiev, Ukraine, 2004.

[Biatov05] K. Biatov and M. Larson, "Speaker Clustering via Bayesian Information Criterion using a Global Similarity Constraint", in Proc. SPECOM 2005, Patras, Greece, 2005.

[Biatov06] K. Biatov and J. Köhler, "Improvement Speaker Clustering Using Global Similarity Features", in Proc. ICSLP 2006, Pittsburgh, USA, 2006.

[Bres95] P. Bresciani, E. Franconi and S. Tessaris, "Implementing and testing expressive description logics: Preliminary report", *In Proc. of the 1995 Description Logic Workshop* (DL'95), pp. 131–139, 1995.

[Bont02] K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan and H. Cunningham, "Shallow Methods for Named Entity Coreference Resolution", *In Proceedings of TALN 2002 Workshop* "Chaines de references et resolveurs d' anaphores", Nancy, France, 24-27 June, 2002.

[Bril95] E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging", *Computational Linguistics*, vol. 21, 1995.

[Buch93a] M. Buchheit, F. M. Donini and A. Schaerf, "Decidable reasoning in terminological knowledge representation systems", *In Proc. of the 13th Int. Joint Conf. on Artificial Intelligence* (IJCAI'93), pp. 704–709, 1993.

[Buch93b] M. Buchheit, F. M. Donini and A. Schaerf, "Decidable reasoning in terminological knowledge representation systems", *Journal of Artificial Intelligence Research*, vol. 1, pp. 109–138, 1993.

[Cali05] J. Calic, N. Campbell, S. Dasiopoulou and Y. Kompatsiaris, "An Overview Of Multimodal Video Representation For Semantic Analysis", *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies (EWIMT 2005)*, IEE, December 2005.

[Calv98] D. Calvanese, G. De Giacomo and M. Lenzerini, "On the decidability of query containment under constraints", *In Proc. of the 17th ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems* (PODS'98), pp. 149–158, 1998.

[Cann86] J. Canny, "A Computational Approach To Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679-714, 1986.

[Cast05] S. Castano and A. Ferrara and S. Montanelli, "Matching Ontologies in Open Networked Systems: Techniques and Applications", *Journal on Data Semantics* (JoDS), 2005.

[Chan01] S. F. Chang, T . Sikora and A. Puri, "Overview of the MPEG-7 standard. Special issue on MPEG-7", *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), pp. 688-695, 2001.

[Chen05] Chen, Lei; zsu, M. Tamer; Oria, Vincent. Robust and Fast Similarity Search for Moving Object Trajectories. In Proc. 2005 ACM SIGMOD Int'l Conf. Management of Data, 491?502, 2005.

[Chen98] S. Chen and P. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with the applications in speech recognition", in Proc. ICASSP 98, 1998.

[Ciep01] L. Cieplinski, W. Kim, J.-R. Ohm, M. Pickering and A. Yamada, "MPEG-7 Visual part of experimentation model version 11.1", *Tech. Rep. M7691, ISO/IEC JTC1/SC29/WG11*, 2001.

[Dam93] F.J. Damerau. Evaluating domain-oriented multiword terms from texts. Information Processing and Management, 29(4):433–447, 1993.

[Dala05] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", *International Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 886-893, June 2005.

[Dasi05] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V-K Papastathis and M. Strintzis, "Knowledge-assisted Semantic Video Object Detection", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15, No. 10, pp. 1210-1224, 2005.

[Daub88] I. Daubechies, "Orthonormal bases of compactly supported wavelets", *Communications on Pure and Applied Mathematics*, vol. XLI, no. 41, pp. 909-996, 1988.

[DelBi99] Del Bimbo, Alberto. Visual Information Retrieval. Morgan Kaufmann, San Francisco, 1999.

[Dese05a] T. Deselaers, T. Weyand, D. Keysers, W. Macherey and H. Ney, "FIRE in ImageCLEF 2005: Combining Content-based Image Retrieval with Textual Information Retrieval", *in Proceedings of the Tof the CLEF Workshop*, NIST Special Publication, Vienna, Austria, September 2005.

[Dese05b] T. Deselaers, D. Keysers, and H. Ney, "Improving a Discriminative Approach to Object Recognition using Image Patches", *In DAGM 2005, Pattern Recognition, 27th DAGM Symposium, Vienna*, Austria, LNCS volume 3663, pages 326-333, August/September 2005.

[Dimi02] M. Dimitrov, K. Bontcheva, H. Cunningham and D. Maynard, "A Light-weight Approach to Coreference Resolution for Named Entities in Text", *In A. Branco, T. McEnery and R. Mitkov (eds.), Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, 2004.

[Dix03] A. Dix, J. Finlay, G. Abowd and R. Beale. *Human-Computer Interaction*. Prentice Hall, 3rd edition, ISBN:0130461091. 2003.

[Doni91] F. M. Donini, M. Lenzerini, D. Nardi and W. Nutt, "The complexity of concept languages", *In Proc. of the 2nd Int. Conf. on the Principles of Knowledge Representation and Reasoning* (KR'91), pp. 151–162, 1991.

[Doni03] F. M. Donini, "Complexity of reasoning", *In Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors, The Description Logic Handbook: Theory, Implementation, and Applications*, pp. 96–136. Cambridge University Press, 2003.

[Drap92a] B. Draper, A. Hanson and E.Riseman, "Learning knowledge-directed visual strategies", *In Proceedings of DARPA Image Understanding Workshop*, pp. 933-940, 1992.

[Drap92b] B. Draper and A. Hanson, "An example of learning in knowledge-directed vision", *In P. Johansen and S. Olsen, editors, Theory and Applications of Image Analysis*, pp. 237–252, World Scientific, 1992.

[Dubo93] D.Dubois, H.Prade and R.R. Yager. *Readings in Fuzzy Sets for Intelligent Systems*. Morgan Kaufmann, San Mateo, 1993.

[Duyg02] P. Duygulu, K Barnard, N de Fretias and D Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary", *In Proceedings of the European Conference on Computer Vision*, pp. 97-112, 2002.

[Eick00] Eickeler, Stefan; Rigoll, Gerhard. "A Novel Measure for the Evaluation of Video Indexing Systems". In Proc. 2000 IEEE Intl'l Conf. Acoustics, Speech, and Signal Processing, 1991?1994, 2000.

[Eickeler03] S. Eickeler, K. Biatov, M. Larson and J. Koehler, "Two Novel Application of Speech Recognition for Robust Spoken Document Retrieval", Workshop on Multimedia Content in Digital Libraries, Greece, 2003.

[Eide03] H. Eidenberger, "How good are the visual MPEG-7 features?" *In Proceedings of the SPIE Visual Communications and Image Processing Conference*, Lugano SPIE Vol. 5150 pp. 476-488, 2003 (online: http://www.ims.tuwien.ac.at/ hme/papers/vcip2003-mpeg7.pdf)

[Eide04] H. Eidenberger, "Statistical analysis of content-based MPEG-7 descriptors for image retrieval", *Multimedia Systems*, 10(2), pp. 84-97, August 2004.

[Eite97] T. Eiter, G. Gottlob and H. Mannilla, "Disjunctive Datalog", *ACM Transactions on Database Systems*, vol. 22(3), pp. 364–418, 1997.

[Faat02] A. Faatz and R. Steinmetz, "Ontology Enrichment with texts from the WWW", *In Semantic Web Mining 2nd Workshop at ECML/PKDD*-2002, Helsinki, Finland, 2002.

