2001

Semantic content-based video retrieval

Lilac A. Al-Safadi
University of Wollongong

SEMANTIC CONTENT-BASED VIDEO RETRIEVAL

A thesis submitted in fulfilment of the requirements for the award of the degree

DOCTOR OF PHILOSOPHY

from

UNIVERSITY OF WOLLONGONG

by

Lilac A. Al-Safadi

MSc, University of Wollongong

Department of Information Technology and Computer Science

2001

To God be the Glory

Dedicated to my family

for their patience & support

DECLARATION

I, Lilac Al-Safadi, declare that this thesis, submitted in partial fulfillment of the requirements for the award of Doctor of Philosophy, in the Department of Informatics and Computer Science, University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Lilac Al-Safadi

ABSTRACT

The development in multimedia technology has brought the use of video documents to personal computers. The increased volume of multimedia data available in everyday lives has dramatically adopted these technologies for storing that multimedia data. Now these everyday live environments demand sophisticated systems for management and effective systems for the search and retrieval of multimedia data.

This thesis presents a semantic content-based video retrieval system. This work focuses on the semantic content of video documents and describes the implementation of a semantic-based video indexing and retrieval system suitable for the video-on-demand style applications.

This thesis addresses issues related to developing a model for describing the semantic content of a video document and representing information *about* this content. It develops a sophisticated semantic video model that expresses the underlying semantic structure of a video document and retrieves video clips among different levels of details. The proposed semantic model is an extension of the traditional conceptual model which will be applied to the video domain. The semantic video model describes how the metadata can be represented. The *metadata* contain information on the semantic video structure, the high-level semantics composition of elementary semantic units, and the video content indexing and storage. The proposed model divides a video document based on its semantic content into a structure of story, events, activities and objects with interrelationships in the various spaces in the video (time, space, context and structure).

Semantic content-based video retrieval demands human and machine understanding of video content. This thesis investigates and suggests a methodology suitable for integrating manual human understanding and automatic machine understanding technologies of video documents. A computer-aided semantic video analyzer, which utilizes the processing techniques for semantic video acquisition, is simulated.

This thesis proposes a video query language based on the first order logic for querying video information, and a design and an implementation for video retrieval. This language will provide operations for utilizing compositional data, description, and contextual, spatial and temporal relationships in the user's queries. This thesis also introduces a graphical conceptual model to describe the relations among semantic units constituting a composite unit which is a step toward an easy-to-grasp graphical user interface.

The results of this thesis lead to the conclusion that:

- A video document has a *rich* internal semantic structure that can be formally expressed and used for semantic content-based video retrieval.
- It is possible to construct a semantic based video indexing system and a computer-aided analyzer to assist in semantic video analysis and acquisition.
- It is possible to retrieve video documents based on their semantic content.

The author considers this work a step toward making video documents searchable as text.

ACKNOWLEDGMENTS

My greatest gratitude goes to God Almighty who has overwhelmed me with His Graces all through my life. Without His will and blessings this work would have not come into existence.

Special thanks to my advisor, Dr. Janusz Getta, whose help, kindness, wise advice and humane understanding made this accomplishment at last possible.

The project would not have been started if I had not received a scholarship from the Kingdom of Saudi Arabia government and an acceptance from the University of Wollongong to conduct studies in multimedia databases supported by ORACLE co.

I am indebted to my family, especially to my dad, who has been a constant source of encouragement and support during the difficult times of this work. My family has helped me understand the real values of life.

My thanks also go to the referees whom have wisely read my thesis and critically examined it. Their comments are highly appreciated and greatly taken into consideration.

Finally, I owe thanks to all those whom directly or indirectly contributed to this work.

CONTENT

LIST OF FIGURES

1

INTRODUCTION

Semantic content-based video retrieval is the selection of a sequence of frames from a collection of video documents on the basis of the content description of these frames represented by semantic units, descriptions and associations. For example, 'get me a video clip of a police chasing a man'. This chapter introduces the semantic content-based video retrieval and ends with a statement of the main problems and the solution strategy.

1.1 Semantic Content-Based Video Retrieval

Advances of multimedia technologies have enabled the electronic processing of information to be recorded in formats that are different from the standard text format. These include image, audio and video formats. The video format is a rich and expressive form of media used in many areas of everyday life, such as in education, medicine and engineering. The expressiveness of video documents will be the main reason for their domination in future information systems. Therefore, effective and efficient access to video information that supports video-based applications has become an important field for researcher. This has led to the development of, for example, new digitizing and compression tools and technologies, video data models and query languages, and video data management systems and video analyzers. With applications of a vast amount of stored video data, such as news archives and digital television, video retrieval has become an active area of research.

