

This introduction course on the role of NLP for Multimedia Applications is resulting from interests shared by Elisabeth André, owner of a chair dedicated to Multimedia Applications at the University of Augsburg, Thierry Declerck, who at DFKI in Saarbrücken was involved in the European MUMIS project investigating the use of NLP-based technologies for the content indexing of Multimedia material, and Yiannis Kompatsiaris, one of the leader of the European SCHEMA Network of Excellence in Content-Based Semantic Scene Analysis and Information Retrieval. This Network is also dedicated to the use, further development and the concrete implementation, in the form of a reference platform, of a standard for Multimedia content indexing and retrieval, the MPEG-7 standard.

The structure of the course will reflect the “origin” of the various contributors to this investigation about the role of NLP for Multimedia Applications. Elisabeth André is giving a broad introductory overview on the various contributions of NLP and other media to the successful deployment of Multimedia applications in our everyday life. Thierry Declerck will go into more details into the use of Information Extraction techniques applied to textual material, which is accompanying video material, for improving the content indexing of the video, and Yiannis Kompatsiaris will offer an introduction to MPEG-7 as well as a description of the SCHEMA network, in which actually the integration of NLP and Media analysis tools will be tried out in the context of a particular application: the TRECVID scenario.

Additional introductory material on MPEG-7 has been added, and we thank Philippe Salembier for his excellent slides on this topic.

1. Introduction and overview: NLP, Multimedia, Multimodality. Elisabeth André. 1-62
2. Content-based Indexing and Searching of Multimedia Documents. Thierry Declerck, Jan Kuper, Horacio Saggion, Anna Samiotou and Peter Wittenburg. 63-110. Published in „Annotation for the Semantic Web“, S. Handschuh and S. Staab (Eds), ISO Press, 2003.
3. Overview of MPEG-7: Multimedia Content Description Interface. Philippe Salembier. 2001. 101-183
4. The SCHEMA NoE Reference System. Yiannis Kompatsiaris. 184-2088.

Introduction

-
- Terminology
 - Advantages of Multimodal/Multimedia Interfaces
 - Contents of the Tutorial

1

What is Multimedia?



-
- Search with Google: 41,1 Mio hits
 - 1995 „ Word of the Year“ in Germany
 - Interesting Word Creations:
 - » Multimedia-Kitchen, Multimedia-Hotel, ...
 - Literal Meaning:
 - » Multus = „much, multiple, ...“
 - » Medium = „Device for conveying information“
 - Problem: Different Use of Terminology in Different Disciplines

2
ESSLLI 04, André

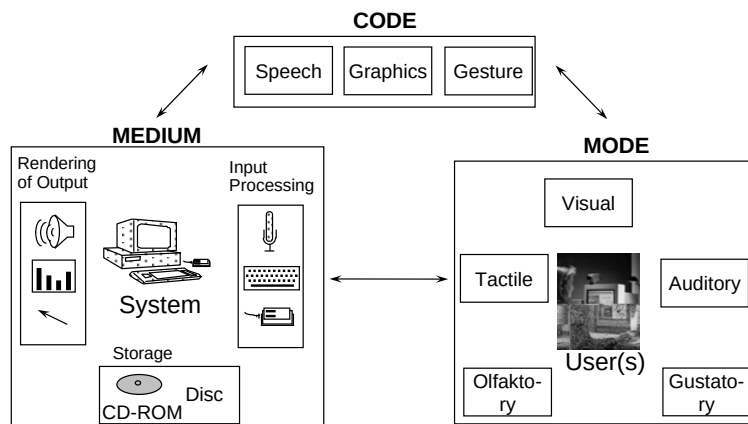
Medium versus Modus or Modality [Maybury 1999]



- **Medium:**
 - » Carrier of information (e.g. paper or CD-ROM)
 - » Different kinds of physical devices (e.g. screens, loudspeakers, microphones and printers)
 - » Information types (e.g. graphics, text and video)
- **Modus or Modality:** Cognition aspect
 - » Perceptible entity (e.g. visual, auditory, haptic or olfactory media)
 - » Interaction style (e.g. natural language dialogue)
- **Code**
 - » Particular means of encoding (e.g. sign languages or pictorial languages)

3
ESSLLI 04, André

Medium, Mode or Code



Source: Maybury 1999

Copyright © 2000 The MITRE Corporation.
All rights reserved.

ESSLLI 04, André

Display Dimensions



- Discrete/time-independent media
 - » Text
 - » Graphics
- Time-dependent/continuous media
 - » Audio
 - » Animations
 - » Video

5
ESSLLI 04, André

Why to Use Multiple Media?



- Increasing the accessibility of applications $\Rightarrow$ Universal Access
 - » Adaptation to the situative context
 - gestures in environments with high noise level
 - » Adaptation to the individual user
 - spoken language for people with impaired vision
 - Sign language for deaf people
 - Gestures for non-native speakers

6
ESSLLI 04, André

Reduction of the Error Rate



- Resolution of ambiguities by analyzing multiple input media in Quickset



Oviatt, S.L. & Cohen, P.R.. "Multimodal Interfaces That Process What Comes Naturally"
Communications of the ACM, Vol. 43, No. 3, March, 2000, pp. 45-53.

ESSLLI 04, André

Expressive Power and Efficiency



- **Shorter and more fluent verbal utterances** [Oviatt 2000]
Example from Quickset User Study:
 - » User with drawing gestures
User: [draws irregular area] "Open space"
 - » Same user with drawing gestures
User: "Add an open space on the north lake to br- include the north lake part of the road and north."
- **Completion Time** [Oviatt 2000]
 - » Multimodal interaction with Quickset (pen and voice input) 9x faster than interaction with a conventional GUI
- **Expressive Power**
 - » 90-100% of the Quickset users prefer multimodal input options

8
ESSLLI 04, André

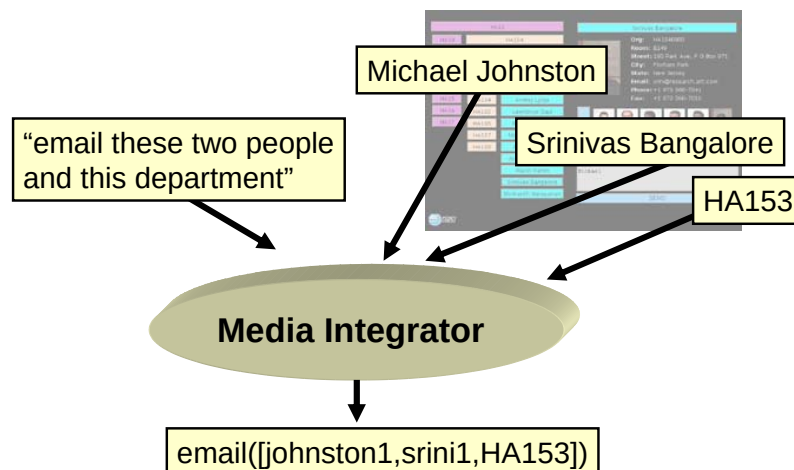
Media Integration, Coordination and Synchronization



- **Media Integration or Fusion**
Transforming multiple media into a uniform representation format
- **Media Coordination**
Tailoring several media during the generation process, e.g. media selection, generation of cross references etc.
- **Media Synchronization**
Spatial and temporal layout

9
ESSLLI 04, André

Example: Media Integration



Source: Johnston, AT&T Labs, 2000.

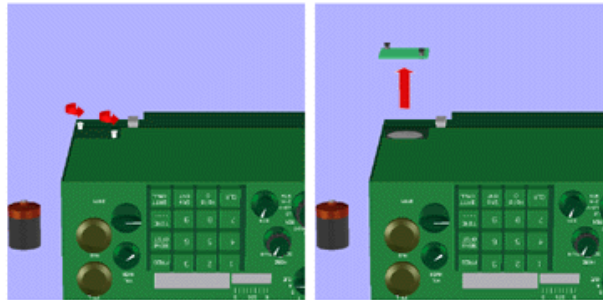
ESSLLI 04, André

Example: Media Coordination



Cross References in COMET

Remove the old holding battery. Step 1 of 2



Step1:
Remove the holding battery cover plate, highlighted in the right picture:
Loosen the captive screws and pull the holding battery cover plate off the radio.

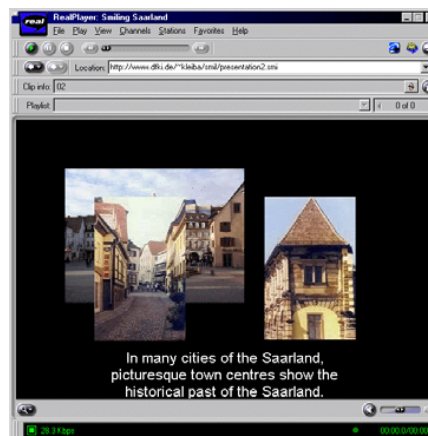
Source: D.D. Seligmann: Interactive Intent-Based Illustration: A Visual Language for 3D Worlds, PhD Thesis, 1993.

ESSLI 04, André

Example: Media Synchronization



Synchronizing running texts, music, voice and image sequences



Source: André et al., DFKI

12
ESSLI 04, André

Multimodal Analysis

-
- Unification methods for analyzing gestures and speech
 - Analysis of gestures and speech with finite state automata

13

Multimodal Analysis



-
- **Modality-Specific Analysis**
e.g. Analysis of spoken language
 - **Crossmedia Analysis**
e.g. Resolution of crossreferences, such as the „the left button in the illustration“
 - **Fusion** of different input into a uniform representation format

14
ESSLLI 04, André

Analysis of Gestures and Speech



• Early Systems:

- » SCHOLAR [Carbonell 70]
- » Put-That-There [Bolt 80]
- » XTRA [Allgayer et al. 1989]
- » CUBRICON [Neal/Shapiro 91]



• Characteristics:

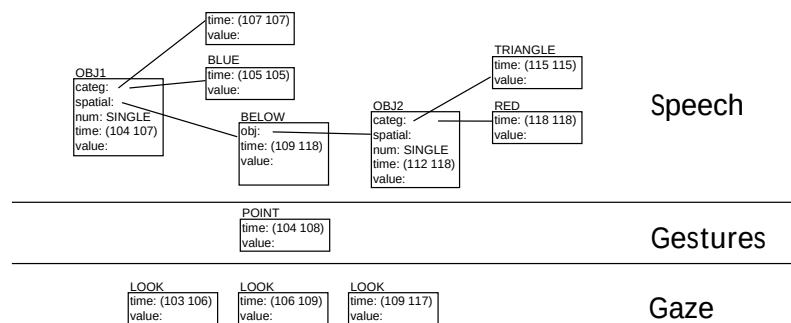
- » Speech-driven, i.e. integration of gestures is triggered by verbal referring expressions, such as „this one“
- » No uniform representation formalism for the single modes
- » No general well-defined multimodal integration mechanism
- » Limited to pointing gestures

15
ESSLLI 04, André

Analysis of Gestures, Speech and Gaze



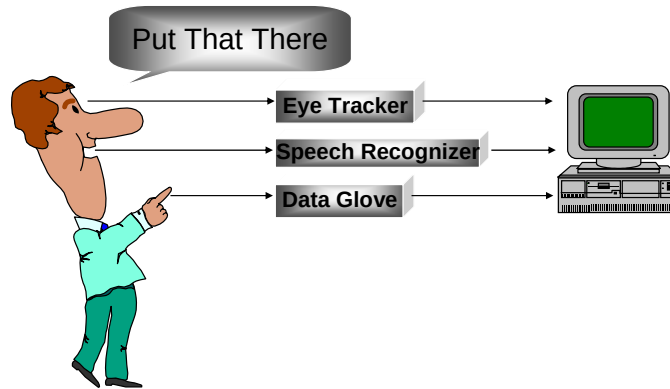
Koons et al. 93: Frames as a uniform representation mechanism for speech, gestures and gaze, but no declarative approach to media integration



Idealized sample frames for the utterance: “ ... that blue square below the red triangle”

16
ESSLLI 04, André

Multimedia Analysis: Example

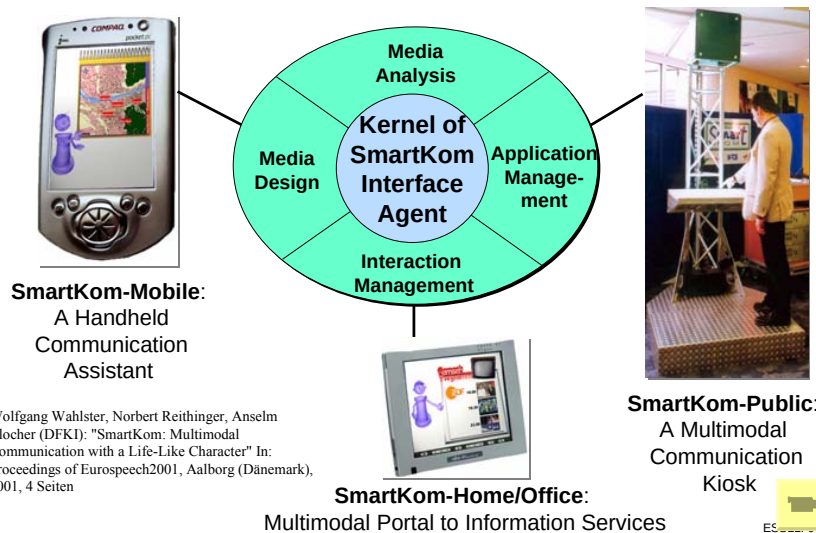


Source: M. Maybury: Tutorial
on Intelligent User Interfaces,
IUI 2000.

Copyright © 2000 The MITRE Corporation. All rights reserved.

ESSLI 04, André

SmartKom: A Transportable Interface Agent



Wolfgang Wahlster, Norbert Reithinger, Anselm
Blocher (DFKI): "SmartKom: Multimodal
Communication with a Life-Like Character" In:
Proceedings of Eurospeech2001, Aalborg (Denemark),
2001, 4 Seiten

ESSLI 04, André

Multimodal Analysis in Quickset



Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. (1997). QuickSet: Multimodal interaction for distributed applications, *Proc. of the Fifth International Multimedia Conference (Multimedia '97)*, ACM Press, pp 31-40.

ESSLLI 04, André

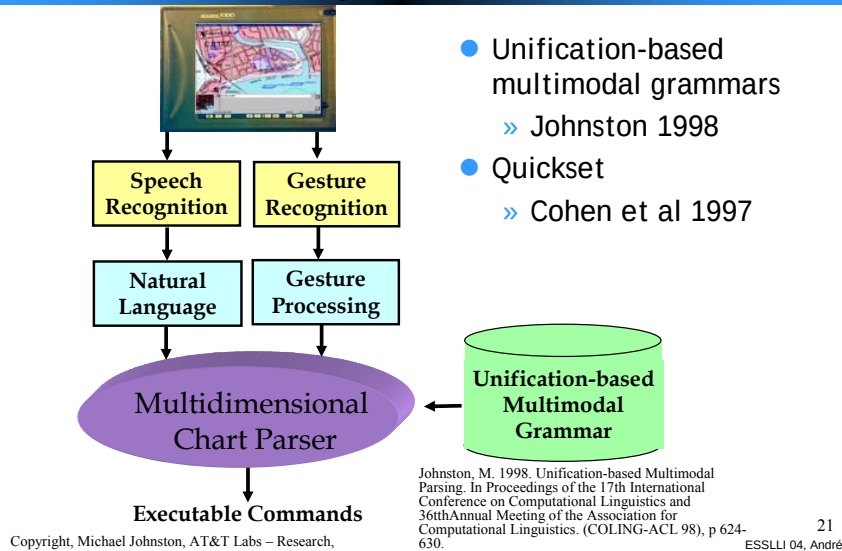
Media Integration



- Converting multiple media into a uniform representation format
- **Questions:**
 - » How should the results of media-specific recognition components be represented to enable the integration of the single modalities?
 - » How far should an input stream be processed before the results of other modalities are integrated?

20
ESSLLI 04, André

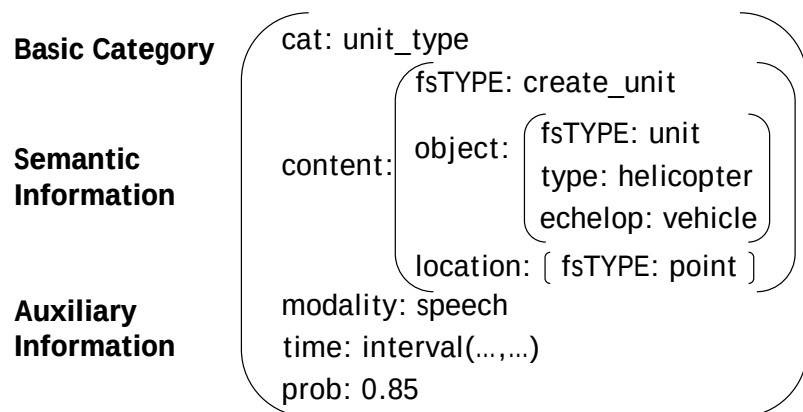
Architecture for multimodal Analysis



Multimodal Grammar Formalism



Representation of the spoken utterance: „Helicopter“
by typed feature structures



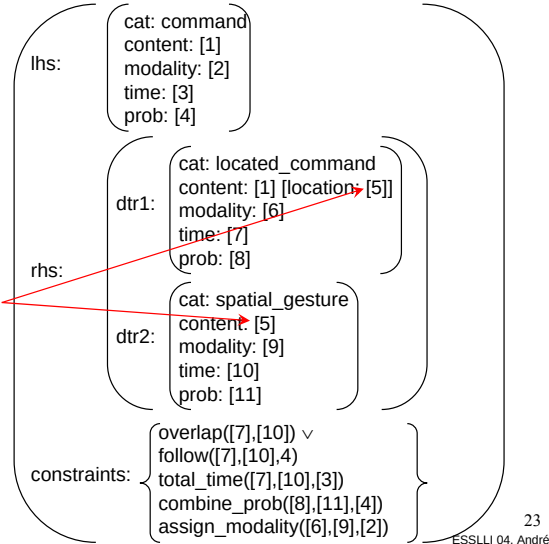
22
ESSLLI 04, André

Coding Grammar Knowledge by Feature Structure Rule Schemata



**LHS → DTR1 DTR2,
where LHS, DTR1
and DTR2 are typed
feature structures**

structure-sharing tags



Constraints



- **Temporal Constraints**

`overlap([7],[10]) ∨ follow([7],[10],4)`

The time of the speech [7] needs either to overlap with or start within four seconds of the time of the gesture [10].

- **Spatial Constraints**

`close_to(X,Y)`

`contact(X,Y)`

- **Additional Constraints for Determining Values**

`assign_modality([6],[9],[2])`

Multimodal Sub Categorization



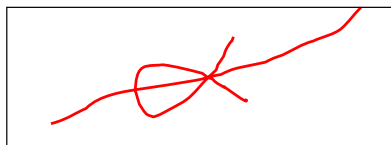
- **Problem:** Treatment of speech input which is combined with several gestures
- **Solution:** Introduction of a Subcat feature which is implemented as a recursive first/rest feature structure.
- **Example:** Utterance: „move this here“ results in a structure which expects two gestures as a complement.

25
ESSLI 04, André

Visual Parsing: Complex Gestures



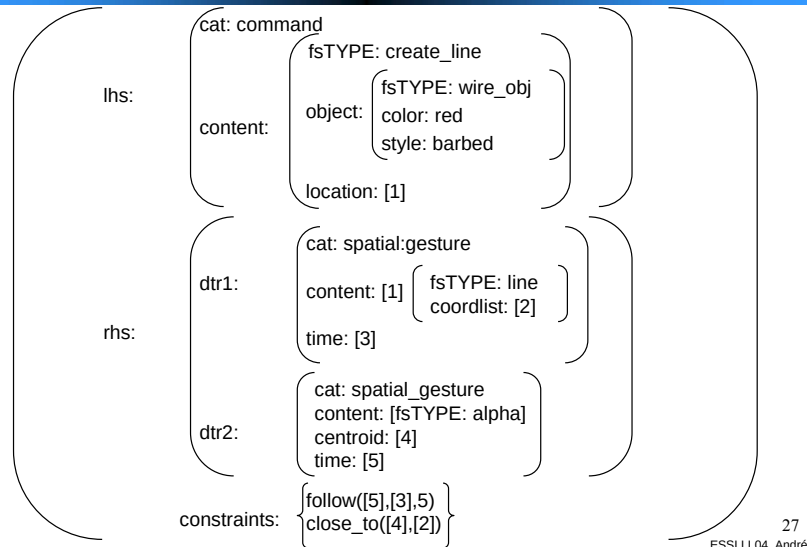
- To create a barbed wire, the user may specify a line which specifies its spatial extension and then draw an alpha to indicate its type.



Complex Gesture for Barbed Wire

26
ESSLI 04, André

Rule Schema for Unimodal Barbed Wire



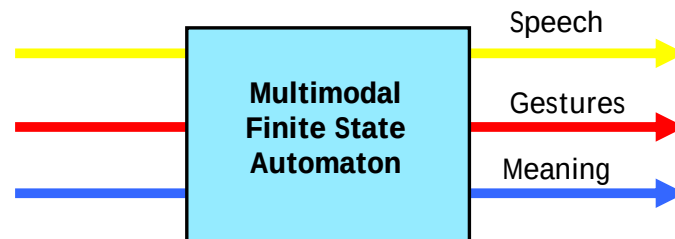
Multimodal Unification



- **Advantages:**
 - » Declarative specification of multimodal integration rules
 - » Separation of integration rules and parsing mechanism
 - » Analysis of complex graphical input $\Rightarrow$ Parsing of visual languages
- **Disadvantages:**
 - » Computational complexity
 - » No probabilistic framework to resolve multimodal ambiguities

28
ESSLLI 04, André

Multimodal Analysis with finite Automata



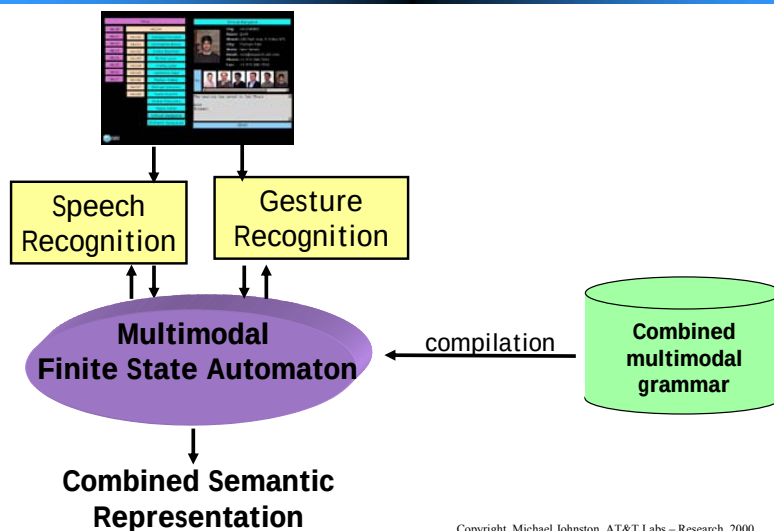
- For a user interface with n modalities, we use an automaton with $n+1$ tapes.

Copyright, Michael Johnston, AT&T Labs – Research, 2000.

Johnston, M. and S. Bangalore. 2000. Finite-state multimodal parsing and understanding. Proceedings of COLING-2000, Saarbruecken, Germany.

29
ESSLI 04, André

Parsing with Multimodal FSA



Copyright, Michael Johnston, AT&T Labs – Research, 2000.

30
ESSLI 04, André

Compiling a Multimodal CFG

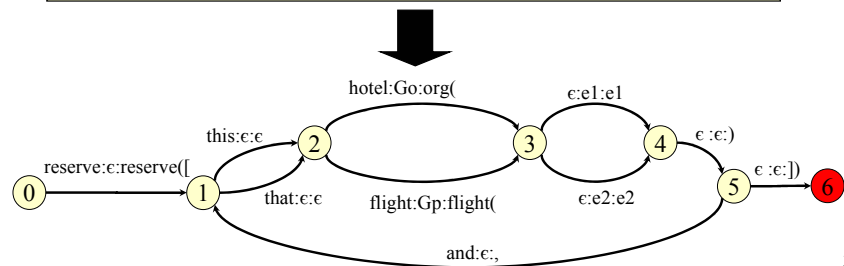


```

S → V NP ε:ε:] )
CONJ → and:ε:,
V → reserve:ε:res([
NP → DET N
NP → DET N CONJ NP
DET → this:ε:ε

N → hotel:Gh:hotel(ENTRY
N → flight:Gf:flight(ENTRY

ENTRY → ε:e1:e1 ε:ε:)
ENTRY → ε:e2:e2 ε:ε:)
ENTRY → ...
    
```



31
ESSLLI 04, André

Summary



- Easy integration into a probabilistic framework → easy to train
- Enable tight integration of media
- Computationally less intensive than unification grammars
- More difficult to read than unification grammars → compilation

32
ESSLLI 04, André

Generation of Multimedia Presentation

-
- Why?
 - Approaches to Multimedia Presentation Planning
 - Coordination of Multiple Media

33

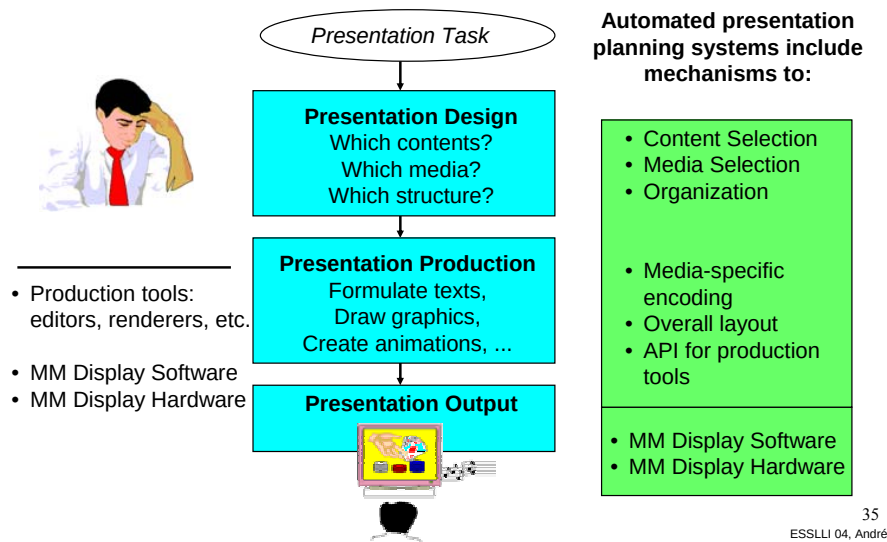
Why Automated Presentation Planning?



-
- Manual creation is error-prone.
 - Manual creation is extremely time-consuming.
 - Not feasible when information dynamically changes.
 - Methods for automated presentation planning enable adaptation to the current situation and the target group.

34
ESSLLI 04, André

Automated Presentation Planning



Multimodal Generation



- Context-sensitive selection of relevant information
- Media allocation
- Media-specific encoding
- Spatial and temporal coordination of the output

36
ESSLLI 04, André

Criteria for Media Allocation



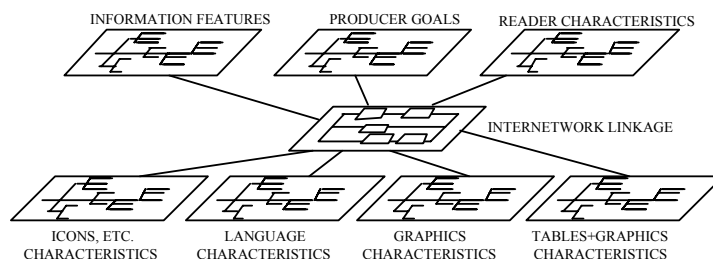
- **Media Characteristics**
 - » Static versus dynamic media ...
- **Information Type**
 - » concrete, spatial, temporal, abstract, ...
- **Communicative function**
 - » Contrast, Summarize, Elucidate
- **User Characteristics**
 - » Perceptual capabilities, prior knowledge, ...
- **Tasks**
 - » Operating a device, comparing values ...
- **Resource Limitations**
 - » Time, Space

37
ESSLLI 04, André

Media Allocation: Required Knowledge



- Systemic Network contains
 - Contents, form, purpose
 - Producer characteristics
 - Recipient characteristics
 - Communicative situation



Yigal Arens, Eduard H. Hovy and Mira Vossers. *On the Knowledge Underlying Multimedia Presentations*. In *Intelligent Multimedia Interfaces*. Mark Maybury, editor. AAAI Press/The MIT Press, pp. 280-306, 1993.

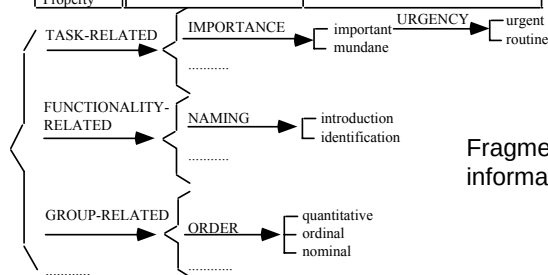
38
ESSLLI 04, André

Information Features



Type	Characteristic	Values
Intrinsic Property	Dimensionality	0D, 1D, 2D, >2D, ∞D
	Transience	live, dead
	Urgency	urgent, routine
Class Property	Order	ordered, nominal, quantitative
	Density	dense, discrete, N/A
Set Property	Naming	identification, introduction
	Volume	singular, little, much

Information features organized by type



Fragment of a network referring to information features

Source: Arens et al. 1993

39
ESSLLI 04, André

Characteristics of Media



Generic Medium	Carrier Dimension	Internal Semantic Dimension	Temporal Endurance	Granularity	Medium Type	Default Detectability	Baggage
Beep	0D		transient	N/A	aural	high	
Icon	0D		permanent	N/A	visual	low	
Map	2D	>2D	permanent	continuous	visual	low	high
Picture	2D	∞D	permanent	continuous	visual	low	high
Table	2D	2D	permanent	discrete	visual	low	high
Form	2D	>2D	permanent	discrete	visual	low	high
Graph	2D	1D	permanent	continuous	visual	low	high
Ordered List	1D	#D	permanent	discrete	visual	low	low
Unordered List	0D	#D	permanent	N/A	visual	low	low
Written Sentence	1D	∞D	permanent	discrete	visual	low	low
Spoken Sentence	1D	∞D	transient	discrete	aural	medhigh	low
Animated Material	2D	∞D	transient	continuous	visual	high	high
Music	1D	∞D	transient	continuous	aural	medhigh	low

Source: Arens et al. 1993

40
ESSLLI 04, André

Media Allocation



- Rules map characteristics of information and characteristics of media
 - » 1. Present data tuples (e.g., locations) on planar media (e.g., graphs, tables, maps)
 - » 2. Present data with specific denotations (e.g., spatial) on media with same denotations (e.g., locations on maps)
 - » 3. “If more than one medium can be used, and there is an existing presentation, prefer the medium/a that is/are already present ...”
 - » 4. Choose medium/a that can accommodate most amount of information to be presented (e.g., location, heading, nationality of ship -> former 2 on a map, all 3 in language if only 1 ship)

41
ESSLLI 04, André

Media Coordination



- Coordination within a medium (e.g. consider the effects of different annotation techniques)
- Coordination of multiple media (e.g. coordinate sentence breaks with picture breaks, cross references)
- Coordination of output (e.g. spatial arrangement of text blocks and images, temporal coordination of animations and speech, distribution of information on multiple windows)

42
ESSLLI 04, André

Coordination of Sentence and Picture breaks



Problem: Both textual and graphical constraints need to be considered when determining sentence and picture breaks

» **Graphical constraints determine image size**

If too many objects are shown in a single image, the single objects are probably rendered too small.

» **Grammatical constraints determine sentence length**

Obligatory case roles of a verb may not simply be left out to save space, e.g. “Hans puts” and “Hans puts the book” are ungrammatical sentences.

43
ESSLLI 04, André

Generation of Referring Expressions



- Distribution of information required for the identification of the referent onto multiple media, e.g. prefer graphics over text for visual properties (such as shape, color etc.), text for quantifying information
- Cross references, e.g. “the highlighted button in the figure”
- Establishing cohesive links between textual descriptions and graphical depictions, e.g. by textually annotating picture elements

44
ESSLLI 04, André

Communication-Theoretic Generation Approach



Basic Idea: Generalization of a speech act theoretic view

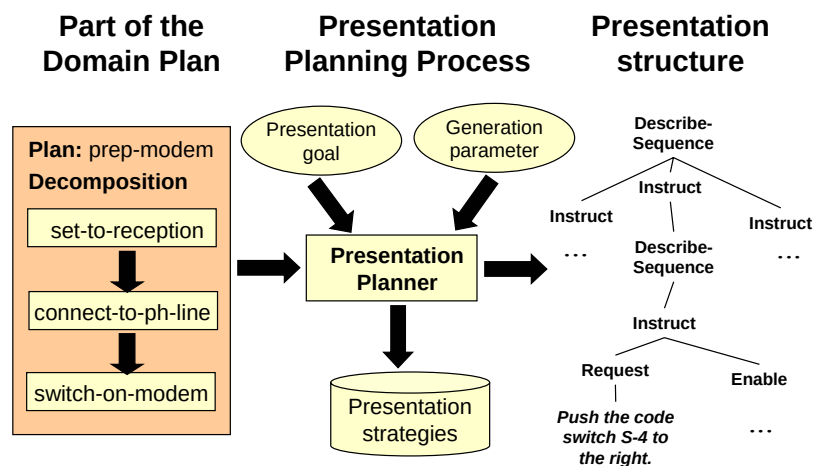
Not only spoken utterances, but also the presentation of multimedia documents can be considered as a communicative act to achieve certain goals.



Exploitation of planning methods

45
ESSLLI 04, André

From a Domain Plan to a Multimedia Presentation



46
ESSLLI 04, André

Hierarchical Task Network Planning



- **Basic Idea:** Decompose complex tasks in less complex ones
- As long as there are unfulfilled goals:
 - » Select a goal
 - » Search for an applicable operator
 - » Set up the actions in the body of the operator as new sub goals
 - » In case a goal cannot be achieved, try the next applicable operator
 - » The planning process terminates as soon as all goals have been transformed into primitive actions

47
ESSLI 04, André

Multimedia Generation in TEXPLAN

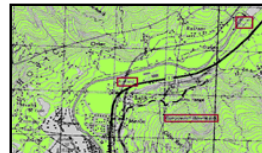
[Maybury, 1991]



Plan Operator

NAME Identify-location-linguistically-&-visually
 HEADER Identify(S, H, entity)
 CONSTRAINTS Cartographic-Entity?(entity)
 PRECONDITIONS Visible(entity) &
 WANT(S, KNOW(H, Location(entity)))
 EFFECTS KNOW(H, Location(entity))
 DECOMPOSITION Indicate-Deictically(S, H, entity)
 Assert(S, H, Location(entity))

Multimedia Presentation



Direction Commu

Explain-Route(SYSTEM, USER, #<Main>)
 ↓
 Display-Region(#<Wiesbaden>)
 ↓
 Indicate-Deictically(SYSTEM, USER, #<Wiesbaden>)
 ↓
 Highlight(#<Wiesbaden>)
 ↓
 Command(SYSTEM, USER, Go(#<Wiesbaden>, #<Road Segment A66>, #<Frankfurt-am-Main>))
 ↓
 Indicate-Deictically(SYSTEM, USER, #<Frankfurt-am-Main>)
 ↓
 Highlight(#<Road Segment A66>)

(Display map region around Wiesbaden)
 (highlight Wiesbaden) From Wiesbaden take
 Autobahn A66 Northeast for thirty-one
 kilometers to Frankfurt-am-Main. (highlight
 Autobahn A66)
 (indicate direction with blinking arrow)
 (highlight Frankfurt-am-Main)
 Frankfurt-am-Main is located at 50.11° latitude
 and 8.66° longitude.

ESSLI 04, André

Planning Material **and** Presentation



```
<?xml version='1.0' encoding='utf-8'?>
<!DOCTYPE Presentation SYSTEM "presentation.dtd">
<Presentation>
  <Slide>
    <Title>Photo of Victor Polay
    </Title>
    <Content>
      <Picture>polay.jpg</Picture>
    </Content>
    <Persona>
      SAY Here is a photo of Victor Polay;;;
      SLIDE++ = SAY Next Slide please.
    </Persona>
  </Slide>
</Presentation>
```

Contents
of slide

Presentation

E. André, K. Concepcion, I. Mani, and L. van Guilder. Autobriefer: A System for Authoring Narrated Briefing, In: O. Stock and M. Zancanaro (eds.) Multimodal Intelligent Information Presentation, Dordrecht: Kluwer, to appear.