[Felz05] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition", *International Journal of Computer Vision*(IJCV), vol. 61(1), pp. 55–79, 2005.

[Fort05] B. Fortuna, D. Mladevic, and M. Grobelnik, "Visualization of Text Document Corpus", *In ACAI 2005 Summer School*, 2005.

[Fran00] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: The c-value/nc-value method", *International Journal on Digital Libraries*, vol. 3(2), pp. 115–130, 2000.

[Free95] J. W. Freeman. *Improvements to Propositional Satisfiability Search Algorithms*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, 1995.

[Gato05] B. Gatos, I. Pratikakis and S.J. Perantonis "Text detection in indoor/outdoor scene images", *In Proceedings of the first International Workshop on Camera-based Document Analysis and Recognition* (CBDAR'05), pp. 127-132, Seoul, Korea, August 2005.

[Gato06] B. Gatos, I. Pratikakis and S. J. Perantonis, "Adaptive Degraded Document Image Binarization", *Pattern Recognition*, vol. 39, pp. 317-327, 2006.

[Glim05] B. Glimm and I. Horrocks, "Handling cyclic conjunctive queries", *In Proc. of the 2005 Description Logic Workshop* (DL 2005), volume 147. CEUR (http://ceur-ws.org/Vol-147/), 2005.

[Gonz02] R. C. Gonzalez, R. E. Woods. *Digital Image Processing*. 2nd edition, Prentice Hall Inc, NJ, 2002, ISBN: 0-13-094650-8.

[Gupt02] K.M. Gupta, D. Aha, E. Marsh, and T. Maney, "An Architecture for engineering sublanguage WordNets", *In Proceedings of the First International Conference On Global WordNet*, Mysore, India: Central Institute of Indian Languages, pp. 207–215, 2002.

[Haar00] V. Haarslev and R. Möller, "Expressive ABox reasoning with number restrictions, role hierarchies, and transitively closed roles", *In Proc. of the 7th Int. Conf. on Principles of Knowledge Representation and Reasoning* (KR 2000), pp. 273–284, 2000.

[Haar01a] V. Haarslev and R. Möller, "Combining tableaux and algebraic methods for reasoning with qualified number restrictions", *In Proc. of the 2001 Description Logic Workshop* (DL 2001), pp. 152–161. CEUR Electronic Workshop Proceedings, http://ceur-ws.org/Vol-49/, 2001.

[Haar01b] V. Haarslev and R. Möller, "High performance reasoning with very large knowledge bases: A practical case study", *In Proc. of the 17th Int. Joint Conf. on Artificial Intelligence* (IJCAI 2001), pp. 161–168, 2001.

[Haar01c] V. Haarslev and R. Möller, "Optimizing reasoning in description logics with qualified number restrictions", *In Proc. of the 2001 Description Logic Workshop* (DL 2001), pp. 142–151. CEUR Electronic Workshop Proceedings, http://ceur-ws.org/Vol-49/, 2001.

[Haar01d] V. Haarslev and R. Möller, "RACER system description", *In Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2001), volume 2083 of Lecture Notes in Artificial Intelligence*, pp. 701–705. Springer Verlag, 2001.

[Haar01e] V. Haarslev, R. Möller and A.-Y. Turhan, "Exploiting pseudo models for tbox and abox reasoning in expressive description logics", *In Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2001), volume 2083 of Lecture Notes in Artificial Intelligence*, Springer Verlag, 2001.

[Haar01f] V. Haarslev, R. Möller and M. Wessel, "The description logic $ALCNH_{R+}$ extended with concrete domains: A practically motivated approach", *In Proc. of the Int. Joint Conf. on Automated Reasoning* (IJCAR 2001), pp. 29–44, 2001.

[Haar04] V. Haarslev and R. Möller, "Optimization techniques for retrieving resources described in OWL/RDF documents: First results", *In Proc. of the 9th Int. Conf. on Principles of Knowledge Representation and Reasoning* (KR 2004), 2004.

[Haas05a] , P. Haase and L. Stojanovic. *Consistent Evolution of OWL Ontologies*. pp. 182–197. Springer Verlag, LNCS 3532, 2005.

[Haas05b] P. Haase, F. Van Harmelen, Z. Huang, H. Stuckenschmidt and Y. Sure, "A Framework for Handling Inconsistency in Changing Ontologies", *In Proceedings of the 2005 International Semantic Web Conference*, ISWC2005, 2005.

[Haas05c] P. Haase and J. Volker, "Ontology Learning and Reasoning- Dealing with Uncertainty and Inconsistency", *In Proceedings of the Workshop on Uncertainty Reasoning for the Semantic Web* (URSW), 2005.

[Hess06] Hesseler, Wolfgang; Eickeler, Stefan. "MPEG-2 Compressed-Domain Algorithms for Video Analysis". In EURASIP J. Applied Signal Processing, 2006.

[Holl90] B. Hollunder and W. Nutt, "Subsumption algorithms for concept languages", *Technical Report RR-90-04*, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Kaiserslautern (Germany), 1990.

[Holl91a] B. Hollunder and F. Baader, "Qualifying number restrictions in concept languages", *In Proc. of the 2nd Int. Conf. on the Principles of Knowledge Representation and Reasoning* (KR'91), pp. 335–346, 1991.

[Holl91b] B. Hollunder, A. Laux, H.-J. Profitlich, and T. Trenz, "KRIS-manual", *Technical report*, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), 1991.

[Holl94] B. Hollunder. *Algorithmic Foundations of Terminological Knowledge Representation Systems*. PhD thesis, University of Saarbrücken, Department of Computer Science, 1994.

[Horr97] I. Horrocks. *Optimising Tableaux Decision Procedures for Description Logics*. PhD thesis, University of Manchester, 1997.

[Horr98a] I. Horrocks, "The FaCT system", *In Harrie de Swart, editor, Proc. of the 7th Int. Conf. on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX'98)*, vol. 1397 of Lecture Notes in Artificial Intelligence, pp. 307–312. Springe Verlag, 1998.

[Horr98b] I. Horrocks, "Using an expressive description logic: FaCT or fiction?", *In Proc. of the 6th Int. Conf. on Principles of Knowledge Representation and Reasoning* (KR'98), pp. 636–647, 1998.

[Horr99] I. Horrocks, U. Sattler and S. Tobies, "Practical reasoning for expressive description logics", *In Harald Ganzinger, David McAllester, and Andrei Voronkov, editors, Proc. of the 6th Int. Conf. on Logic for Programming and Automated Reasoning* (LPAR'99), number 1705 in Lecture Notes in Artificial Intelligence, pp. 161–180. Springer Verlag, 1999.

[Horr00a] I. Horrocks, U. Sattler and S. Tobies, "Reasoning with individuals for the description logic SHIQ", *In David McAllester, editor, Proc. of the 17th Int. Conf. on Automated Deduction* (CADE 2000), volume 1831 of Lecture Notes in Computer Science, pp. 482–496. Springer Verlag, 2000.

[Horr00b] I. Horrocks and S. Tessaris, "A conjunctive query language for description logic ABoxes", *In Proc. of the 17th Nat. Conf. on Artificial Intelligence* (AAAI 2000), pages 399–404, 2000.

[Horr00c] I. Horrocks and S. Tobies, "Reasoning with axioms: Theory and practice", *In Proc. of the 7th Int. Conf. on Principles of Knowledge Representation and Reasoning* (KR 2000), pp. 285–296, 2000.