Retrieving video clips

Current video retrieval systems, such as those used in libraries and news archives, return a whole video document by means of search criteria. In video retrieval, it may not be suffiecient to know that a video document contains a given piece of information; it is also important to return the *parts* extracted from many video documents that contain the required information. For example, in searching the news archives, a user may be interested in a video clip where the president is making a speech about the peace process but not the whole event.

The traditional way of retrieving video clips

The traditional way of searching for a part of a video based on search criteria, lacks expressiveness and precision. The user starts with searching the video database for a video document that contains the specified search criteria. The process results in making a reference to the matching video document. Users view the video document sequentially to locate the required clip. This approach is imprecise, time consuming and inefficient in applications with a vast amount of video data.

Content-based video retrieval

A number of approaches are currently in use for determining search criteria for retrieving digital video clips. These approaches are based on:

- *Media description*, such as type, format and compression techniques. For example, 'get me all video clips stored in MPEG format';

- *Content classification*, such as a user's level of expertise and program category. For example, 'get me video clips of romance type';

- *Subjective description*, such as keywords, title and producer. For example, 'get me video clips produced by Warner Brothers';

- *Technical description*, such as length, recording speed, frame number and time. For example, 'get me video clips of 130 frames';

- *Content description*, such as casts and their descriptions, actions and relationships. For example, 'get video clips of a car chase'.

In searching a document, end users think of ideas contained in the document rather than its title or its technical details. Users surfing the Internet often search web sites using keyword contained in the required site. Hence, to make a video document searchable as text and web sites, we must focus our attention on the video content rather than on titles or attributes irrelevant to the content. *Content-based video retrieval* is characterized by the ability of the system to retrieve a video clip from a collection of video documents based on the content rather than on attributes irrelevant to the content.

Semantic content-based retrieval

> *" ... what distinguishes one movie from another is the sequence of the events, the story, but not necessarily the sequence of color histograms or edge maps"*
>
> *Dimitrova (1995)*

The world of database and computer technologies is becoming more human-oriented. *Human-oriented video retrieval* is a retrieval system that is based on the way a human views a video document, extracts and addresses its content, and builds a mental model to describe the video content in order to comprehend it. Humans tend to address a video on the basis of meanings or its semantics. During retrieval, humans seek to find information in response to spontaneous worded requests. This information, in a way, meets their perception of the content of a video document. Hence, the new trend of video retrieval systems aims at retrieving video clips on the basis of semantic content, which is referred to as *semantic content-based video retrieval.*

1.2 Applications of Semantic Content-Based Retrieval

With the explosion of video information and the advancement of storage and digital technologies, semantic content-based video retrieval can be used in a number of areas. The following are some examples:

- Movies, concerts, TV programs or other events delivered on demand;
- News on demand: retrieving and watching items from news archives;
- Biomedical applications: searching organs and pathologies;
- Security films where the investigator looks through archives for an event;
- Education: searching digital libraries, museums and art galleries;
- Shopping for specific products by description;
- Geographical information systems: searching a territory by name or population, streets and maps;
- Structure, interior design, real estates, etc.

As the Internet is typically oriented toward delivery of digital video over public Internet pages, and with the increasing availability of DSL and high-speed connections, there are many innovative uses for Internet-based delivery of video, providing powerful and efficient search engines which answers the user needs.

1.3 The Problem

Video data provide users with a wealth of information. This information needs to be addressed by the machines in order to facilitate retrieval. Much work research in image processing has been devoted to understanding and analyzing the *perceptual* content of a video document. However, the fact that such work has succeeded in extracting perceptual video content does not mean that it offers much help in the semantic-based video retrieval, since the *semantic* content has been ignored in the analysis. This has fed to the negligence of a vast amount of semantic content which was left to be on the whole merely addressed by human perception and expected to take place in users' queries. Consequently, current content-based video retrieval systems based on processing techniques may not fully meet needs or answer queries. The conclusion is that technologies are needed for video documents to support content-based searching and retrieval of video information, and overcome the limitations of processing techniques.

1.4 Strategy of Solution

Video data are a complex, unstructured media type. A great deal of effort has been put into multimedia retrieval. Yet, the main question that needs to be asked is how a video retrieval system can be developed if a video document is not understood. It becomes evident here that a rich model to represent the different aspects of the information contained is needed.