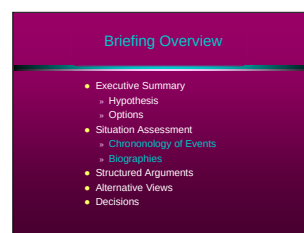
49

ESSLLI 04, André

Contents of a Slide



```
<Slide>
<Title> Briefing Overview</Title>
<Content>
  <S>Executive Summary
    <S>Hypothesis</S>
    <S>Options</S></S>
  <S>Situation Assessment
    <S>Chronology of Events</S>
    <S>Biographies</S></S>
  <S>Structured Arguments</S>
  <S>Alternative Views</S>
  <S>Decisions</S>
</Content>
<Persona>...</Persona>
</Slide>
```



50

ESSLLI 04, André

Representation of Presentation Knowledge



- **Example of Plan Operator**

```
(define-plan-operator
  :header (A0 (PresentPersona ?id ?SlideNumber))
  :constraints (Label ?id ?label)
  :inferiors (
    (A1 (sAddPPCode ("PLAY OneHand;;")))
    (A2 (SAddPPCode ("SAY Here is a " ?label ";;"))))
)
```

- **Entry in Knowledge Base:**

(Label "I123" "Photo of Martin Meier")

51
ESSLLI 04, André

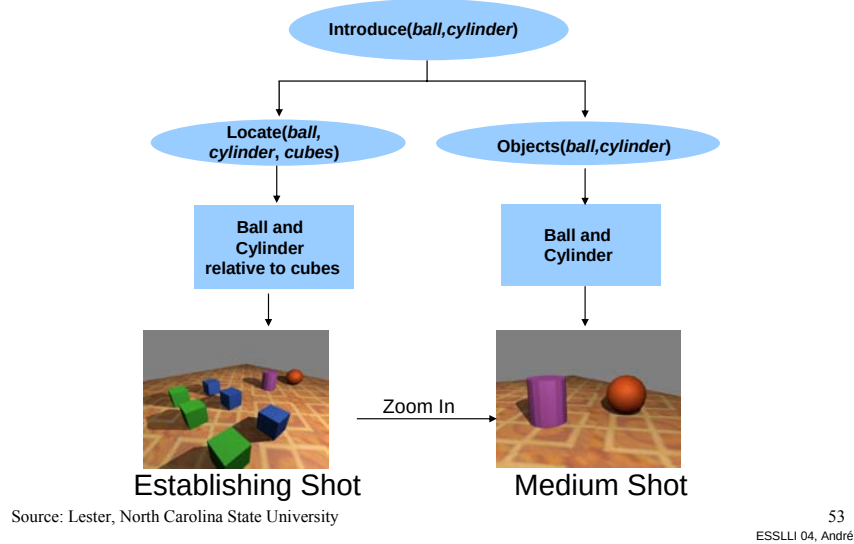
Temporal Coordination of Media



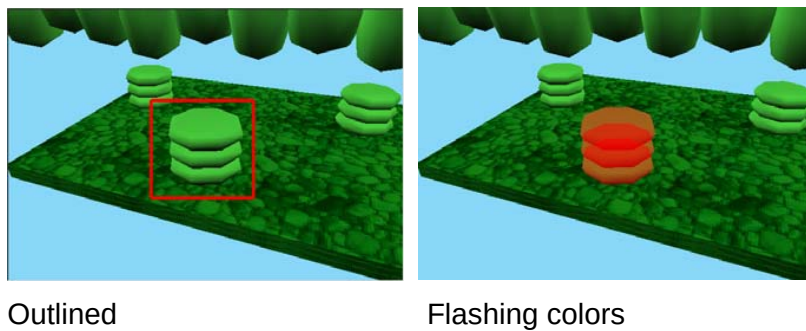
- Coordination of video, speech, text and image elements
- Coordination of camera movements, object movements and speech
- Coordination of speech and annotations
- Synchronization of speech and lip movements

52
ESSLLI 04, André

Shot Compositions and Transitions



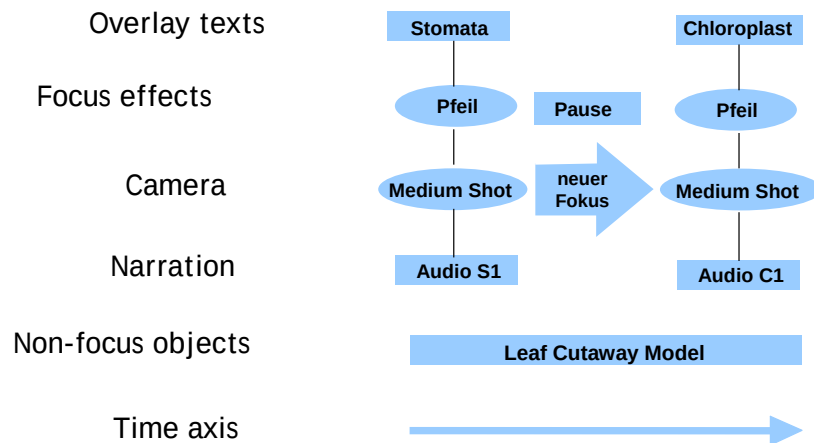
Focus Effects



Source: Lester, North Carolina State University

Brent Daniel, William Bares, Charles Callaway, and James Lester Student-Sensitive Multimodal Explanation Generation for 3D Learning Environments *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pp. 114-120.
54
ESSLLI 04, André

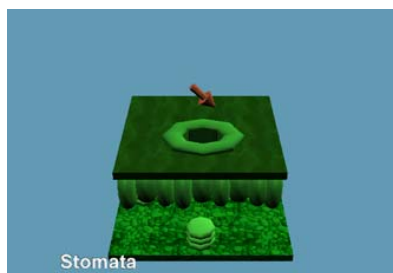
Animation Timeline Example



Source: Lester, North Carolina State University

55
ESSLI 04, André

Animation Explanation Example



Identify the stomata

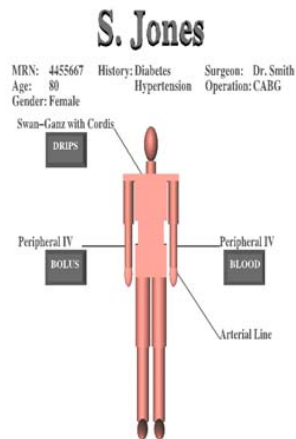


Identify the chloroplast

Source: Lester, North Carolina State University

56
ESSLI 04, André

Coordination of Speech and Annotations



S. Jones

MRN: 4455667 History: Diabetes Surgeon: Dr. Smith
Age: 80 Hypertension Operation: CABG
Gender: Female

Ms. Jones is an 80 year old, diabetic, hypertensive female patient

S. Jones

MRN: 4455667 History: Diabetes Surgeon: Dr. Smith
Age: 80 Hypertension Operation: CABG
Gender: Female

of Dr. Smith undergoing CABG.

Mukesh Dalal, Steven Feiner, Kathleen McKeown, Shimei Pan, Michelle X. Zhou, Tobias Höllerer, James Shaw, Yong Feng, Jeanne Fromer: Negotiation for Automated Generation of Temporal Multimedia Presentations. ACM Multimedia 1996: 55-64

57
ESSLI 04, André

Synchronization of Intonation, Gestures and Gaze in BEAT



- Extraction of linguistic and contextual information from text
- Generation of proposals for intonation, gestures and gaze
- Synchronization of selected behaviors



You just have to type in some text ...

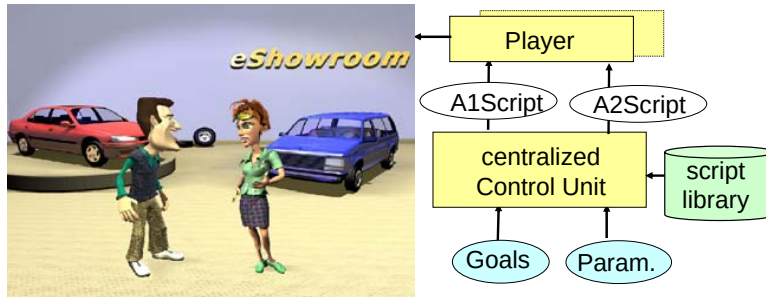


... this is a good or a bad thing

Cassell, J., Vilhjálmsón, H., Bickmore, T. (2001) "BEAT: the Behavior Expression Animation Toolkit." Proceedings of SIGGRAPH '01, pp. 477-486. August 12-17, Los Angeles, CA.

58
ESSLI 04, André

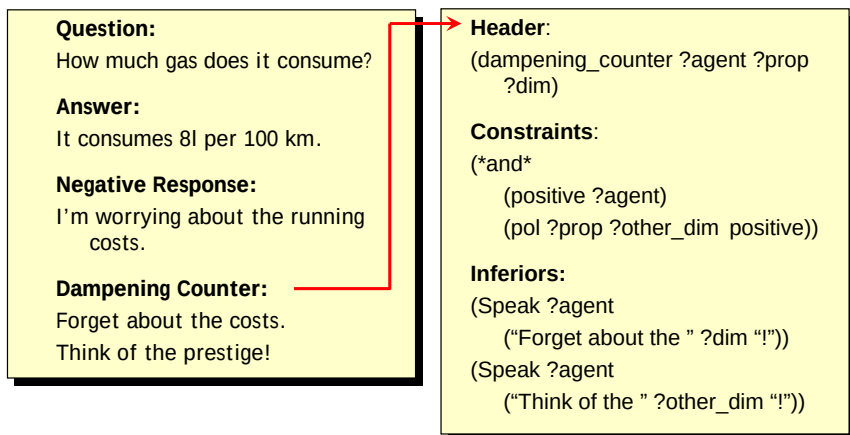
Automated Generation of Simulated Dialogues



System as a screen writer which allocates multimodal dialogue contributions to the single agents.

Elisabeth André, Thomas Rist, Susanne van Mulken, Martin Klesen, and Stephan Baldes , "The Automated Design of Believable Dialogues for Animated Presentation Teams", Embodied Conversational Agents , J. Cassell, S. Prevost, J. Sullivan, and E. Churchill, The MIT Press, Cambridge , 2000 59
ESSLLI 04, André

Example of a Dialogue Strategy



60
ESSLLI 04, André

Summary



- Media Coordination via
 - » Uniform Representation formats
 - » Common data structure which multiple media-specific components may access
 - » Declarative specification of coordination rules, e.g. for media allocation
- Heavily influenced by methods for natural language generation

61

ESSLLI 04, André

Overview Articles and Books



- André, E. "Natural Language in Multimedia/Multimodal Systems", Handbook of Computational Linguistics , R. Mitkov, 650-669, Oxford University Press, 2003
- Declerck, T. and André, E. L'indexation conceptuelle de documents multilingues et multimédias, Multilinguisme et traitement de l'information (Traité des sciences et techniques de l'information) , Lavoisier, 2002
- Maybury, M. T. ed. 1993. Intelligent Multimedia Interfaces. AAAI/MIT Press.
- Maybury, M. T. ed. 1997. Intelligent Multimedia Information Retrieval. AAAI/MIT Press. 350 pages, ISBN 0-262-63179-2.
- Maybury, M. T. and Wahlster, W. eds. 1998. Readings in Intelligent User Interfaces. Morgan Kaufmann Press. ISBN: 1-55860-444-8.
- André, E. The Generation of Multimedia Documents, in: R. Dale, H. Moisl and H. Somers: A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text, Marcel Dekker Inc., 305-327, 2000.

62

ESSLLI 04, André

Content-based Indexing and Searching of Multimedia Documents

Thierry Declerck † Jan Kuper ‡ Horacio Saggion Ł Anna Samiotou § and Peter Wittenburg ¶

DFKI GmbH, Language Technology Department †

Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany

Department of Computer Science, University of Twente ‡

PO Box 217, 7500 AE Enschede, The Netherlands

Department of Computer Science, University Sheffield Ł

Regent Court, 211 Portobello St., Sheffield S1 4DP, UK

ESTeam AB §

Sikelianou 8-10, 146 71 Athens, Greece

Max Planck Institute for Psycholinguistics ¶

Wundtlaan 1, PB 310, 6500 AH Nijmegen, The Netherlands

Abstract. This chapter describes the role *natural language processing* (NLP) can play for multimedia applications. As an example of such an application, we present an approach dealing with the automatic conceptual indexing of multimedia documents which can then be searched by semantic categories instead of key words. This work has been investigated and developed in the EU funded project MUMIS. A novelty of the approach is to exploit multiple sources of information relating to video content (for example the rich range of textual and transcribed sources covering soccer games).

1 Introduction

This chapter describes the role *natural language processing* (NLP) can play in the automatic conceptual indexing of multimedia documents which can then be searched by semantic categories instead of key words. This topic was a key issue in the MUMIS project¹. A novelty of the approach developed in MUMIS is to exploit multiple sources of information relating to video content (for example the rich range of textual and transcribed sources covering soccer games).

In the first part of the paper (section 2) we will propose a general discussion on the role that can be played by NLP for multimedia application. This overview is widely based on [2] and [3]. In the rest of the paper we will exemplify the general topic with the detailed presentation of the MUMIS project that is concerned with the topic of multimedia indexing and searching, where we will also reserve a large part to the description of methods in use in IE. The presentation of this project is an extension and update of [11].

¹MUMIS is a project within the Information Society Program (IST) of the European Union, section Human Language Technology (HLT). See for more information <http://parlevink.cs.utwente.nl/projects/mumis/>.

2 The Role of NLP for Multimedia Applications

[2] and [3] give an overview of the role that can be played by NLP in multimodal and multimedia systems. We summarize here the central points of those studies.

2.1 Multimodal and multimedia systems

The terms *multimedia* and *multimodal* are often source of confusion. [2] adopts the definitions as proposed in [23], which establishes a distinction between the terms *medium*, *mode* and *code*. The term *mode* (or *modality*) refers to the type of perception concerned, being for example visual, auditory or olfactory perception. The term *medium* refers to the carrier of (CD-ROM, paper etc.), to the devices (microphone, screen, loudspeakers etc.), as well to the distinct types of information (texts, audio or video sequences). The term *code* refers to the particular means of encoding information (sign languages or pictorial languages). One can speak of a multimedia system if this allows to *generate* and/or to *analyse* multimedia/multimodal information or provide some *access* to archives of multiple media. In existing applications, often the process of analysis applies only to multimodal data, whereas generation is concerned with the production of multimedia information.

2.2 Integration of modalities

We speak in the case of analysis of a process of *integration of modalities*, since all the available modalities need to be merged at a more abstract level in order to take the maximal advantage of every modality involved in the application. Certain representation formalisms, as they have been defined in the context of advanced NLP (see for example [22]), can play a central role in this process of fusion. So the well-defined technique of *unification* of typed feature structures, combined with a chart parser, is used in a system described in [18]. Using this formalism allows to build a semantic representation that is common to all modalities involved in the application, unifying all the particular semantic contributions on the base of their representation in typed feature structures

2.3 Media Coordination

In the case of the *generation* of multimedia material including natural language, for the purpose of *multimedia presentation*, one can speak of a process of *media coordination*: it is not enough to merge various media in order to obtain a coherent presentation of the distinct media involved. The information contained in the various media has to be very carefully put into relation if one wants to obtain real complementarities of media in the final presentation of the global information. And since systems for *natural language generation* have been always confronted with this problem of selecting and organising various contributions for the generation of an utterance, they can provide for a very valuable model for the coordination of media in the context of the generation of multimedia presentations. A lot of systems for natural language generation are therefore said to be *plan-based*.

Essentially two concepts from the field of automatic natural language generation are playing a role in systems for multimedia presentation: the concept of *schema* (see [24]) and the

concept of *operator* (see [28]). The first concept describes a set of standard patterns of natural language discourse by means of predicates that indicate relationships between some parts of the presentation. The second concept deals more specially with the particular contributions of the distinct parts of the presentation and allows local revisions during the process of generation of the presentation.

2.4 Natural Language access to multimedia

Multimedia repositories of moving images, texts, and speech are becoming increasingly available. This together with the needs for 'video-on-demand' systems require fine-grain indexing and retrieval mechanisms allowing users access to specific segments of the repositories containing specific types of information.

It turns out that natural language can play a multiple role. It is first easier to access information contained in the multimedia archive using queries addressed to (transcript of) audio sequences or to the subtitles (if available) associated to the videos as to analyse the pictures themselves. It is further more appealing to access visual data by means of natural language, since the latter supports more flexible and efficient queries as the query based on image features. And ultimately natural language offers a good means for condensing visual information. The selected list of projects concerned with video indexing we give below is stressing this fact: at some point always some language data will be considered to support retrieval of images or videos.

In order to support this kind of natural language access, video material was usually manually annotated with 'metadata' such as people involved in the production of the visual record, places, dates, and keywords that capture the essential content of what is depicted. Still, there are a few problems with human annotation. First, the cost and time involved in the production of "surrogates" of the programme is extremely high; second, humans are rather subjective when assigning descriptions to visual records; and third, the level of annotation required to satisfy user's need can hardly be achieved with the use of mere keywords.

2.5 NLP techniques for indexing multimedia material

Many research projects have explored the use of parallel linguistic descriptions of the images (either still or moving) for automatic tasks such as indexing [56], classifying [58], or understanding [59] of visual records, instead of using only content-based (or visually-based) methods in use [54]. This is partly also due to the fact that NLP technologies are more mature for extracting meaning as the technologies in use in the field of image. Content-based indexing and retrieval of visual records is based on features such as colour, texture, and shape. Yet visual understanding is not well advanced and is very difficult even in closed domains. For example, visual analysis of the video of a football match can lead to the identification of interesting "content" like a shooting scene (i.e., the ball moving towards the goal) [60], but this image analysis approach will hardly ever detect who is the main actor involved in that scene (i.e., the shooter).

For accessing visual information with the help of natural language, certain systems make use of a shallow analysis of linguistic data associated with pictures, like the transcripts of audio comments or subtitles. In most of the case this is already enough in order to provide for a first classification and indexation of the visual data (see for example [20], [46] or [47]). Other

systems use more sophisticated linguistic analysis, like *information extraction* (IE): the detection of *named entities* and of standard linguistic patterns can help the multimedia retrieval systems to filter out non-relevant sequences (for example the introduction of speakers in news broadcasting). An example of such systems is given in the "Broadcast News Navigator" developed at MITRE (see [26]). The MUMIS projet, described in details below, is going even further, since full IE systems are analysing a set of so-called "collateral" (parallel) documents and produce unified conceptual annotations, including metadata information that is used for indexing the video material and supports thus concept-based queries on a multimedia archive. A similar approach is described in [37], where the domain of application is classical dance. In this work only a small set of textual documents is considered, in a monolingual setting.

3 Related approaches and projects

MUMIS differs in many significant ways from technologies developed in former or running projects. Most closely related to the thematic focus of MUMIS are the HLT projects Pop-Eye ([47], [20]) and OLIVE ([46], [19]). Pop-Eye used subtitles to index video streams and offered time-stamped texts to satisfy a user query, on request displaying a storyboard or video fragment corresponding to the text hit. OLIVE used automatic speech recognition to generate transcriptions of the sound tracks of news reports, which were then indexed and used in ways similar to the Pop-Eye project; both projects used fuzzy matching *information retrieval* (IR) algorithms to search and retrieve text, offering limited multilingual access to texts. Instead of using IR methods to index and search the transcriptions, MUMIS generates formal annotations building on and abstracting from a whole set of available documents, and merges information annotations from different media sources. The fusion result is then used to direct retrieval, through interface techniques such as pop-up menus, keyword lists, and so on. Search takes the user direct to the storyboard and video clippings.

The Informedia project at Carnegie-Mellon-University [42] has a similar conceptual baseline to MUMIS. The innovative contribution of MUMIS is that it uses a variety of multilingual information sources and fuses them on the basis of formal domain-specific annotations. Where Informedia primarily focuses on special applications, MUMIS aims at the advancement and integrability of *human language technology* (HLT) enhanced modules to enable information filtering beyond the textual domain. Therefore, MUMIS can be seen as complementary to Informedia with extensions typical for the multilingual European context.

The THISL project [49] is about spoken document retrieval, i.e., automatic speech recognition is used to auto-transcribe news reports and then information retrieval is carried out on this information. One main focus of THISL is to improve speech recognition. Compared to MUMIS it lacks the strong language processing aspects, the fusion of multilingual sources, and the multimedia delivery.

Columbia University was running a project [53] to use textual annotations of video streams to indicate moments of interest, in order to limit the scope of the video processing task, which requires extreme CPU capacities. So the focus is on finding strategies to limit video processing. The University of Massachusetts (Amherst) is also running projects about video indexing [50], but these focus on the combination of text and images. Associated text is used to facilitate indexing of video content. The aspect of generating formal annotations from textual documents is not part of those projects.

Much work has been done on video and image processing (Virage [42], the EUROMEDIA

project [40], Surfimage [48], the ISIS project [43], IBM's Media Miner, projects funded under the NSF Stimulate program [45], and many others). This technology in general can indicate, for example, scene changes using very low-level cues and to extract key frames at those instances to form a storyboard for easy video access. The project ASSAVID (see [5] and [6]) is giving an excellent example of such a technology for the sport domain, and is giving fruitful hints on the still unsolved issue of precisely adjusting time-coded events extracted from textual documents and the places in the video streaming, as detecting on the base of the analysis of picture features. The use of MPEG-7 ([9], [1] and [36]) as a standard for multimedia annotation, including textual annotation, can probably help a lot in providing future solutions to this issue.

Some institutions are running projects to detect subtitles in the video scene and create a textual annotation. This task is very difficult, given a sequence of real scenes with moving backgrounds and so on. Even more ambitious tasks such as finding real patterns in real movies (tracing the course of the ball in a soccer match, for example) have still some way to go.

4 MUMIS: a Multimedia Indexing and Searching Environment

MUMIS is proposing an integrated solution to the problem of multimedia indexing and search. The solution consists of using information extracted from different sources (structured, semi-structured, free, etc.), modalities (text, speech), and languages (English, German, Dutch) all describing the same event to carry out data-base population, indexing, and search.

MUMIS makes an intensive use of linguistic and semantic based annotations (see also the chapter on linguistic annotation for the web in this volume), coupled with domain-specific information, in order to generate formal annotations of events that can serve as index for videos querying. MUMIS applies IE technologies on multilingual and multimedia information from multiple sources. The novelty of the approach is not only the use of these 'heterogeneous' sources of information but also the combination or cross-source fusion of the information obtained from each source. Single-document, single-language information extraction is carried out by independent systems that share a semantic model and multi-lingual lexicon of the domain. The result of all information extraction systems is merged by a process of alignment and rule-based reasoning that also uses the semantic model.

For this purpose the project makes use of data from different media (textual documents, radio and television broadcasts) in different languages (Dutch, English and German) to build a specialized set of lexicons and an ontology for the selected domain (soccer). It also digitizes non-text data and applies speech recognition techniques to extract text for the purpose of annotation.

Audio material is being analysed by Phicos [61], an HMM-based recognition system, in order to obtain transcriptions of the football commentaries (spontaneous speech). It uses acoustic models, word-based language models (unigram and bigram) and a lexicon. For Dutch, English, and German different recognition systems have been developed, i.e. different phone sets, lexicons, and language models are used. Transcriptions for 14 German, 3 Dutch, and 8 English matches have been produced. [39] gives more details on the *automatic speech recognition* (ASR) and the transcription work done in the context of MUMIS.

The core linguistic processing for the annotation of the multimedia material consists of advanced information extraction techniques for identifying, collecting and normalizing significant text elements (such as the names of players in a team, goals scored, time points or

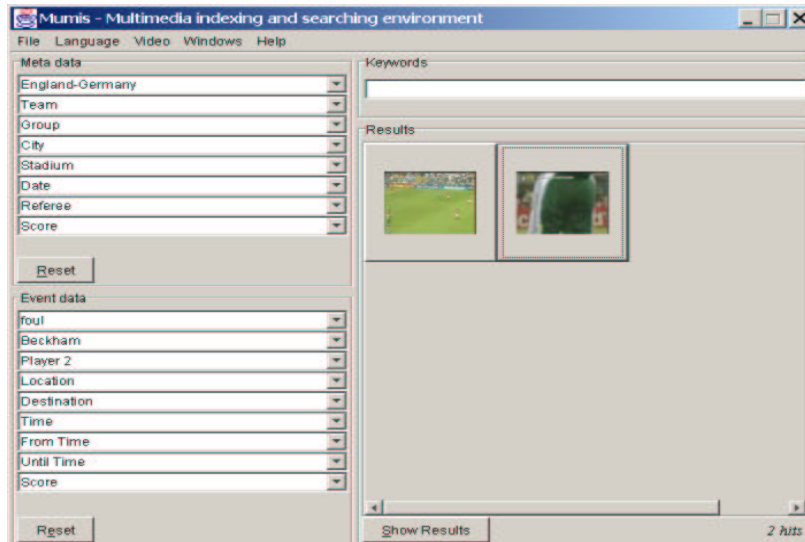


Figure 1: MUMIS User Interface. Thumbnails proposed for the query “Show me the fouls committed by Beckham”

sequences etc.) which are critical for the appropriate annotation of the multimedia material in the case of soccer. One system per language has been used or developed.

Each system delivers a XML aoutput, which serves as the input of a *merging component*, whose necessity in the project is due to the fact that MUMIS is accessing and processing multiple sources from distinct media in distinct languages. The merging tool is combining the semantically related annotations generated from those different data sources, and detect inconsistencies and/or redundancies within the combined annotations. The merged annotations are then stored in a database, where they will be combined with relevant metadata that are also automatically extracted from the textual documents.

Keyframes extraction from MPEG movies around a set of pre-defined time marks - result of the information extraction component - is being carried out to populate the database. JPEG keyframes images are extracted that serve for quick inspection in the user interface. The software used for off-line keyframe extraction is Spikes: it takes a movie file, a list of times stamps, and the size of the keyframe and produces a list of keyframes.

The on-line part of MUMIS consists of a state of the art user interface allowing the user to query the multimedia database (e.g., “The fouls committed by Beckham”). The user is first presented with selected video key-frames as thumbnails that can be played obtaining the corresponding video and audio fragments. Figure 1 shows the results obtained when selecting the fouls committed by Beckham in the England-Germany match.²

4.1 Formal Annotations for the Soccer Domain

Soccer has been chosen as the domain to test and apply the algorithms to be developed. There are a number of reasons for this choice: availability of people willing to help in analyzing user

²A web version of the demonstrator is under construction at the time of the redaction of this chapter.

requirements, existence of many information sources in several and possibly large economic and public interest.

The principles and methods derived from this domain can be applied to other as well. This has been shown already in the context of text-based information extraction, for which methodologies for a fast adaptation to new domains have been developed ([29], [30] and [10]). And generally speaking the use of IE for automatic annotation of multimedia document has the advantage of providing, besides the results of the (shallow) syntactic processing, accurate semantic (or content/conceptual) information (and thus potential annotation) for specific pre-defined domains, since a mapping from the linguistically analyzed relevant text parts can be mapped onto an unambiguous conceptual description. Thus in a sense it can be assumed that IE is supporting the word sense disambiguation task (see the chapter on linguistic annotation for the web in this volume).

It is also commonly assumed (see among others [8]) that IE occupies an intermediate place between *information retrieval* (with few linguistic knowledge involved) and *text understanding* (involving the full deep linguistic analysis and being still not realized for the time being.). IE being robust but offering only a partial (but mostly accurate) syntactic and content analysis, it can be said that this language technology is actually filling the gap between available low-level annotated/indexed documents and corpora and the desirable full content annotation of those documents and corpora. This is the reason why MUMIS has chosen this technology for providing automatic annotation (at distinct linguistic and domain-specific levels) of multimedia material, allowing thus to add queryable "content information" to this material. And if MUMIS was not explicitly designed for supporting knowledge management tasks, it nevertheless suggests that the meaningful organization of domain-specific multimedia material proposed by the project can be adapted to the organization of the distributed information of an enterprise and thus support the sharing and access to companies expertise and know-how.

4.2 The Multimedia Material in MUMIS

The MUMIS project is about automatic indexing of videos of soccer matches with formal annotations and querying that information to get immediate access to interesting video fragments. For this purpose the project chose the European Football Championships 2000 in Belgium and the Netherlands as its main database. A major project goal is to merge the formal annotations extracted from textual and audio material (including the audio part of videos) on the EURO 2000 in three languages: English, German, Dutch. The material MUMIS has to process can be classified in the following way:

1. Reports from Newspapers (reports about specific games, general reports) which is classified as free texts (FrT)
2. Tickers, close captions, Action-Databases which are classified as semi-formal texts (SFT)
3. Formal descriptions about specific games which are classified as formal texts (FoT)
4. Audio material recorded from radio and TV broadcasts
5. Video material recorded from TV broadcasts

1. to 4. is being used for automatically generating formal annotations in order to index 5. MUMIS is investigating the precise contribution of each source of information for the overall goal of the creation of merged formal annotations on events and other relevant information. Examples of German FrT and SFT are given in figures 2 and 3.

England - Deutschland 1:0 (0:0)

England: Seaman (2,5) - G. Neville (3,5), Keown (3), Campbell (2), P. Neville (4,5) - Ince (3,5), Wise (5) - Beckham (4), Scholes (3) - Shearer (3), Owen (5) -
Trainer: Keegan

Deutschland: Kahn (2) - Matthaeus (3) - Babbel (3,5), Nowotny (2,5) - Deisler (3), Hamann (2,5), Jeremies (3,5), Ziege (3,5) - Scholl (5) - Jancker (4), Kirsten (5) -
Trainer: Ribbeck

Eingewechselt: 61. Gerrard fuer Owen,
72. Barmby fuer Scholes - 70. Rink fuer Kirsten,
72. Ballack fuer Deisler, 78. Bode fuer Jeremies

Tore: 1:0 Shearer (53., Kopfball, Vorarbeit Beckham)

Schiedsrichter: Collina, Pierluigi (Viareggio), Note 2 -
bis auf eine falsche Abseits-Entscheidung souverain und sicher

Zuschauer: 30000 (ausverkauft)

Gelbe Karten: Beckham - Babbel, Jeremies

Figure 2: Example of a so-called formal text, where one can see that only 5 distinct time points can be extracted, concerning the player substitutions (“Eingewechselt”) and one goal (“Tore”).

Gruppe A: England - Deutschland 1:0 (0:0)

7. Ein Freistoss von Christian Ziege aus 25 Metern geht ueber das Tor.

12. Ziege flankt per Freistoss in den Strafraum und Jeremies versucht es per Kopfball, verfehlt den Kasten jedoch deutlich.

16. Scholes flankt gefaehrlich von der Torauslinie in den Fuenfmeterraum, doch Ziege hat aufgepasst und kann klaeren.

...

53. Beckham flankt per Freistoss an der deutschen Abwehr vorbei auf den Kopf von Alan Shearer, der voellig freistehend zum 1:0 fuer die Englaender verwandelt.

...

Figure 3: Example of a so-called semi-formal text, where one can see that here more time points are available, and that those can be complementary to the time points to be extracted from formal texts.

Since the information contained in formal texts can be considered as a database of true facts, they play an important role within MUMIS. But nevertheless they contain only little information about a game: the goals, the substitutions and some other few events (penalties, yellow and red cards). So there are only few time points available for indexing videos. Semi-formal texts (SFT), like live tickers on the web, are offering much more time points sequences, related with a higher diversity of events (goals scenes, fouls etc.) and seem to offer the best textual source for our purposes. Nevertheless the quality of the texts of online tickers is often quite poor. Free texts, like newspapers articles, have a high quality but the extraction of time points and their associated events in text is more difficult. Those texts also offer more background information that might be interesting for the real fans.

4.3 Processing Steps in MUMIS

4.3.1 Media Pre-Processing

Media material has been delivered in various formats (AudioDAT, AudioCassettes, Hi-8 video cassettes, DV video cassettes etc) and qualities. All audio signals (also those which are part of the video recordings) are digitized and stored in an audio archive. Audio digitization is done with 20 kHz sample frequency, the format generated is according to the de-facto "wav" standard. For digitization any available tool can be used such as SoundForge.

Video information (including the audio component) of selected games have been digitized into MPEG1 streams first. Later it will be encoded in MPEG2 streams. While the quality of MPEG1 is certainly not satisfying to the end-user, its bandwidth and CPU requirements are moderate for current computer and network technology. The mean bit rate for MPEG1 streams is about 1.5 Mbps. Current state-of-the-art computers can render MPEG1 streams in real time and many network connections (Intranet and even Internet) can support MPEG1. MPEG2 is specified for about 3 to 5 Mbps. Currently the top-end personal computers can render MPEG2, but MPEG2 is not yet supported for the most relevant player APIs such as JavaMediaFramework or Quicktime. When this support is given the MUMIS project will also offer MPEG2 quality.

For all separate audio recordings as for example from radio stations it has to be checked whether the time base is synchronous to that one of the corresponding video recordings. In case of larger deviations a time base correction factor has to be estimated and stored for later use. Given that the annotations cannot be created with too high accuracy a certain time base deviation will be accepted. For part of the audio signals manual transcriptions have to be generated to train the speech recognizers. These transcripts will be delivered in XML-structured files.

Since keyframes will be needed in the user interface, the MUMIS project will develop software that easily can generate such keyframes around a set of pre-defined time marks. Time marks will be the result of information extraction processes, since the corresponding formal annotations is referring to specific moments in time. The software to be written has to extract the set of time marks from the XML-structured formal annotation file and extract a set of keyframes from the MPEG streams around those time marks. A set of keyframes will be extracted around the indicated moments in time, since the estimated times will not be exact and since the video scenes at such decisive moments are changing rapidly. There is a chance to miss the interesting scene by using keyframes and just see for example spectators. Taking a number of keyframes increases the chance to grab meaningful frames.

4.3.2 Multilingual Automatic Speech Recognition

Domain specific language models have been trained. The training can be bootstrapped from written reports of soccer matches, but substantial amounts of transcribed recordings of commentaries on matches are also required. Novel techniques have been investigated to interpolate the base-line language models of the Automatic Speech Recognition (ASR) systems and the domain specific models. Moreover, techniques must be developed to adapt the vocabularies and the language models to reflect the specific conditions of a match (e.g., the names players have to be added to the vocabulary, with the proper bias in the language model). In addition, the acoustic models have been adapted to cope with the background noise present

in most recordings.

Automatic speech recognition of the sound tracks of television and (especially) radio programmes can make use of closed caption subtitle texts and information extracted from formal texts to help in finding interesting sequences and automatically transcribing them. Further, the domain lexicons is helping with keyword and topic spotting. Around such text islands ASR is used to transcribe the spoken soundtrack. The ASR system has been enriched with lexica containing more keywords, to increase the number of sequence types that can be identified and automatically transcribed. Preliminary results and problems have been documented in [39]. Due to the extreme background noise present in the audio files, the results of ASR can not be expected to be very good. But transcripts are giving a very valuable: the time of the utterance is given in the encoding.

An example of a (manually improved) transcription from English TV comment is given below. We can not go in details here, but one can see that transcription of audio materials can be quite challenging for NLP, since a lot of utterances are causing problems, like the values of the word feature *w* show: “Sh”, “eh”.

```
<tau ref="407" s="EngDldE2" tb="4724.396" te="4728.254" tt="eq" tq="man">
  <tw ref="407.1" tb="4724.396" te="4727.373" ... w="I"/>
  <tw ref="407.2" tb="4724.396" te="4727.373" ... w="don't"/>
  <tw ref="407.3" tb="4724.396" te="4727.373" ... w="know"/>
  <tw ref="407.4" tb="4724.396" te="4727.373" ... w="if"/>
  <tw ref="407.5" tb="4724.396" te="4727.373" ... w="it"/>
  <tw ref="407.6" tb="4724.396" te="4727.373" ... w="was"/>
  <tw ref="407.7" tb="4724.396" te="4727.373" ... w="Jeremies"/>
  <tw ref="407.8" tb="4724.396" te="4727.373" ... w="or"/>
  <tw ref="407.9" tb="4724.396" te="4727.373" ... w="Sh"/>
  <tw ref="407.10" tb="4724.396" te="4727.373" ... w="eh"/>
  <tw ref="407.11" tb="4727.373" te="4728.254" ... w="Scholes"/>
  <tw ref="407.12" tb="4727.373" te="4728.254" ... w="one"/>
  <tw ref="407.13" tb="4727.373" te="4728.254" ... w="of"/>
  <tw ref="407.14" tb="4727.373" te="4728.254" ... w="the"/>
</tau>
....
```

4.3.3 Domain Modelling

Any domain-specific NLP applications need a model of the domain under consideration. This allow to process the documents with the help of generic NLP tools (see also here the chapter on linguistic annotation for the web in this volume), which can thus be used for every type of applications. The results of the NL analysis can be mapped onto a domain-specific organization of knowledge and produce a corresponding set of annotations, which can be very helpful, for example, for knowledge management applications.

Domain Ontology: An analysis of the domain and a user study led the project to propose 31 types of event for a football match (kick-off, substitution, goal, red card, etc.) that need to be identified in the sources in order to produce a semantic index. The elements to be extracted that are associated with these events are: players, teams, scores, and locations on the field. Time codes need to be extracted from the sources as well. All events and their participants have been organized in an ontology that has more than 300 concepts related as a 'is-a' hierarchy. This ontology is giving the base for the reasoning component of the prototype. The link between the concept and the three languages consists of a flexible XML

format which maps concepts into lexical entries, as show below.

Multilingual Domain Lexicon Building: All the collected textual data for the soccer domain are used for building the multilingual domain lexicons. This data can be in XML, HTML, plain text format, etc. A number of automatic processes are used for the lexicon building, first on a monolingual and secondly on a multilingual level. Manual browsing and editing is taking place, mainly in order to provide the semantic links to the terms, but also for the fine-tuning of the lexicon according to the domain knowledge.

Domain lexicons are built for three languages, namely English, German and Dutch. The lexicons are delivered in a fully structured, XML-compliant, TMX-format (Translation Memory eXchange format, see [62]) An example of the MUMIS lexicon generated so far:

```
<lex-element id="3.9" concept="Corner">
  <term-lang lang="DE" type="main">Ecke</term-lang>
  <term-lang lang="EN" type="main">corner</term-lang>
  <term-lang lang="NL" type="main">hoekschoop</term-lang>
  <definition>A corner kick is awarded when the whole ball has:
  Crossed the whole goal line in the air or on the ground, was last
  touched by a defender and goal was not scored</definition>
  <term-lang lang="DE" type="synonym">Eckball</term-lang>
</lex-element>
```

This lexicon is being revisited to conform to more formal representation languages (as discussed in the chapter on linguistic annotation in this volume).

4.3.4 Generation of Formal Annotations

The formal annotations are generated by the IE technology and are reflecting the typical output of IE systems, i.e. instantiated domain-specific templates or event tables. The slots to be filled by the systems are basically entities (player, teams etc.), relations (player_of, opponents etc.) and events (goal, substitution etc.), which are all derived from the current version of the domain ontology and can be queried for in the online component of the *MUMIS* system. All the templates associated with an event are including a time slot to be filled if the corresponding information is available in a least one of the sources consulted during the IE procedure. This time information is necessary for the indexing of the video material.

The IE systems are applying to distinct sources (FoT, FrT etc.) but they are not concerned with achieving consistency in the IE result on distinct sources about the same event (game): this is the task of the merging tools, described below.

Since the distinct textual sources are differently structured, from "formal" to "free" texts, the IE systems involved have adopted a modular approach: regular expressions for the detection of Named Entities in the case of formal texts, full chunk parsing for the free texts. On the base of the factual information extracted from the formal texts, the IE systems are also building dynamic databases on certain entities (like name and age of the players, the clubs they are normally playing for, etc.) or certain metadata (final score), which can be used at the next level of processing. Examples of annotations generated from formal and semi-formal texts are given in figure 4

4.3.5 Merging

Merging here is the process of checking whether various linguistic annotations from different sources can refer to the same entity or event. The merging component combines in fact the

```

<DOM=SOCCER TYPE=SUBSTITUTION TIME=61:00 SYN=NP
  PLAYER_OUT=Owen PLAYER_IN=Gerrard>

<DOM=SOCCER TYPE=GOAL GOAL_TYPE=Kopfball SYN=NP TIME=53:00
  SCORE=1:0 PLAYER=Shearer PREPARATION=Beckham>

```

Figure 4: Two examples of time-marks and their associated events (substitution and goal) automatically extracted from the formal text above.

```

7. Ein Freistoss von Christian Ziege aus 25 Metern geht ueber das Tor.
(7. A 25-meter free-kick by Christian Ziege goes over the goal.)

<EVENTS>
  1-event_TYPE="Free-kick" 1-event_DIST="Meter-from_(25)" 1-event_1_PLAYER="Ziege"
  1-event_EVENT_CLASS="goal_scene_fail" 1-event_ART="Goal" 1-event_TIME="7:00"
</EVENTS>

<META> DOM_NAME="SOCCER"</META>

<PLAYER>
  <PLAYER_NAME>Ziege</PLAYER_NAME>
  <PLAYER_NOTE>#(3,5)</PLAYER_NOTE>
  <PLAYER_POS>#4 ##3</PLAYER_POS>
  <PLAYER_NUMBER>##17</PLAYER_NUMBER>
  <PLAYER_OLD>##28</PLAYER_OLD>
  <PLAYER_CLUB>##FC Middlesbrough</PLAYER_CLUB>
  <PLAYER_NO_PLAYS>##52</PLAYER_NO_PLAYS>
</PLAYER>

```

Figure 5: Example of the extraction of event with its associated relations entities and relations, taken from the same game. The information about the player is dynamically included from the processing of the formal texts. In this case from the merging from 2 different formal texts, marked with # and ## respectively.

annotated partial information as extracted from various sources, such that more complete (and consistent) and less redundant annotations for the indexing of video can be obtained. The merging component makes use of inferential mechanisms for supporting the fusion of information. The use of general representation languages supporting all kind of operations on (typed) features as well as multiple inheritance (see the chapter on linguistic annotation in this volume for a description of such formalisms), combined with inference mechanisms, supports a better sequential organization of some events in larger scenarios. It allows also the induction of some events that are not explicitly mentioned in the sources under consideration (or which the IE systems might not have detected).

An example of merging of annotations extracted from 2 different formal texts on one game is given in figure 5, where merged information about a player (*Ziege*) has been dynamically integrated into the processing of semi-formal texts. New information about a player available in the semi-formal texts can be backward added into the merged annotations. This makes merging also quite interesting for the automatic and dynamic updating of information encoded in databases.

4.3.6 User Interface Building

The user first interacts with a web-portal to start a MUMIS query session. An applet is being down-line loaded in case of showing the sc MUMIS demonstration. This applet mainly offers a query interface. The user then enters a query that either refers to metadata, formal annotations, or both. The on-line system searches for all formal annotations that meet the criteria of the query. In doing so it will find the appropriate meta-information and/or moments in some media recording. In case of meta-information it simply offers the information in scrollable text widgets. This is done in a structured way such that different type of information can easily be detected by the user. In the case that scenes of games are the result of queries about formal annotations the user interface first presents selected video keyframes as thumbnails with a direct indication of the corresponding metadata. The user can then ask for more metadata about the corresponding game or for more media data. The actual state of the demonstrator is shown in figure 1 above.