[Horr01] I. Horrocks and U. Sattler, "Ontology reasoning in the SHOQ(D) description logic", *In Proc. of the 17th Int. Joint Conf. on Artificial Intelligence* (IJCAI 2001), pp. 199–204, 2001.

[Horr02] I. Horrocks and U. Sattler, "Optimised reasoning for SHIQ", *In Proc. of the 15th Eur. Conf. on Artificial Intelligence* (ECAI 2002), pp. 277–281, July 2002.

[Horr05] I. Horrocks and U. Sattler, "A tableaux decision procedure for SHOIQ", *In Proc. of the 19th Int. Joint Conf. on Artificial Intelligence* (IJCAI 2005), pp. 448–453, 2005.

[Horr06] I. Horrocks, O. Kutz and U. Sattler, "The even more irresistible SROIQ", *In Proc. of the 13th Int. Conf. on Principles of Knowledge Representation and Reasoning* KR-06, 2006.

[Hunt01] J. Hunter, "Adding multimedia to the Semantic Web – Building an Mpeg-7 ontology", *In Proceedings of the International Semantic Web Working Symposium*, Stanford University, California, USA, 2001.

[Hust04] U. Hustadt, B, Motik and U. Sattler, "Reducing SHIQ-description logic to disjunctive datalog programs", *In Proc. of the 9th Int. Conf. on Principles of Knowledge Representation and Reasoning* (KR 2004), pp. 152–162, 2004.

[ISO] International Organization for Standardization, Coding of Moving Pictures and Audio, ISO/IEC JTC1/SC29/WG11N6828, "MPEG-7 Overview", online at: http://www.chiariglione.org/MPEG/standards/mpeg-7/mpeg-7.htm.

[Kaly05] A. Kalyanpur, B. Parsia, E. Sirin and J. Hendler, "Debugging unsatisfiable classes in owl ontologies", *Journal of Web Semantics - Special Issue of the Semantic Web Track of WWW2005*, vol. 3(4), 2005.

[Ke05] Ke, Yan; Sukthankar, Rahul; Hebert, Martial. "Efficient Visual Event Detection Using Volumetric Features". In Proc. 10th IEEE Int'l Conf. Computer Vision, 166?173, 2005.

[Kiet00] J.U. Kietz, A. Maedche, and R. Volz, "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.

[Kosk02] M. Koskela, J. Laaksonen and E. Oja, "Using MPEG-7 Descriptors in Image Retrieval with Self-Organizing Maps", *In Proceedings of the 16th Interantional Conference on Pattern Recognition*, vol. 2, pp. 1049-1052, 2002.

[Koti06] K. Kotis, G. A. Vouros, K. Stergiou, "Towards Automatic Merging of Domain Ontologies: The HCONE-merge approach", *Journal of Web Semantics*, (2006).

[Laak04] J. Laaksonen, M. Koskela and E. Oja, "Class distributions on SOM surfaces for feature extraction and object retrieval", *Neural Networks*, 17(8-9), pp. 1121-1133, 2004.

[Lamb96] P. Lambrix. *Part-Whole Reasoning in Description Logic*. PhD thesis, Dep. of Computer and Information Science, Linköping University, Sweden, 1996.

[Lavr03] V. Lavrenko, R Manmatha, and J Jeon, "A model for learning the semantics of pictures", *In Proceedings of the 16th Conference on Advances in Neural Information Processing Systems NIPS*, 2003.

[Lenz91] M. Lenzerini and A. Schaerf, "Concept languages as query languages", *In Proc. of the 9th Nat. Conf. on Artificial Intelligence* (AAAI'91), pp. 471–476, 1991.

[Li03] R. Li and W. K. Leow, "From region features to semantic labels: A probabilistic approach", *In Proceedings of the International Conference on Multimedia Modeling*, pp 402-420, 2003.

[Liu06] H. Liu, C. Lutz, M. Milicic and F. Wolter, "Description logic actions with general TBoxes: a pragmatic approach", *In Proceedings of the 2006 International Workshop on Description Logics* (DL 2006), 2006. [Lowe-99] David

[Lowe99] G. Lowe, "Object recognition from local scale-invariant features", *International Conference on Computer Vision*, Corfu, Greece, pp. 1150-1157.

[Lowe04] D. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision* (IJCV), vol. 60(2), pp. 91–110, 2004.

[Lutz99a] C. Lutz, "The complexity of reasoning with concrete domains", *LTCS-Report 99-01, LuFg Theoretical Computer Science*, RWTH Aachen, Germany, 1999.

[Lutz99b] C. Lutz, "Complexity of terminological reasoning revisited", *In Proc. of the 6th Int. Conf. on Logic for Programming and Automated Reasoning (LPAR'99), volume 1705 of Lecture Notes in Artificial Intelligence*, pp. 181–200, Springer Verlag, 1999.

[Lutz99c] C. Lutz, "Reasoning with concrete domains", *In Proc. of the 16th Int. Joint Conf. on Artificial Intelligence* (IJCAI 1999), pp. 90–95, 1999.

[Lutz01a] C. Lutz, "Interval-based temporal reasoning with general TBoxes", *In Proc. of the 17th Int. Joint Conf. on Artificial Intelligence* (IJCAI 2001), pp. 89–94, 2001.

[Lutz01b] C. Lutz, "NEXPTIME-complete description logics with concrete domains", *In Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2001), volume 2083 of Lecture Notes in Artificial Intelligence*, pp. 45–60. Springer Verlag, 2001.

[Lutz03] C. Lutz, "Description logics with concrete domains: A survey", *In Philippe Balbiani, Nobu-Yuki Suzuki, Frank Wolter, and Michael Zakharyaschev, editors, Advances in Modal Logics*, volume 4. King's College Publications, 2003.

[Lutz04] C. Lutz, "Nexptime-complete description logics with concrete domains", *ACM Transactions on Computational Logic*, vol. 5(4), pp. 669–705, 2004.

[Lutz05] C. Lutz and M. Milicic, "A tableau algorithm for description logics with concrete domains and gcis", *In Proceedings of the 14th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods* (TABLEAUX 2005), LNAI, Koblenz, Germany, 2005. Springer.

[Mao92] J. Mao, A.K. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models", *Pattern Recognition* 25, pp. 173–188, 1992.

[Metz04] D. Metzler and R Manmatha, "An inference network approach to image retrieval", *In Proceedings of the International Conference on Image and Video Retrieval*, pp. 42-50, 2004.

[Meye90] F. Meyer and S. Beucher, "Morphological segmentation", *Journal of Visual Communication and Image Representation*, no. 1, vol. 1, Oct. 1990.

[Meza04] V. Mezaris, I. Kompatsiaris and M. Strintzis, "Region-based Image Retrieval using an Object Ontology and Relevance Feedback", *EURASIP Journal on Applied Signal Processing*, vol. 2004, No. 6, pp. 886-901, 2004.

[Moll00] R. Möller. *Expressive description logics: Foundations for applications*, 2000. Habilitation Thesis.

[Moll03] R. Möller and V. Haarslev, "Description logic systems" *In Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors, The Description Logic Handbook: Theory, Implementation and Applications*, chapter 8, pp. 282–305. Cambridge University Press, 2003.

[Moll06] R. Möller, V. Haarslev and M. Wessel, "On the scalability of description logic instance retrieval", *In Chr. Freksa, Kohlhase, and K. M. Schill, editors, 29. Deutsche Jahrestagung für Künstliche Intelligenz, Lecture Notes in Artificial Intelligence*, Springer Verlag, 2006.