To develop a semantic content-based video retrieval system, it is necessary to follow this procedure:

- Developing a formal description for semantic video content;
- Setting indexes that are efficient in terms of storage and search time, conforming to the human perspective, and addressing as much information as possible in a video document;
- Studying the capability of current signal processors and the method of integration with the proposed semantic model to maximize procedures that can be automatically conducted;
- Developing an efficient structure for the semantic video acquisition and retrieval in light of the proposed semantic model;
- Designing querying methods for video documents that meet human needs;
- Eliminating semantic and schematic heterogeneity between query content and video content.

1.5 Outline of the Thesis

This thesis is divided into seven parts.

- Chapter 1 **Introduction**

  This chapter provides a brief definition of the semantic content-based video retrieval, and states the main problems and the strategy for a solution.

- Chapter 2 **Current Works in Content-Based Video Retrieval**

  This chapter lists current approaches to video indexing and related works.

- Chapter 3 **Content-Based Video Retrieval System Architecture**

  This chapter describes the overall architecture for the recommended solution. The main components of a semantic content-based video retrieval are repositories, the semantic video model, semantic video acquisition and retrieval.

- Chapter 4 **Semantic Video Model**

  This chapter describes the suggested approach towards semantic video structuring and how this model is to be organized and stored in databases. A graphical conceptual model is proposed for representing video content interrelationship.

- Chapter 5 **Semantic Video Acquisition**

  This chapter suggests a semantic computer-aided analyzer as a tool for the user to insert video documents into the database and annotates their semantic content. Moreover, this chapter explains how this thesis can utilize the current processing techniques to serve the acquisition of video semantic.

- Chapter 6 **Semantic Video Retrieval**

  This chapter proposes a structure for retrieving video documents stored in databases based on their semantic content. This chapter introduces a query language to show the possible queries that could be answered by the proposed retrieval system.

- Chapter 7 **Semantic Heterogeneity**

  This chapter discusses the possible heterogeneity between the user model and the defined semantic video model, and studies the possibility of eliminating that semantic heterogeneity.

Three indexes are attached to this thesis for further readings

- Appendix A **Published Papers**

  Lists published works related to this thesis.

- Appendix B **Review of IMAQ Vision**

  A signal processing technique studied and tested in the laboratory.

- Appendix C **Review of MPEG-7**

  The multimedia content description interfaces standard.

- Appendix D **Mapping the Semantic video Model into Relational Databases**

  Represents the proposed semantic model in relational databases.

2 CURRENT WORKS IN CONTENT-BASED VIDEO RETRIEVAL

This chapter documents major approaches followed during video content-based indexing highlighting the approach adopted in this thesis. Also, incorporates a review of related works into discussion.

2.1 Approaches in Content-Based Video Retrieval

Video documents contain two categories of content: perceptual and semantic.

- *Perceptual* content, sometimes referred to as low-level content, is what is seen and heard which is represented visually by visual features, such as pixels, colors, texture and shape; aurally by audio features, such as loudness, pitches, brightness and frequencies, and textually by alphabets and symbols;

- *Semantic* content is the *meaning* of what has been seen or heard conveyed by the perceptual content.

Throughout this work, *content* will refer to both the perceptual and the semantic content in a video unless specified otherwise.

Consecutively, the following two main approaches have been followed in content-based video retrieval:

- *Perceptual-based* retrieval utilizes processing techniques in matching the video database indexes and content of queries, both represented in terms of visual or aural samples or features;

- *Semantic-based* retrieval searches the video database for meanings similar to those occurring in the user's query.

2.2 Perceptual Content-Based Retrieval

Most of the researches in content-based retrieval are based on the perceptual content of a video. In CONIVAS (Abdel-Mottaleb, Dimitrova, Desia and Martino; 1996), EXCALIBUR technology for searching and retrieving images and videos (http://www.excalib.com), JAISR (Iannizzotto, Puliafito and Vita; 1997), JACOB (http://wwwcsai.diepa.unipa.it), QBIC (Flickner et al.; 1995), VIRAGE technique for searching video documents (http://www.virage.com), VIMSYS (Yeung, Yeo and Liu; 1996) and WebSeek (Smith and Chang; 1996), users retrieve video clips based on contained colors, textures, shapes and sketches. VIOLONE (Yoshitaka, Hosoda, Yoshimitsu and Ichikawa; 1996) enables the users to retrieve an object's motion by a drawn example. One of current commercial perceptual content-based retrieval systems widely used by major TV networks in the United States of America is Virage.