5 Standards for Multimedia Content

MUMIS is looking for a compliance with existing standards in the context of the processing of multimedia content on the computer and so will adhere to standards such as MPEG4, which defines how different media objects will be decoded and integrated at the receiving station, and the emerging standard MPEG7, which is about defining standards for annotations which can be seen as multimedia objects. Further, MUMIS will also maintain awareness of international discussions and developments in the area of multimedia streaming (RTP, RTSP, JMF...), and are following the discussions within the W3C consortium and the EBU that are also about standardizing descriptions of media content.

6 Role of MUMIS for the Annotation of Multimedia Content

To conclude on the project, we would like to list the points where we think MUMIS is contributing, directly or indirectly, to the extraction and to the access of multimedia content. MUMIS:

1. Uses multimedia (MM) and multilingual information sources;
2. Carries out multimedia indexing by applying information extraction to a well-delineated domain and using already existing information as constraints;
3. Uses and extends advanced language technology to automatically create formal annotations for MM content;
4. Merges information from many sources to improve the quality of the annotation database;
5. Applies advanced IE to the output of ASR and the combination of this with already existing knowledge;
6. Defines a complex information annotation structure, which is stored in a standard document type definition (DTD);
7. Integrates new methods into a query interface that is guided by domain knowledge (ontology and multilingual lexicons).

In this sense MUMIS is contributing in defining semantic structures of multimedia contents, at the level proposed by domain-specific IE analysis. The full machinery of IE, combined with ASR (and in the future with Image Analysis, so for example with the results of the ASSAVID project) can be used for multimedia contents development and so efficiently support cross-media (and cross-lingual) information retrieval and effective navigation within multimedia information interfaces. There seems thus that this technology can play a highly relevant role for the purposes of knowledge detection and management. This is probably specially valid for the merging component, which is eliminating redundancies in the annotations generated from sets of documents and establishing complex reference resolutions, thus simplifying the access to content (and knowledge) distributed over multiple documents and media, which are also increasingly available on the Web.

7 Conclusions

We have seen that NLP methods and techniques can be extended in order to be useful also for multimedia. Unification grammars can be helpful for the organisation and analysis of multimodal and multimedia data. The plan-based organisation of the content to be delivered in automatic natural language generation can be applied to the selection and structure of content in the context of the automatic generation of multimedia presentation. And the use of IR analysis techniques can support a complex but flexible access to digital multimedia archives. But NLP has also to open itself to new forms of communication made possible through the developments in multimedia and multimodal applications. NLP no longer can confine itself to the analysis or the generation of textual documents, but has to go for a level of analysis such that its results can be used for the organisation of a multimedia and multimodal complex around natural language. Only in this sense, one can really speak of human language technology (HLT), taking into account various communication means in the creation, distribution and access of content and knowledge over the Web.

8 Acknowledgements

This research been supported by EC grant IST-1999-10651 for the MUMIS project.

References

- [1] Adani N., Bugatti A., Leonardi R., and Migliorati P. Semantic description of multimedia documents: the Mpeg-7 approach. In *Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001*, Brescia, 2001.
- [2] André E. Natural Language in Multimedia/Multimodal Systems. In Mitkov R. (ed.), *Handbook of Computational Linguistics*, Oxford, 2000.
- [3] André E. The Generation of Multimedia Presentations. In *Handbook of Natural Language Processing*, Marcel Dekker, 2000.
- [4] Appelt D.E. An Introduction to Information Extraction. In *AI Communications* (vol. 12), 1999.
- [5] Assfalg J., Bertini M., Colombo C., and Del Bimbo A. Semantic annotations of sports videos. In *Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001*, Brescia, 2001.
- [6] Bertine M. and del Bimbo A. Multimedia content extraction from video streams. In *Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001*, Brescia, 2001.
- [7] Brugman H., Bontcheva K., Wittenburg P. and Cunningham H. Integrating Multimedia and Textual Software Architectures for Language Technology. Technical report MPI-TG-99-1, MPI for Psycholinguistics, Nijmegen, Netherlands, 1999.
- [8] Cunningham H. Information Extraction: A user Guide, Research Report CS-99-07, Department of Computer Science, University of Sheffield, May 1999.
- [9] Day N. MPEG-7 Applications: Multimedia Search and Retrieval. In *Proceedings of the First International Workshop on Multimedia Annotation, MMA-2001*, 2001.
- [10] Declerck T. and Neumann G. Using a parameterisable and domain-adaptive information extraction system for annotating large-scale corpora? *Proceedings of the Workshop Information Extraction meets Corpus Linguistics, LREC-2000*, 2000.
- [11] Declerck T., Wittenburg P., Cunningham H. The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment. *Proceedings of the Workshop on Human Language Technology and Knowledge Management, ACL-2001*, 2001.
- [12] Declerck T. A set of tools for integrating linguistic and non-linguistic information. *Proceedings of SAAKM 2002, ECAI 2002*, Lyon.
- [13] Djoerd H., de Jong F., Netter K. (Eds). 14th Twente Workshop on Language Technology, Language Technology in Multimedia Information Retrieval, TWLT 14, Enschede, Universiteit Twente, 1998.
- [14] Grishman R., Sundheim B. Message Understanding Conference – 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, 1996.
- [15] Heflin J., Hendler J., and Luke S. SHOE: A Knowledge Representation Language for Internet Applications. Technical Report CS-TR-4078. Department of Computer Science, University of Maryland, 1999.
- [16] Hellwig P. NATURAL LANGUAGE PARSERS A "Course in Cooking. *COLING-ACL '98– Pre-Conference Tutorial*. 1998.
- [17] Humphreys K., Gaizauskas R., Azzam S., Huyck C., Mitchell B., Cunningham H., Wilks Y. Description of the LaSIE-II System as used for MUC-7, in *Seventh Message Understanding Conference (MUC-7)*, <http://www.muc.saic.com/>, SAIC Information Extraction. 1998.
- [18] Johnston M. Unification-based Multimodal Parsing, In *Proceedings of the 17th International Conference on Computational Linguistics, COLING-98*, 1998.
- [19] de Jong F., Netter K. Olive: Speech-Based Video Retrieval. In Hiemstra D., de Jong F., Netter K. (Eds), *Language Technology in Multimedia Information Retrieval (Proceedings of the 14th Twente Workshop on Language Technology, TWLT 14)*, Enschede, Universiteit Twente, 1998.
- [20] de Jong F., Gauvin J., Hiemstra D., Netter K. Language-Based Multimedia Information Retrieval. In *Proceedings of the 6th Conference on Recherche d'Information Assistée par Ordinateur, RIAO-2000*, 2000.

- [21] de Jong F. and Westerveld T. MUMIS: Multimedia Indexing and Searching. Proceedings of the Content-Based Multimedia Indexing Workshop (CBMI2001), 2001.
- [22] Krieger H.-U., Schaefer U. TDL – a type description language for constraint-based grammars. In Proceedings of the 15th International Conference on Computational Linguistics, COLING-94, 1994.
- [23] Maybury M. Multimedia Interaction for the New Millenium. In Proceedings of Eurospeech 99, 1999.
- [24] McKeown K. Text generation, Cambridge University Press, 1985.
- [25] McNaugh J. (Ed), Information Extraction meets Corpus Linguistics. LREC-2000, 2000.
- [26] Merlino A., Morey D., Maybury M. Broadcast News Navigation using Story Segments. ACM International Multimedia Conference, 1997.
- [27] Miller, G.A. WordNet: A Lexical Database for English. Communications of the ACM 11. 1995.
- [28] Moore J., Paris C. Planning Text for Advisory Dialogues. In Proceedings of the 27th ACL, Vancouver, 1989.
- [29] Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann, 1995.
- [30] Seventh Message Understanding Conference (MUC-7), <http://www.muc.saic.com/>, SAIC Information Extraction, 1998.
- [31] Naphade Milid R. and T.S. Huang. Recognizing high-level concepts for video indexing. In Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001, Brescia, 2001.
- [32] Neumann G., Backofen R., Baur J., Becker M., Braun C. An Information Extraction Core System for Real World German Text Processing. In Proceedings of the 5th Conference on Applied Natural Language Processing, ANLP-97, 1997.
- [33] Piskorski J., Neumann G.. An Intelligent Text Extraction and Navigation System. In Proceedings of the 6th Conference on Recherche d'Information Assistee par Ordinateur, RIAO-2000, 2000.
- [34] Saggion H. , Cunningham H., Bontcheva K., Maynard D, Ursu C. Hamza O. and Wilks Y. Access to Multimedia Information through Multisource and Multilanguage Information Extraction. 7th Workshop on Applications of Natural Language to Information Systems (NLDB 2002), 2002.
- [35] Saggion H., Cunningham H., Maynard D., Bontcheva K., Hamza, O., Ursu, C. and Wilks, Y. Extracting Information for Information Indexing of Multimedia Material. Proceedings of 3rd Language Resources and Evaluation Conference (LREC'2002), 2002.
- [36] Salembier P. An overview of Mpeg-7 multimedia description schemes and of future visual information challenges for content-based indexing. In Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001, Brescia, 2001.
- [37] Salway A., Talking Pictures: Indexing and Representing Video with Collateral Texts. In Hiemstra D., de Jong F., Netter K. (Eds), Language Technology in Multimedia Information Retrieval (Proceedings of the 14th Twente Workshop on Language Technology, TWLT 14), Enschede, Universiteit Twente, 1998.
- [38] Staab S., Maedche A., Handschuh S. An Annotation Framework for the Semantic Web. In The First International Workshop on Multimedia Annotation, Tokyo, Japan, 2001.
- [39] Wester M., Kessens J.M. and Strik H. Goal-directed ASR in a Multimedia Indexing and Searching Environment (MUMIS). Proceedings of the 7th International Conference on Spoken Language Processing (ICLSP2002), 2002.
- [40] EUR: <http://www.foyer.de/euromedia/>
- [41] GDA: <http://www.csl.sony.co.jp/person/nagao/gda/>
- [42] INF: <http://www.informedia.cs.cmu.edu/>
- [43] ISI: <http://www.wins.uva.nl/research/isis/isisNS.html>
- [44] ISLE: http://www.ilc.pi.cnr.it/EAGLES/ISLE_Home_Page.htm

- [45] NSF: <http://www.nsf.gov/od/lpa/news/press/pr9714.htm>
- [46] OLI: <http://twentyone.tpd.tno.nl/olive>
- [47] POP: <http://twentyone.tpd.tno.nl/popeye>
- [48] SUR: <http://www-rocq.inria.fr/nastar/MM98/node1.html>
- [49] THI: <http://www.dcs.shef.ac.uk/research/groups/spandh/projects/thisl>
- [50] UMA: <http://ciir.cs.umass.edu/research/>
- [51] UNL: http://www.ias.unu.edu/research_prog/science_technology/universalnetwork_languge.html
- [52] VIR: <http://www.virage.com/>
- [53] COL: <http://www.cs.columbia.edu/hjing/sumDemo>
- [54] Veltkamp R. and Tanase M. Content-based Image Retrieval Systems: a survey. Technical report UU-CS-2000-34, Utrecht University, 2000.
- [55] Chang S.F., Chen, W., Meng H.J., Sundaram H. and Zhong D. A Fully Automated Content-based Video Search Engine Supporting Spatio Temporal Queries. IEEE Transactions on Circuits and Systems for Video Technology, 1998.
- [56] Netter K. Pop-Eye and OLIVE. Human Language as the Medium for Cross-lingual Multimedia Information Retrieval. Technical report, Language Technology Lab. DFKI GmbH, 1998.
- [57] Hampapur A. Semantic video indexing: approach and issues. In SIGMOD-Record, 1998.
- [58] Sable C. and Hatzivassiloglou V. Text-based approaches for the categorization of images. Proceedings of ECDL, 1999.
- [59] Srihari R.K. Automatic Indexing and Content-Based Retrieval of Captioned Images, Computer 28/9, 1995.
- [60] Gong Y., Sin L.T., Chuan C.H., Zhang H. and Sakauchi M. Automatic Parsing of TV Soccer Programs. Proceedings of the International Conference on Multimedia Computing and Systems (IEEE), 1995.
- [61] Steinbiss V., Ney H., Haeb-Umbach R., Tran B.-H., Essen U., Kneser R., Oerder M., Meier H.-G., Aubert X., Dugast C. and Geller D. The Philips Research System for Large-Vocabulary Continuous-Speech Recognition. In Proc. of Eurospeech '93, 1993.
- [62] <http://www.lisa.org/tmx/tmx.htm>.

July 2001

Overview of MPEG-7: Multimedia Content Description Interface

Philippe Salembier

Universitat Politecnica de Catalunya

Barcelona, Spain

philippe@gps.tsc.upc.es

(Thanks to Sylvie Jeannin for
contribution to Visual Ds)

Outline

- Objective, goals, requirements and applications, basic components of the MPEG-7 standard
- Systems tools and Description Definition Language
- Multimedia Description Schemes
- Visual Tools
- Audio Tools
- Relation with other standards

Motivation

- The multimedia context:
 - ◆ More information is in digital form and is on-line.
 - ◆ AV content covers: still pictures, audio, speech, video, graphics, 3D models, etc.
 - ◆ AV content is available at all bitrates and on all networks.
 - ◆ Increasing number of multimedia applications, services.

- Necessity of describing content:
 - ◆ Increasing amount of information.
 - ◆ More needs to have “information about the content”.
 - ◆ Difficult to manage (find, select, filter, organize, etc) content.
 - ◆ User: human or computational systems.

What is MPEG

- MPEG: Moving Picture Experts Group (Created in 1988)
- ISO (Int. Standards Organization) / IEC (Int. Electro-technical Commission)
 - ◆ ISO/IEC JTC 1 / SC 29 / WG 11
- Develop standards for the coded representation of moving pictures and associated audio

MPEG Standards

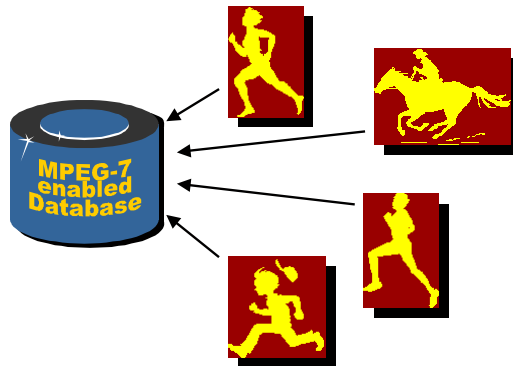
- **MPEG-1**: Storage of moving picture and audio on storage media (CD-ROM) 11 / 1992
- **MPEG-2**: Digital television 11 / 1994
- **MPEG-4**: Coding of natural and synthetic media objects for multimedia applications
v1: 09 / 1998
v2: 11 / 1999
- **MPEG-7**: Multimedia content description for AV material 08 / 2001
- **MPEG-21**: Digital audiovisual framework: Integration of multimedia technologies (identification, copyright, protection, etc.) 11 / 2001

Objective of MPEG-7

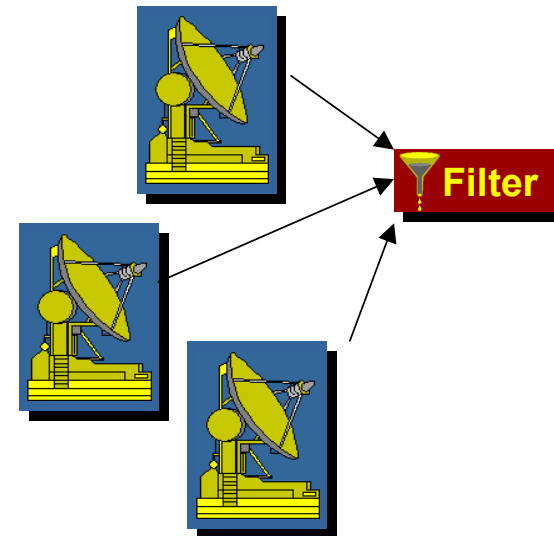
- Standardize content-based description for various types of audiovisual information
 - ◆ Enable fast and efficient content searching, filtering and identification
 - ◆ Describe several aspects of the content (low-level features, structure, semantic, models, collections, creation, etc.)
 - ◆ Address a large range of applications
($\Rightarrow$ user preferences)
- Types of audiovisual information:
 - ◆ Audio, speech
 - ◆ Moving video, still pictures, graphics, 3D models
 - ◆ Information on how objects are combined in scenes
- Descriptions independent of the data support
- Existing solutions for textual content or description

Type of applications

- Pull Applications: “Search Browsing”
 - ◆ Example: Search engines for Internet and databases



- Push Applications: “Filtering”
 - ◆ Example: Broadcast of video, Interactive TV



- Universal Multimedia Access: Adapt delivery to network / terminal characteristics
- Specialized Professional and Control Applications

Example of application areas

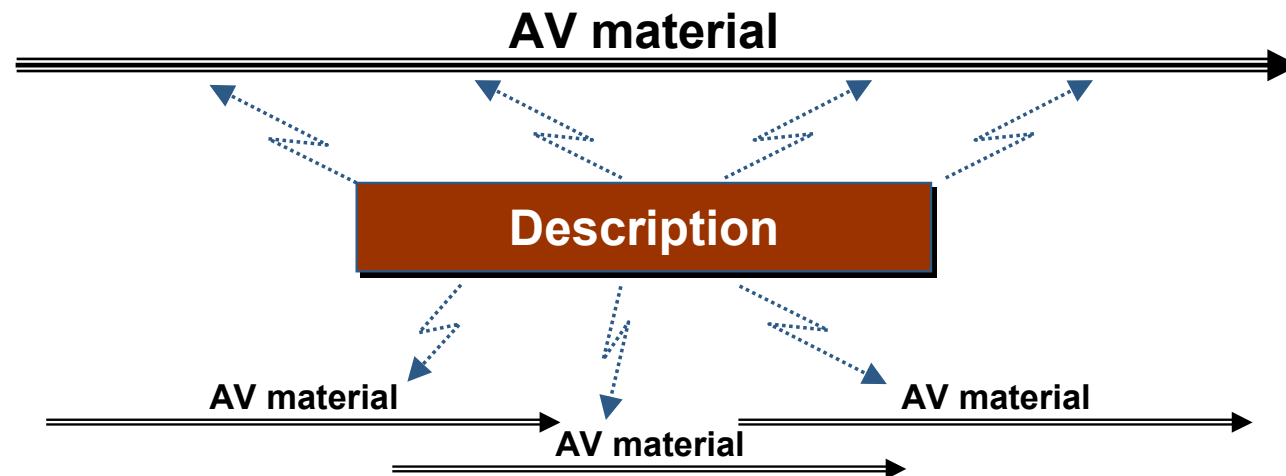
- Storage and retrieval of audiovisual databases (image, film, radio archives)
- Broadcast media selection (radio, TV programs)
- Surveillance (traffic control, surface transportation, production chains)
- E-commerce and Tele-shopping (searching for clothes / patterns)
- Remote sensing (cartography, ecology, natural resources management)
- Entertainment (searching for a game, for a karaoke)
- Cultural services (museums, art galleries)
- Journalism (searching for events, persons)
- Personalized news service on Internet (push media filtering)
- Intelligent multimedia presentations
- Educational applications
- Bio-medical applications

Example of queries

- Text (keywords):
 - ◆ Find AV material with subject corresponding to some keywords
- Semantic description:
 - ◆ Find AV material corresponding to a specified semantic
- Image as an example:
 - ◆ Find an image with similar characteristics (global or local)
- A few notes of music:
 - ◆ Find corresponding musical pieces or movies
- Low level features (example: motion):
 - ◆ Find video with specific object motion trajectories

Relation content / description

- Description may be separated from the content



- Description may be multiplexed with the content



Type of description

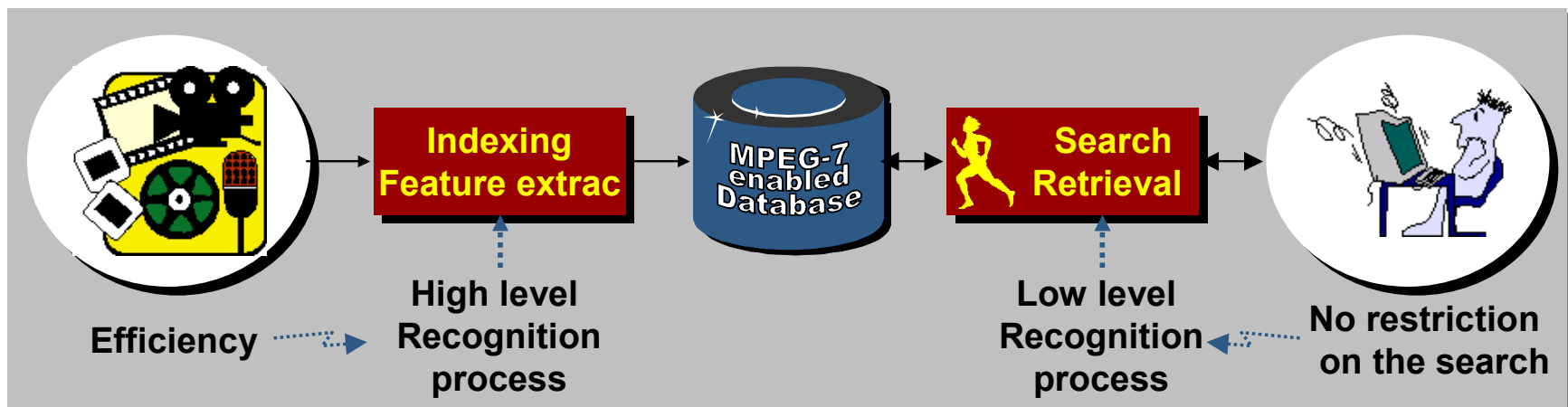
- **Information about the content:** recording date & conditions, title, author, copyright, coding format, classification, etc.
- **Information present in the content:** Combination of low level and high level descriptors

- ◆ **High level description:**

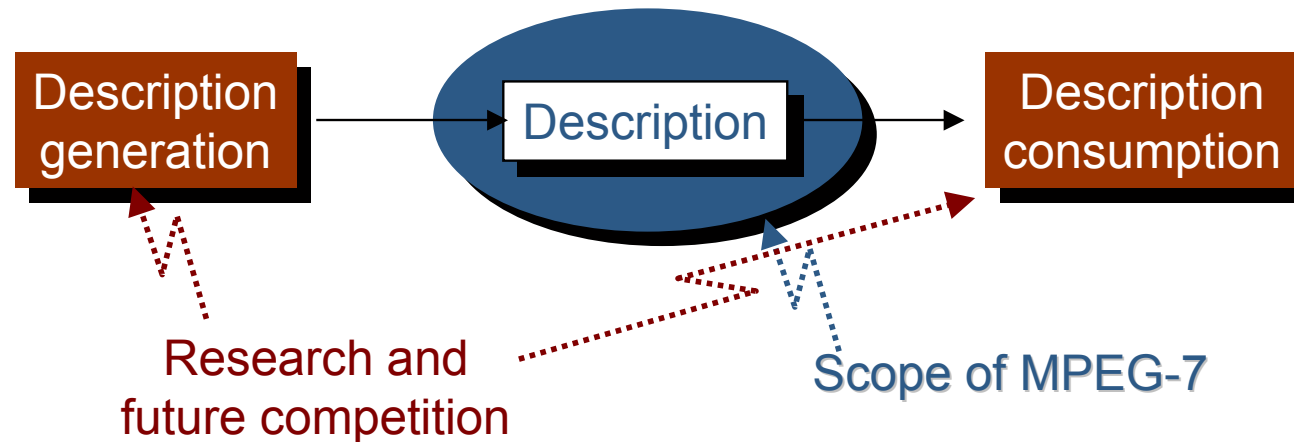
- ☞ Efficient and powerful
- ☞ Lack of flexibility

- ◆ **Low level description**

- ☞ Generic and flexible
- ☞ Intelligent / efficient search engine

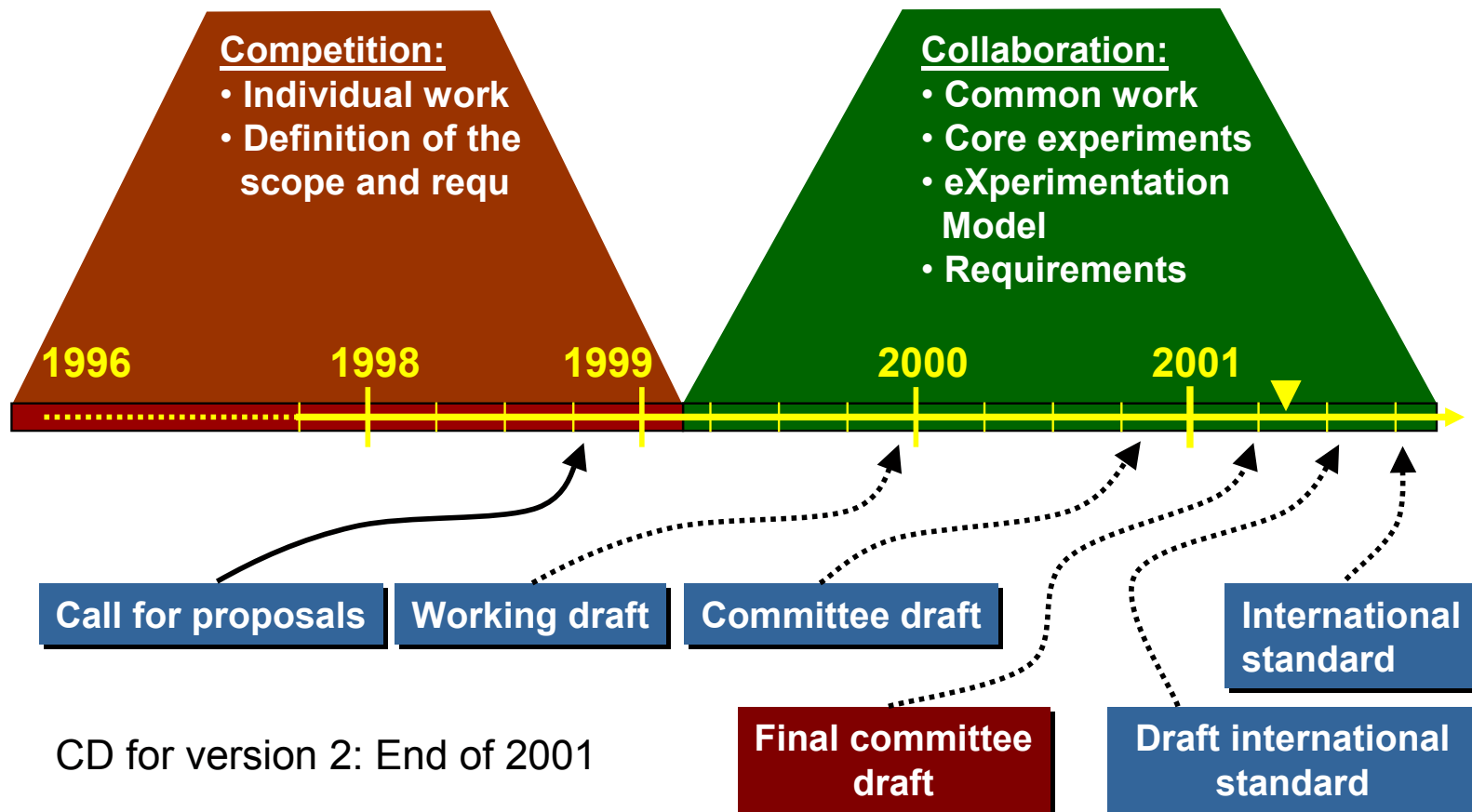


Scope of MPEG-7

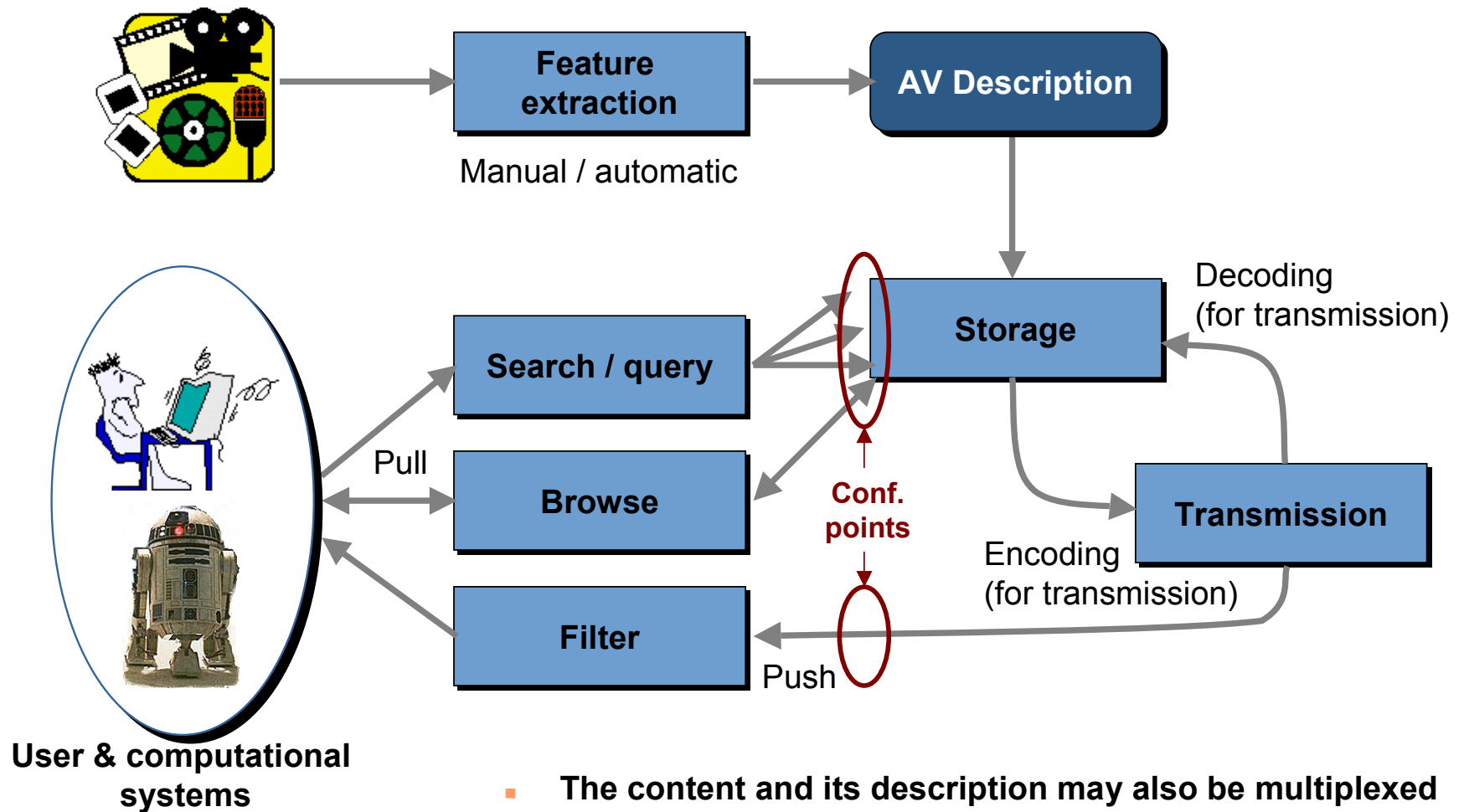


- The description generation (feature extraction, indexing process, annotation & authoring tools,...) and consumption (search engine, filtering tool, retrieval process, browsing device, ...) are non normative parts of MPEG-7.
- The goal is to define the minimum that enables interoperability (syntax and semantic of description tools).

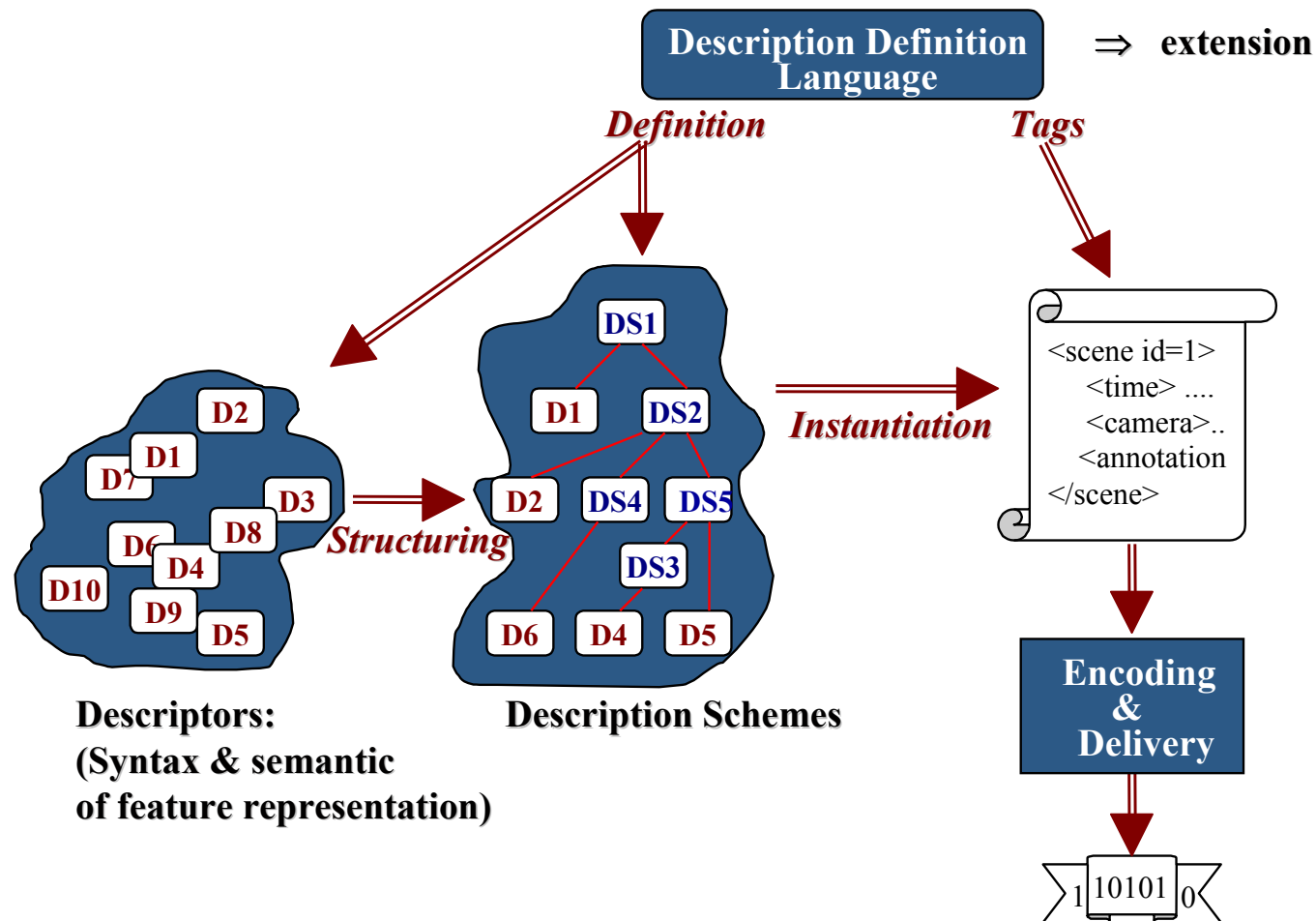
MPEG-7: The Workplan



Information Flow



MPEG-7 working areas



Parts of the MPEG-7 Standard

- ISO / IEC 15938 - 1: Systems
- ISO / IEC 15938 - 2: Description Definition Language
- ISO / IEC 15938 - 3: Visual
- ISO / IEC 15938 - 4: Audio
- ISO / IEC 15938 - 5: Multimedia Description Schemes
- ISO / IEC 15938 - 6: Reference Software
- ISO / IEC 15938 - 7: Conformance

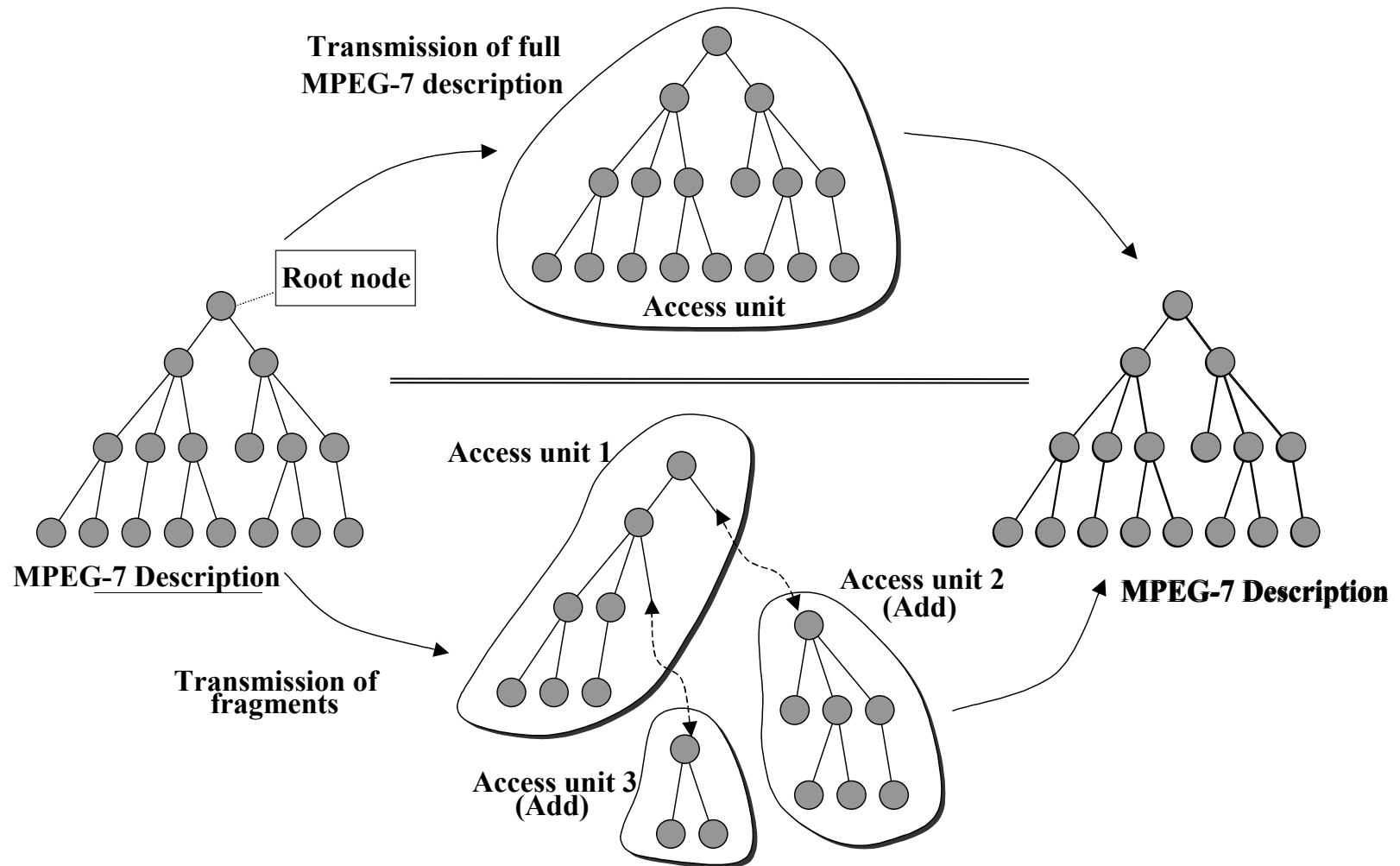
Outline

- Objective, goals, requirements and applications, basic component of the MPEG-7 standard
- Systems tools and Description Definition Language
- Multimedia Description Schemes
- Visual Tools
- Audio Tools
- Relation with other standards