[Moti06] B. Motik. *Reasoning in Description Logics using Resolution and Deductive Databases*. PhD thesis, Univ. Karlsruhe, 2006.

[Nebe90] B. Nebel, "Terminological reasoning is inherently intractable", *Artificial Intelligence*, vol. 43, pp. 235–249, 1990.

[Nech91] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senatir, and W.R. Swartou, "Enabling technology for knowledge sharing", *AI Magazine*, Vol. 12, No. 3, Fall 1991.

[Ng02] V. Ng and C. Cardie, "Improving Machine Learning Approaches to Coreference Resolution", *In Proceedings of the 40^{th} Annual Meeting of the Association for Computational Linguistics* (ACL), Philadelphia, July 2002, pp. 104-111.

[Ohlb99] H. J. Ohlbach and J. Koehler, "Modal logics, description logics and arithmetic reasoning", *Artificial Intelligence*, vol. 109(1-2), pp. 1–31, 1999.

[Pear88] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of Plausible Beliefs*. Morgan Kaufmann, San Mateo, 1988.

[Pert03] M. Pertselakis, N. Tsapatsoulis, S. Kollias and A. Stafylopatis "An Adaptive Resource Allocating Neural Fuzzy Inference System", *in Proceedings of the IEEE Intelligent Systems Application to Power Systems* (ISAP'03), Lemnos, Greece, August 2003.

[Polk96] L. Polkowski and A. Skowron, "Rough mereology: A new paradigm for approximate reasoning", *International Journal of Approximate Reasoning*, vol. 15(4), pp. 333–365, 1996.

[Plat91] J. Platt, "A resource-allocating network for function interpolation", *Neural Computing*, vol. 3, no. 2, pp. 213-225, 1991.

[Prat06a] I. Pratikakis, B. Gatos and S. Thomopoulos, "Scene categorisation using low-level visual features", *International Conference on Computer Vision Theory and Applications* (VISAPP'06), Setubal, Portugal, Feb. 25-28, 2006, ISBN: 972-8865-40-6, pp. 155-160, 2006.

[Prat06b] I. Pratikakis, I. Vanhamel, H, Sahli, B, Gatos and S. Perantonis, "Unsupervised watershed-driven region-based image retrieval", *IEE Proceedings on Vision, Image and Signal Processing, Special Issue on Knowledge-based Digital Media Processing*, vol. 153, Issue 3, pp. 313-322, 2006.

[Salt75] A. Saltion, G.Wong and C.S. Yang, "A vector space model for automatic indexing", *Communications of the ACM*, vol. 18(11), pp. 613–620, 1975.

[Satt96] U. Sattler, "A concept language extended with different kinds of transitive roles", *In Günter Görz and Steffen Hölldobler, editors, Proc. of the 20th German Annual Conf. on Artificial Intelligence (KI'96), number 1137 in Lecture Notes in Artificial Intelligence*, pp. 333–345, Springer Verlag, 1996.

[Satt98] U. Sattler. *Terminological Knowledge Representation Systems in a Process Engineering Application*. PhD thesis, LuFG Theoretical Computer Science, RWTH-Aachen, Germany, 1998.

[Scha93] A. Schaerf, "Reasoning with individuals in concept languages", *In Proceedings of the 3rd Conference of the Ital. Assoc. for Artificial Intelligence (AI*IA'93), Lecture Notes in Artificial Intelligence*, Springer Verlag, 1993.

[Schm91] M. Schmidt-Schauß and G. Smolka, "Attributive concept descriptions with complements", *Artificial Intelligence*, vol. 48(1), pp. 1–26, 1991.

[Schn04] H. Schneiderman, "Feature-centric evaluation for efficient cascaded object detection", *International Conference on Computer Vision and Pattern Recognition* (CVPR), 2004.

[Siko95] T. Sikora and B. Makai, "Shape-adaptive DCT for generic coding of video", *IEEE Transactions on Circuits Systems for Video Technology*, vol. 5, pp. 59-62, Feb. 1995.

[Siri06] E. Sirin, B. C. Grau and B. Parsia, "From wine to water: Optimizing description logic reasoning for nominals", *In International Conference on the Principles of Knowledge Representation and Reasoning* (KR-2006), 2006.

[Smit01] J. R. Smith, S. Srinivasan, A. Amir, S. Basu, G. Iyengar, C. Lin, M. Naphade, D. Poncelon and B. Tseng, "Integrating features, models, and semantics for trec video retrieval", *in Proceedings of the Tenth Text REtrieval Conference (TREC10)*, NIST Special Publication, 2001.

[Snoe05] C. G.M. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art", *Multimedia Tools and Applications*, 25 (1), pp. 5-35, Jan. 2005.

[Spee95] P.-H. Speel, F. van Raalte, P. van der Vet and N. J. I. Mars, "Runtime and memory usage performance of description logics", *In G. Ellis, R. A. Levinson, A. Fall, and V. Dahl, editors, Knowledge Retrieval, Use and Storage for Efficiency: Proc. of the 1st Int. KRUSE Symposium*, pp. 13–27, 1995.

[SP06] Special Issue: Semantic Retrieval of Multimedia, *IEEE Signal Processing Magazine*, vol. 23, no. 2, March 2006.

[Tobi01] S. Tobies. *Complexity Results and Practical Algorithms for Logics in Knowledge Representation*. PhD thesis, LuFG Theoretical Computer Science, RWTH-Aachen, Germany, 2001.

[Tsar04] D. Tsarkov and I. Horrocks, "Efficient reasoning with range and domain constraints", *In Proceedings of the 2004 Description Logic Workshop* (DL 2004), vol. 104, CEUR (http://ceur-ws.org/), 2004.

[Tsar05a] D. Tsarkov and I. Horrocks, "Optimised classification for taxonomic knowledge bases", *In Proceedings of the 2005 Description Logic Workshop* (DL 2005), vol. 147, CEUR (http://ceur-ws.org/), 2005.

[Tsar05b] D. Tsarkov and I. Horrocks, "Ordering heuristics for description logic reasoning", *In Proc. of the 19th International Joint Conference on Artificial Intelligence* (IJCAI 2005), pp. 609–614, 2005.

[Tsar06] D. Tsarkov and I. Horrocks, "FaCT++ Description Logic Reasoner: System Description", *In Proceedings of the International Joint Conference on Automated Reasoning* (IJCAR 2006), 2006.

[Unse95] M. Unser, "Texture Classification and Segmentation Using Wavelet Frames," *IEEE Trans. Image Processing*, vol. 4, no. 11, pp. 1549-1560, 1995.

[Vail01] A. Vailaya, M. A. T Figueiredo, A. K. Jain and H.-J. Zhang, "Image classification for content-based indexing," *IEEE Transactions on Image Processing*, vol. 10, No. 1, pp. 117-130, 2001.

[Vala04] A. Valarakos, G. Paliouras, V. Karkaletsis and G. Vouros, "A Name-Matching Algorithm for Supporting Ontology Enrichment", *In Proceedings of the Pan-hellenic Conference in Artificial Intelligence* (SETN), Lecture Notes in Artificial Intelligence, n. 3025, pp. 381-389, Springer Verlag, 2004.

[Vali84] L. Valiant, "A theory of the learnable", *Communications of the ACM*, vol. 27, 1984.

[Viol01] P. Viola and M. Jones, "Robust Real-time Object Detection", *ICCV 2001 – 2^{nd} Int. Workshop on Statistical and Computation Theories of Vision–modeling, learning, computing and sampling*, Vancouver, Canada, July 2001.