Virage

Virage provides owners of video content with the end-to-end solution for publishing, managing and distributing video assets over the Internet. The Virage platform allows content owners to perform the following:

1. Index and encode

   Virage allows simultaneous, automatic encoding and indexing in real time. Virage contains a number of media analysis software plug-ins. These plug-ins allow content owners to enhance automatic indexing capabilities. Plug-ins include: face and on-screen text recognition, and audio recognition, which identifies spoken words, speaker names and audio types.

   With automatic, real-time recognition of faces and text in the video content, users no longer have to manually enter the name of a political figure, celebrity, corporate executive or other person who appears in a scene. Nor do they need to stop for important textual information such as anchor names, sport scores and product information.

   Virage audio plug-ins automatically transforms the video's audio content into searchable text in real time. By intelligently "listening" to the audio track, it identifies spoken words, speaker names and audio types, virtually eliminating the expensive and labor intensive manual annotation process traditionally used to log video.

2. Manage, share, publish and distribute

   Virage provides a web-based interface that gives enterprises an easy and efficient way to manage and administer video assets. Virage provides developers and systems integrators access to the full range of server functions for custom integration of video to suit any online publishing environment such as video libraries, content syndication, video-enhanced corporate training and e-commerce.

3. Synchronize, assemble and present video and PowerPoint slides

   Virage provides the fully integrated, end-to-end solution for rapidly assembling, synchronizing and publishing streaming video with PowerPoint to website and audience.

Limitations of perceptual content-based retrieval

Perceptual content-based retrieval has a number of limitations as under:

- Processing techniques, which are essential in the perceptual analysis, are still immature, and contain problems, illustrated later in chapter 5, that have not yet been fully solved;
- End users are, mostly, not interested in *how* a video sounds or looks but in *what* a clip is about. For instance, instead of posing a query 'brown triangular object' or 'an image that looks like this sample', end users are most likely to ask for a 'mountain';
- Similar views may sometimes give different semantics. The view of a person carrying a book could refer to a student or a teacher;
- There are elements of meanings beyond the perceptual level as for instance, generalized and specialized concepts such as mammal or postgraduate student, classification such as a particular kind of bone, and subjective information such as the video title and the cast name;
- The perceptual content-based retrieval provides not only both exact and similar matches on the perceptual level but also deals with a precise conceptual entry, while users often have unclear descriptions of their own needs or may seek conceptually *related* clips;
- A search in video databases can be computationally intensive, and requires large storage;

- For large video databases, end users do not always have visual or audio samples to drive the search, which is essential in the query-by-example types of perceptual-based retrieval, such as QBIC (Flickner et al.; 1995); and
- The perceptual-based approach does not address the semantics implied in a video.

While a great deal of effort has been undertaken in the perceptual-based approach, making it efficient in some applications, such as medical images, still relatively little has been claimed in the semantic-based area. The semantic-based retrieval of video documents, ignored by many researchers mainly because it is based on manual annotation, is believed to be imprecise and impractical in the application field (Dimitrova; 1995), whereas the perceptual retrieval has the advantage of automating the video analysis. In semantic-based retrieval, *annotation* is a process of assigning semantics to video content. Having a standard data model for video documents, which is a trend with the MPEG-7 standard, could be an essential step towards automating the video analysis and the annotation process in the semantic retrieval. *MPEG-7* (Nack & Lindsay; 1999) is a common high-level description language for multimedia documents. It is hoped that this work can possibly contribute to the future of MPEG-7 video indexing standards.

2.3 Semantic Content-Based Retrieval

Semantically, videos are a type of unstructured media. Some semantic-based retrieval systems have focused on specific well-structured application domains, such as television news (Swabnerg, Shu and Jain; 1992, and Zhang, Tan and Smoliar; 1995) and sports (Sudhir, Lee and Jain; 1998, Saur, Tan, Kulkarni and Ramadge; 1997, and Yow, Yeo, Yeung and Liu; 1995). These approaches have succeeded - to some extent - in automating semantic video analysis though with limited query capabilities. Unstructured application domains are so far manually indexed.

Current works in semantic content-based retrieval

In order to index a video, a video document need to be logically segmented. The *segmentation* process partitions the video stream into segments and assigns annotations to each segment. The notion of *stratification* was proposed (Davenport, Smith and Pincever; 1991) where layered information is used to describe the cinematic content other than the traditional segmentation process. This work structures video media into physical elements (shots) and a hierarchy of logical elements of scenes, sequence, and segments. Stratification has been adopted later by many workers.