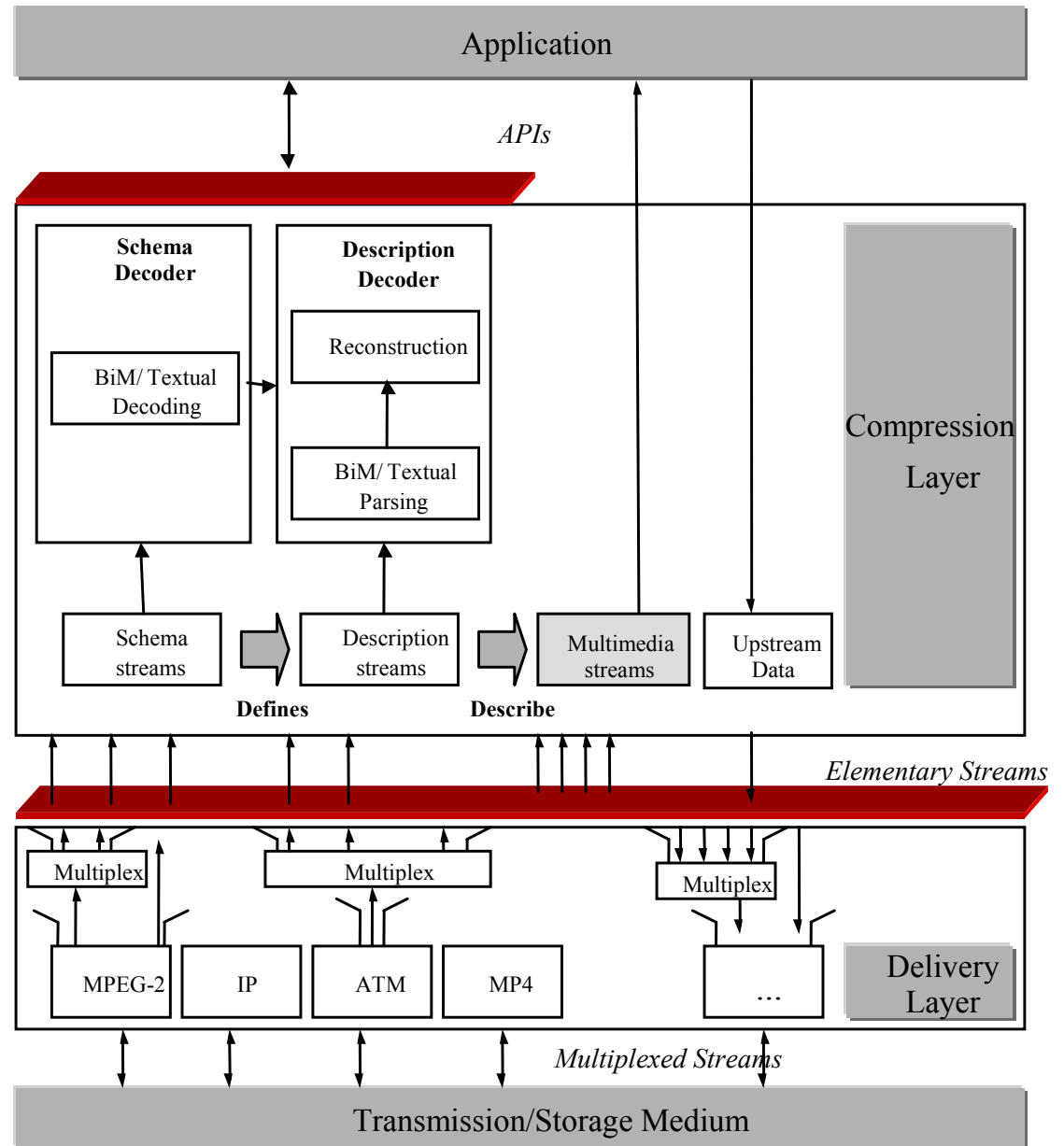
System tools

- System tools:
 - ◆ Streaming and delivery:
 - ☞ Split the description in pieces
 - ☞ Encapsulate the pieces of description in “access units”
 - ☞ Transmit the access units
 - ☞ Dynamic description
 - ◆ Two formats:
 - ☞ Textual: XML
 - ☞ Binary: BiM (Compression and Error resilience)
 - ◆ Terminal architecture

Access Units

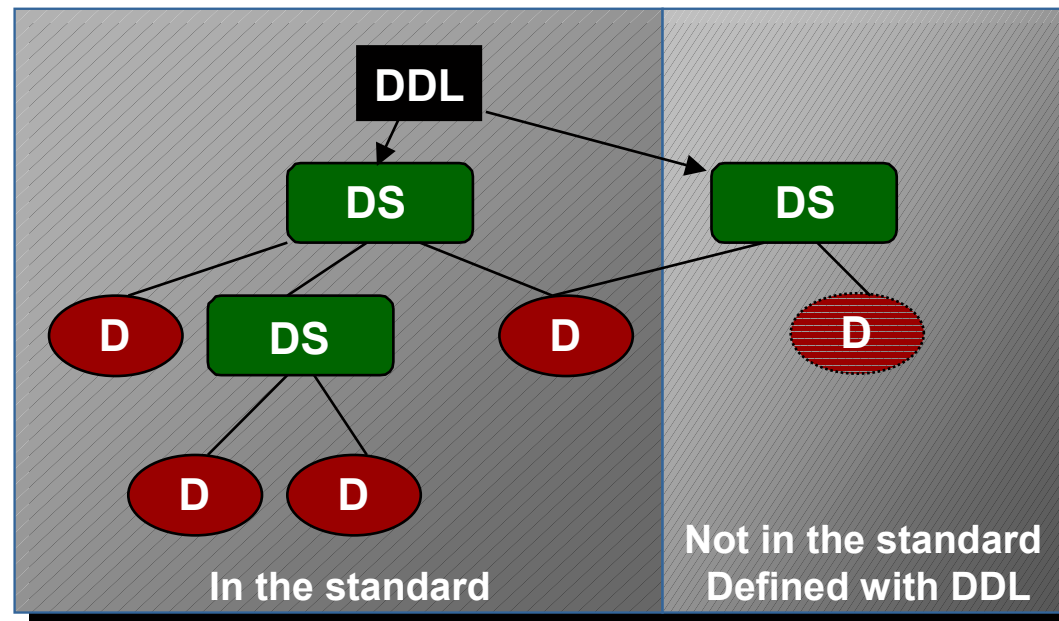


MPEG-7 terminal architecture



Description Definition Language

- ◆ Definition of the Ds and DSs: XML Schema
+ MPEG-7 extensions
- ◆ Instantiation (description): XML
- ◆ Allow to define new entities



DDL: Schema definition

- XML Schema:
 - ◆ Datatypes
 - ◆ Simple and Complex types
 - ◆ Elements
 - ◆ Inheritance, Abstract types

- MPEG-7 extensions:
 - ◆ Array and Matrix datatype
 - ◆ Enumerated datatypes for MimeType, CountryCode, RegionCode, CurrencyCode and CharacterSetCode
 - ◆ Typed references

Example of Schema definition

Simplified example of the Segment DS:

```
<complexType name="Segment" abstract="true">
  <element name="MedialInformation" type="mds:MedialInformation" minOccurs="0" maxOccurs="1"/>
  <element name="CreationInformation" type="mds:CreationInformation" minOccurs="0" maxOccurs="1"/>
  <element name="UsageInformation" type="mds:UsageInformation" minOccurs="0" maxOccurs="1"/>
  <element name="StructuredAnnotation" type="mds:StructuredAnnotation" minOccurs="0" maxOccurs="unbounded"/>
  <element name="SegmentDecomposition" type="mds:SegmentDecomposition" minOccurs="0" maxOccurs="unbounded"/>
  <attribute name="id" type="ID" use="required"/>
</complexType>

<element name="VideoSegment" type="mds:VideoSegment" equivClass="Segment"/>
<complexType name="VideoSegment" base="mds:Segment" derivedBy="extension">
  <element name="MediaTime" type="mds:MediaTime" minOccurs="1" maxOccurs="1"/>
  <element ref="ColorSpace" minOccurs="0" maxOccurs="1"/>
  <element ref="ColorQuantization" minOccurs="0" maxOccurs="1"/>
  <element ref="GofGopColorHistogram" minOccurs="0" maxOccurs="1"/>
  <element ref="CameraMotion" minOccurs="0" maxOccurs="1"/>
  <element ref="MotionActivity" minOccurs="0" maxOccurs="1"/>
  <element ref="Mosaic" minOccurs="0" maxOccurs="1"/>
</complexType>
```

DDL: Example of instantiation

Simplified example of VideoSegment instantiation:

```
<VideoSegment id = "VS1" >
  <MediaTime>
    <MediaTimePoint> T0:0:0  </MediaTimePoint>
    <MediaDuration>  PT10M  </MediaDuration>
  </MediaTime>

  <StructuredAnnotation>
    This is an example of video segment
  </StructuredAnnotation>

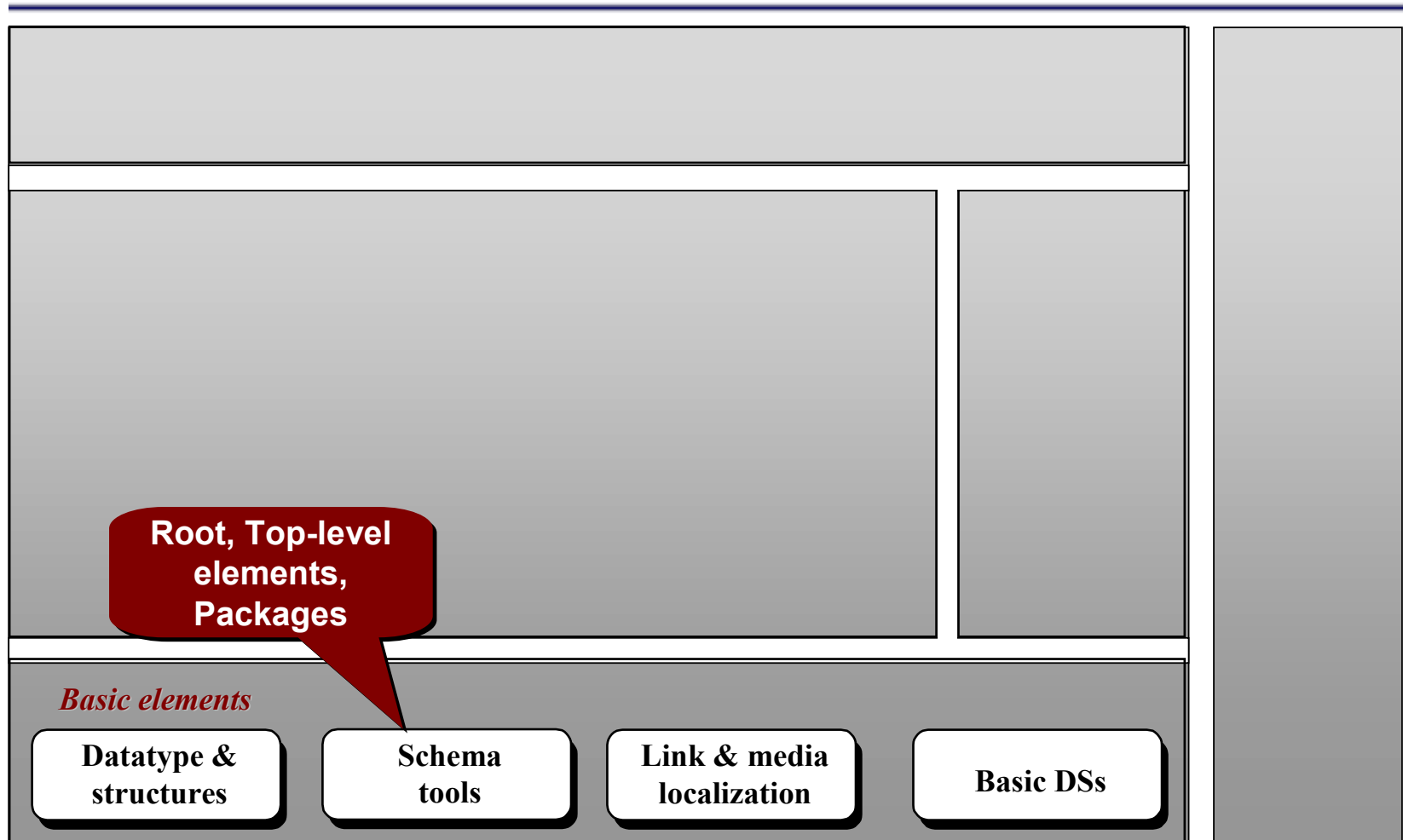
  <GofGopColorHistogram HistogramTypeInfo = "Average">
    ....
  </GofGopColorHistogram>

  <SegmentDecomposition Gap = "true" Overlap = "true" DecompositionType = "temporal">
    <VideoSegment id = "VS2" > .... </VideoSegment>
    <VideoSegment id = "VS3" > .... </VideoSegment>
    <VideoSegment id = "VS4" > .... </VideoSegment>
  </SegmentDecomposition>
</VideoSegment>
```

Outline

- Objective, goals, requirements and applications
- Basic component of the MPEG-7 standard
- Systems tools and Description Definition Language
- Multimedia Description Schemes
- Visual Tools
- Audio Tools
- Relation with other standards

Basic elements

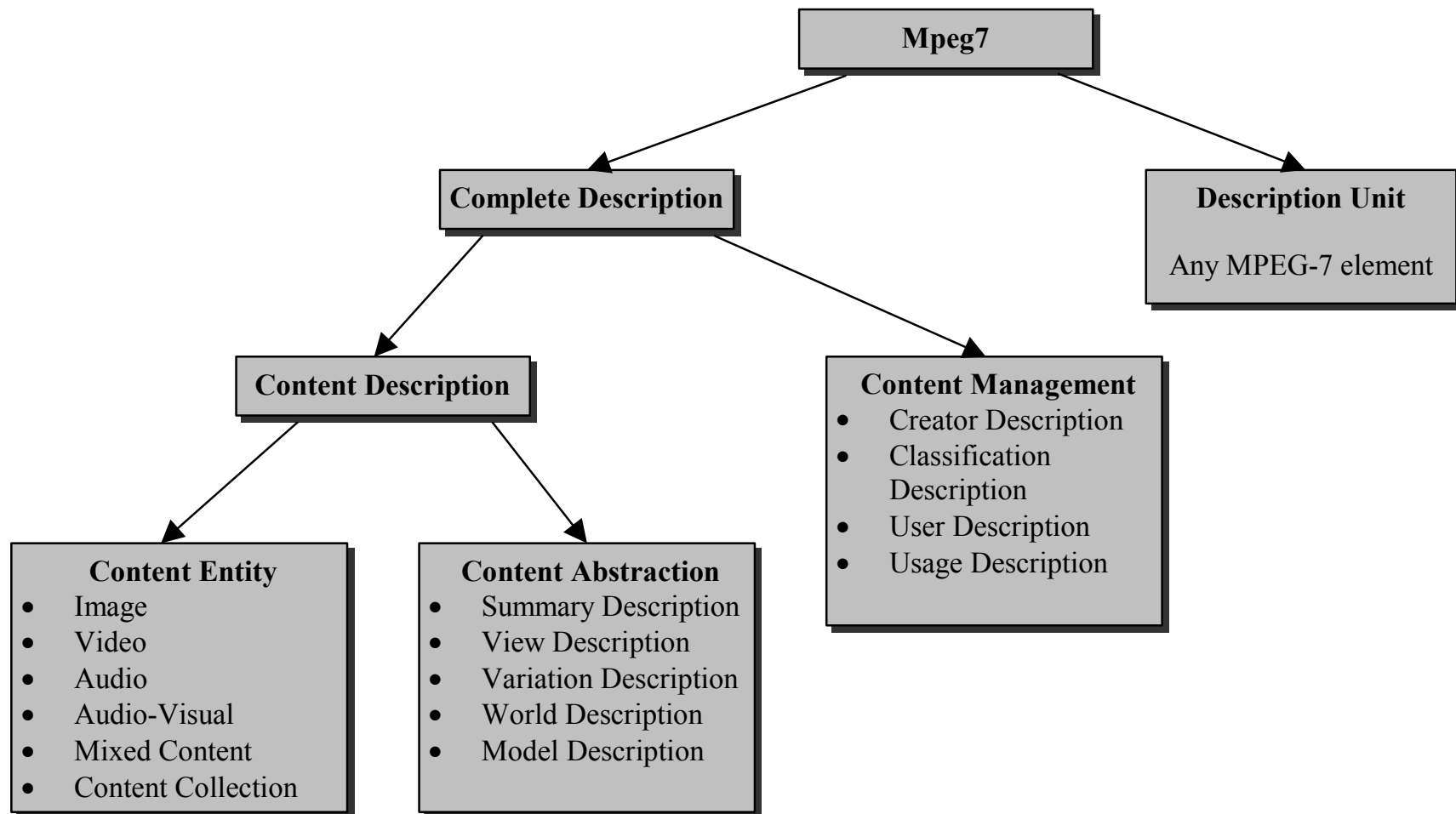


Schema tools

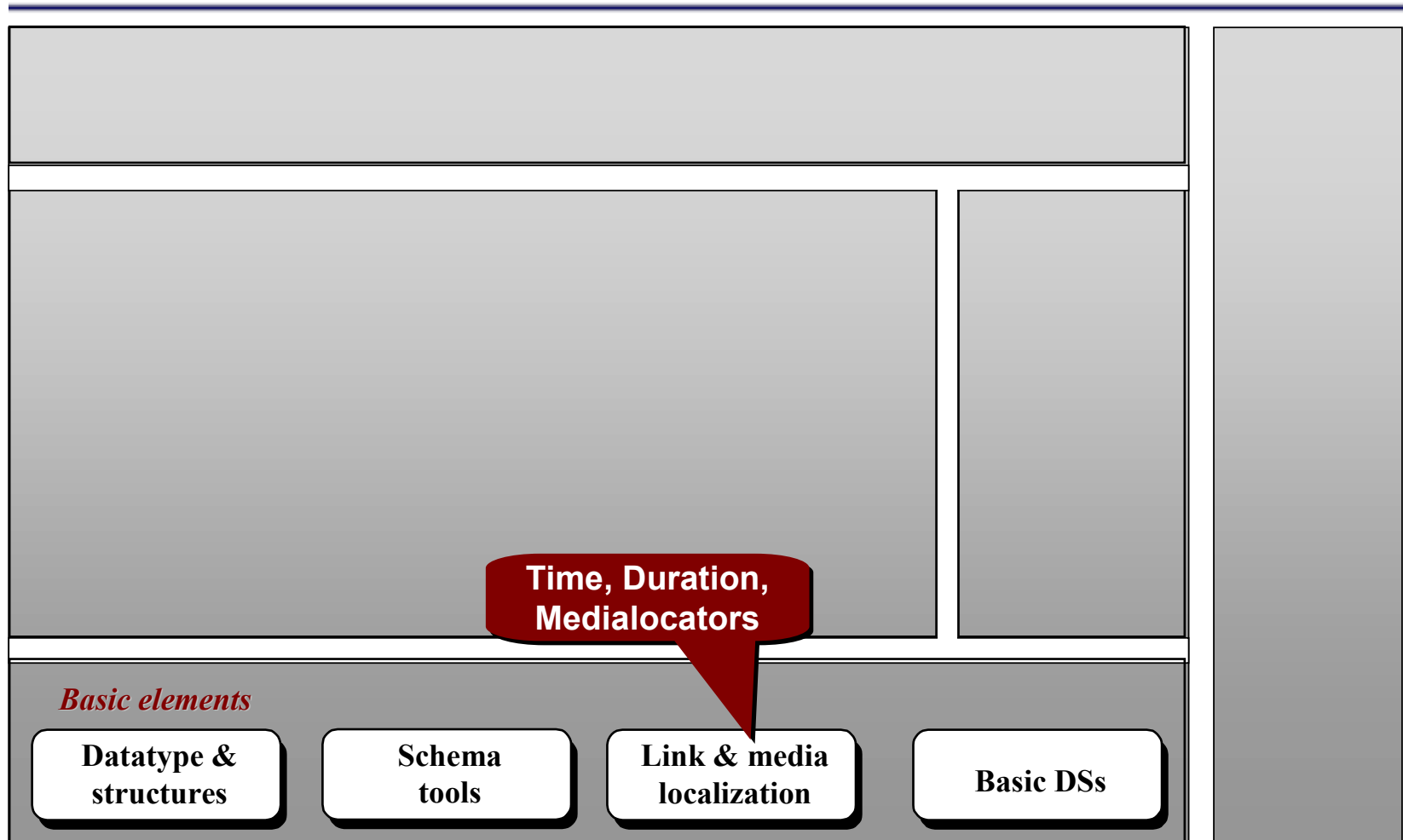
- Root elements:
 - ◆ Def.: Starting element of the description (wrapper)
 - ◆ Two roots: full description or fragments

- Top-level elements:
 - ◆ Def.: Possible child of the root element
 - ◆ Full description: Image, Video, Audio, Collection, etc.
 - ◆ Fragments: Any global element (DDL sense)

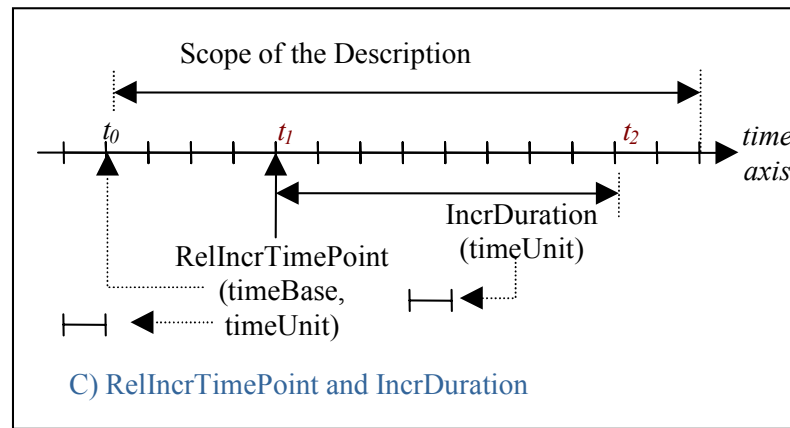
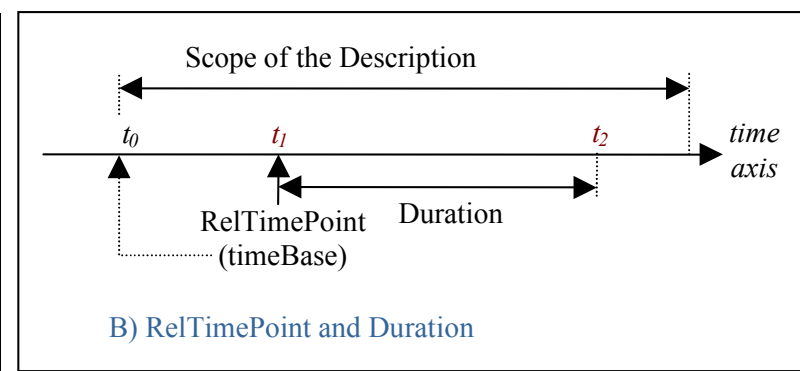
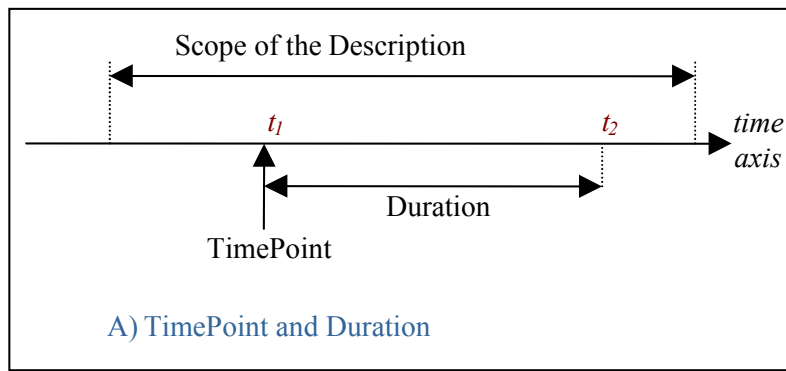
Root and Top level elements



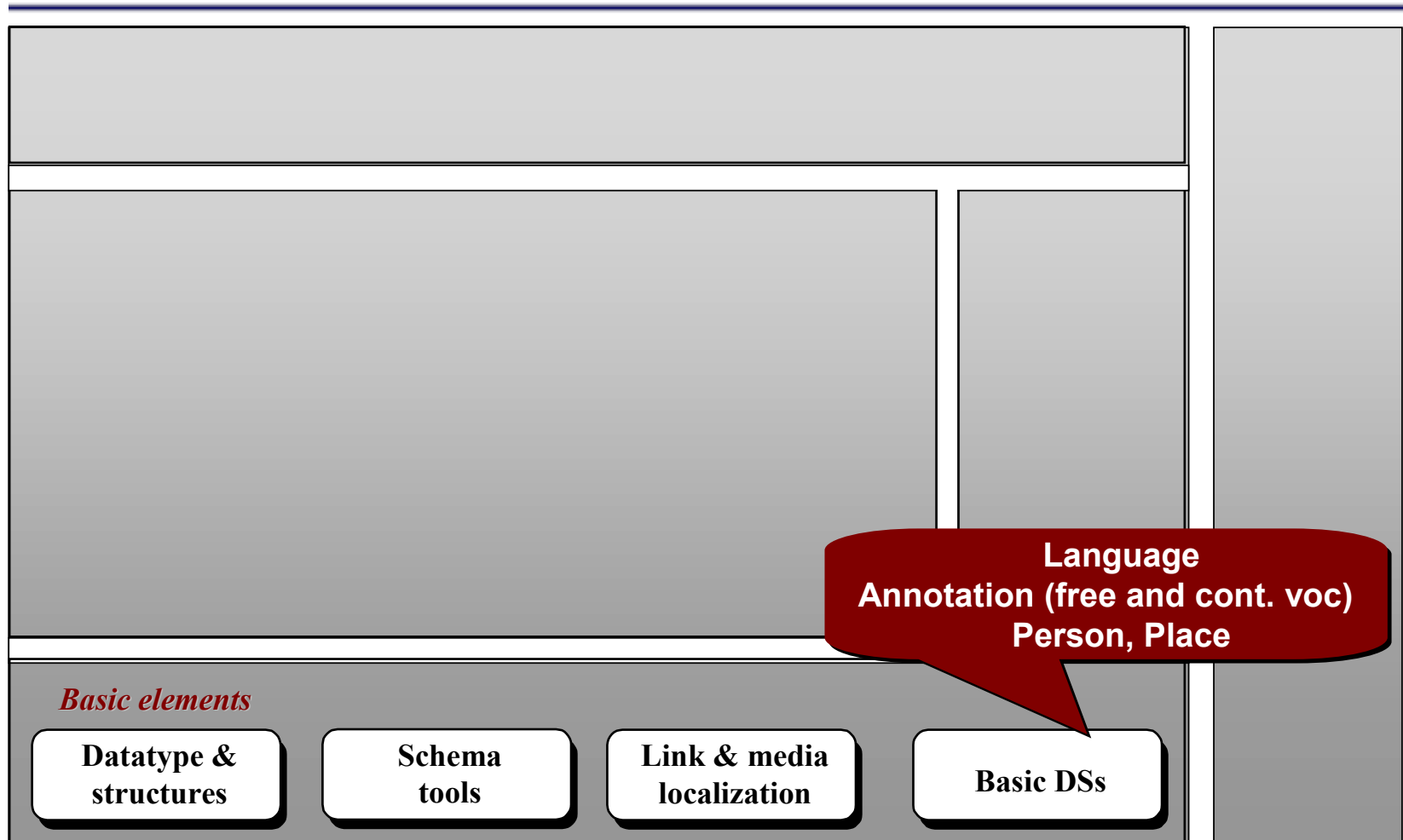
Basic elements



Time description



Basic elements



Textual Annotation

Free text:

```
<TextAnnotation>
  <FreeTextAnnotation xml:lang="en">
    Spain scores a goal against Sweden.
    The scoring player is Morientes.
  </FreeTextAnnotation>
</TextAnnotation>
```

Structured Annotation:

```
<TextAnnotation>
  <StructuredAnnotation>
    <Who><Name>Spain</Name></Who>
    <WhatAction><Name>score goal</Name></WhatAction>
    <Where><Name>A Coruña, Spain</Name></Where>
    <When><Name>March 25, 1998</Name></When>
  </StructuredAnnotation>
</TextAnnotation>
```

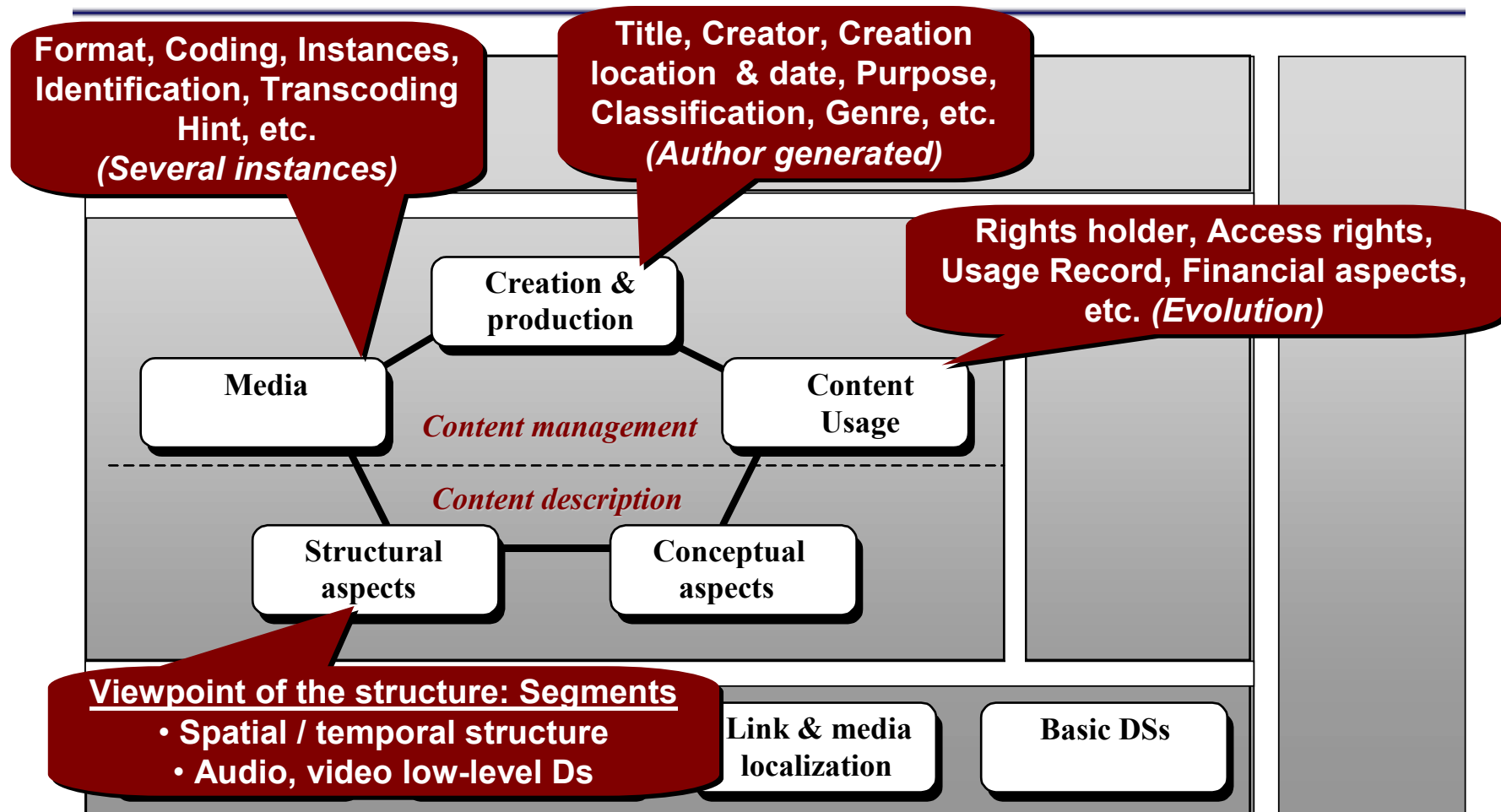
Key-words:

```
<TextAnnotation>
  <KeywordAnnotation>
    <Keyword>score</Keyword>
    <Keyword>Sweden</Keyword>
    <Keyword>Spain</Keyword>
    <Keyword>Morientes</Keyword>
  </KeywordAnnotation>
</TextAnnotation>
```

Dependency Structure:

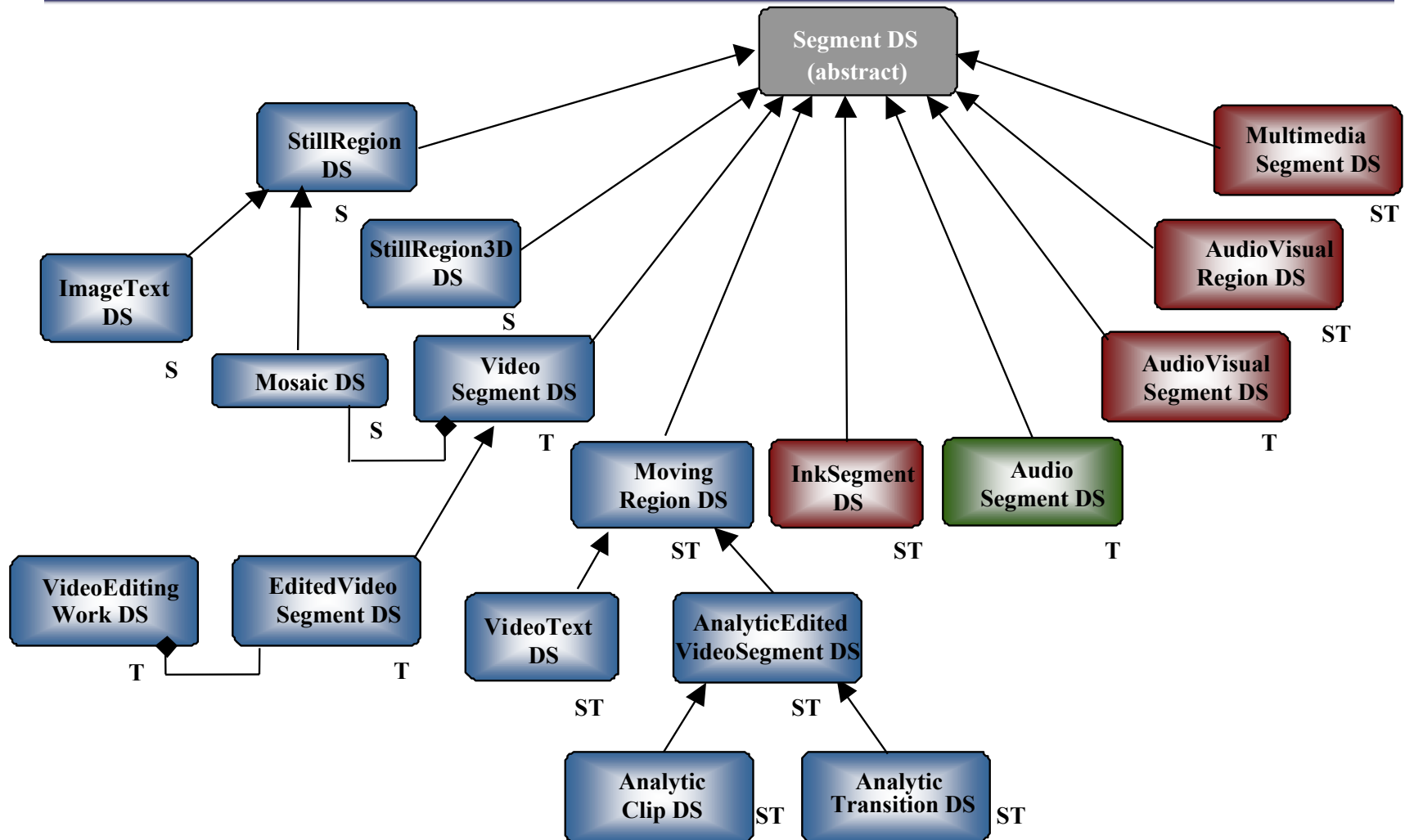
```
<TextAnnotation>
  <DependencyStructure>
    <Sentence>
      <Phrase operator="subject">
        <Head type="noun">Spain</Head>
      </Phrase>
      <Head type="verb" baseForm="score">scored</Head>
      <Phrase operator="object">
        <Head type="article noun">a goal</Head>
      </Phrase>
      <Phrase>
        <Head type="preposition">against</Head>
        <Phrase><Head>Sweden</Head></Phrase>
      </Phrase>
    </Sentence>
  </DependencyStructure>
</TextAnnotation>
```

Content Management & Description



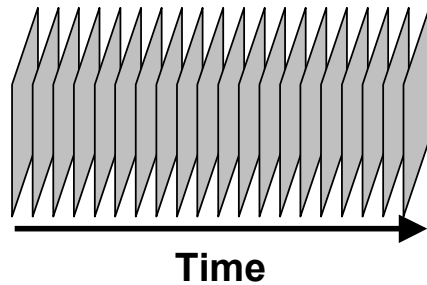
Segment DS

Common properties of:

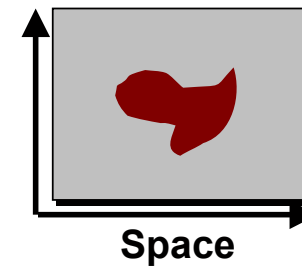


Examples of Segments

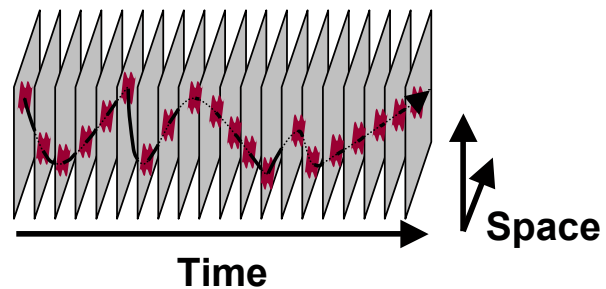
Video segments



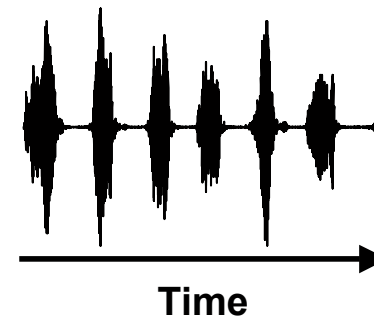
Still regions



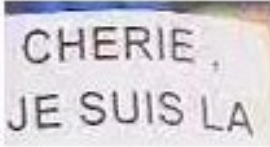
Moving regions



Audio segments



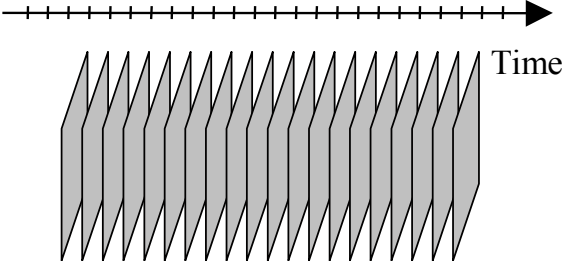
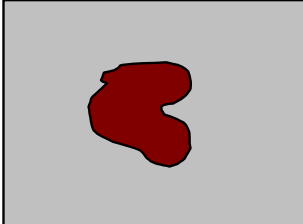
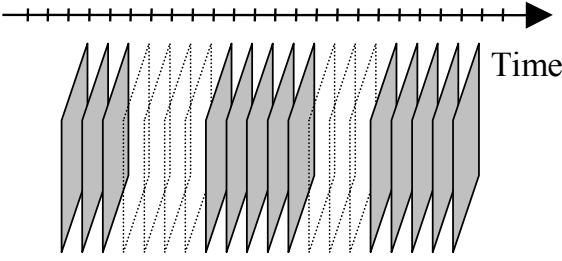
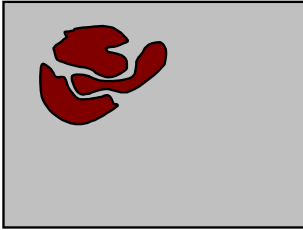
Example of Segments