[Viola] Viola, Paul; Jones, Michael. "Rapid object detection using a boosted cascade of simple features", *In Proc. 2001 IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[Wall05] M. Wallace, N. Tsapatsoulis, S. Kollias, "Intelligent Initialization of Resource Allocating RBF Networks", *Neural Networks*, vol. 18 (2), pp. 117-122, 2005.

[Yavl05] A. Yavlinsky, E Schofield and S Ruger, " Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation", *International Conference on Image and Video Retrieval*, CIVR05, Singapore, July 2005.

[Yi03] H. Yi, D. Rajan and L.-T. Chia, "Automatic Generation of MPEG-7 Compliant `xml` Document for Motion Trajectory Descriptor in Sports Video". In Proc. First ACM Int'l Workshop Multimedia Databases, 10?17, 2003.

[Zaan04] M. van Zaanen, G. de Croon. Multi-model Information Retrieval Using FINT.", *in Proceedings of the 5^{th} Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, Springer: Multilingual Information Access for Text, Speech and Images, Vol. 3491 of LNCS, pp. 728739, 2004.

A Description Logics: The $\mathcal{SH}$ Family

Ontology languages such as OWL, which has been adopted in BOEMIE, are often based on description logics (DLs). Since the beginning of description logic research, the tradeoff between expressivity of the logic and complexity of decision problems has been investigated in the research community [Doni03, Doni91]. DLs are distinguished by the set of concept and role constructors they provide. The $\mathcal{SH}$ family of description logic is inspired by demands from applications in the sense that the language facilities of description logics of this family are designed in such a way that (i) application requirements w.r.t. expressivity can be fulfilled and (ii) reasoners can be built that are efficient in the average case. As we will see in this section, work on ontology languages based on the $\mathcal{SH}$ family revealed that there are intricate interactions between language constructs required for practical applications. Thus, reasoning gets harder (from a theoretical point of view) if the expressivity is increased. We start our discussion of expressive description logics of the $\mathcal{SH}$ family with the core logic $\mathcal{ALC}$.

A.1 The Foundation: $\mathcal{ALC}$

The term "expressive description logic" is usually used for a logic that provides for full negation. A cornerstone logic with this feature is $\mathcal{ALC}$ [Baad03], which is the backbone of all logics of the $\mathcal{SH}$ family. Let A be a concept name and R be a role name. Then, the set of $\mathcal{ALC}$ concepts (denoted by C or D) is inductively defined as follows:

$$C, D \rightarrow A|\neg C|C \sqcap D|C \sqcup D|\forall R.C|\exists R.C$$

Concepts can be written in parentheses. The semantics is defined in the standard form using a Tarskian interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ such that

- $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}, R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$
- $(\neg C)^{\mathcal{I}} = \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$ (complement)
- $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$ (conjunction)
- $(C \sqcup D)^{\mathcal{I}} = C^{\mathcal{I}} \cup D^{\mathcal{I}}$ (disjunction)
- $(\exists R.C)^{\mathcal{I}} = \{x \in \Delta^{\mathcal{I}} | \exists y.(x.y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$ (existential restriction)
- $(\exists R.C)^{\mathcal{I}} = \{x \in \Delta^{\mathcal{I}} | \forall y.(x.y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}}\}$ (value restriction)

A concept C is satisfiable if there exists an interpretation such that $C^{\mathcal{I}} \neq 0$. The concept satisfiability problem of the logic $\mathcal{ALC}$ was shown to be PSPACE-complete in [Schm91]. For $\mathcal{ALC}$ the finite model property holds.

A (generalised) terminology (also called TBox, $\mathcal{T}$) is a set of axioms of the form $C \sqsubseteq D$ (generalised concept inclusion, GCI). A GCI is satisfied by an interpretation $\mathcal{I}$ if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. An interpretation which satisfies all axioms of a TBox called a model of the TBox. A concept C is satisfiable w.r.t. a TBox if there exists a model $\mathcal{I}$ of $\mathcal{T}$ such that $C^{\mathcal{I}} \neq \emptyset$. A concept D subsumes a concept C w.r.t. a TBox if for all models $\mathcal{I}$ of the TBox it holds that $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. D is called the subsumer, C is the subsumee. A concept name A_1 mentioned in a TBox called a most-specific subsumer of a concept name A_2 (mentioned in the TBox and different from A_1) if A_1 subsumes A_2 and there is no other concept name A_3 (mentioned in the TBox and different from A_1 and A_2) such that A_1 subsumes A_3 and A_3 subsumes A_2. The least general subsumee of a concept name is defined analogously. The classification problem for a TBox is to find the set of most-specific subsumers of every concept name mentioned in the TBox (or knowledge base). The induced graph is called the subsumption hierarchy of the TBox.

The semantics for generalised terminological axioms (which do not impose any restrictions on the concepts on both sides of a GCI) is called descriptive semantics (see, e.g., [Baad03] for a discussion of the consequence of this semantics and for an investigation of other possible semantics for GCIs). The problem of verifying satisfiability or checking subsumption w.r.t. generalised TBoxes is EXPTIME-hard [Nebe90, Lutz99b]. However, this result holds for the worst case, which does not necessarily occur in

practical applications (see, e.g., [Spee95]), and practical work on reasoners for languages of the $\mathcal{SH}$ family exploits this insight.

As of now, we have covered axioms for expressing restrictions for subsets of the domain $\Delta^{\mathcal{I}}$ (or subsets of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$). For BOEMIE, restrictions about particular elements of the domain are also important, since it is of interest to be able to classify the multimodal enquired information according to a generalised terminology of the sports domain. For this task, early description logics of the $\mathcal{SH}$ family provide specific assertions that are collected in a so-called ABox. An ABox is a set of assertions of the form $C(a)$, $R(a, b)$, $a = b$, or $a \neq b$. An ABox is satisfied by an interpretation $\mathcal{I}$ if $a^{\mathcal{I}} \in C^{\mathcal{I}}$, $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$, $a^{\mathcal{I}} = b^{\mathcal{I}}$ and $a^{\mathcal{I}} \neq b^{\mathcal{I}}$ respectively. The ABox satisfiability problem (w.r.t. a TBox) is to check whether there exists an interpretation (a model of the TBox) that satisfies all ABox assertions. As usual, we define a knowledge base (KB) to be a pair $(\mathcal{T}, \mathcal{A})$ of a TBox $\mathcal{T}$ and an ABox $\mathcal{A}$. A model of a KB is an interpretation that is a model of $\mathcal{T}$ and $\mathcal{A}$. The instance problem $instance_{\mathcal{T},\mathcal{A}}(i, C)$ w.r.t. a knowledge base $(\mathcal{T}, \mathcal{A})$ is to test whether $i^{\mathcal{I}} \in C^{\mathcal{I}}$ for all models of the knowledge base. We say $instance_{\mathcal{T},\mathcal{A}}(i, C)$ is entailed. The knowledge base is often omitted in the notation if clear from context. The instance retrieval problem $instance_{\mathcal{T},\mathcal{A}}(i, C)$ w.r.t. a KB $(\mathcal{T}, \mathcal{A})$ and a query concept C is to determine all individuals i mentioned in the ABox for which $instance(i, C)$ is entailed. A role filler for a role R w.r.t. an individual i is an individual j (mentioned in the ABox) such that for all models $\mathcal{I}$ it holds that $(i^{\mathcal{I}}, j^{\mathcal{I}}) \in R^{\mathcal{I}}$ (we say $related(i, j, R)$ is entailed).