- OVID (Object-oriented Video Information) (Oomoto & Tanaka; 1993) is an object-oriented data model for video retrieval. Video objects in OVID corresponds to sets of arbitrary portions of time sequential (video frame sequences). Each video object has a set of attributes and a unique identifier. OVID allows the sharing of a common description among video objects. Unlike the main objective of the work presented in this thesis, OVID's video data model does not explicitly support modeling of the video document structure.

  OVID provides the user with an SQL-based query language VideoSQL which gives the user the ability to retrieve video objects by specifying some attribute values. OVID does not consider the interrelationship between video data. Hence, VideoSQL does not contain language expressions for specifying relations between video objects, not even temporal relations which are considered an important aspect of video contents. One of the key advantages of the VideoSQL is that it can allow the user to specify how many frames he would like to see in a presentation.

  OVID suggests manual keyword annotations and textual description for the video based on a generalization hierarchy. An important aspect of this work illustrates how related semantic entities form high-level concepts. However, the OVID has no schema. And semantic units are limited to objects.

- VideoStar (Hjelsvold & Midtstraum; 1994) is a generic data model for capturing video content and structure based on stratification. The model is built upon an enhanced Entity-Relationship (ER) model and includes video as a data type in relational databases.

  The structure part of the video model is a hierarchy of Shots (frames recorded contiguously, representing as continous action in time and space), Scenes (containing shots which are related in time ans space), Sequences (containg scenes which together give a meaning) and CompoundUnits.

  VideoStar allows the user to define temporal relationship between frame sequences. Additionally, one of the innovations in the approach presented in this thesis is the consideration of other relationahips that may exist in a video document rather than temporal. The proposed model is quite complex and no logical elements have been formally identified. Even the basic element of the hierarchy structure (action) is not defined. The semantic indexing was only based on annotations that gives a description for the content of a frame.

- VODM (Video Object Description Model) (Change, Lin and Lee; 1995) is an ER model for the database conceptual level organization. Entities in VODM are defined as a sequence of frames referred to as video objects. Another basic element is relationship, which is an association between objects. The representation of a video object and relationships can be described by attributes of different data types, such as keywords, paragraphs, images and other objects. VODM is considered one of the leading works to take into consideration attributed of relationships. A free-text annotation mechanism has been used.

A two-step query processing method is introduced that can reduce the processing time of each video object query. First step finds objects for each query selection condition, and the second step search decriptive elements. The author adopt this two-step query processing and extends to serve the model proposed and illustrated later in this thesis.

- CVOT (Common Video Object Model) (Li, Goralwalla, Özsu and Szafron; 1996) is capable of automatic video segmentation and incorporates temporal relationships among video objects. The only semantic units considered in this work are frame-based objects. The video is structured into clips and frames. The basic idea of the model is simple and aims at finding all common objects among clips and at grouping clips according to contained objects with only temporal relationships taken into consideration.

- VIRON (Video Information Retrieval On Notation) (Kim, Kim and Kim; 1996) is a video data model that shares and reuses annotations. Annotated video units (objects) are mapped into a unified video annotation system. Objects are used to refer to video segments and textual annotations are used for the objects' description. Like all works already reviewed in this section, the semantic structure of the video document is so simple and does not suite a medium with high semantic complexity like videos. It fails to provide rich description of video contents, which is the original contribution of this thesis.

- VideoText (Jiang, Montesi and Elmagarmid; 1997) is a simple semantic video model based on logical video segment used in layering, video annotations and associations between them. The logical segment is represented by a varying number of frames. The indexing is based on free text annotations rather than a fixed set of keywords. The model is generic and no formal structure has been defined.

  VideoText allows querying based on the temporal and interval relationship between annotated logical video segments. Results are ranked based on their relevence to the semantic content of the video data.

- VIDAM (VIdeo DAta Model) (Srinivasan & Riessen; 1997) is a video data model that represents concepts in a video as semantic objects and spatio-temporal information as structural objects. Objects are defined as any description of catalogue, segment, and what is seen and heard. No formal definitions of semantic objects, relationships or description schema has been defined. The system is based on manual notations of keywords.

- CAETI IML (Computer Assisted Education & Training Initiative/ Internet Multimedia Liberary) (Yu & Wolf; 1997) is an automatic video library retrieval. It supports the subject-based retrieval that allows that system to be retrieved by visual objects. It is well suited for extracting visual content which can be matched with information.