<i>Mosaic</i> 	<i>3D still region</i> 	<i>Image text</i> 
<i>Ink segment</i> 	<i>Multimedia segment</i> 	<i>Analytic clips/transitions</i> 

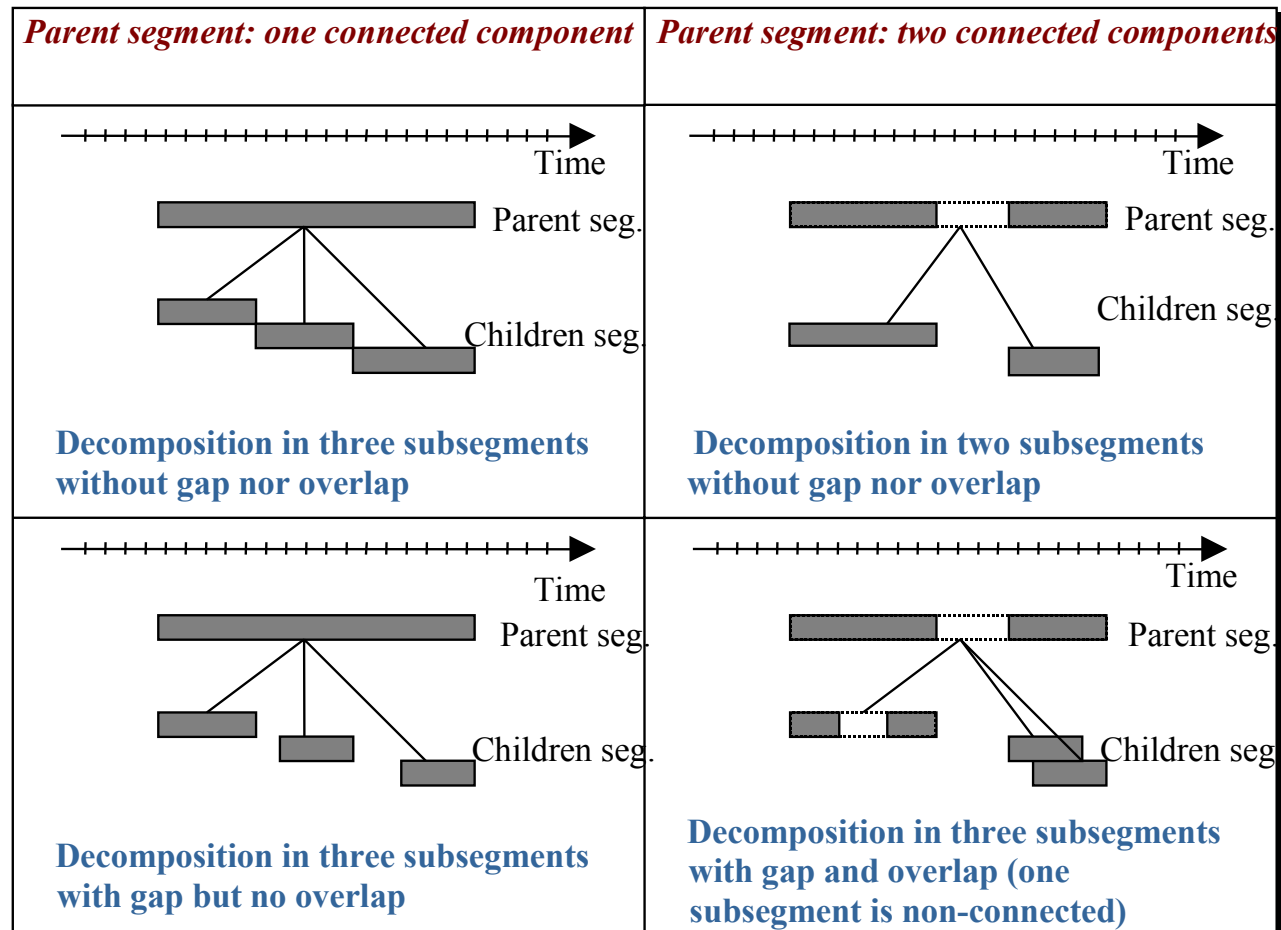
Common segment properties

- Identification
- Creation Information
- Usage Information
- Media Information
- Annotation
- Semantic
- Segment Decomposition

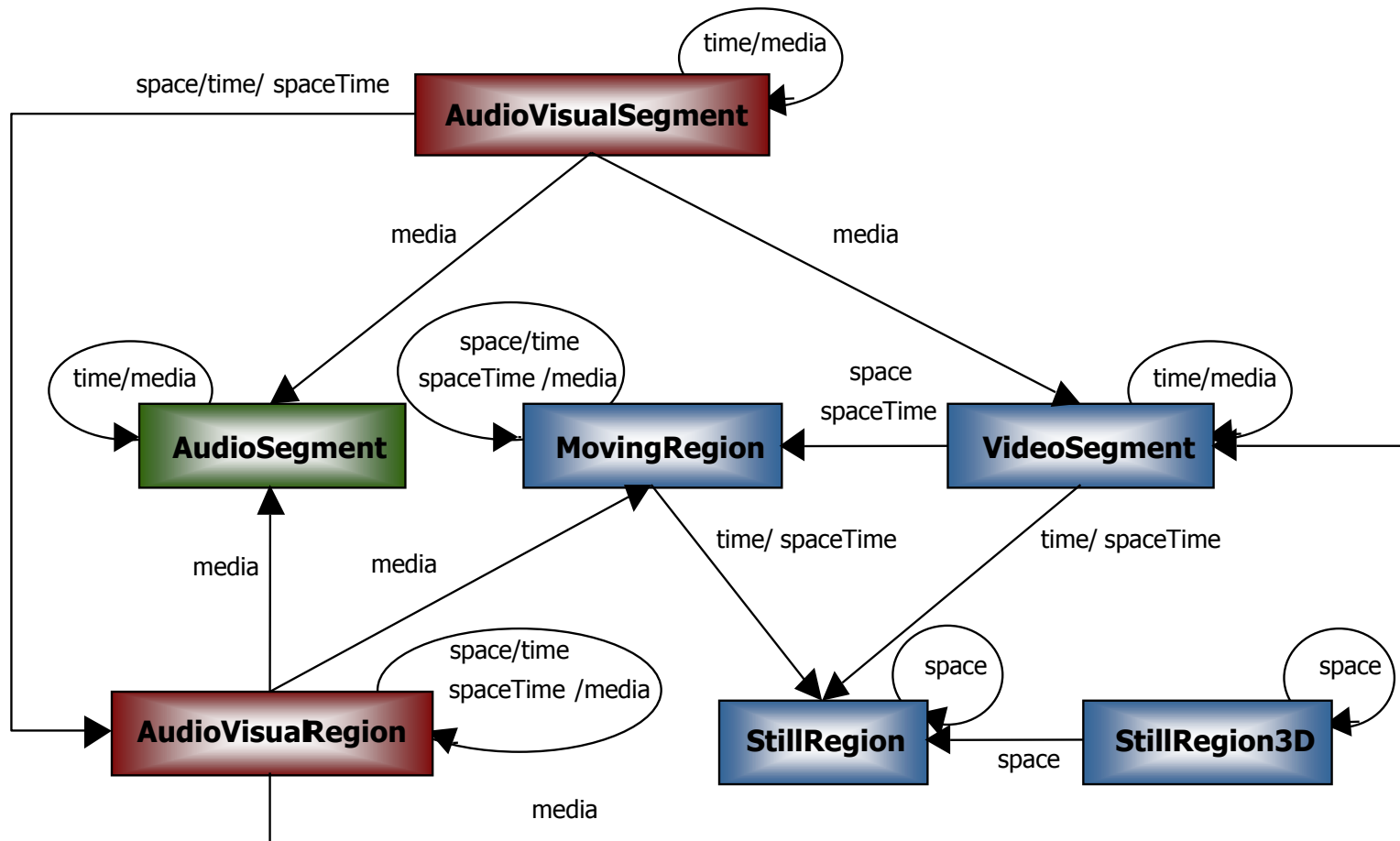
Segment connectivity

<i>Temporal segment</i> (AudioVisual, Video, Audio Segments)	<i>Spatial segment</i> (StillRegion)
 one connected component	 one connected component
 three connected components	 three connected components

Segment decomposition

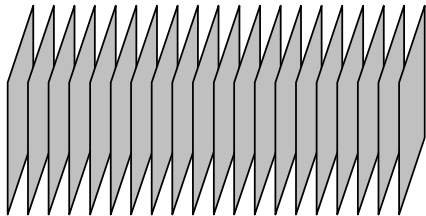


Examples of valid decompositions



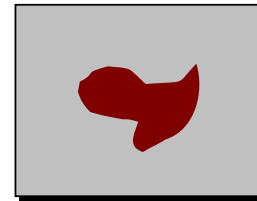
Specific description of segments

Video segments



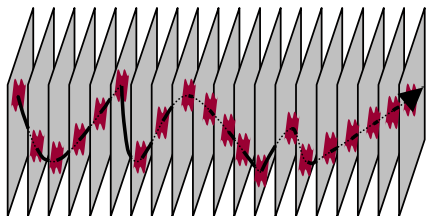
- Color
- Camera motion
- Motion activity
- Mosaic

Still regions



- Color
- Shape
- Position
- Texture

Moving regions



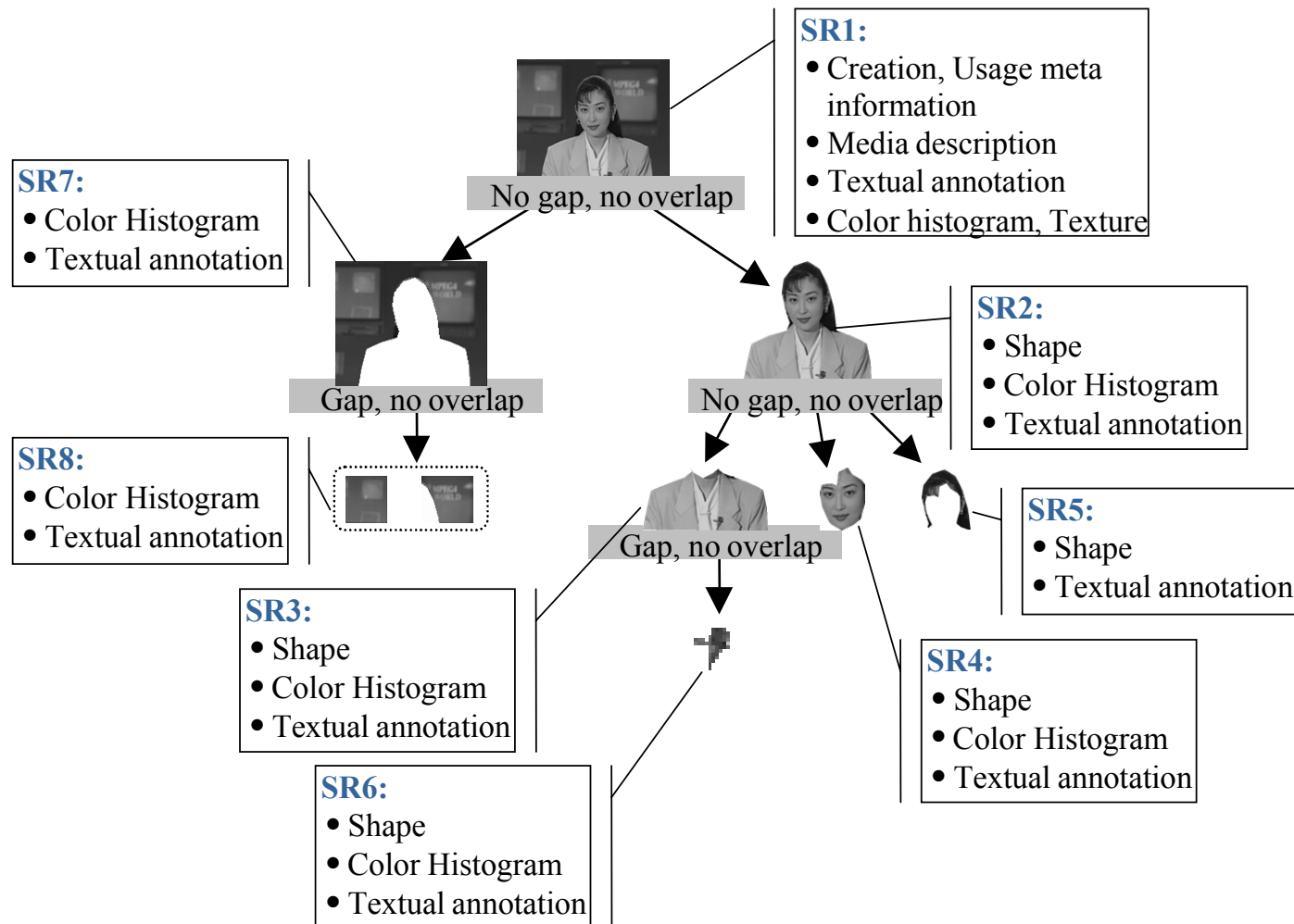
- Color
- Motion trajectory
- Parametric motion
- Spatio-temporal shape

Audio segments



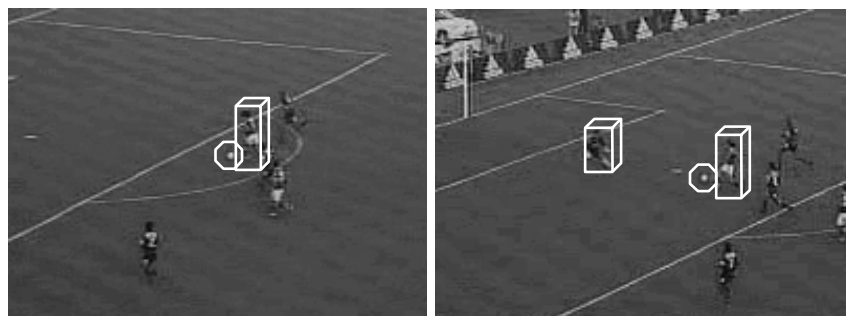
- Spoken content
- Spectral characterization
- Music: timbre, melody

Example of Segment trees

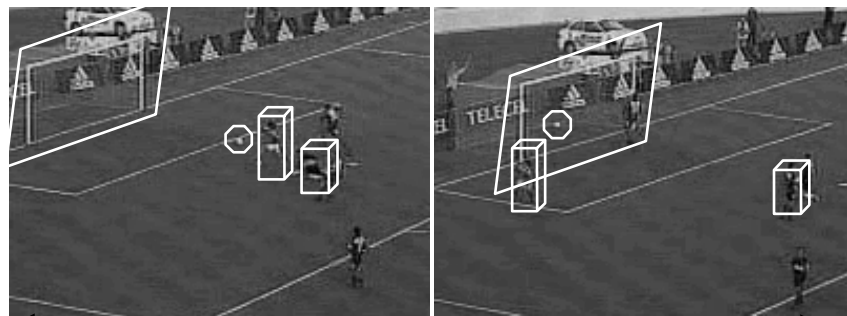


Graph

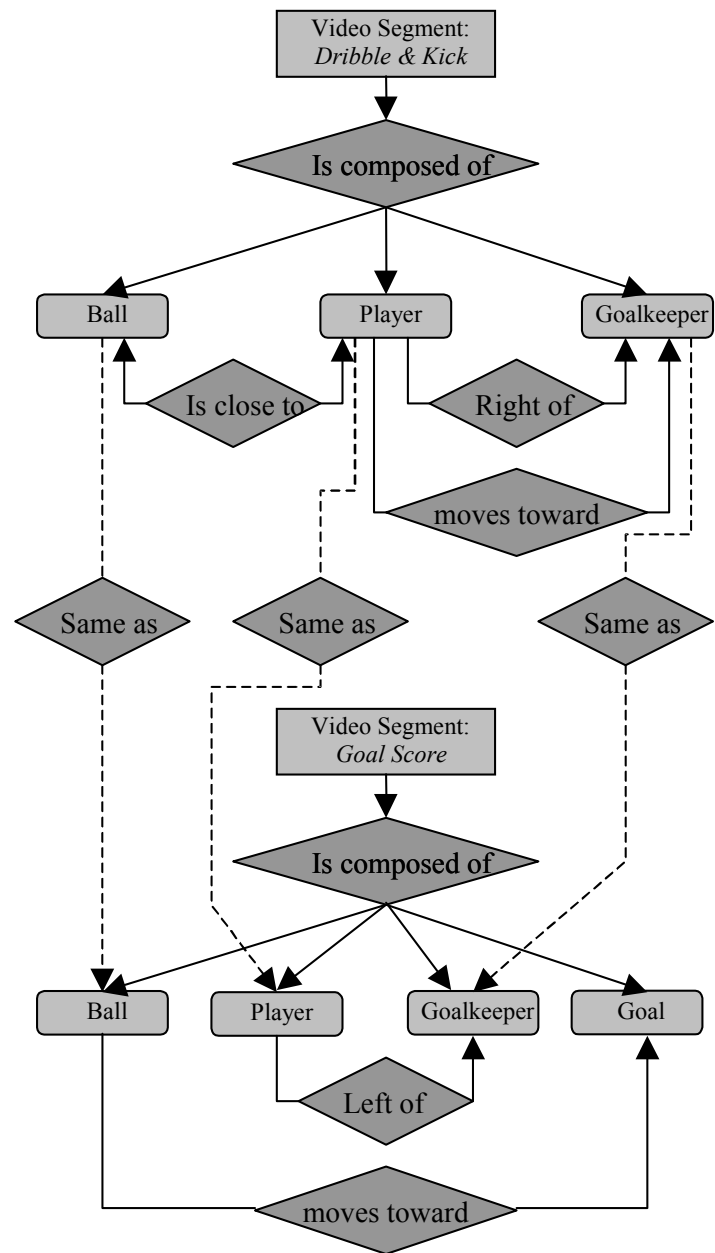
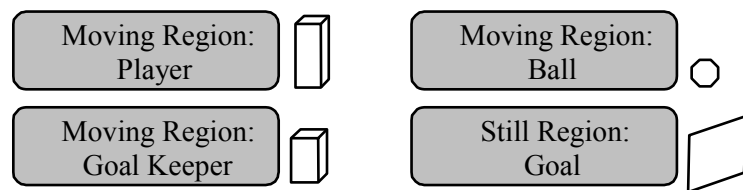
- Goal:
 - ◆ The segment DS allows the construction of tree structures
 - ☞ Efficient for access, retrieval, compression.
 - ☞ Lack of flexibility
 - ◆ Graph structure to improve flexibility
- Outline of the approach:
 - ◆ Definition of entity nodes representing segments
 - ◆ Definition of relationships: space, time, visual