The inference problems mentioned in this section are called standard inference problems for TBoxes and ABoxes, respectively. Reasoners of the $\mathcal{SH}$ family support standard inference problems either for TBoxes and ABoxes or for TBoxes only. As we have seen, $\mathcal{ALC}$ inference problems are not tractable in the worst case, and, at the beginning, incomplete algorithms were used in concrete system implementations for expressive DLs. However, at the end of the eighties it became clear that incomplete algorithms for expressive description logics cause all kinds of problems for applications. For instance, more often than not, the addition of an axiom or assertion to the knowledge base led to the situation that previously obtained entailments were no longer computed due to peculiarities of the inference algorithm.

The beginning of the $\mathcal{SH}$ family started with work on the system KRIS [Baad91b, Holl91b, Achi91], which provides a sound and complete reasoner based on the tableau calculus presented in [Schm91]. KRIS supports $\mathcal{ALC}$ plus number restrictions (plus some additional language constructs). The reasoner implements optimisation techniques for the concept and ABox satisfiability problem w.r.t. TBoxes (e.g., lazy unfolding, trace technique). The main achievement of this work is that the architecture of KRIS is tailored towards specific services for TBoxes, namely TBox classification. Specific optimisation techniques for the classification problem developed for KRIS are used by all contemporary reasoners of the $\mathcal{SH}$ family (see below). The idea is to classify a TBox using a top-down and bottom-up search phase for computing the most-specific subsumers and least-specific subsumees based on subsumption tests. KRIS avoids unnecessary subsumption tests using marker propagation tests [Baad92b, Baad94].

A.2 Concrete Domains

Another achievement of the work on description logics that is also important for ontology languages is the treatment of specific domains with fixed (concrete) semantics. To denote for instance that `Running_60_Meters` is a running modality conducted on a 60 meters track $\texttt{Running\_60\_meters} \sqcap \exists\texttt{meters} \leq 60$, applications require constraints over the reals, the integers, or a domain of strings. A concrete domain $\mathcal{D}$ is a tuple $(\Delta^{\mathcal{D}}, \Phi)$ of a non-empty set $\Delta^{\mathcal{D}}$ and a set of predicates ϕ. Predicates are defined in a certain language (e.g., linear equations over polynomials or equations over strings). The integration of concrete domains into $\mathcal{ALC}$ is investigated in [Baad91a, Baad92a]. The idea of the new language, $\mathcal{ALC}(\mathcal{D})$, is that the axioms for capturing the concrete semantics of the objects in $\Delta^{\mathcal{D}}$ is not captured with description logic axioms but somehow represented separately. The tableau calculus in [Baad91a, Baad92a] treats the satisfiability problem w.r.t. to conjunctions of concrete domain predicates as separate subproblems. The concrete domain satisfiability problems must be decidable (admissibility criterion for concrete domains).

With concrete domains, so-called attributes are introduced, which are partial functions that map individuals of the abstract domain $\Delta^{\mathcal{I}}$ to elements of $\Delta^{\mathcal{D}}$ of the concrete domain $\mathcal{D}$. For attributes a, the interpretation is extended as follows: $a^{\mathcal{I}} : \Delta^{\mathcal{I}} \rightarrow \Delta^{\mathcal{D}}$.

It is important to note that in the original approach [Baad91a, Baad92a] it is possible to relate (multiple) attribute values of different individuals of the domain $\mathcal{I}$. One can represent, for instance, structures such as lists of numbers with decreasing value where each value is at most half as large as the predecessor. If the language provides concrete domains such as, for instance, linear inequations over the reals, GCIs cannot be supported by description logic part due to undecidability of major inference problems. This follows from a result in [Lutz99a] (a direct proof was developed at the same time and is given in [Moll00]). In a restricted form where no feature compositions can be used, it is only possible to relate attribute values of a single element of $\mathcal{I}$. We use the notation $\mathcal{ALC}(\mathcal{D})^-$ to indicate that feature chains are omitted. Concrete domains are part of many specific description logics of the $\mathcal{SH}$ family that we cover in the next sections.

A.3 Transitive Roles

For many applications, part-whole relations are important. A characteristic of some part-whole relations is that they are transitive (see, e.g., [Lamb96]). In order to cope with these modeling demands, for instance, in process engineering applications, an investigation about extensions of $\mathcal{ALC}$ with means to express transitivity was carried out [Satt96, Satt98]. $\mathcal{ALC}$ was extended with a transitive closure operator, with transitive roles, and with so-called transitive orbits. As discussed in other sections, $\mathcal{ALC}$ extended with a transitive closure operator causes the concept satisfiability problem to move from PSPACE to EXPTIME.

Syntactically, transitive roles are indicated as a subset of all role names. It turned out that $\mathcal{ALC}$ extended with transitive roles remains in PSPACE [Satt96]. Transitive roles have the semantics that for all transitive roles R the models must satisfy $R^{\mathcal{I}} = (R^{\mathcal{I}})^+$. Thus, transitive roles are "globally" transitive and cannot be used in a transitive way in a local setting only (as possible with a specific operator for the transitive closure of a role).

Inspired by work on modal logics, [Satt96] introduces an elegant way to integrate reasoning about transitive roles into the $\mathcal{ALC}$ tableau calculus by a special rule for transitive roles in value restrictions. Additionally, in order to enforce termination, *blocking conditions* were defined such that the calculus terminates. A blocking condition involves a test whether two sets of concepts are in a certain relation (for $\mathcal{ALC}_{\mathcal{R}+}$, the relation is $\subseteq$, for details see [Satt96]).

The logic was initially called $\mathcal{ALC}_{\mathcal{R}+}$. As more language constructs were added later on, and acronyms became hard to read, $\mathcal{ALC}_{\mathcal{R}+}$ was renamed $\mathcal{S}$ in [Horr99].[7]

A.4 Role Hierarchies and Functional Restrictions

Inspired by work on medical domains in which it became important to represent that some relations are subrelations (subsets) of other relations, so-called role inclusions axioms of the form $R \sqsubseteq S$ (with R and S being role names) were investigated in [Horr97] as an extension to $\mathcal{ALC}_{\mathcal{R}+}$. A set of role inclusion axioms is called a role hierarchy. Models for role hierarchies are restricted to satisfy $R^{\mathcal{I}} \subseteq S^{\mathcal{I}}$ for all $R \sqsubseteq S$. The description logic is called $\mathcal{ALCH}_{\mathcal{R}+}$ or $\mathcal{SH}$.

Role hierarchies introduce explicit names for so-called subroles. In [Horr98a] it is argued that role hierarchies provide for adequate expressivity while still allowing for efficient practical implementations at the same time. Another possibility would have been to consider a role-forming operator for constructing role conjunctions ($R \sqcap S$). However, except for inverse roles (see below) the $\mathcal{SH}$ family includes no role-forming operators in order to provide for practically efficient implementations (see also the discussion about a transitive closure operator for roles in the previous subsection).

Additionally, in [Horr97, Horr98a] global functional restrictions on roles were investigated. In the corresponding description logic $\mathcal{ALCH}f_{\mathcal{R}+}$ so-called features were introduced as a specific subset of the role names.[8] Features must not be transitive. The semantics of a feature f is a (single valued) partial function $f^{\mathcal{I}} : \Delta^{\mathcal{I}} \rightarrow \Delta^{\mathcal{I}}$.