  CAETI IML classifies the video key frame using neural network algorithms that utilize color, shape and texture features. The classification index is an organized set of terms which corresponds to visual objects. Only predefined objects can be captured. Resulting in a set of tags describing the key frame. The system accepts object-based queries and only returns key frames which contain objects in question.

  CAETI IML is a frame-based retrieval system. Therefore, it is well suited for images but not video documents where information is spread over a sequence of frames. Also, because it purely searches by key frame contents, the only semantic unit to be extracted from frames are objects and no relationships among those objects are addressed by this work not even in space.

  Although this system provides a fully and precise automated object retrieval, it is still simple and convey little about the semantic content of video doucments. This work seeks to define a rich and a powerful model enough to describe the semantic content of video documents.

- ToC (The table of content) (Zhuang, Rui, Huang and Mehrotra; 1998) enables users to query based on both the visual content and subjective keywords. It presents a semantic-level ToC construction. It concentrates on videos having story lines. Video stream is structured into a hierarchy of video, scene, group, shot and key frame.

  ToC consideres scene as a semantic entity that conveys the smenatic meaning of the video to the viewers. In the model proposed and described later in this thesis, scene is a collection of partially ordered semantic units appearing within the same context and represents no meaningful unit. The work proposes an intelligent unsupervised clustering technique to perform scene structure construction.

  Group is an intermediate entity between the physical shots and semantic scenes. ToC proposes an approach for creating groups to facilitate scene construction based on visual similarity and time locality.

  The aim of this work is to avoid one of the major limitation of semantic analysis which is the manual annotations and maximize the use of procedures that can be automatically conducted.

  Although ToC claims providing smeantic structure based on the video story line, it still convery a little about the semantic structure and semantic entities of the video. Semantics are limited to scenes only with no formal semantic underlying structure of a scene provided.

- UVRS (Hee, IK and Kim; 1999) is a Unified Video Retrieval System that provides content-based query integrating feature-based queries and annotation-based queries of indefinitly formed video data. UVRS segments video document into documents, sequences, scenes and objects. Each of them is considered a unit for retrieval. The only semantic entity is object.

  UVRS suggests three layered Hybrid Object-oriented Metadata Model which is composed of the raw-data layer for the physical video stream, the metadata layer to support the annotation-based retrieval, content-based retrieval, and similarity retrieval and the semantic layer to reform the query.

  Retrieval conditions are based on attributes, color, spatial-temporal relations between objects and similarity. This work does not include the video segment process and video indexing. The main objective of this work is the video query process.

- (Decleir, Hacid and Kouloumdjian; 1998, 1999) presents a simple generic data model and a rule-based query language for content-based video access. This model allows user-defined attributes as well as explicit relations between objects. This model is based on the notion of objects of interest that can be annotated using attributes. Objects can be linked together by means of explicit relation names. The different types of relationships are not distinguished in this model and keyword annotations and descriptions are assigned to objects.

- (Day, Khokhar, Dagtas and Ghafoor; 1999) proposes a multi-level abstraction mechanism for capturing the spatial and temporal semantics associated with various objects in video frames. At the finest level of granularity, video data can be indexed based on mere appearance of objects and faces. At higher levels of indexing of events, an object-oriented paradigm is proposed which is capable of supporting domain-specified views.

- AVIS (Advanced Video Information System) (Adali, Candan, Chen, Erol and Subrahmanian; 1996), a work close to the system presented later in this thesis, that studies methods of indexing video databases so as to store and retrieve video data efficiently in of diverse ways. AVIS is a semantic content-based retrieval system that has been designed in the University of Maryland which structures the video document into objects and activities, and provides an elegant way of storing data. In addition, the primary contributions of this work are following:

    - Shows that the problem of storing objects occurring in certain frames may be viewed as a problem equivalent to that of storing line segments.
    - Shows how a combination of spatial database technology and relational database technology can be merged to solve user queries efficiently.

- Describes how updates to information about video data can be implemented efficiently with the data structures.
- Describes a prototype implementation of the data structures and algorithms.

In AVIS model, two types of information are being queried: a set of *entities* – things of interest to us in the movie, and the video frames in which these entities are present. Three types of entities are listed in the proposed model: video objects, activities and events.