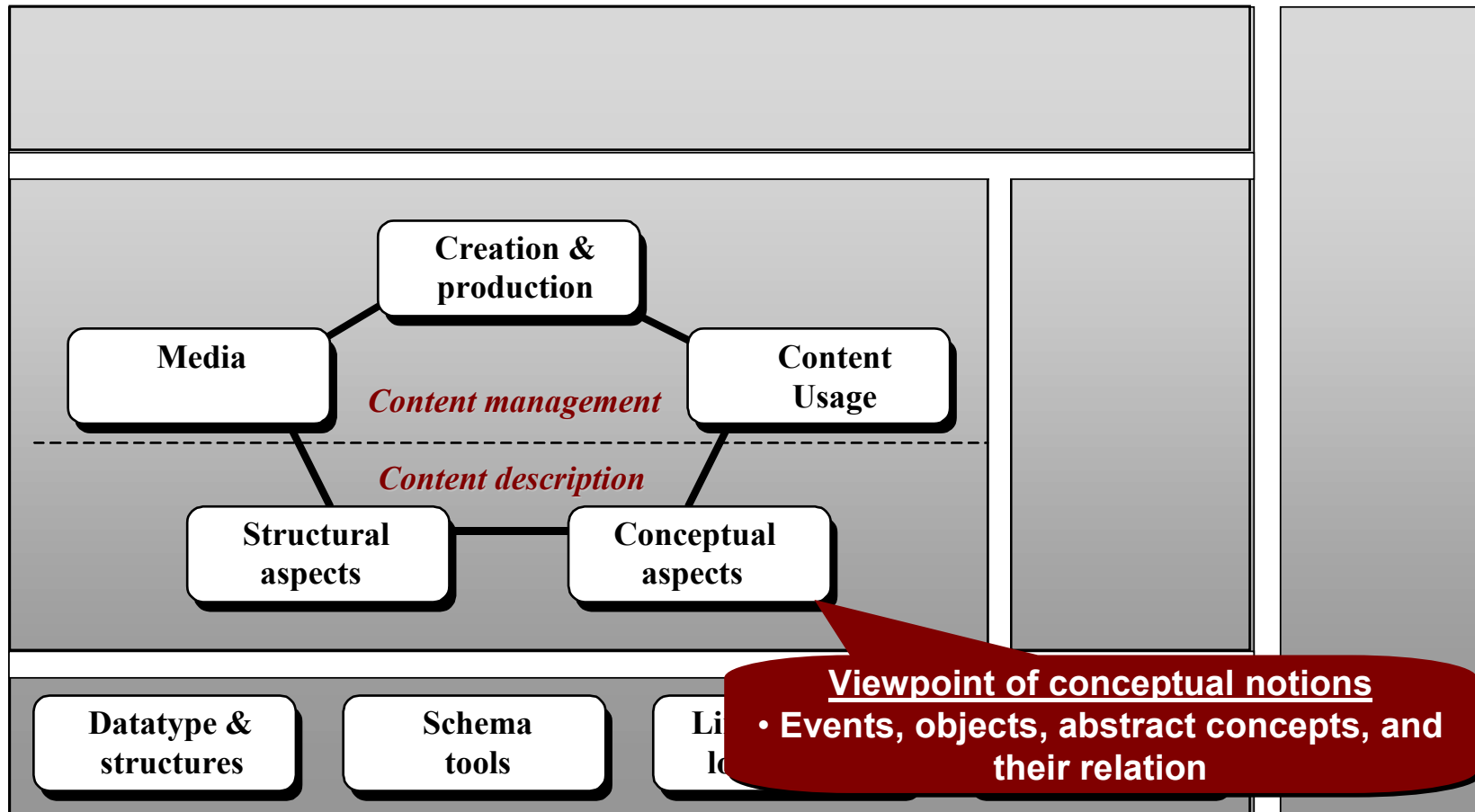
Video Segment 1: Dribble & Kick



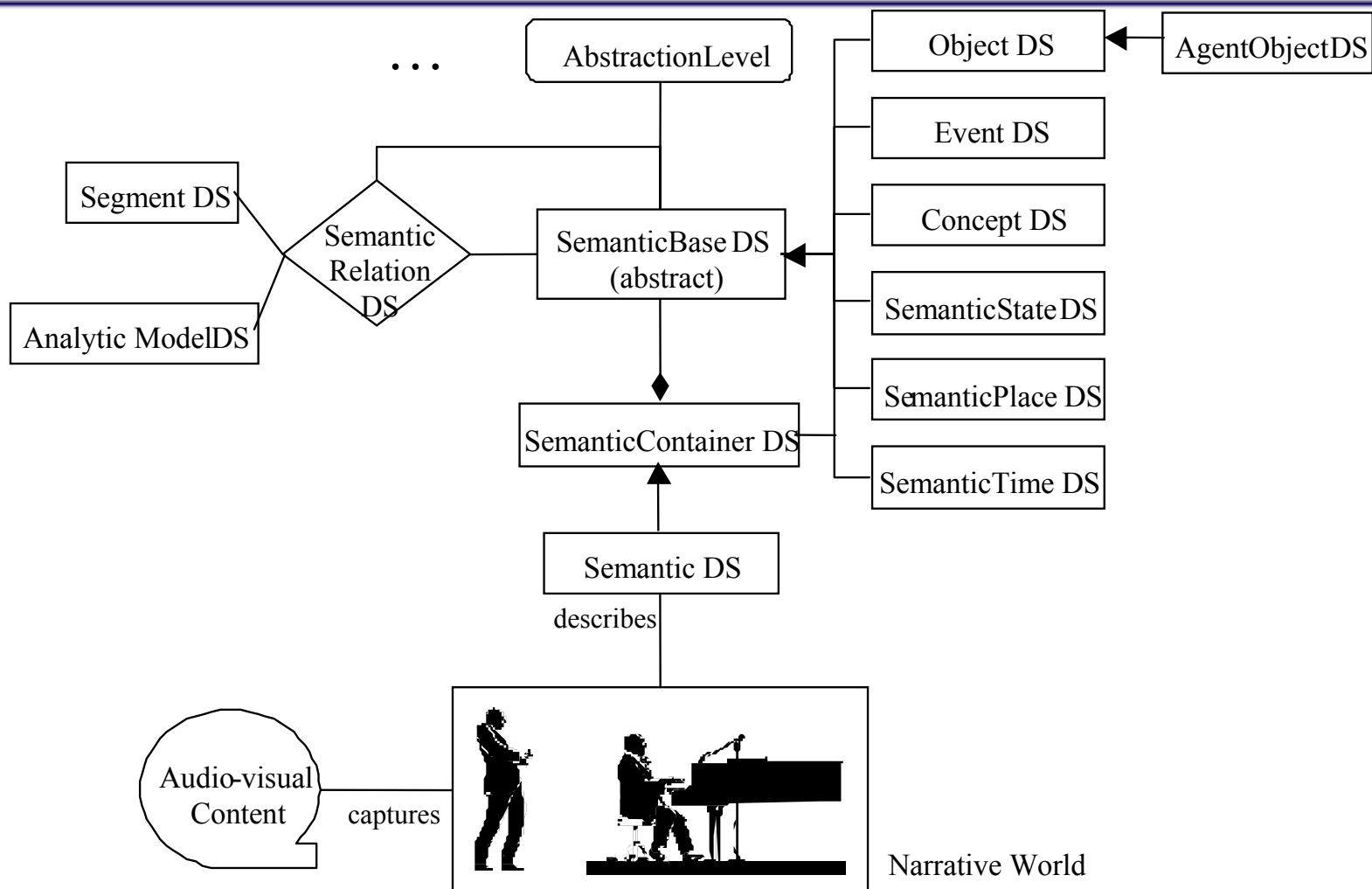
Video Segment 2: Goal Score

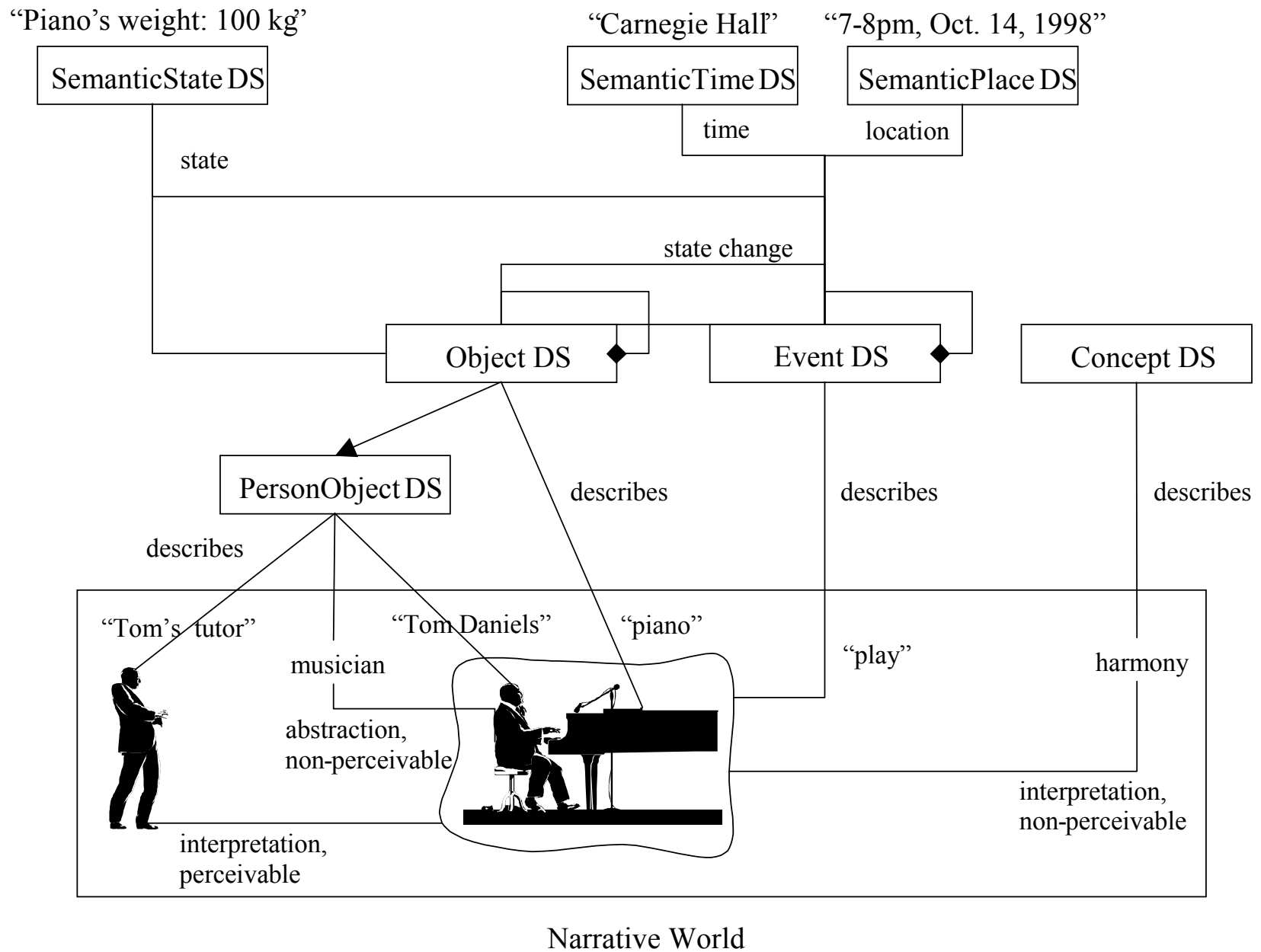


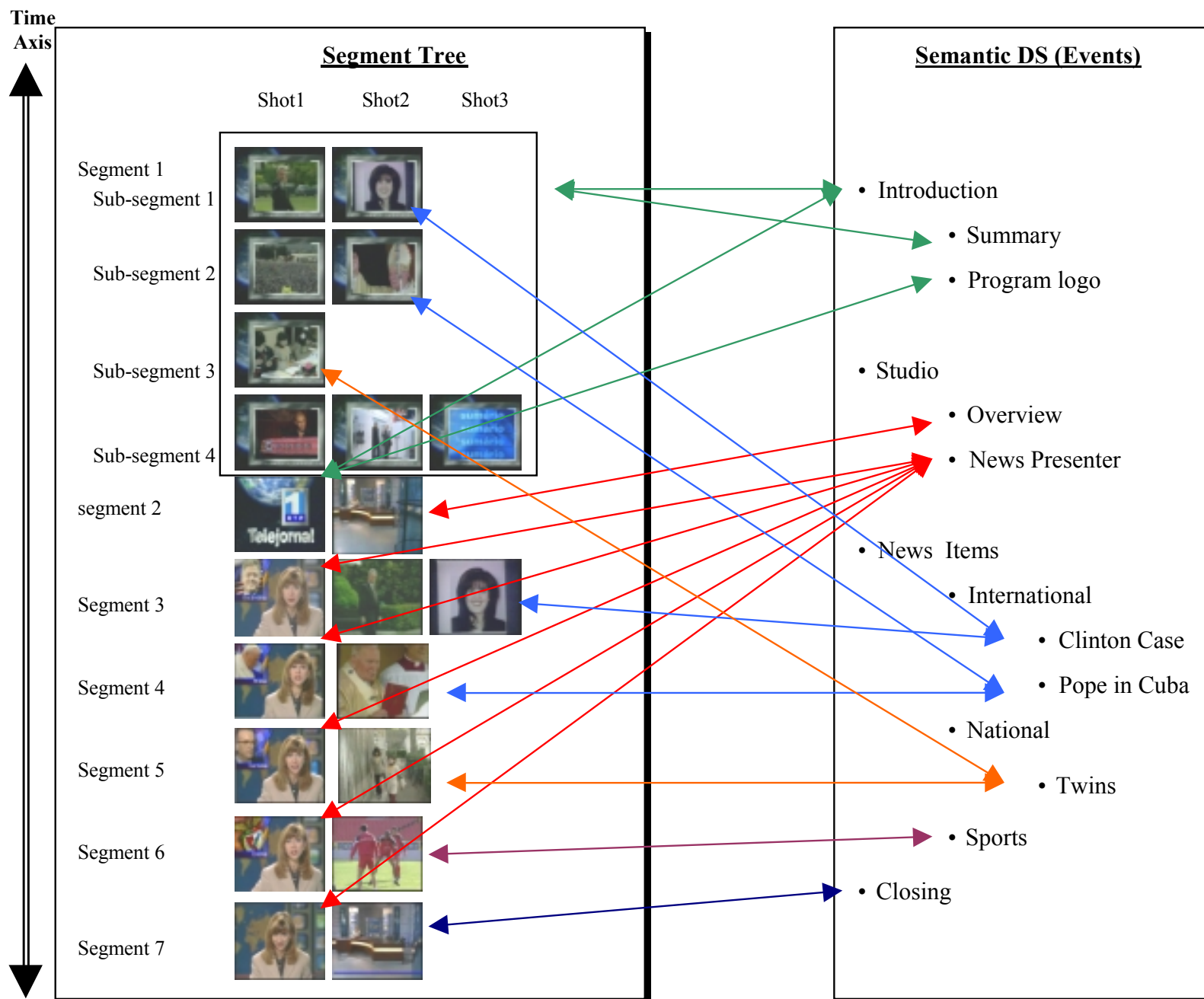
Content Management & Description



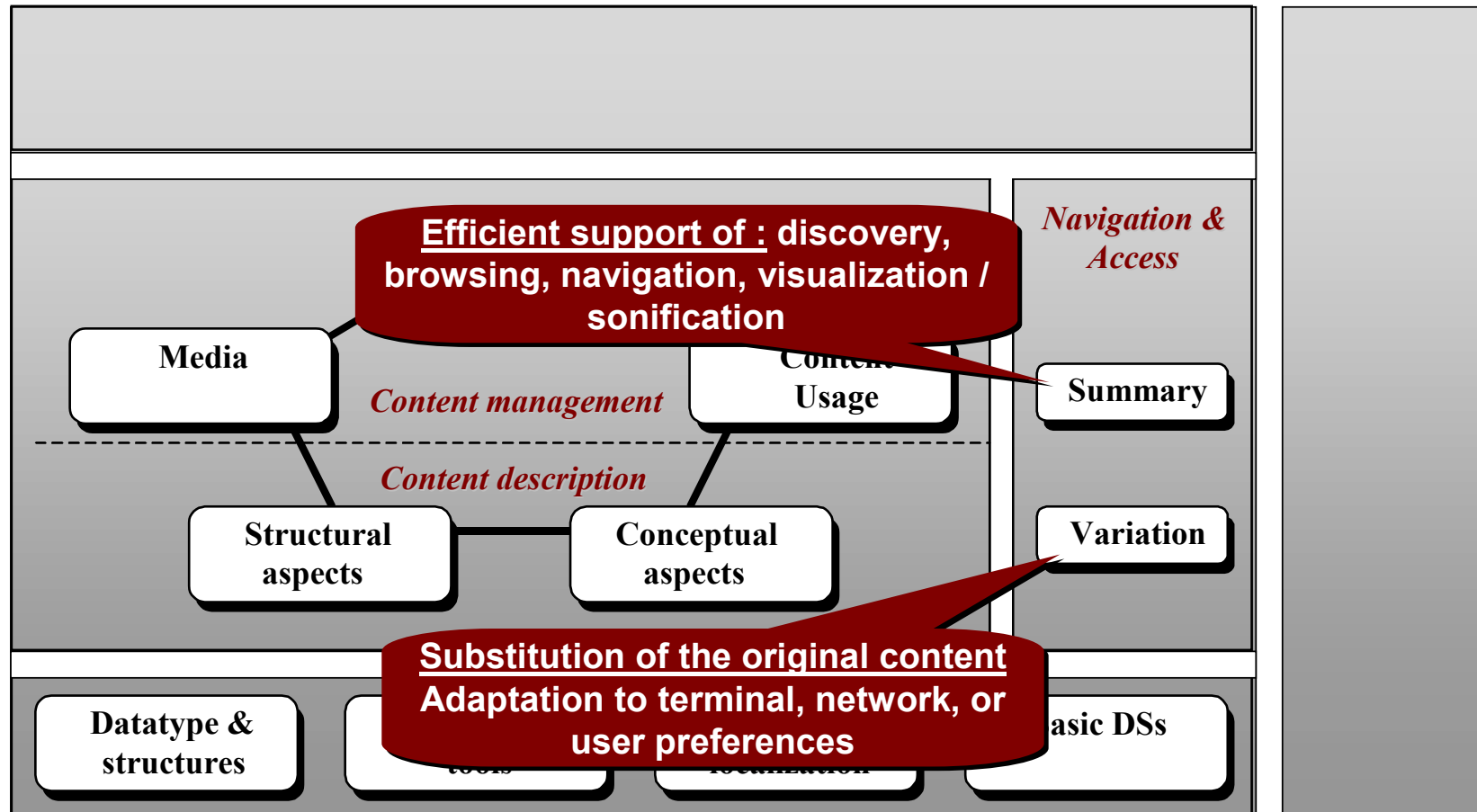
Semantic DS



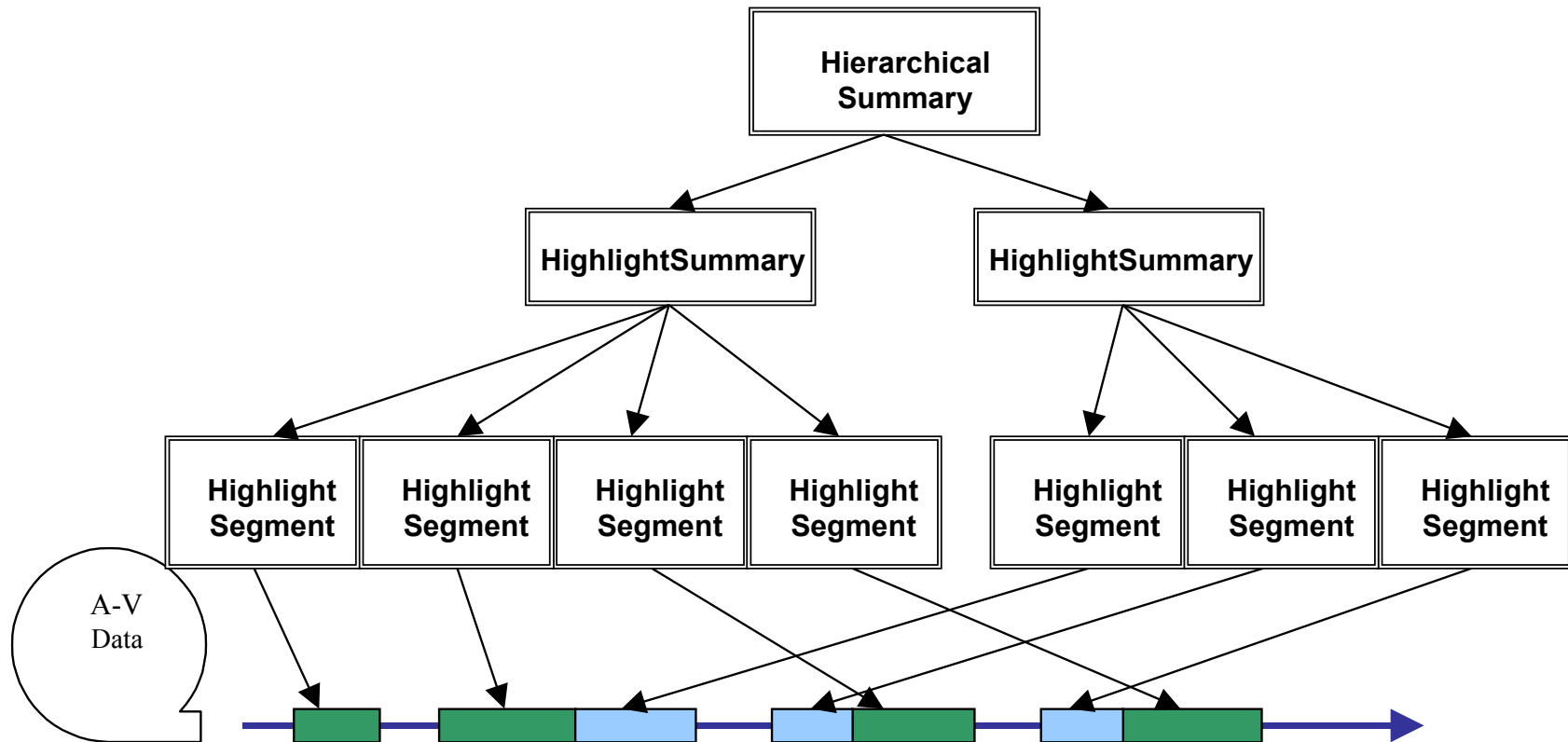




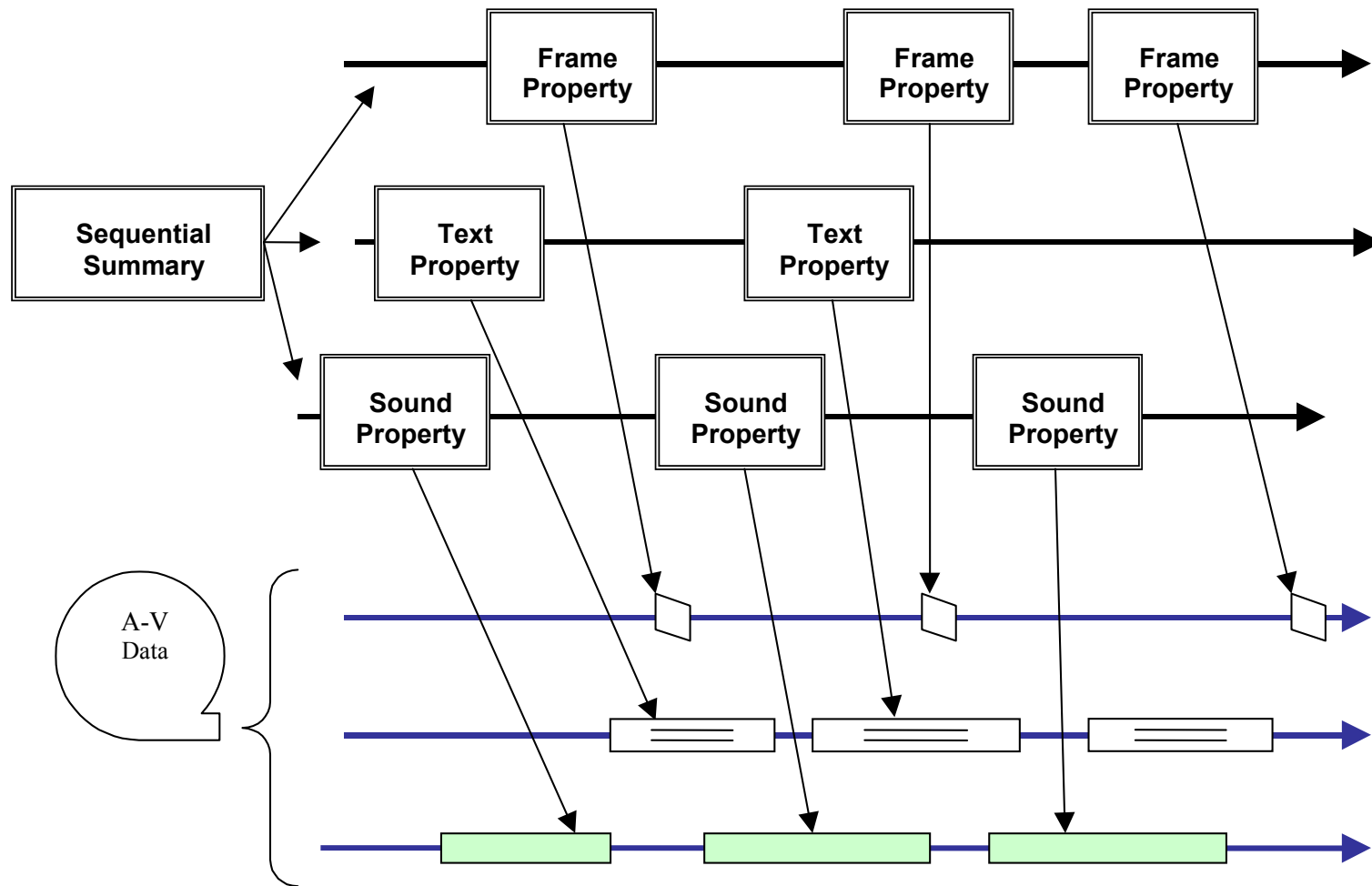
Navigation and Access



Hierarchical summary

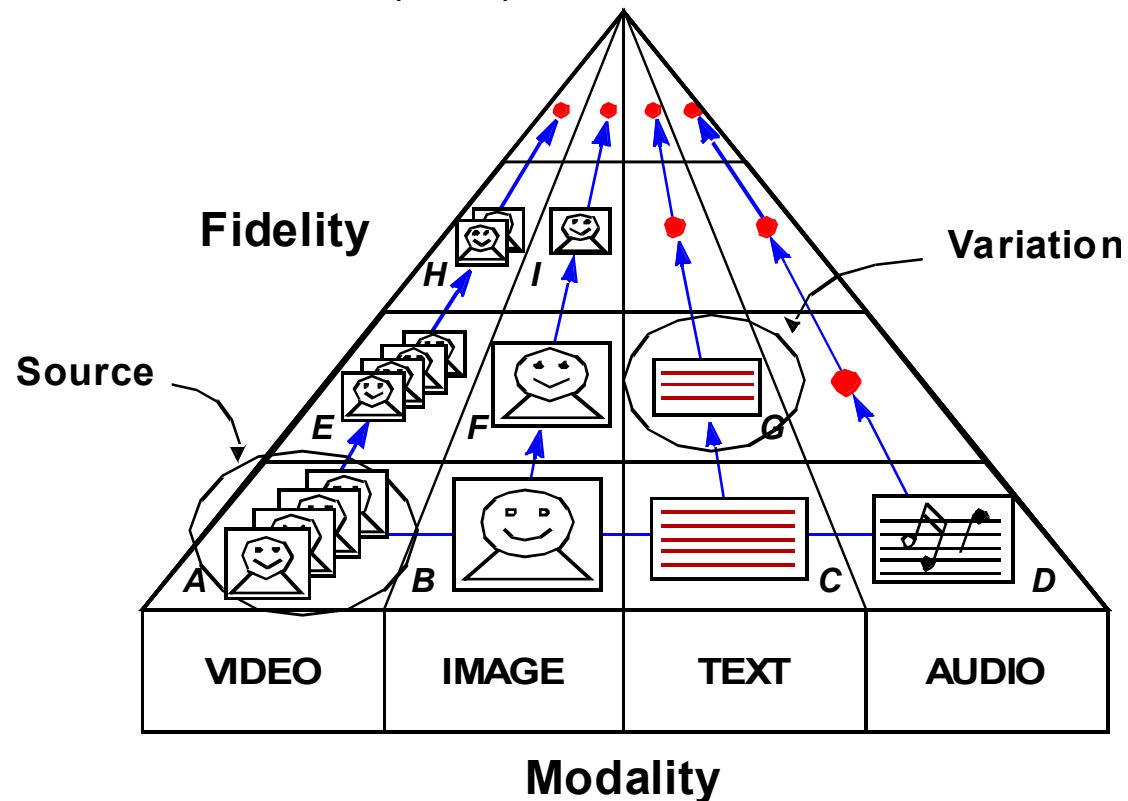


Sequential summary

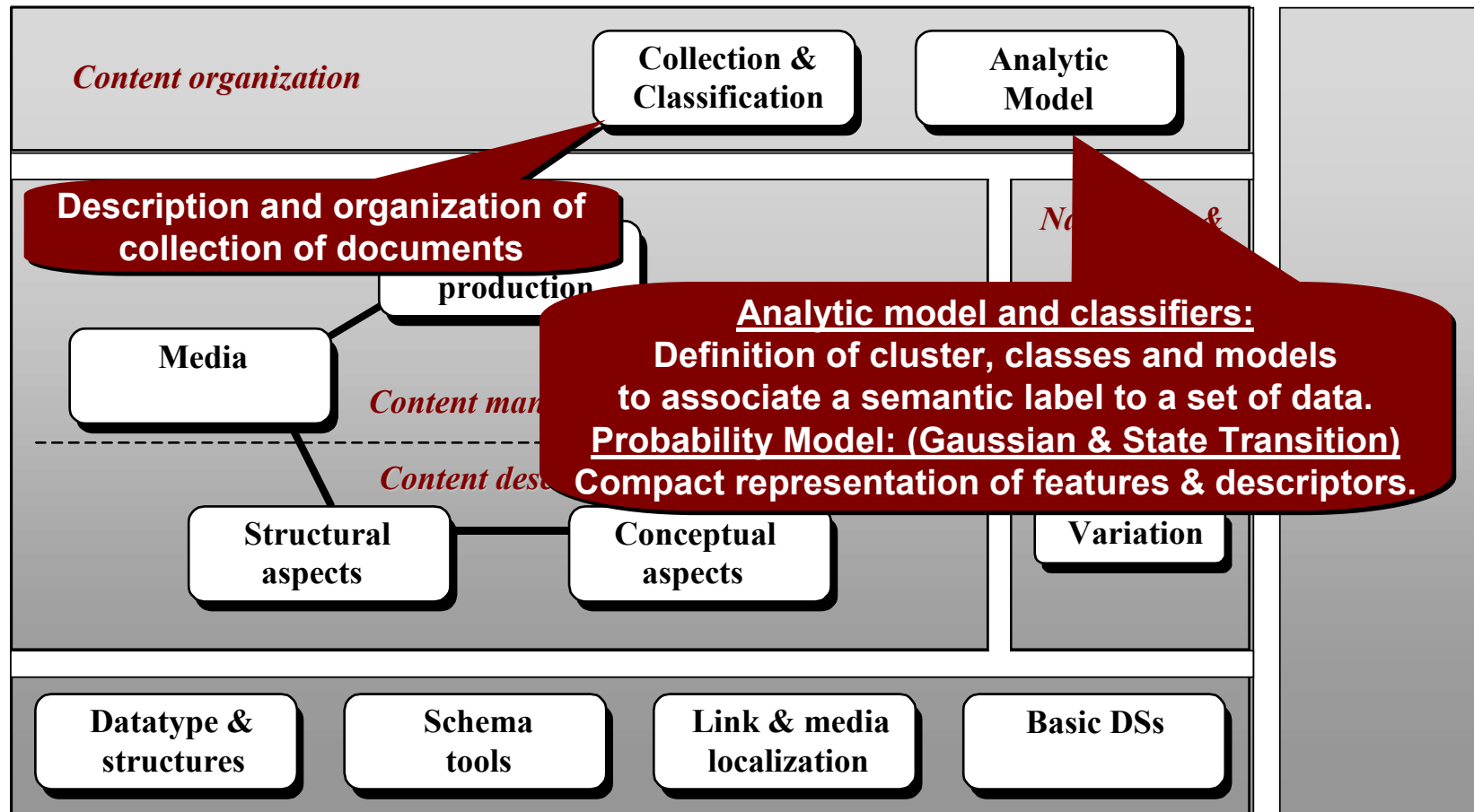


Variation

Universal Multimedia Access: Adapt delivery to network and terminal characteristics (QoS)

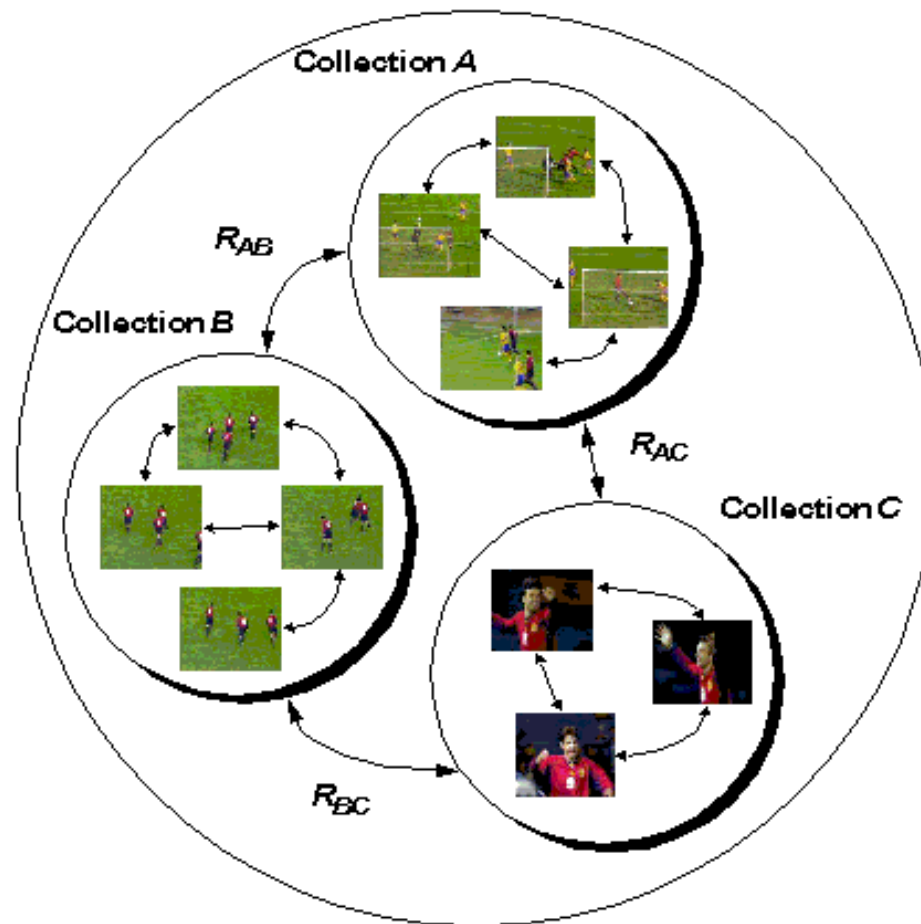


Content Organization

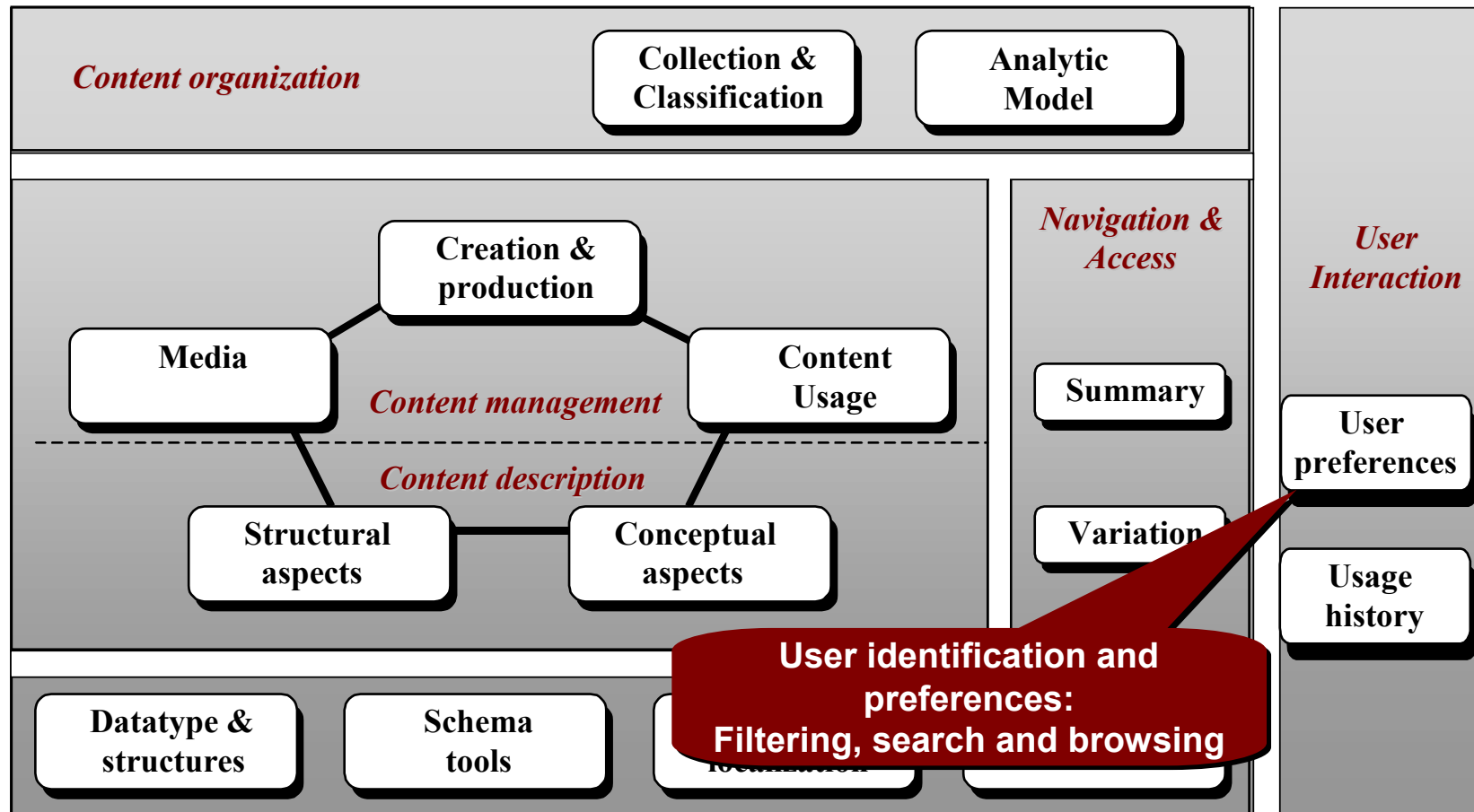


Collection

Collection Structure



User Interaction



Outline

- Objective, goals, requirements and applications
- Basic component of the MPEG-7 standard
- Systems tools and Description Definition Language
- Multimedia Description Schemes
- Visual Tools
- Audio Tools
- Relation with other standards

MPEG-7 Visual Part

MPEG-7 Visual part contains 25 Ds/DSs

Basic Elements (2)

Motion (4)

Color (7)

Localization (2)

Texture (3)

Face (1)

Shape (3)

Containers (3)

Basic Elements

- Coordinates:

Spatial2DCoordinates

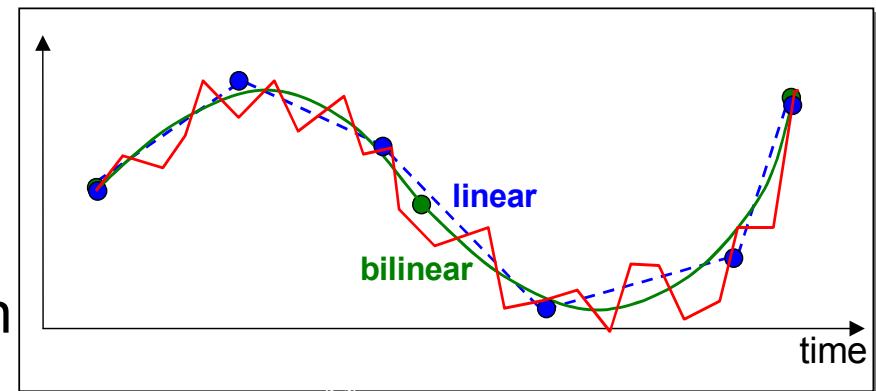
- Units: meters, pixels, pixels normalized by image size
- Spatial Reference: fixed, varying (to follow global motion)

➡ *independence of descriptions from display format*

- Temporal Interpolation:

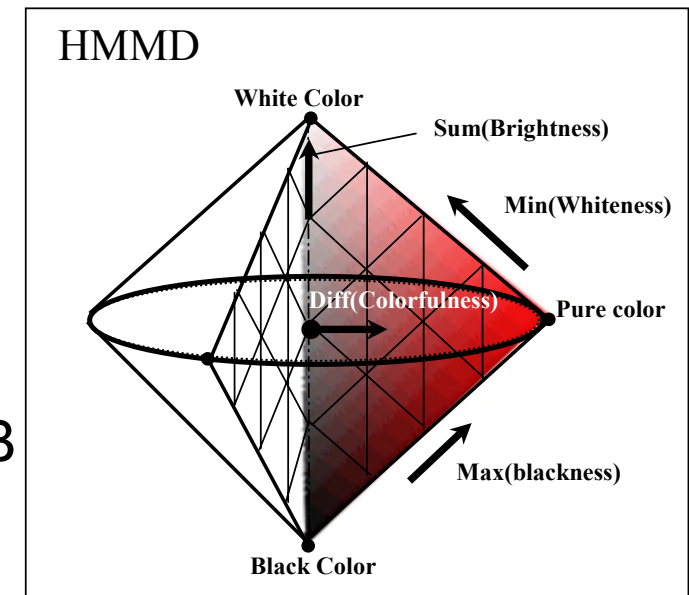
- (n-Dimension + time) values
- optionally parameters define bilinear interpolation

TemporalInterpolation



Color (1)

- Color Space: *ColorSpace*
 - R, G, B
 - Y, Cr, Cb
 - H, S, V
 - Monochrome
 - Linear transformation of R, G, B
 - HMMD

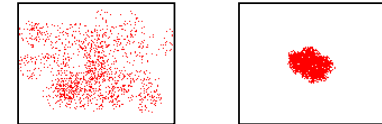


- Quantization: *ColorQuantization*

uniform quantization scheme to be applied on any of the color spaces defined above

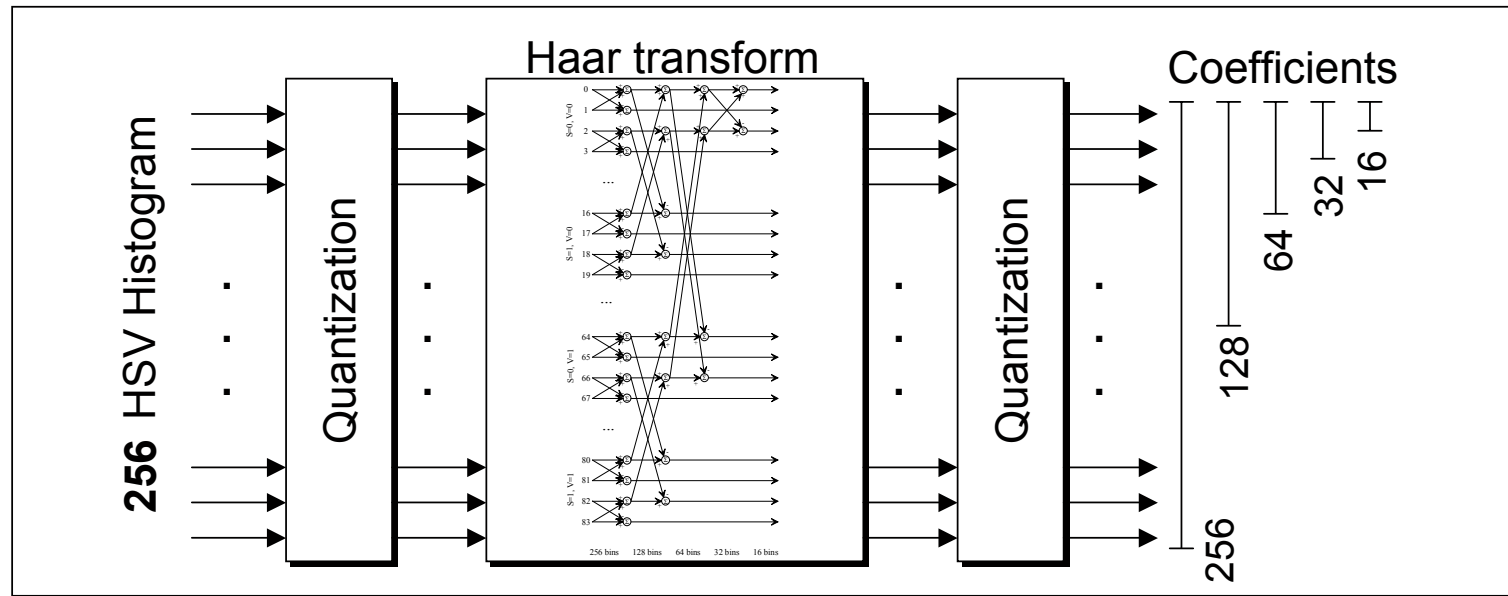
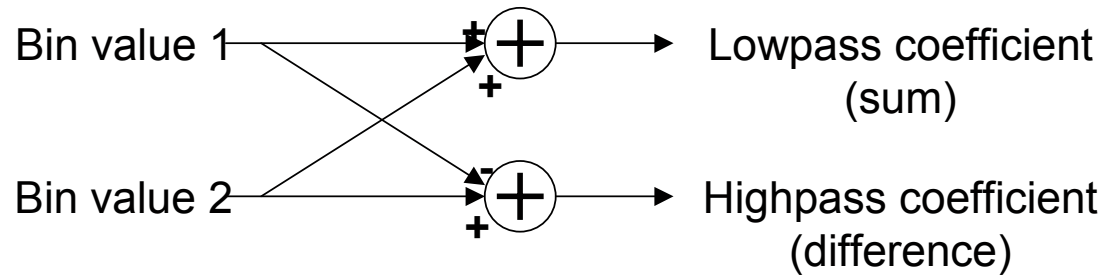
Color (2)

- Dominant Color(s): *DominantColor*
 - ✓ 1-8 dominant colors in image / region
 - ✓ color space, quantization, dominant color(s) value(s)
 - ✓ variance of color value, percentage of pixels of this color, spatial coherency of color repartition



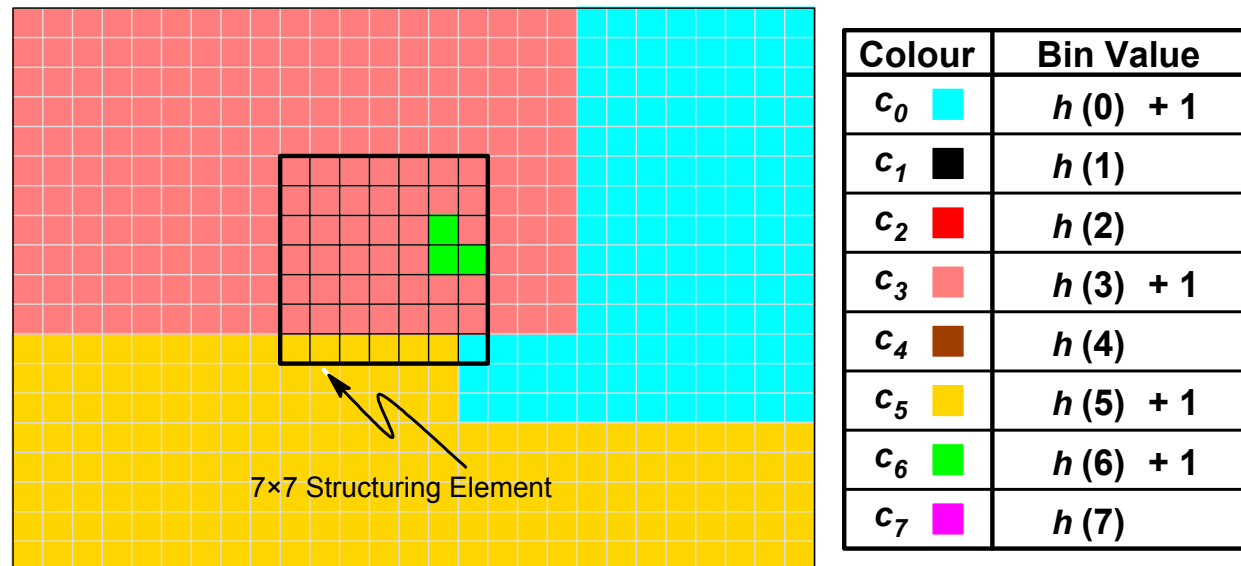
- Color Content (histogram): *ScalableColor*
 - ✓ Color histogram transformed by Haar transform
 - ✓ scalable in number of coefficients kept for representation
 - ✓ scalable in number of bits per coefficients
 - ✓ lower end: 60 bits, very fast matching

Scalable color: Haar transform



Color (3)

- Color content + coherence of colors repartition: *ColorStructure*
Histogram of struct. elements that contain a particular color.
➡ *Enhanced retrieval (in conjunction with HMMD)*



Color (4)

- Color content + its layout:

ColorLayout

Based on the DCT coefficients. (size: about 160 bits)

➡ *Layout sensitive retrieval, sketch-to-image matching*

- Color content of Group of Pictures / Frames:

GoFGoPColor

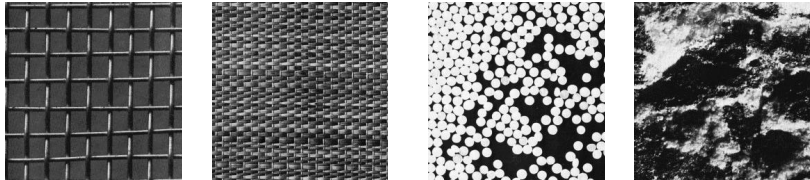
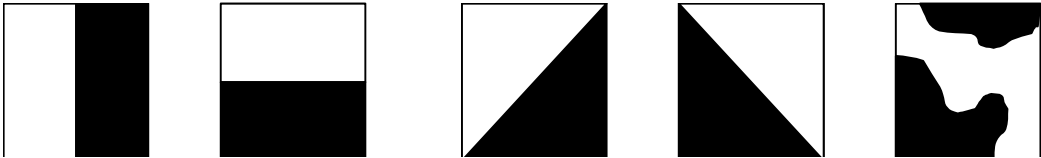
Aggregation of color histograms (average, median, etc)

➡ *Clustering of data for browsing / retrieval*

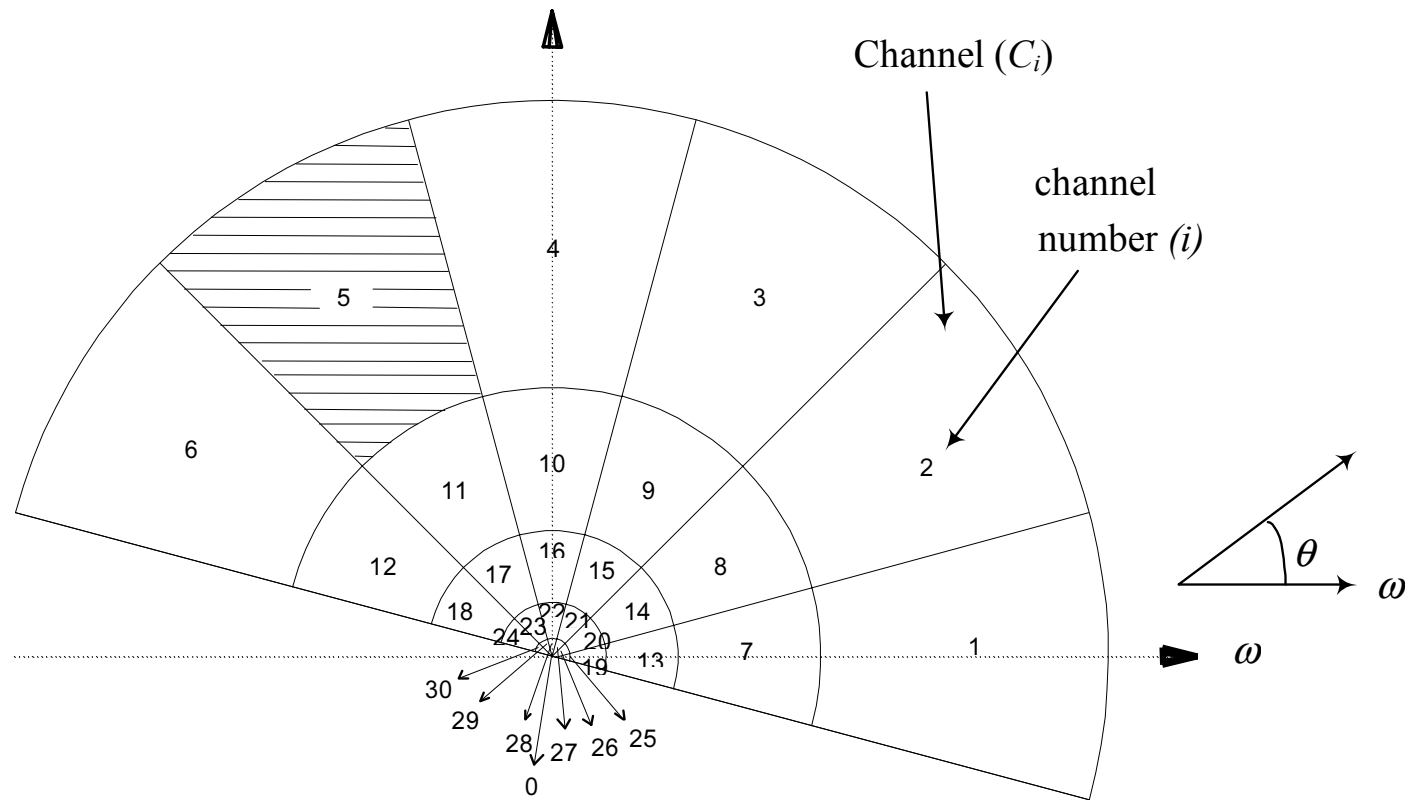
Texture (1)

- Characterization of homogeneous textures:
 - low-level: *HomogeneousTexture* → *retrieval*
 - high-level: *TextureBrowsing* → *browsing*
- Characterization of structures in generic images:
 - edges content and layout: *EdgeHistogram*

Texture (2)

Low-level: <i>HomogeneousTexture</i>	High-level: <i>TextureBrowsing</i>
In the frequency domain: • Decomposition into 30 channels (5 scales, 6 angles) using Gabor filters • Energy and energy dev.	• Regularity (1 to 4) • Main direction(s) • Coarseness (1 to 4) 
<i>EdgeHistogram</i> • Fixed size: 240 bits	• Edge types histograms on 16 sub-images 

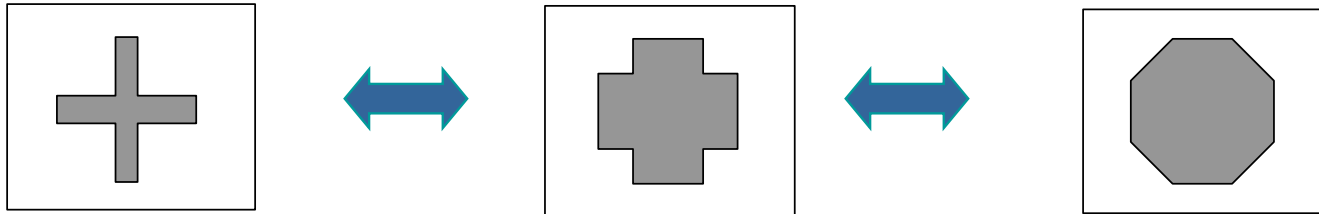
Homogeneous Texture: Filter bank



Shape (1)

- 2D:

ContourShape and *RegionShape*



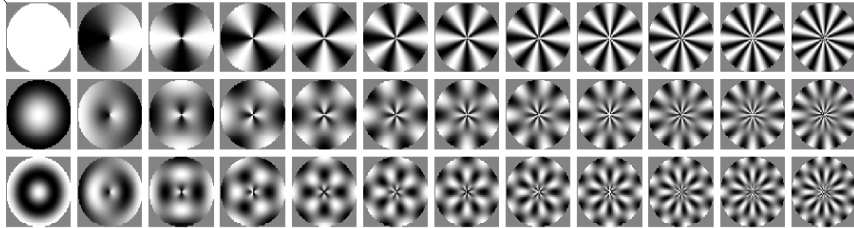
- 3D:

Shape3D

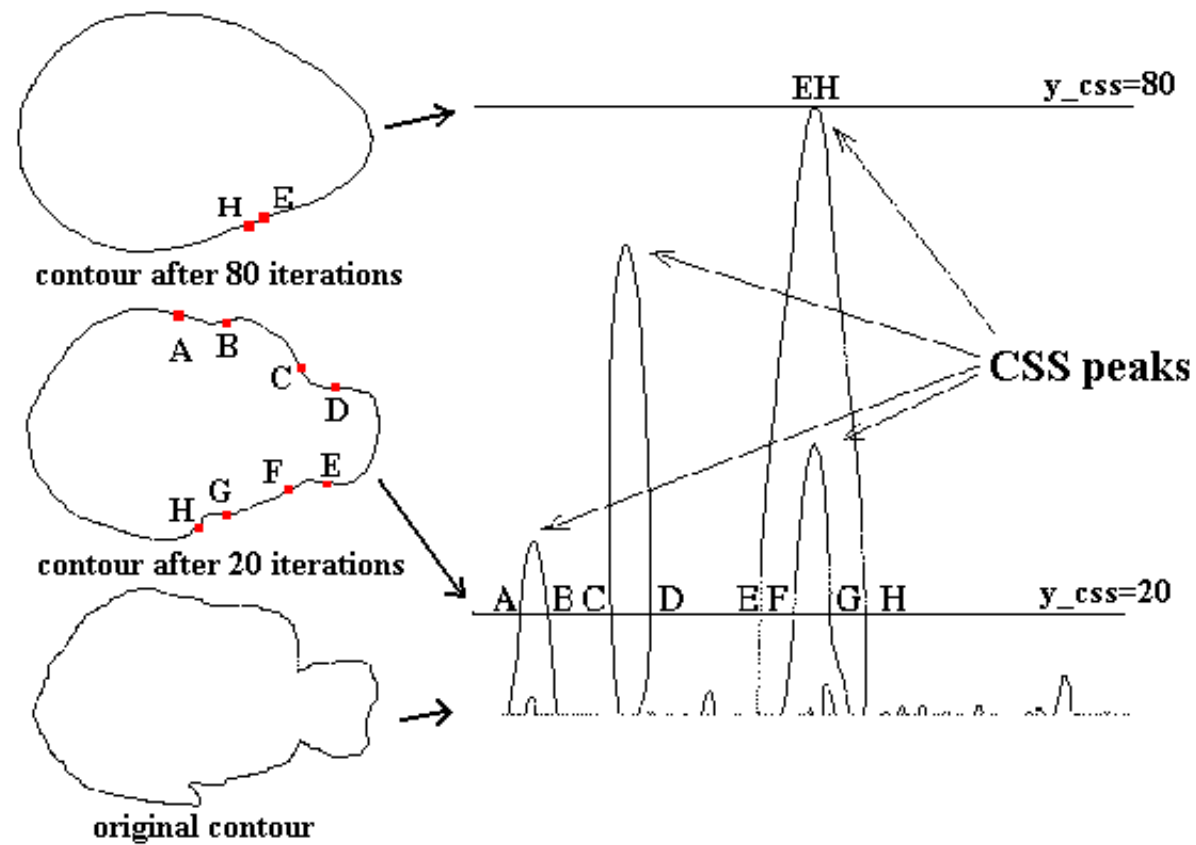
- Some 2D similarity-based retrieval results examples