With several examples, the interactions of role hierarchies and functional restrictions on roles were demonstrated in [Horr97]. A sound and complete tableau calculus for $\mathcal{ALCH}f_{\mathcal{R}+}$ is described in

[7] The name is inspired by modal logic $\mathcal{S}\triangle_{\Updownarrow}$ but, obviously, it is a misnomer. However the name is kept for historical reasons.

[8] Note that $\mathcal{ALCH}f_{\mathcal{R}+}$ does not provide role-value maps as supported by $\mathcal{ALCF}$

[Horr98b]. This tableau calculus provided the basis for the enormous success of the $\mathcal{SH}$ family of ontology languages. Based on an optimised implementation of this calculus in the system FACT [Horr97, Horr98a] it was shown that description logics could provide a solid basis for practical application of ontology languages. Role hierarchies and transitive roles allow one to somehow "simulate" GCIs (by constructing an equisatisfiable knowledge base). However, the system also included full support for GCIs.

The contribution of the $\mathcal{ALCH}f_{\mathcal{R}+}$ reasoner FACT is (at least) threefold. First, improvements to propositional satisfiability search algorithms [Free95] were incorporated into description logic systems (backjumping, boolean constraint propagation, semantic branching, etc.) and, second, classification operations were dramatically increased by the invention of a so-called model merging operation [Horr97], which exploits that most subsumption tests for concept names A_1 and A_2 used to compute the subsumption hierarchy return false. The idea of a model merging operation is to compute (small) data structures for concept names (and their negations) such that it is more or less directly obvious that the conjunction $A_1 \sqcap \neg A_2$ is satisfiable (i.e., there is no subsumption relation). Third, using algebraic transformation, FACT showed that, in many practical applications, corresponding TBox axioms can be converted into a form such that lazy unfolding is maximally exploited in the tableau calculus (GCI absorption [Horr00c]). The system FACT initiated the breakthrough of description logics as the basis for practically used ontology languages. FACT was designed for TBoxes.

A.5 Number Restrictions and Inverse Roles

The need for restrictions on the number of role fillers of an individual, to express for instance that a first place athlete is one which has at least 1 gold medal $\leq 1 has\_medal.GoldMedal$ are also apparent. Number restrictions are concept construction operators of the form $(\leq n\ R)$ or $(\geq n\ R)$ (simple number restrictions, indicated with letter N in language names) and $(\leq n\ R.C)$ or $(\geq n\ R.C)$ (qualified number restrictions [Holl91a], indicated with letter $\mathcal{Q}$ in language names). For simple number restrictions, interpretations must satisfy $(\leq n\ R)^{\mathcal{I}} = \{x|\#\{y|(x,y) \in R^{\mathcal{I}}\} \leq n\}$ and $(\geq n\ R)^{\mathcal{I}} = \{x|\#\{y|(x,y) \in R^{\mathcal{I}}\} \geq n\}$ For qualified number restrictions, interpretations must satisfy $(\leq n\ R.C)^{\mathcal{I}} = \{x|\#\{y|(x,y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\} \leq n\}$ and $(\geq n\ R.C)^{\mathcal{I}} = \{x|\#\{y|(x,y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\} \geq n\}$.

KRIS supported simple number restrictions in a system implementation at the end of the eighties. With only simple number restrictions and no role inclusions, it is possible to use a single placeholder in the tableau calculus for an arbitrary large number of role fillers required by a number restrictions. Results on the interaction of number restrictions and role conjunctions were developed with $\mathcal{ALCNR}$ [Buch93a, Buch93b]. Simple reasoning with placeholder is no longer possible. The same holds for number restrictions in combination with role hierarchies as used in the $\mathcal{SH}$ family.

In addition to problems w.r.t. placeholder reasoning in the presence of number restrictions, it was shown that there is a strong interaction between number restrictions and transitive roles (and role hierarchies). Allowing number restrictions with transitive roles (or roles which have transitive subroles) leads to undecidability [Horr99]. As a consequence, so-called *simple* roles were introduced into the $\mathcal{SH}$ family. In (qualified) number restrictions, only simple roles are allowed. With this restriction, inference problems become decidable [Horr99].

Another demand from practical applications was the support for inverse roles (letter $\mathcal{I}$ in language names). In [Horr99] the research on a corresponding role-forming operator $.^{-1}$ in the context of the $\mathcal{SH}$ family is summarised. Again, a subtle interaction between number restrictions (or features), inverse roles as well as transitive roles and role hierarchies (or GCIs) was discovered. If all these constructs are used, the finite model property does no longer hold. First, due to inverse roles, the trace technique is no longer applicable, and, second, the application of the blocking condition introduced in the work about $\mathcal{ALCR}_+$ had to be made considerably more complex. Blocking must be dynamic [Horr99]. This makes the implementation of concrete reasoning systems much more difficult. An additional source of complexity is that blocking must not occur too early (thus, the blocking condition involves a test for set equality), and, furthermore, due to infinite models, the blocking condition involves a pair of two sets of concepts (pairwise blocking [Horr99]).

Although ABoxes were also investigated in the context of $\mathcal{ALCNR}$, work on the $\mathcal{SH}$ family of description logic languages initially considered TBoxes only.

A.6 Number Restrictions, ABoxes and Concrete Domains

Inspired by work on the $\mathcal{SH}$ family and work on ABoxes in $\mathcal{ALCNR}$ as well as work on concrete domains [Baad91a], a tableau calculus for ABox reasoning in the language $\mathcal{ALCNH}_{\mathcal{R}+}$ was presented in [Haar00] and concrete domain reasoning was investigated in this context in [Haar01f]. The insights of this work are that in the presence of ABoxes, (i) models are no longer (quasi) trees but forests, (ii) individuals mentioned in the ABox must not block each another, and (iii) on the concrete domain part of the language, feature chains cannot be supported (for all kinds of concrete domains) in order to preserve decidability. A tableau calculus for reasoning about ABoxes in the language $\mathcal{SHIQ}$ (aka $\mathcal{ALCQHI}_{\mathcal{R}+}$) with ABoxes was presented shortly afterwards in [Horr00a] (but concrete domains were not considered).

The latter work led to a new version of the FACT system (iFACT) for supporting TBoxes with inverse roles. The above-mentioned research contributions also provided the basis for the implementation of the RACER reasoner [Haar01d], a DL system for TBoxes and ABoxes with concrete domains for linear inequations over the reals and the cardinals as well as inequations over strings and booleans. First versions of both systems, iFACT and RACER, appeared at the end of the nineties, i.e. both systems support the language $\mathcal{SHIQ}$.[9] Both RACER and FACT use the TBox classification techniques developed for KRIS [Baad92b, Baad94].

Optimised reasoning techniques for $\mathcal{SHIQ}$ w.r.t. blocking [Horr02] were developed for later versions of the iFACT system, and also included in RACER. The idea is to relax the blocking condition for inverse roles (see above) and retain the subset tests for some parts of the concept set involved in the blocking test (see [Horr02] for details).

With RACER, optimised reasoning for qualitative number restrictions [Haar01c, Haar01a] was investigated. The work is based on [Ohlb99]. Due to the continuous semantic extraction and ontology evolution process in BOEMIE, it is expected to cope with huge amounts of data continuously increasing. Thus in order to classify huge terminologies with RACER, a refinement of the techniques introduced in [Baad92b, Baad94] is presented in [Haar01b]. Topological sorting of transformed GCIs to classify concepts in definition order allows to skip the bottom-up search phase. Optimisations for concrete domains in terms of extended model merging operations and incremental concrete domain satisfiability testing during a tableau proof are described in [Haar01e]. GCI absorption strategies are also investigated with RACER, e.g., absorption of domain and range restrictions (see also [Tsar04] for similar techniques in FACT).