- *Video objects* are the entities present in the video frames such as `Philip`. Video objects are media-independent and may be invisible, but nonetheless present.
- *Activity* describes the subject of a given frame sequence, such as `murder`. Multiple activities may simultaneously occur in a video clip.
- *Event* is an instantiation of activity, for instance, `opening the chest` may refer to two separate events – Philip opening the chest and Mr. Wilson opening the chest. the activity types are general groups containing many events, and they will be stored implicitly in the form of a set of the same activity type.

- *Role* is the description of certain aspects of an activity. For example, `victim` and `murdered` are roles in the activity `murder`.
- *Team* is a set of descriptions that jointly describe an event. For instance, the event `murder` involves a team consists of David in the role of `victim` and Philip in the role of `murdered`.

Unlike the model presented later in this thesis, AVIS captures no high-level semantics for a detailed description of the movie, no description for video entities and provides no formal definition for video entities. Also, AVIS does not consider the interrelationship between video entities except for those given implicitly through the description of events and manually annotates detailed information.

The presented data structure facilitates the execution of various types of queries: elementary object, elementary activity-type, event, object-occurrence and conjunctive queries. No relational queries answered by the proposed system.

AVIS develops algorithms for updating video databases using the data structure defined. This includes the insertion and deletion of an entity into the database, the insertion and deletion of a set of frame sequence for an object, and the insertion and deletion of a set of frame sequence for an event.

The implementation of AVIS shows that the proposed video database can be stored electronically, and furthermore, they have designed query processing algorithms that traverse these data structure. Methods for updating video databases have been implemented.

2.4 Concept-Based versus Keyword-Based Match

These works are based on keywords match. Keywords match is imprecise and expects users to be aware of the annotations stored in the video database. This is not always true. Often, the information seeker fails to find what is wanted because the words used in the request are different from those stored in databases. Besides, a complete and precise keyword-based video description is impossible. To overcome the limitations of keyword match, many works have addressed concept-based match. These works include Croft & Thompson; 1987, Tong, Appelbaum, Askman and Cunningham; 1987, Djeraba, Bouet and Briand; 1998, Koh, Lee and Chen; 1999, Ambroziak, J. and Woods, W.; January 1999, Sun Microsystems; January 2000, and Wang, Chua and Al-Hawamdeh; 1992.

(Chua, Pung, Lu and Jong; 1994) describes the use of a concept model of the image collection as the basis to guide the retrieval and updating of image content. The system uses concept terms in image indexing and concept-based search engine for accurate retrieval.

Knowledge-base is used in (Yoshitaka, Kishida, Hirakawa and Ichikawa; 1994) as an aid in video retrieval. An object-oriented data model and a query language are proposed for content-based retrieval. The database schema is represented through a hierarchy of *is-a* and *part-of* relationships among classes. A class is associated with domain knowledge to represent a certain concept.

(Smoliar & Zhang; 1994) utilize a frame-base knowledge base to support content-based video retrieval. Slot's type knowledge is translated into knowledge of how to search it for retrieval purpose. The system is based on manual textual annotation. It essentially utilizes the spatial information for indexing the representative frames and ignores the temporal information in the video.

MOODS (Griffioen, Yavatkar and Adams; 1996) integrates an enhanced object-oriented data model, multimedia database and a dynamic semantic information extraction engine. Each semantic object is coupled with a database that stores information about it. A processing engine with semantic inference rules is supported in the system to express high-level semantic concepts. The only semantic unit supported in the model includes objects with a set of description identifying semantic concepts and entities. Relationships between objects are neglected.

(Amato, Mainetto and Savino; 1998) presents an object-oriented multimedia data model for content-based retrieval of multimedia objects (basic and complex). Each object is represented by the values of its physical feature and semantic content represented in terms of concepts. An open set of features and concepts can be defined in the model, where each concept can be extracted through the use of feature values and background knowledge.

(Koh, Lee and Chen; 1999) propose a five-level layered semantic model for video data: frame, chunk, sequence, scene and video level. A uniform semantic representation is proposed to represent the semantic data level. The approach is coordinated with a concept knowledge database where, for each semantic unit, a concept vector with a semantic degree of concepts is used to present the implied semantics. In this system, semantic items (objects and events) are identified manually and relationships are neglected. Semantic items appear in chunks and are semantically represented. An event is not clearly defined. It is used to represent an action or any concept that appears on consecutive frames. The model assumes a concept that may appear a number of times in a video but does not consider concept descriptions.