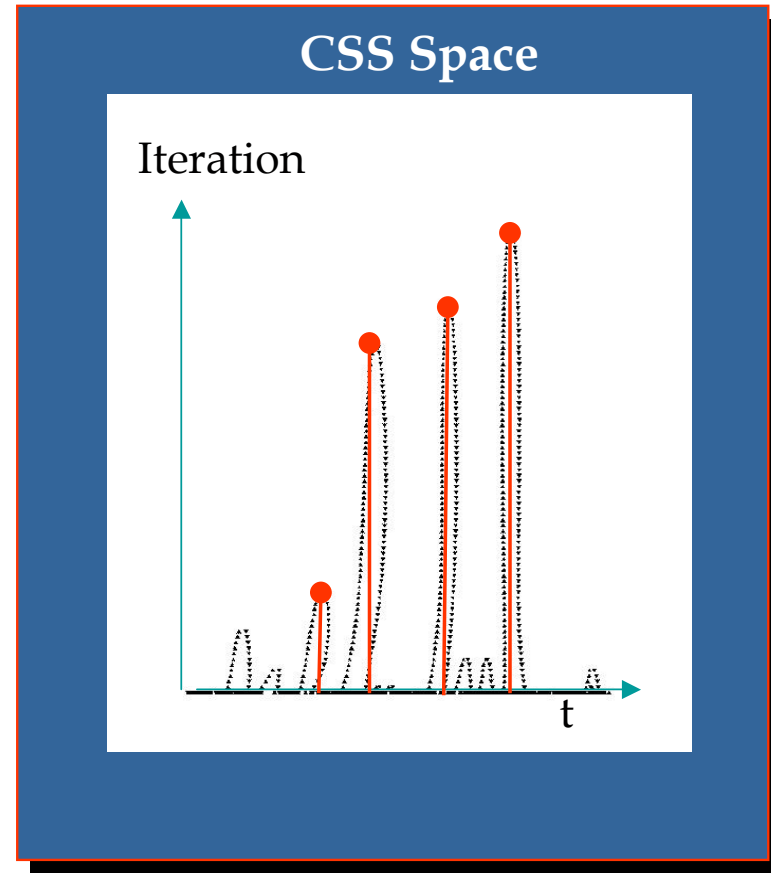
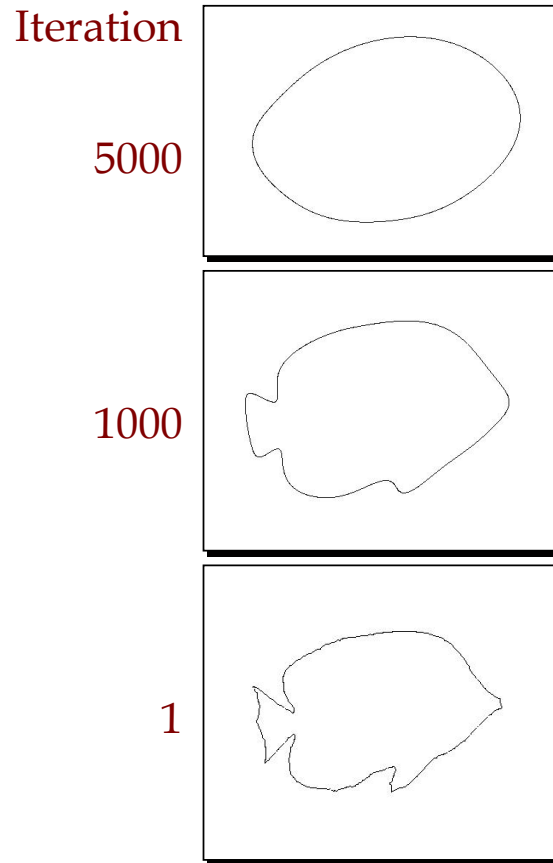
Shape (2)

Contour Based: <i>ContourShape</i>	Region Based: <i>RegionShape</i>
Curvature Scale Space: • curvature points importance and relative positions • Variable size: < 15 Bytes	• Angular Radial transform  • Fixed size: 17.5 Bytes
<i>Shape3D</i> • Based on 3D meshes: histogram of 3D shape indexes (Koenderink) representing local curvature properties of the 3D surface	

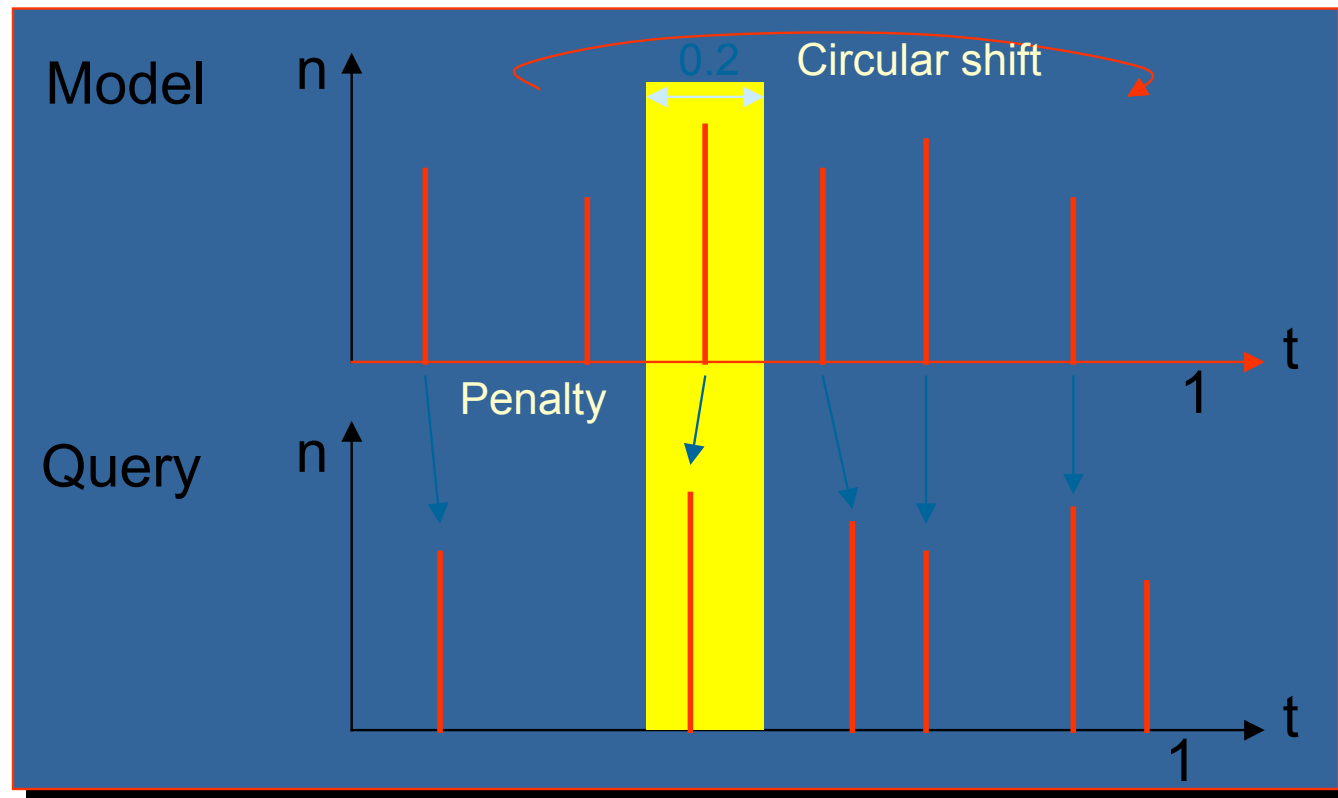
Curvature Scale Space



ContourShape: Curvature Scale Space Descriptor



CSS Descriptor



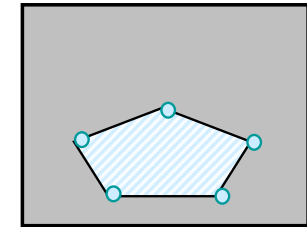
Localization

- Spatial Localization:

RegionLocator

Approximation of region by box or polygon

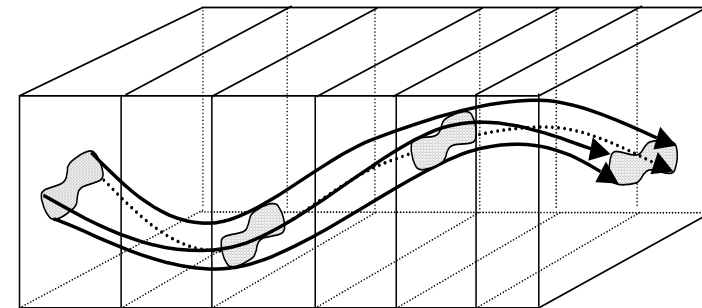
Position of vertices in image



- Spatio-Temporal Localization:

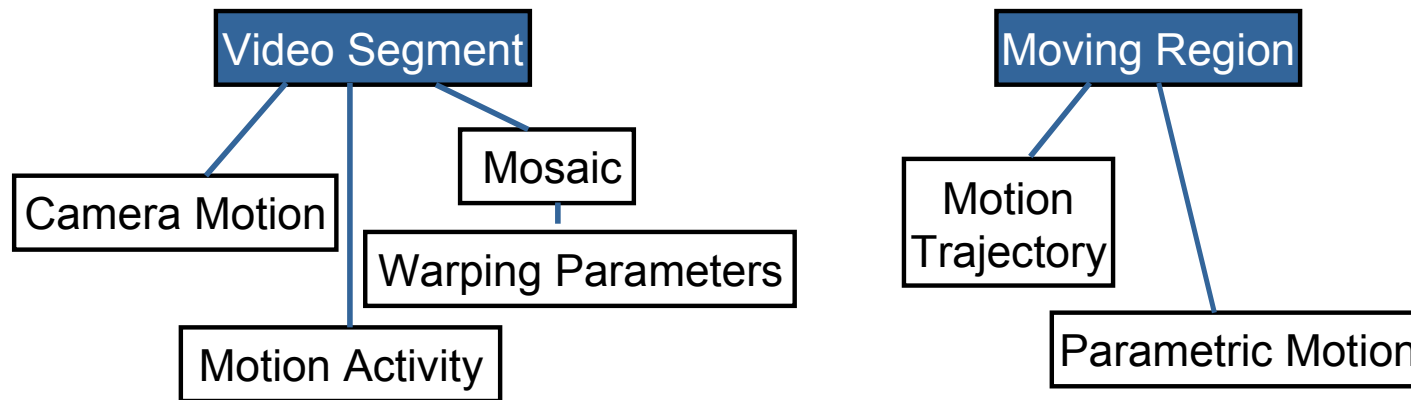
SpatioTemporalLocator

- Reference region, approximated by box, ellipse or polygon.
- Trajectories of representative points
- Temporal interpolation of parametric motion parameters



➡ *hyperlinking*

Motion (1)



MotionActivity

**browsing, repurposing
queries**

CameraMotion

browsing, high level queries

MotionTrajectory

retrieval, high level

ParametricMotion

mosaic, retrieval

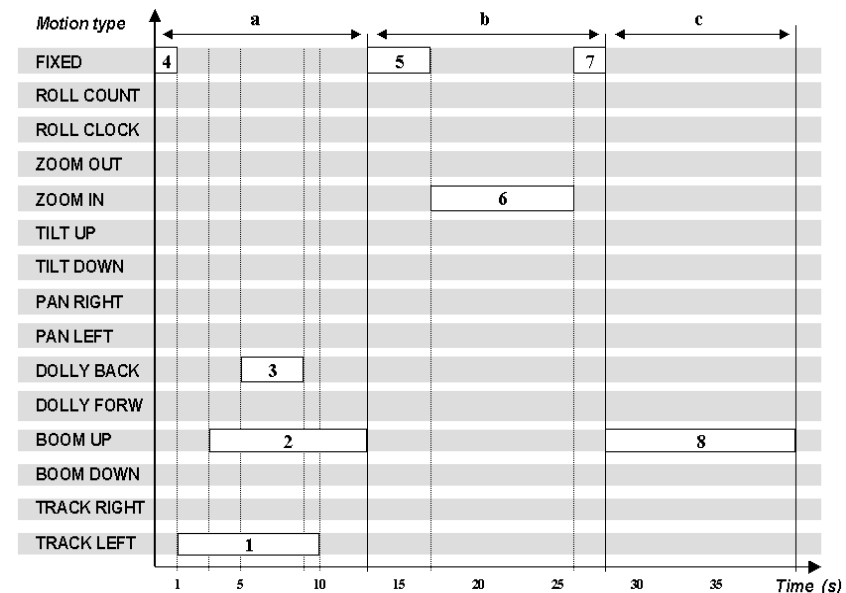
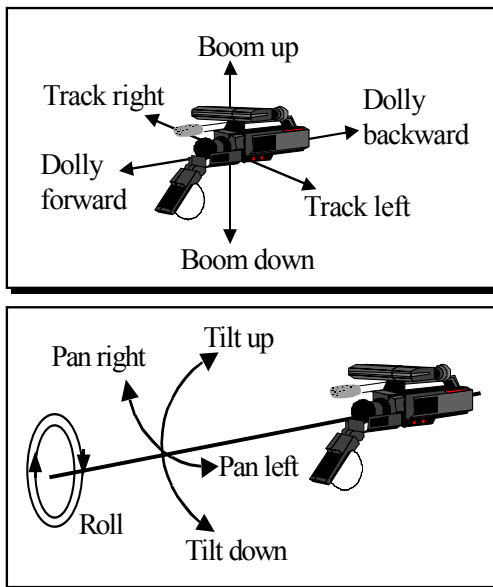
Motion (2)

- Motion Activity:
Intensity of motion (1 to 5)
main direction(s)
spatial and temporal distribution

MotionActivity

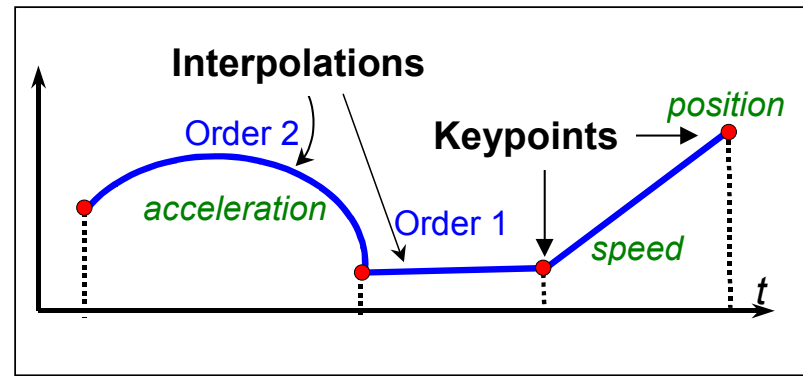
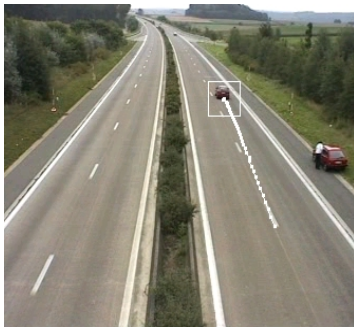
- Camera Motion:

CameraMotion



Motion (3)

- Motion Trajectory:

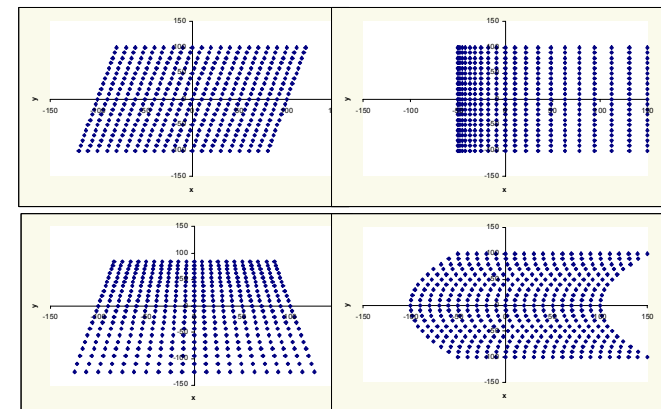
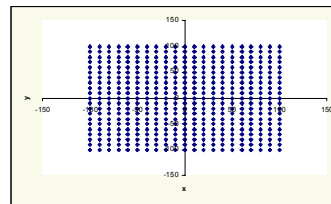


Queries:

- similarity
- high level

- Parametric Motion:
 - translational
 - rotation/scaling
 - affine
 - planar perspective
 - parabolic

Parametric Motion



Face

- Face Characterization: *FaceRecognition*
 - Size: 238 bits
 - Based on eigenfaces (vector of 2576 values, extracted from normalized faces)
 - 49 basis vectors which span the space of possible face vectors
 - projection of the face vector on the 49 eigenfaces

Outline

- Objective, goals, requirements and applications
- Basic component of the MPEG-7 standard
- Systems tools and Description Definition Language
- Multimedia Description Schemes
- Visual Tools
- Audio Tools
- Relation with other standards

Audio descriptors

- Low level audio features:
 - ◆ Waveform and spectrum envelopes
 - ◆ Power, spectrum centroid, spectrum spread
 - ◆ Fundamental frequency, harmonicity
 - ◆ Independent spectral component representation
- Spoken content
 - ◆ Lattice of hypothesis
- Music:
 - ◆ Timbre, Melody
 - ◆ Genre
- Silence description

Use of description tools

- Library of tools!
- The description tools are presented on the basis of the functionality they provide.
- In practice, they are combined into meaningful sets of description units.
- Furthermore, each application will have to select a sub-set of descriptors and DSs.
- DDL can be used to handle specific needs of the application.

Relation with other standards

- AHG on “Metadata integration” :
 - ◆ SMPTE: Metadata dictionary, KLV encoding
 - ◆ Dublin Core Metadata Initiative
 - ◆ European Broadcast Union
 - ◆ NewsML
- AHG on TV AnyTime Application
- Large number of Liaisons:
 - ◆ SMPTE
 - ◆ Dublin Core
 - ◆ W3C (XML Schema)
 - ◆ etc.

Conclusions

- MPEG-7:
 - ◆ AV content description for interoperable applications
- Description Definition Language:
 - ◆ XML Schema (flexibility) + Binary version (efficiency)
- Description Schemes and Descriptors:
 - ◆ Library of description tools
 - ◆ Covers a wide range of generic needs
 - ◆ Structural aspects close to Signal / Image Processing (segment trees and graph).
 - ◆ Low-level Descriptors characterize Segments.

Further information

- Major MPEG-7 documents are public:
 - ◆ MPEG Home page: <http://www.cselt.it/mpeg/>
 - ◆ Public documents: http://www.cselt.it/mpeg/working_documents.htm
 - ◆ Also check: <http://www.mpeg-7.com>



The SCHEMA NoE Reference System

Yiannis Kompatsiaris

ikom@iti.gr

<http://www.iti.gr/~ikom>

Informatics and Telematics Institute
Centre for Research and Technology Hellas
GREECE



Overview

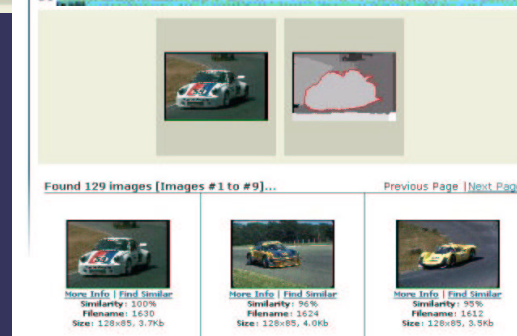
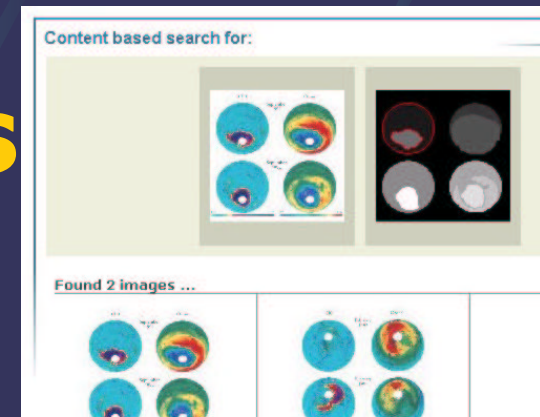
- Introduction
- Reference S/W
 - Analysis Modules
 - MPEG-7 XM
 - Conceptual search
 - Results
- TRECVID
- SCHEMA Description



Introduction

- 1-2 exabytes (millions of terabytes) of new information produced world-wide annually
- 80 billion of digital images are captured each year
- Over 1 billion images related to commercial transactions are available through the Internet
- This number is estimated to increase by ten times in the next two years.
- 4 000 new films are produced each year
- 300 000 world-wide available films
- 33 000 television stations and 43 000 radio stations
- 100 billions of hours of audiovisual content

- Content production, adaptation and consumption
 - Organize and share personal content
 - (Multimedia) Semantic web (multimedia search engines, directories, e-commerce)
 - Cultural Heritage
 - Medicine
 - Content filtering
 - Face detection
 - Transcoding
 - Augmented reality
 - The only “content-based” customer of FAST Norway is dealing with “inappropriate” content
- 
- A collage of three images. The top left shows a baby in a white shirt. The top right shows a person with a red face. The bottom right shows a man in a suit and tie.

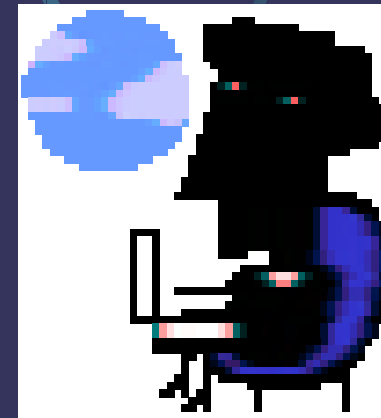


Objective

"The value of information depends on how easily can be found, retrieved, accessed, filtered or managed"

Approaches

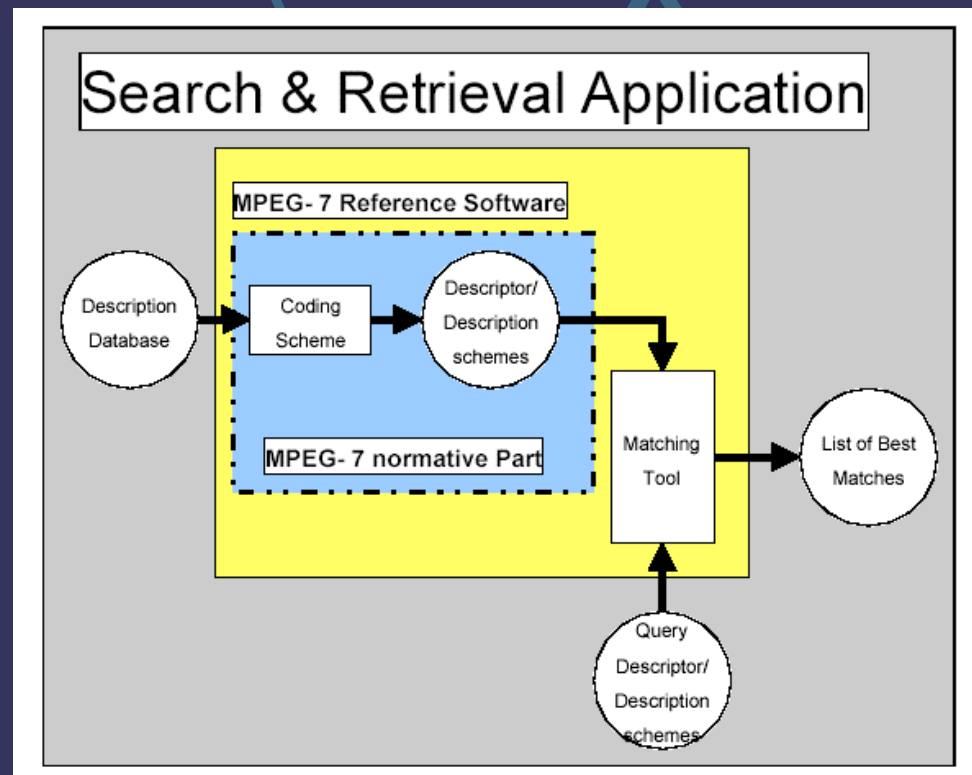
- **Manual text & caption based annotation**
 - + Straightforward
 - + High-level
 - + Efficient during content creation
 - Most commonly used
 - - Time consuming
 - - Operator-application dependent
 - - captions must exist
 - - difficult to express visual-specific queries (e.g. round object)



Approaches

- **Low-level features / Query by Example (color, texture, shape, edges, motion, etc)**
 - + automatic
 - + computation
 - Suitable for many applications
 - - low-level
 - - irrelevant results
 - - “visual” input is needed
- representation
 - features
 - color, texture space
 - invariance
 - compactness
- indexing (MPEG-7)
- database
- matching – distance
- global – local features (segmentation)

Query by Example



Approaches

- **Conceptual clustering – search (categorisation into broad classes, such as people-no people / indoors – outdoors)**
- + automatic
- + computation
- + semantic meaning
- Suitable for many applications
- - depend on training and detection efficiency
- - semantics only for the predefined classes

People – No People



Outdoor - Indoor

Approaches

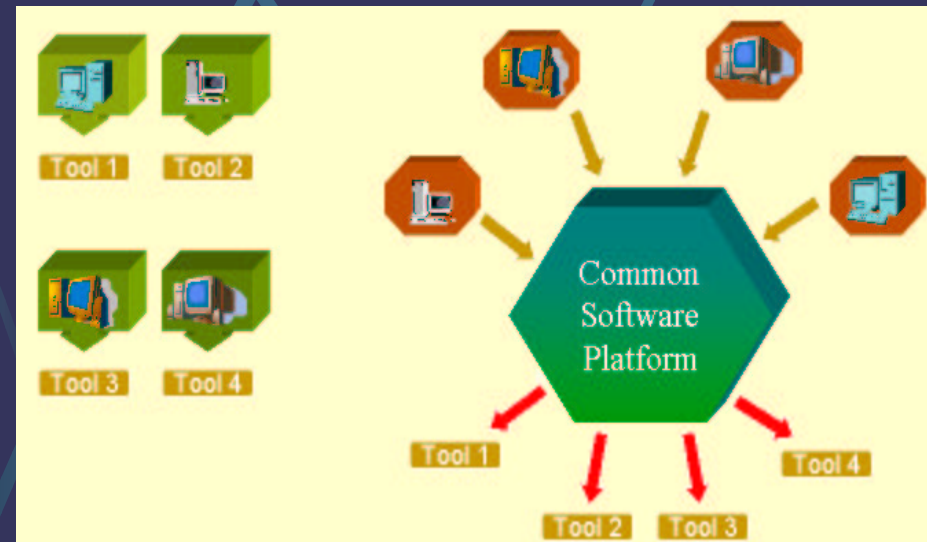
- **Semantic annotation of content**
- + High-level
- + Allows natural queries
- A-priori knowledge is usually needed
- - Domain specific
- - Computation
- - (semi) automatic



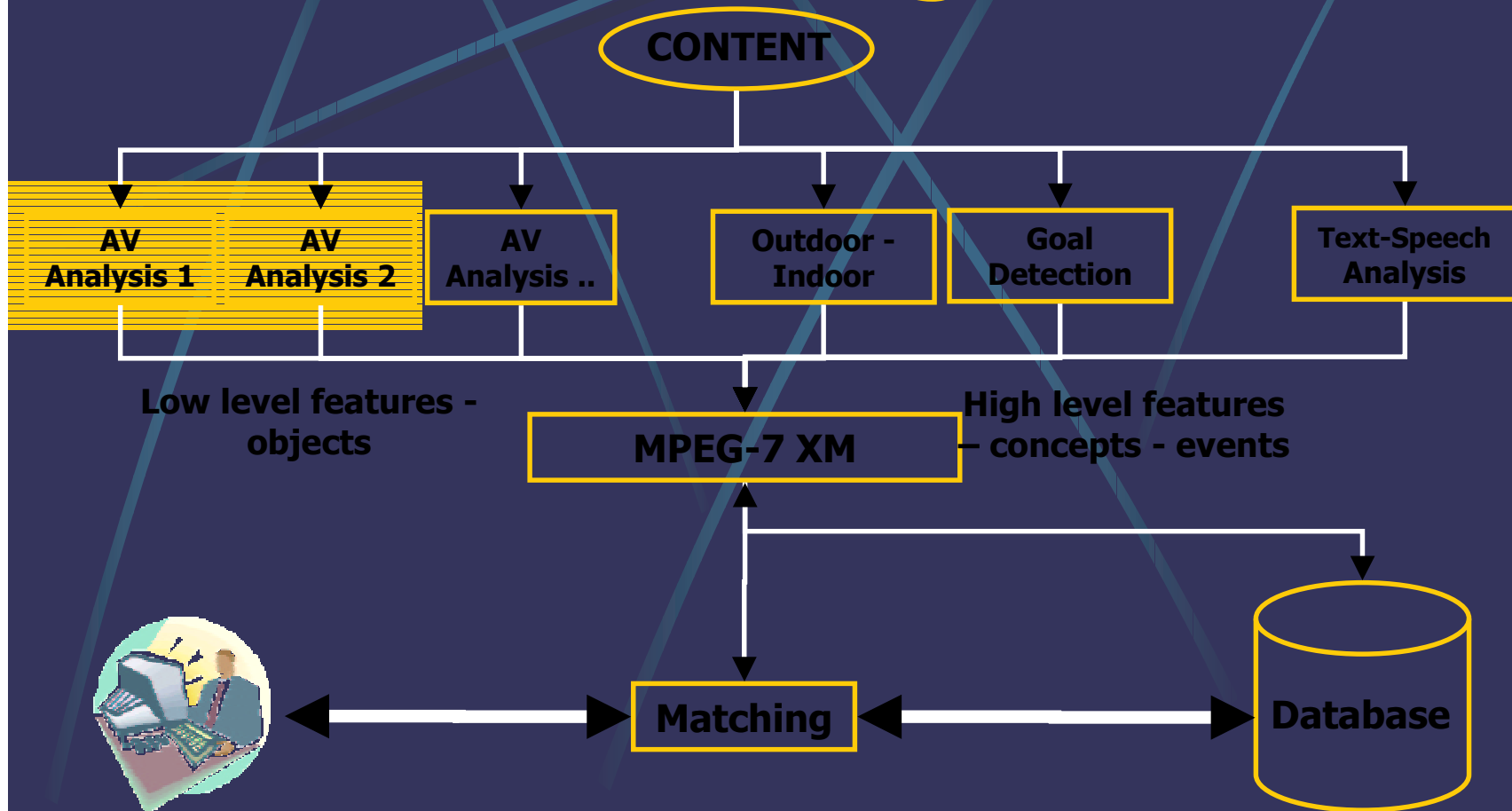
*"I want **video clips** of the **Greek football national team** containing **goals**"*

SCHEMA Reference System

- Develop a reference system for content-based indexing and retrieval
- Integrate individual analysis modules provided by different partners
- Unify multiple visual content (still image, video) and other modalities (e.g. text, audio) indexing and retrieval
- Combine low-level and high-level descriptors (clustering to classes)
- Take advantage of recent advances on standardization (MPEG-7)
- Provide a test-bed and a common dataset for the evaluation of different modules, descriptors and query types



Overall Diagram



Reference S/W Modules: Visual Analysis

- Video and image object segmentation is a crucial pre-processing step for:
 - Content-based functionalities for visual data
 - Users willing to find images or videos containing particular objects
 - Users allowed to see the segmentation of the query image and specify which aspects are central to the query
 - First step for knowledge extraction for semantic-based indexing and retrieval

Example
object-based
image search

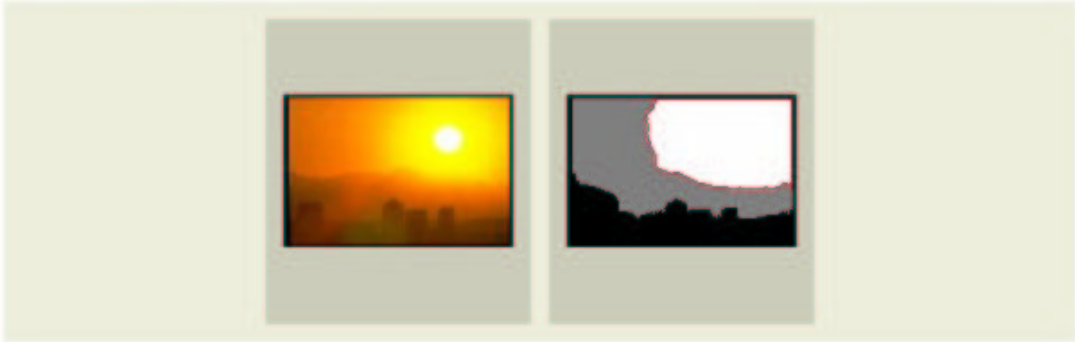
The user can
search for the
sun object
only

istrama
CONTENT BASED IMAGE RETRIEVAL OVER THE INTERNET

Home | About | Contact

Image Search

Content based search for: **Results**



Found 3 images ...

 More Info Find Similar Similarity: 100% Filename: 777 Size: 128x85, 2.3Kb	 More Info Find Similar Similarity: 92% Filename: 1795 Size: 128x85, 1.6Kb	 More Info Find Similar Similarity: 88% Filename: 1292 Size: 128x96, 1.9Kb
---	--	--

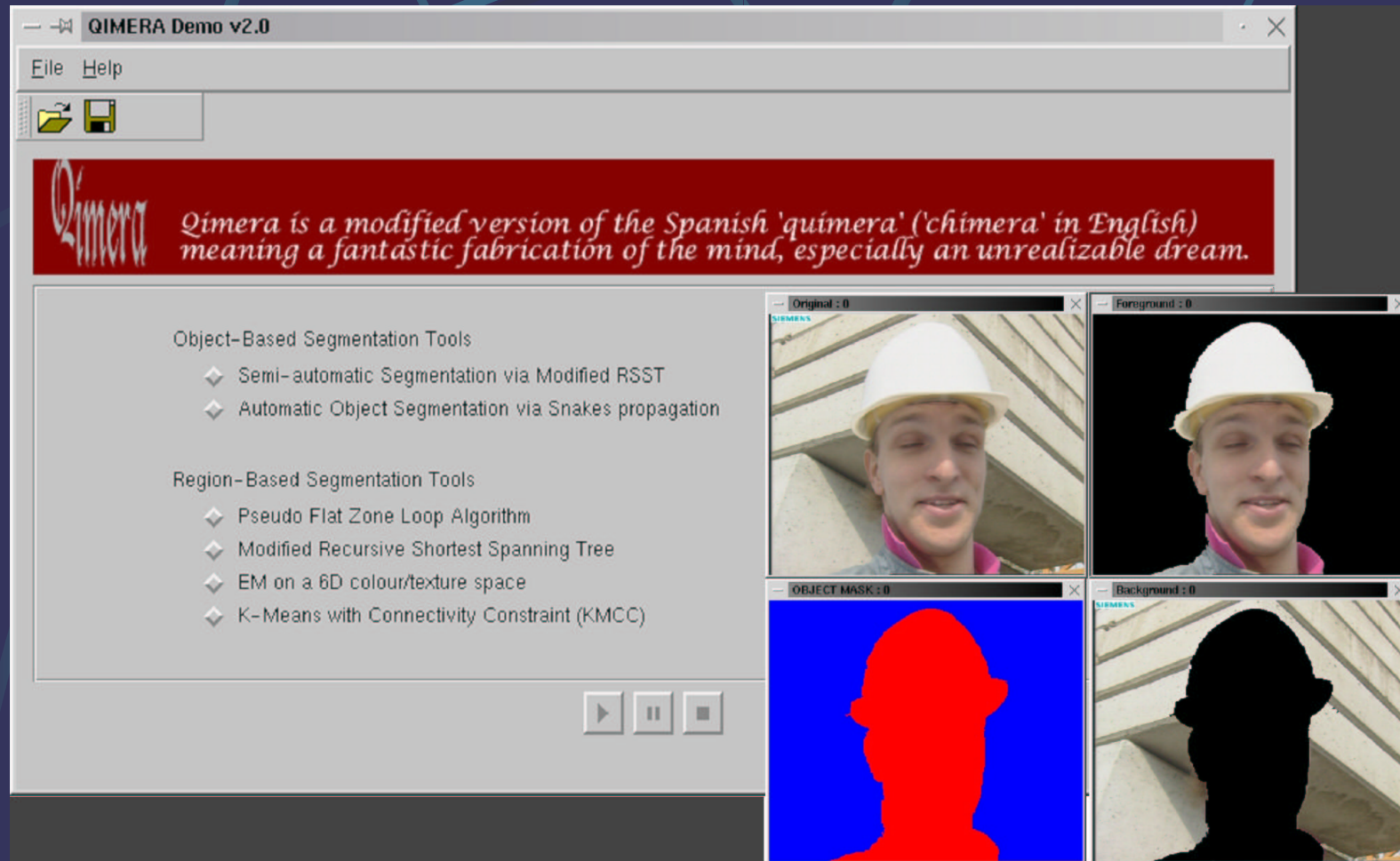
Analysis in the SCHEMA System

- Objective:
 - To develop a common software framework for video and image object segmentation
- Benefits:
 - a common basis for evaluating and testing segmentation algorithms
 - facilitates software exchange and collaborative algorithm development

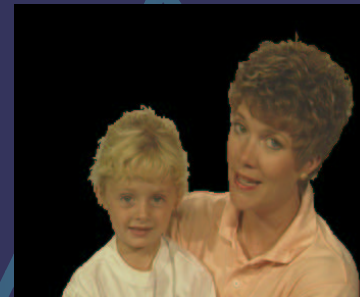
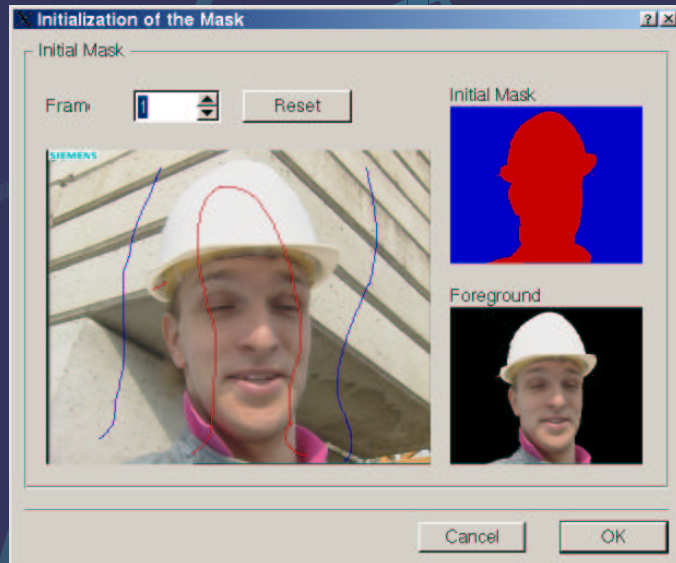
Schema Analysis Modules

- Region-based:
 - Modified Recursive Shortest Spanning Tree (RSST)
 - K-Means-with-connectivity-constraint (KMCC)
 - EM-based segmentation in 6D colour/texture space
 - Pseudo Flat Zone Detection
- Object-based:
 - Semi-automatic segmentation via modified RSST
 - Level-set based snake segmentation

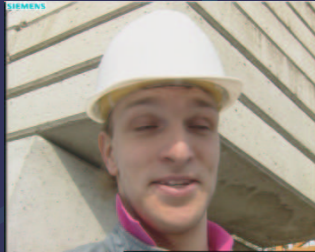
Results - QIMERA v2.0



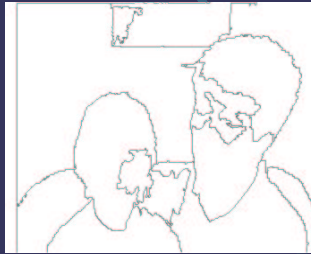
Results: Semi-automatic segm.