Reasoning systems for the $\mathcal{SH}$ family are successful because of research on average-case behaviour and appropriate optimisation techniques. These systems analyse the language of the input problem and select appropriate optimisations to answer queries as fast as possible, moreover, they are based on sound and complete algorithms.

Optimisations for instance retrieval w.r.t. ABoxes is investigated in [Haar04]. An important property of the $\mathcal{SHIQ}$ language is that the subsumption hierarchy of the TBox part of a knowledge base $(\mathcal{T}, \mathcal{A})$ is stable w.r.t. additions to the ABox part. Stability means that the subsumption relation between concepts C and D depends only on axioms in $\mathcal{I}$. This property is exploited in practical ABox systems such as RACER (and also older systems such as KRIS). Multiple knowledge bases $(\mathcal{T}, \mathcal{A}_k), \ldots, (\mathcal{T}, \mathcal{A}_k)$ with the same TBox can be efficiently supported in the sense that computations for the TBox can be reused for answering queries on any of the ABoxes $\mathcal{A}_i$. Unfortunately, the stability property is lost with the introduction of cardinalities for concepts or with the inclusion of so-called nominals, which became necessary in order to further increase the expressivity of $\mathcal{SHIQ}$ for some applications.

A.7 Nominals

A nominal denotes a singleton concept. The syntax is $\{o\}$ and the semantics w.r.t. the interpretation is $\{o\}^{\mathcal{I}} = \{o^{\mathcal{I}}\}$. With nominals it is possible to relate all individuals of a certain concept to a particular individual (e.g., all athletes that come from a particular country called Italy). Nominals were first investigated in [Scha93] and are related to cardinality restrictions on concepts [Baad93, Baad96]. The first system with support for nominals was CRACK [Bres95].

[9] In RACER initially the unique name assumption was always employed, in later versions the assumption could be activated on demand.

Although nominals in the context of $\mathcal{SHIQ}$ were proven to be decidable (see [Tobi01]) it took some time until the first tableau calculus was presented for the language $\mathcal{SHOQ}(\mathcal{D})$ [Horr01]. This work also introduced so-called datatype roles [Horr01], which must not be confused with concrete domain attributes. Datatype roles map object from the domain to sets of objects from a concrete domain. In $\mathcal{SHOQ}(\mathcal{D})$ concrete domain predicates apply to (multiple) datatype properties of single object of the interpretation domain $\Delta^{\mathcal{I}}$. It is not possible to enforce constraint on datatype values of multiple objects from $\Delta^{\mathcal{I}}$. The insight gives rise for corresponding optimisation techniques but it should be noted that some expressivity is lost.

The distinction between TBoxes and ABoxes are no longer required for languages with nominals. Instead of using $C(a)$ or $R(a, b)$ as ABox assertion one can just write GCIs such as $\{a\} \sqsubseteq C$ or $\{a\} \subseteq R.b$ respectively.[10] Even if ABoxes would be supported by practical systems, it is obvious that the subsumption relation is not stable for languages with nominals.

Intricate interactions of nominals in $\mathcal{SHOQ}(\mathcal{D})$ with inverse roles were investigated $\mathcal{SHOIQ}(\mathcal{D})$ in [Horr05]. Indeed, it was shown that concept satisfiability in $\mathcal{SHOIQ}(\mathcal{D})$ is NEXPTIME-complete.

A.8 The research Frontier of SH Family

Further results on optimised classification [Tsar05b, Tsar05a] has opened up additional application areas for ontology languages. And, although much has been achieved by dedicated optimisation techniques developed for the $\mathcal{SH}$ family of description logic languages, still there are hard knowledge bases known (e.g., [Bera01]). New languages features with respect to specific kinds of role axioms involving role composition have been proposed for medical domains. A tableau calculus for the new language $\mathcal{SROIQ}(\mathcal{D})$ is presented in [Horr06]. To the best of our knowledge, there is no system implementation at the time of this writing that supports all features of this language.

Concrete domain reasoning is also actively explored. Starting with investigation involving new combination operators ([Lutz99c]), in [Lutz01a, Lutz01b] it is shown that for specific concrete domains, feature chains can indeed be allowed in the presence of GCIs (see also [Lutz03, Lutz04]). The language investigated ($\mathcal{Q} - \mathcal{SHIQ}$) provides predicates for linear inequalities between variables (but no polynoms). A more modular approach is described in [Lutz05, Liu06] where the notion of admissibility (see above) is extended to a so-called w-admissibility. No system implementation exists at the time of this writing.

New versions of the description logic systems discussed in the previous section have been developed. These systems are FACT++ (for $\mathcal{SHOIQ}(\mathcal{D})$) [Siri06] and RACERPRO (at the time of this writing the latter only provides an approximation for nominals). FACT++ is written in C++ whereas RACERPRO is implemented in COMMONLISP. A new JAVA-based description logic system for $\mathcal{SHOIQ}(\mathcal{D})$ (and OWL DL) is PELLET. As FACT++, PELLET is based on a tableau reasoning algorithm and integrates various optimisation techniques in order to provide for a fast and efficient practical implementation. New developments also tackle the problem of "repairing" knowledge bases in case an inconsistency is detected [Kaly05]. In addition, with PELLET, optimisation techniques, for instance, for nominals have been investigated [Siri06]. Other description logic systems are described in [Moll03].

Compared to initial approaches for query languages (see [Lenz91]), recently, more expressive languages for instance retrieval have been investigated (conjunctive queries [Calv98, Horr00b, Glim05]). To the best of the authors' knowledge, algorithms for answering conjunctive queries for expressive description logics such as $\mathcal{SHIQ}$ are not known. In practical systems such as RACER implementations for a restricted form of conjunctive queries is available (variables are bound to individuals mentioned in the ABox only). Database-inspired optimisation techniques for this language in the context of a tableau prover are presented in [Moll06]. In addition, RACER supports the incremental computation of result sets for restricted conjunctive queries. The demand for efficient instance retrieval has led to the development of new proof technique for languages of the $\mathcal{SH}$ family. A transformation approach using disjunctive datalog [Eite97], resolution techniques as well as magic-set transformations to support reasoning for $\mathcal{SHIQ}$ is described in [Hust04, Moti06] with encouraging results. In this context, a new system, KAON2 has demonstrated that techniques from the database community can be successfully used also for implementing description logic systems. Although at the time of this writing, KAON2 is

[10] Equality and inequality of individuals can also easily be specified using negation.

a very recent development and not quite as expressive as FACT++, PELLET, or RACERPRO (e.g., w.r.t. datatypes, nominals, large numbers in qualified number restrictions, etc.).

The synergistic approach of BOEMIE, realized by the integration of different components in an open architecture can profit from the recent advances in the development of standards for description logic reasoning systems (such as DIG [Bech03]) have contributed to the fact that DL systems can be interchanged such that the strength of particular reasoning systems can be exploited for building practical applications. Since semantic web applications have become interesting from a business point of view, commercial DL systems have appeared (e.g., from Cerebra Inc.) and commercial versions of above-mentioned systems became available (e.g., KAON2 from Ontoprise or RACERPRO from Racer-Systems and Franz Inc.).