Results: Region-based segm.



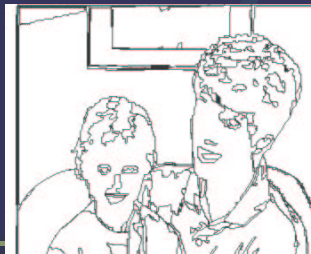
Modified Recursive Shortest Spanning Tree segmentation



K-means with Connectivity Constraint segmentation



EM -based 6D Space segmentation

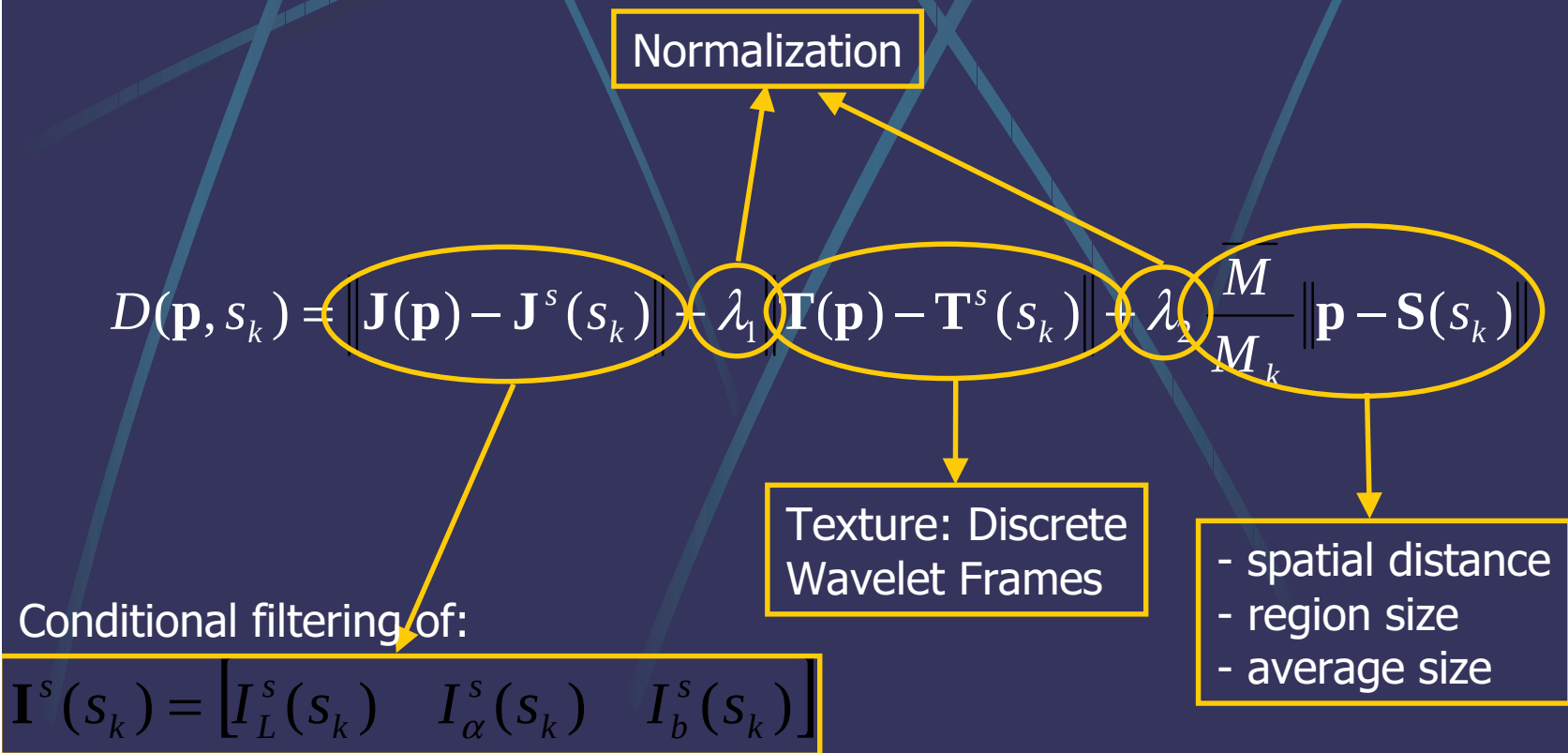


Pseudo-Flat Zone Colour segmentation

The K-Means with Connectivity Constraint Algorithm (KMCC)

- **Stage 1.** Extraction of the colour and texture feature vectors corresponding to each pixel
- **Stage 2.** Estimation of the initial number of regions and their spatial, intensity and texture centers
- **Stage 3.** Conditional filtering using a moving average filter
- **Stage 4.** Final classification of the pixels, using the KMCC algorithm

The K-Means with Connectivity Constraint Algorithm (KMCC)



The K-Means with Connectivity Constraint Algorithm (KMCC)



Experimental Results



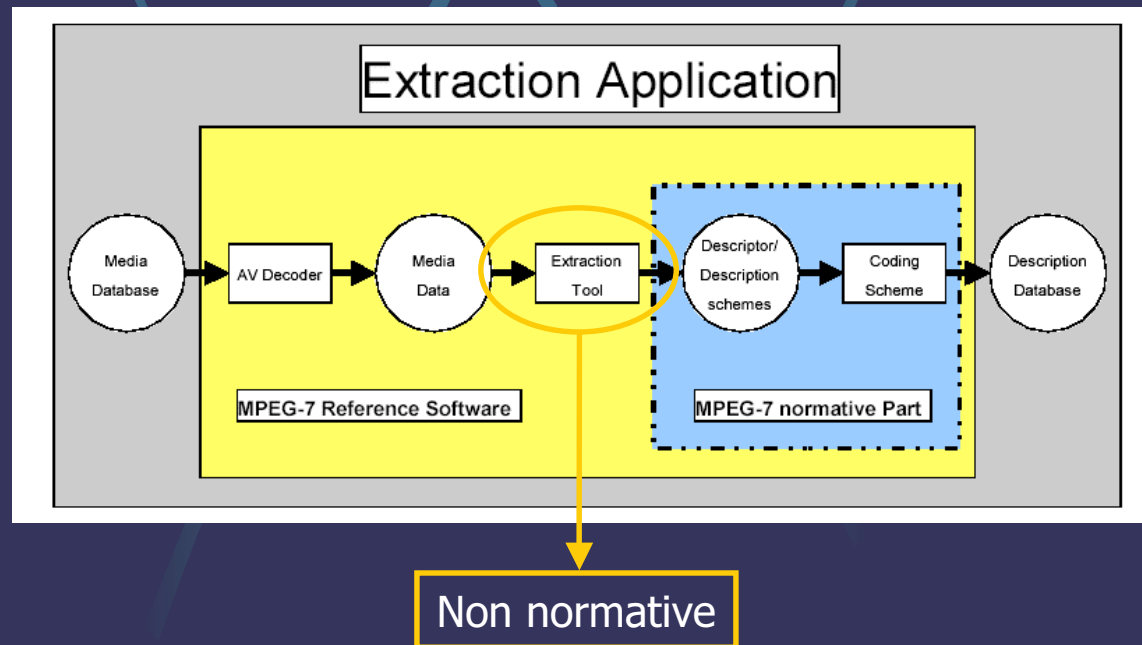
Experimental Results



MPEG-7 XM

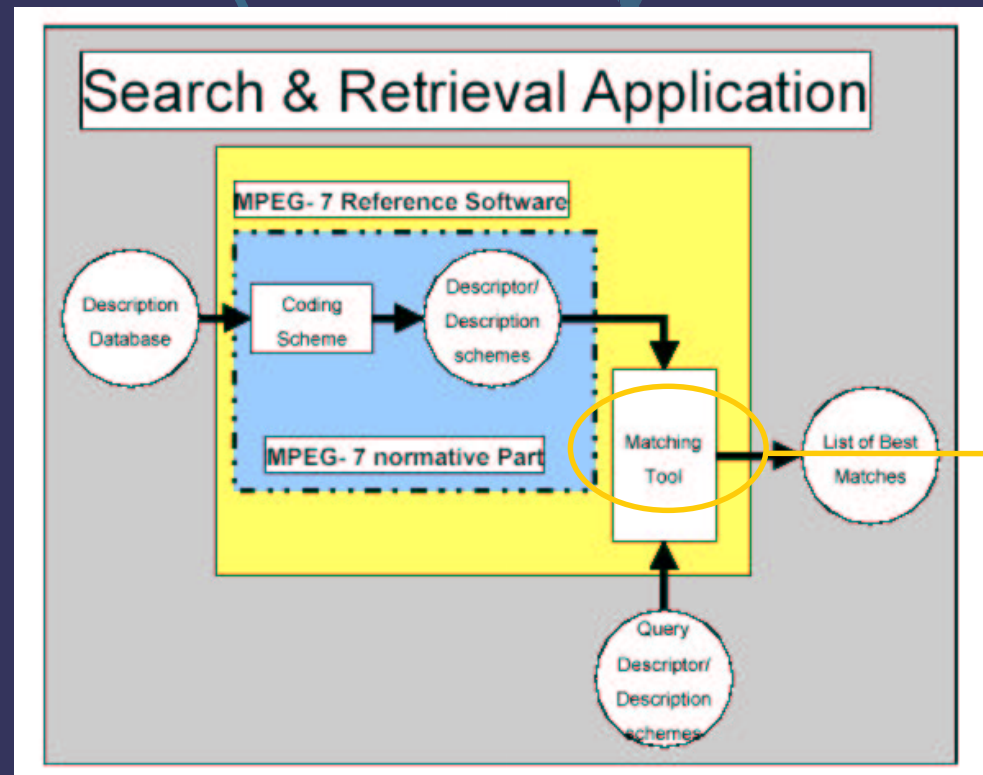
MPEG-7 eXperimentation Model, a non-normative part of the Standard realizing the normative descriptors

- A description database is built from a media database
- MPEG-7 descriptors extracted for each spatiotemporal object



MPEG-7 XM

- Compute distances between descriptions
 - indexing and retrieval
 - transcoding



Non normative

XM MultiImage Module

- The MultiImage extraction application performs extraction of several MPEG-7 Visual descriptors
 - Color Layout
 - Color Structure
 - Dominant Color
 - Scalable Color
 - Edge Histogram
 - Homogeneous Texture
 - Contour Shape
 - Region Shape

Conceptual search in the SCHEMA System

- Categorization of the database into predefined classes
 - Image containing faces – no faces
 - Outdoor images – indoor images
- Enables queries like
 - *I want images like this one with faces*



Conceptual search in the SCHEMA System

- The 'global' approach
 - apply a pattern recognition technique globally
- The region based approach
 - segment the image and apply classification to regions
 - For face – no face classification, the region based approach is followed in SCHEMA



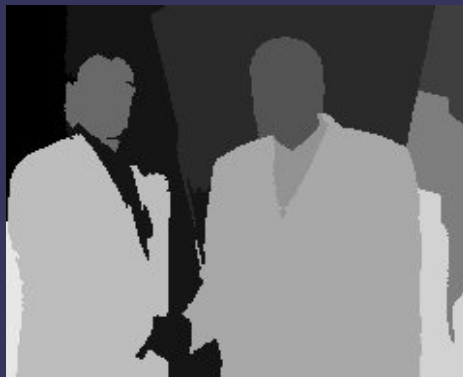
Learning procedure

- extract features from images (or regions)
- manually annotate a subset of the images available
- construct a 'concept model', i.e. a dense description of what you have 'taught' to the computer

Feature Extraction

- Currently the following features are available:
- Color Features (MPEG-7 Dominant Color, Lab centers, RGB centers, Ohta color space centers, color histogram)
- Edge Direction Features (modified MPEG-7 edge direction descriptor, edge direction histogram)
- Contour Shape Features (MPEG-7 Curvature Scale Space)
- Position Features (only for regions)

Region Feature Extraction and Classification



- The features are extracted for each region after the image has been segmented.
- A n – dimensional feature vector $f = \{f_1, f_2, \dots, f_n\}$ is obtained for each region.
- The MAP criterion (maximum a posteriori probability), is used for classification

System v1.0 GUI



Welcome

SCHEMA
Network of Excellence in
Content-Based Semantic
Scheme Analysis and
Information Retrieval

One of the main results
of SCHEMA will be a
general architecture for
content-based analysis,
representation, content
protection, indexing and
retrieval systems,
designed under WP3.



The employed image
analysis algorithms have
been developed under
the Qimera Initiative.



© 2002 SCHEMA
(IST-2001-32795)
A project financed by the
European Community

Search

1
Provide an image

2
Select a region

3
Define desired features

In order to initiate a search you must first **provide an image**.
You can either search for an image using a keyword, browse
through our categories or upload an image you already have.

Keyword Search

Search for:

or Browse Categories





eagles horses roses

More Categories: [airplanes](#), [fireworks](#), [cars](#), [sunsets](#)

or

Se

Browse Categories - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home | About | Contact

Browse Categories

Category: cars

Results

Found 5 images ...

 More Info Find Similar Filename: 29002 Size: 127x84, 3.4Kb	 More Info Find Similar Filename: 29015 Size: 84x127, 3.4Kb	 More Info Find Similar Filename: 29020 Size: 127x84, 3.4Kb
 More Info Find Similar Filename: 29021 Size: 127x84, 3.4Kb	 More Info Find Similar Filename: 29036 Size: 127x84, 3.4Kb	

[works](#) | [sunsets](#)

Information Retrieval
Informatics and Telematics Institute (CERTH)
The images are copyright [Coral Corp.](#)

The user can select a category to browse

The user can select an image to start the query

System v1.0 GUI

Image Search - Microsoft Internet Explorer

File Edit View Favorites Tools Help

ICHENA

About | Contact

Search

Step 1

Region Selection

Select a region from the masks that is of interest to you.

Original

MRSST (DCU)

KMCC (ITI)

EM (QMUL)

PFZL (TUM)

Features Selection

How important

Color: Not at all | Very

Shape/Size: Not at all | Very

Location: Not at all | Very

Go →

The user can select from five different algorithms

Automatic segmentation of the example image

The user can adjust the features weights

http://iti-62.itl.gr/site/Schema/en/results/29021.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

ICHENA

Home | About | Contact

Image Search

Content based search for:

Results

Found 25 images [Images #1 to #9]...

Previous Page | [Next Page](#)

 More Info Find Similar Similarity: 100% Filename: 29021 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 96% Filename: 29095 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 95% Filename: 29078 Size: 127x84, 3.4Kb
 More Info Find Similar Similarity: 95% Filename: 29076 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 95% Filename: 29043 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 95% Filename: 29033 Size: 127x84, 3.4Kb



Bald Eagle Query

Image Search - Microsoft Internet Explorer

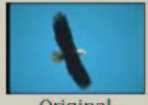
File Edit View Favorites Tools Help

Home | About | Contact

Image Search

Region Selection Step 1

Select a region from the masks that is of interest to you.

		
Original	MRSST (DCU)	KMCC (ITI)
		
EM (QMUL)		
		
PFZL (TUM)		

Features Selection Step 2

How important are the features of the region you've selected?

Color:	Not at all	Very
Shape/Size:	Not at all	Very
Location:	Not at all	Very

Go →

http://iti-62.iti.gr/site/Schema/en/results/135060.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home | About | Contact

Image Search

Content based search for: Results

Found 39 images [Images #1 to #9]... [Previous Page](#) [Next Page](#)

 More Info Find Similar Similarity: 100% Filename: 135060 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 98% Filename: 135068 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 97% Filename: 135032 Size: 127x84, 3.4Kb
 More Info Find Similar Similarity: 95% Filename: 135028 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 94% Filename: 135023 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 94% Filename: 135047 Size: 127x84, 3.4Kb



Brown Horse Query

Image Search - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home | About | Contact

Image Search

Region Selection Step 1

Select a region from the masks that is of interest to you.

 Original	 MRSST (DCU)	 KMCC (ITI)
	 EM (QMUL)	 PFZL (TUM)

Features Selection Step 2

How important are the features of the region you've selected?

Color:	Not at all	Very
Shape/Size:	Not at all	Very
Location:	Not at all	Very

Go →

http://iti-62.iti.gr/site/Schema/en/results/113068.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home | About | Contact

Image Search

Content based search for: Results

Found 149 images [Images #1 to #9]... Previous Page | [Next Page](#)

 More Info Find Similar Similarity: 100% Filename: 113068 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 99% Filename: 113089 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 97% Filename: 113092 Size: 127x84, 3.4Kb
 More Info Find Similar Similarity: 97% Filename: 113063 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 97% Filename: 113098 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 96% Filename: 113070 Size: 127x84, 3.4Kb



Rose Query

Image Search - Microsoft Internet Explorer

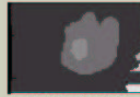
File Edit View Favorites Tools Help

Home | About | Contact

Image Search

Region Selection Step 1

Select a region from the masks that is of interest to you.

 Original	 MRSST (DCU)	 KMCC (ITI)
	 EM (QMUL)	 PFZL (TUM)

Features Selection Step 2

How important are the features of the region you've selected?

Color:	Not at all	Very
Shape/Size:	Not at all	Very
Location:	Not at all	Very

Go →

http://iti-62.iti.gr/site/Schema/en/results/84037.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home | About | Contact

Image Search

Content based search for: Results

Found 10 images [Images #1 to #9]... Previous Page | [Next Page](#)

 More Info Find Similar Similarity: 100% Filename: 84037 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 94% Filename: 84026 Size: 84x127, 3.4Kb	 More Info Find Similar Similarity: 93% Filename: 84012 Size: 127x84, 3.4Kb
 More Info Find Similar Similarity: 92% Filename: 84070 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 91% Filename: 84068 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 91% Filename: 84038 Size: 127x84, 3.4Kb



Red Car Query

Image Search - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home | About | Contact

Image Search

Region Selection Step 1

Select a region from the masks that is of interest to you.

 Original	 MRSST (DCU)	 KMCC (ITI)
	 EM (QMUL)	 PFZL (TUM)

Features Selection Step 2

How important are the features of the region you've selected?

Color:	Not at all	Very
Shape/Size:	Not at all	Very
Location:	Not at all	Very

Go →

http://iti-62.iti.gr/site/Schema/en/results/29020.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home | About | Contact

Image Search

Content based search for: Results

Found 10 images [Images #1 to #9]... [Previous Page](#) [Next Page](#)

 More Info Find Similar Similarity: 100% Filename: 29020 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 97% Filename: 29077 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 97% Filename: 29078 Size: 127x84, 3.4Kb
 More Info Find Similar Similarity: 96% Filename: 29075 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 96% Filename: 29017 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 96% Filename: 29091 Size: 127x84, 3.4Kb



Airplane Query

Image Search - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home About Contact

Image Search

Region Selection Step 1

Select a region from the masks that is of interest to you.

 Original	 MRSST (DCU)	 KMCC (ITI)
	 EM (QMUL)	 PFZL (TUM)

Features Selection Step 2

How important are the features of the region you've selected?

Color:	Not at all	Very
Shape/Size:	Not at all	Very
Location:	Not at all	Very

Go →

http://iti-62.iti.gr/site/Schema/en/results/34001.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Home About Contact

Image Search

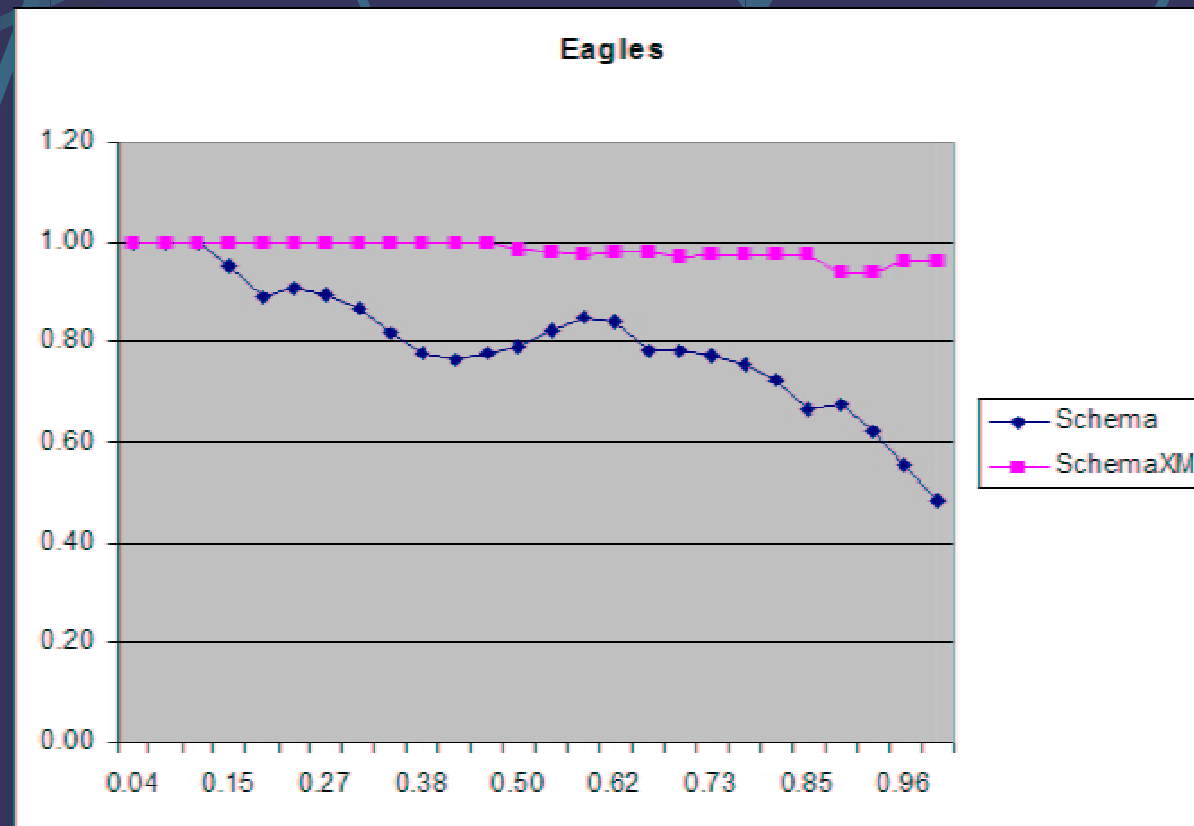
Content based search for: Results

Found 42 images [Images #1 to #9]... Previous Page | [Next Page](#)

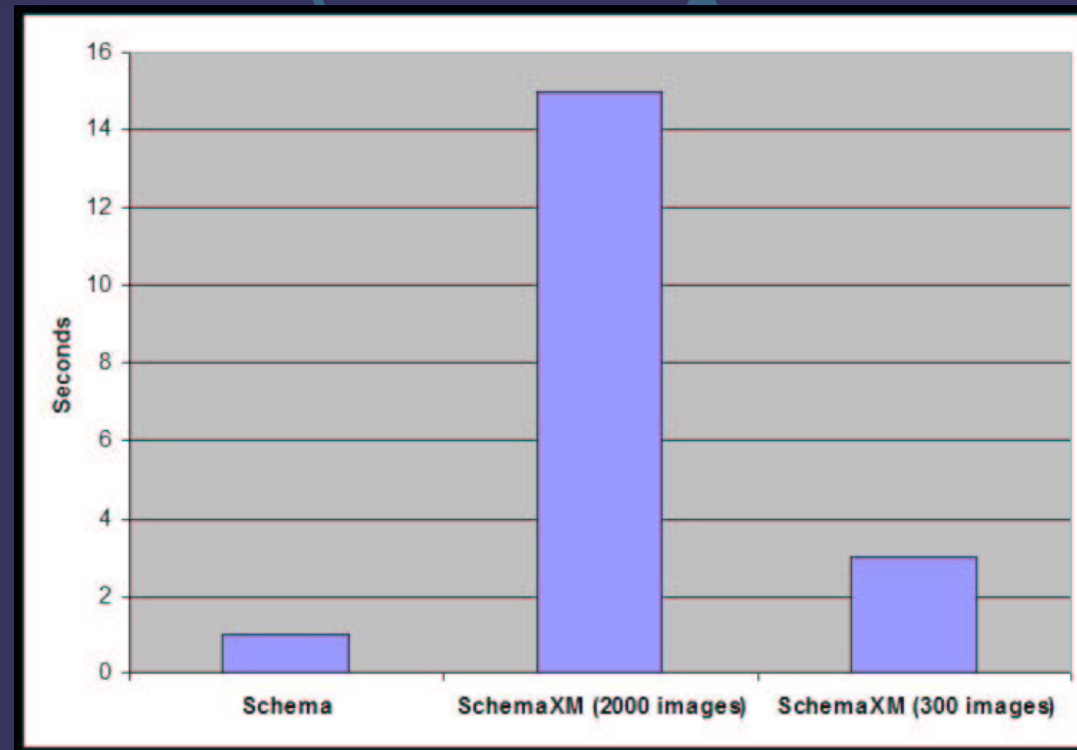
 More Info Find Similar Similarity: 100% Filename: 34001 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 97% Filename: 34087 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 96% Filename: 34034 Size: 127x84, 3.4Kb
 More Info Find Similar Similarity: 96% Filename: 34098 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 95% Filename: 29005 Size: 127x84, 3.4Kb	 More Info Find Similar Similarity: 95% Filename: 360091 Size: 127x84, 3.4Kb



Evaluation and Comparison



Evaluation and Comparison



TRECVID Overview

- Goal: to benchmark participants' video IR systems based on:
 - a specified set of tasks
 - a commonly available test corpus
 - a commonly agreed ground truth
- 4 tasks
 - Shot boundary detection
 - News story segmentation
 - High-level feature extraction
 - Search (manual and interactive)

Tasks

- Shot boundary detection
 - Given test corpus, identify shot boundaries and type
- News story segmentation
 - Given test corpus and shot bounds, identify story bounds and type (news/misc)
- High-level feature extraction
 - Given test corpus and shot bounds, identify which shots contain specified features
- Search
 - Given test corpus, shot bounds, story bounds, extracted features and a **topic**, return shots which satisfy the information need.

System Overview

- TRECVID provides shots, keyframes and associated text
- Based on a keyword query, the user will retrieve the most relevant shots (includes text retrieval and matching)
- The user can select a keyframe (shot)
- The SchemaXM visual query by example will follow

Example

Home | About | Contact

Keyword Search

Search for: "karamanlis"

Found 18 images [Images #1 to #9]... Previous Page

 More Info Find Similar Filename: 182546 Size: 84x127, 3.4Kb	 More Info Find Similar Filename: 324171 Size: 127x84, 3.4Kb	 More Info Find Similar Filename: 127x84 Size: 127x84
 More Info Find Similar Filename: 422127 Size: 127x79, 3.4Kb	 More Info Find Similar Filename: 422128 Size: 127x84, 3.4Kb	 More Info Find Similar Filename: 127x84 Size: 127x84
 More Info Find Similar	 More Info Find Similar	 More Info Find Similar Filename: 422133

Home | About | Contact

Image Search

Step 1

Select a region from the masks that is of interest to you.

 Original	 MRSST (DCU)	 KMCC (ITI)
	 EM (QMUL)	 PFZL (TUM)

Ολοκληρώθηκε

Internet



Example

Execution time: 1.16462278366 seconds

Home | About | Contact

Image Search

Content based search for: **Results**



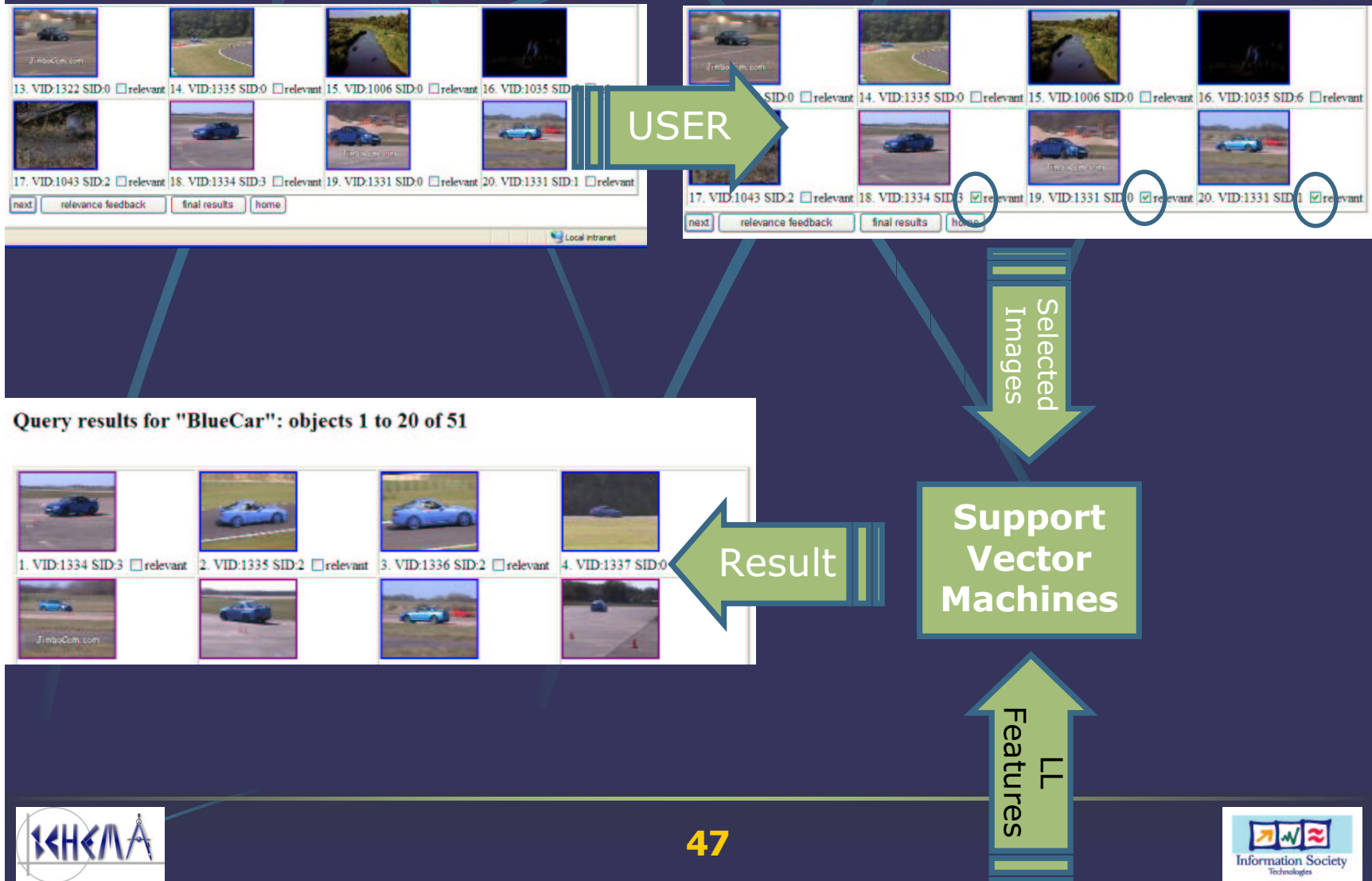
Found 18 images [Images #1 to #9]... [Previous Page](#) | [Next Page](#)

 More Info Find Similar Similarity: 100% Filename: 422131 Size: 94x127, 3.4Kb	 More Info Find Similar Similarity: 97% Filename: 422133 Size: 127x89, 3.4Kb	 More Info Find Similar Similarity: 97% Filename: 422132 Size: 97x127, 3.4Kb
 More Info Find Similar Similarity: 96% Filename: 422134	 More Info Find Similar Similarity: 95%	 More Info Find Similar Similarity: 95%

Internet



Relevance feedback



SCHEMA NoE Aim

- Bring together a critical mass of industrial partners, end users, universities and research centres
- Improve the systematic exchange of information by the forging of links between the partners
- Create a synergy and a multiplier-effect enriching the capabilities of the group of partners by bringing together complementary skills and know-how
- Continuous expansion of the network
- Develop common European R&D agendas
- Strengthen the European expertise and research excellence
- Disseminate the results worldwide



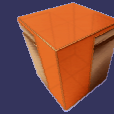
SCHEMA Research Topics

- Content-based multimedia analysis
- Access to the information using query structures that come naturally to human beings
- Semantic web technologies
- Copyright issues of multimedia
- New methods for multimedia access and delivery
- MPEG-7 and MPEG-21 standards
- User interfaces and human factors

Consortium

Research Institutes - Universities:

- Informatics and Telematics Institute
- Tampere University of Technology
- Munich University of Technology
- Université Catholique de Louvain
- Centre National de la Recherche Scientifique, Université de Nice – Sophia Antipolis
- Dublin City University - Centre for Digital Video Processing



Consortium

- Queen Mary, University of London
- Universitat Politècnica de Catalunya
- Fondazione Ugo Bordoni

- University of Brescia

Companies:

- Fratelli Alinari
 - BText Technologies
- ## End users:
- Macedonian Press Agency



Affiliated members

JOANNEUM
RESEARCH



UNIVERSITY OF
TRIESTE

UNIVERSITY
OF
QUEENSLAND,
AUSTRALIA



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

ARISTOTLE
UNIVERSITY OF
THESSALONIKI



LTU
TECHNOLOGIES

THOMSON
MULTIMEDIA R&D



KANGWON NATIONAL
UNIVERSITY, KOREA



ITALIAN NATIONAL AGENCY
FOR NEW TECHNOLOGIES,
ENERGY AND THE
ENVIRONMENT



INSTITUTE FOR
LANGUAGE AND
SPEECH PROCESSING



강원대학교

Kangwon National University

MOTOROLA UK
RESEARCH LAB

UNIVERSITY
OF PATRAS



UNIVERSITÀ
DI FIRENZE



MOTOROLA
intelligence everywhere™

HEWLETT-PACKARD
LABORATORIES, USA



UNIVERSITÄT
DES
SAARLANDES



INESC PORTO

COMPUTATIONAL
LINGUISTICS
DEPARTMENT, UNIVERSITY
OF SAARLAND

ORTA DOĞU TEKN

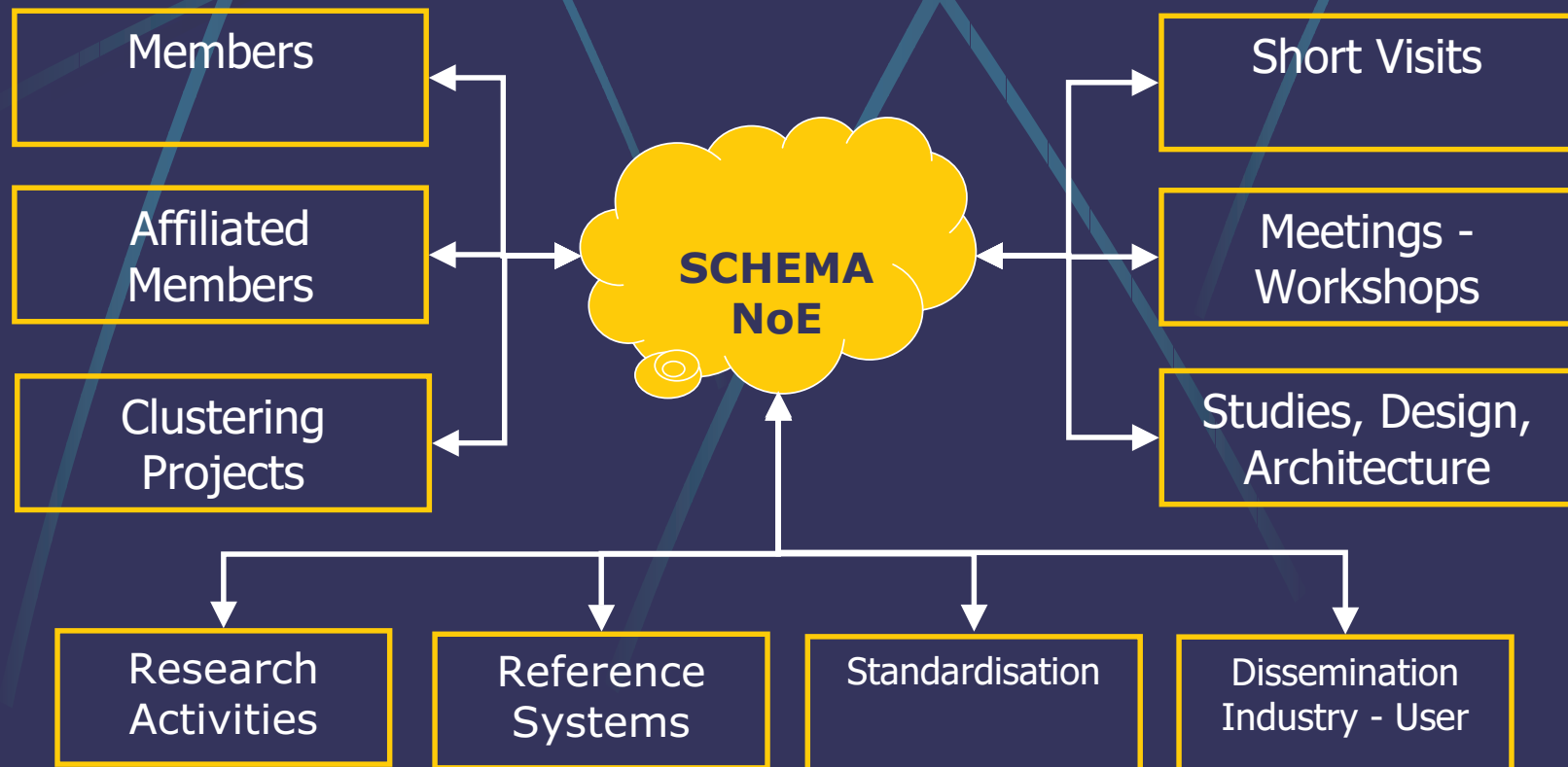


MIDDLE EAST
TECHNICAL UNIVERSITY

NATIONAL
TECHNICAL
UNIVERSITY
OF ATHENS



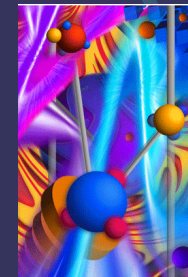
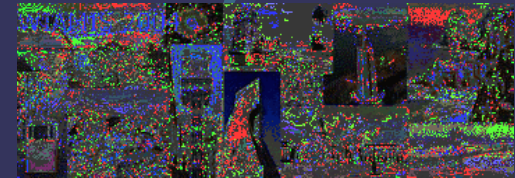
Description of Work



Networking Activities

Future events organized by SCHEMA:

- “Semantic-based Multimedia Analysis and Access”, special session in FP6 IST projects during WIAMIS 2004, April, Lisbon
 - aceMedia, MediaNet, VISNET, Direct Info, Presto Space, CHIL
- International Conference on Image and Video Retrieval, July 21-23, 2004 (CIVR2004), Dublin
- Special session, Multimedia processing and applications, 8th International Conference INFORMATION VISUALISATION, July 2004, LONDON
- 12th European Signal Processing Conference (EUSIPCO 2004), September 2004, Vienna, Austria



Thanks for your attention!

- Home page: <http://www.iti.gr/~ikom>
- Lab: <http://media.iti.gr>
- Schema: <http://www.schema-ist.org>