(Liou, Hjelsvold, Depommier and Hsu; 1999) includes tools for extracting structure information from video, interfaces for integrated multimedia logging, and tools for content-based query. The system segments the video stream into shots. It automatically generates a video table of content to facilitate the manual augmentation of multimedia descriptions while allowing for correction and verification by motion, trying to capture the temporal information inherent in videos. These descriptions are managed through the establishment of structured thesauri, thus ensuring the integrity of the database.

WordNet (Miller, Beckwith, Fellbaum, Gross and Miller; 1990, and Miller; 1995) is an electronic lexical system developed at Princeton University. WordNet lexical database expands concepts by indicating synonyms, hypernyms or hypinyms of the original searched concept. A number of works adopted WordNet in their models, such as SCORE and TOC.

SCORE (System for Content Based Retrieval of pictures) (Aslandogan, Their, Yu, Liu and Nair; 1997) presents techniques for improving retrieval effectiveness based on semantic content of images. The system uses an extended ER model to represent image content. SCORE uses WordNet to expand both user queries and metadata associated with the images stored. The result of experiments indicates that specific uses of an electronic thesaurus can provide significant improvement over the non-utilization of.

TOC (The table of content) (Zhuang, Rui, Huang and Mehrotra; 1998) enables users to query based on both the visual content and subjective keywords. ToC has been discussed previously in section 2.3. Searched keywords extracted from close-captions are posed against WordNet for keywords not found in the video database.

2.5 Dynamic Objects and Motion

Previous works do not address complex semantic units or high-level semantics implicit in video documents. Most of the works are limited to static objects and only a little progress has been achieved on indexing actions or activities represented by motion. Yet, the author believes (as will be illustrated later) that there are more semantics in a video beyond objects and actions. Among the works conducted on the motion of objects there is AVI (Automatic Video Indexing) (Courtney, J.; 1996). AVI performs automatic content-based video indexing from object's motions to assist human analysis of digital video data. The indexing method proceeds in three stages: motion segmentation, object tracking and motion analysis. Users may select an object and the AVI returns all video clips where the object is involved in specified actions.

(Lee & Koa; 1993) develops a mechanism and a prototype for indexing video data based on the concept of objects and object motion with interactive annotation. A motion representation for the track of a moving object is presented.

2.6 Video Multi-media Content

A video data, composed of several contextually related streams; including visual, speech, non-speech and textual. Information related to video documents, is extracted from all streams simultaneously and delivered in such a way that describes the content. Most current video retrieval deals with the visual stream only and ignores the information presented in other media streams (speech, non-speech and textual). A number of works have been conducted considering the various media streams in a video, such as the Informedia project (Smith & Christel; 1995, Wactlar, Kanade, Smith and Stevens; 1996, Wactlar et al.; 2000, Nakamura & Kanade; 1997, BNATM (Maybury, Merlino and Rayson; 1997), Hauptmann & Witbrock; 1998, VISION (Gauch, Gauch, Bouix, and Zhu; 1999), MAESTRO; 2000 August, and Boykin & Merlino; 2000).

2.7 Summary

This chapter discusses current trends in content-based video retrieval and outlines the approach to be considered during the video indexing and the developing of a content-based video retrieval system. A review of related works has been presented and discussed.

3
CONTENT-BASED VIDEO RETRIEVAL SYSTEM ARCHITECTURE

This chapter aims at providing an overview of the architecture of the semantic content-based video retrieval system and describing its main components. An approach for video content representation is discussed at the end of this chapter.

3.1 The Architecture of Semantic Content-Based Video Retrieval System

Figure 1 shows the architecture of the semantic content-based retrieval system framework proposed in this thesis for indexing and retrieving video clips. While listing the main components, the figure illustrates how the different components may share a common base of video document and their metadata, and the data flow from one component to the other. Semantic content-based video retrieval system is an integration of four main components: semantic video acquisition, semantic video retrieval, semantic video model and repositories.

Semantic video acquisition is the process of analyzing a video document, and detecting and extracting its content. *Semantic video retrieval* is the process of accepting the user's requests, processing and returning a set of matching video clips. *Semantic video modeling* is the process of developing a semantic structure for video documents and indexes for storing their content. *Repositories* store video documents and semantic video content for retrieval.

A video retrieval system is as good as the indexing system defined by the semantic model. This thesis aims at studying a method for representing semantic video content. Based on the proposed semantic model, this thesis investigates the content acquisition and proposes a retrieval system. As illustrated in Figure 1, semantic video acquisition and retrieval are built on the top of the semantic video model, which is defined on top of data repositories. Semantic content-based video retrieval system framework provides an interface to repositories through annotation and retrieval components.

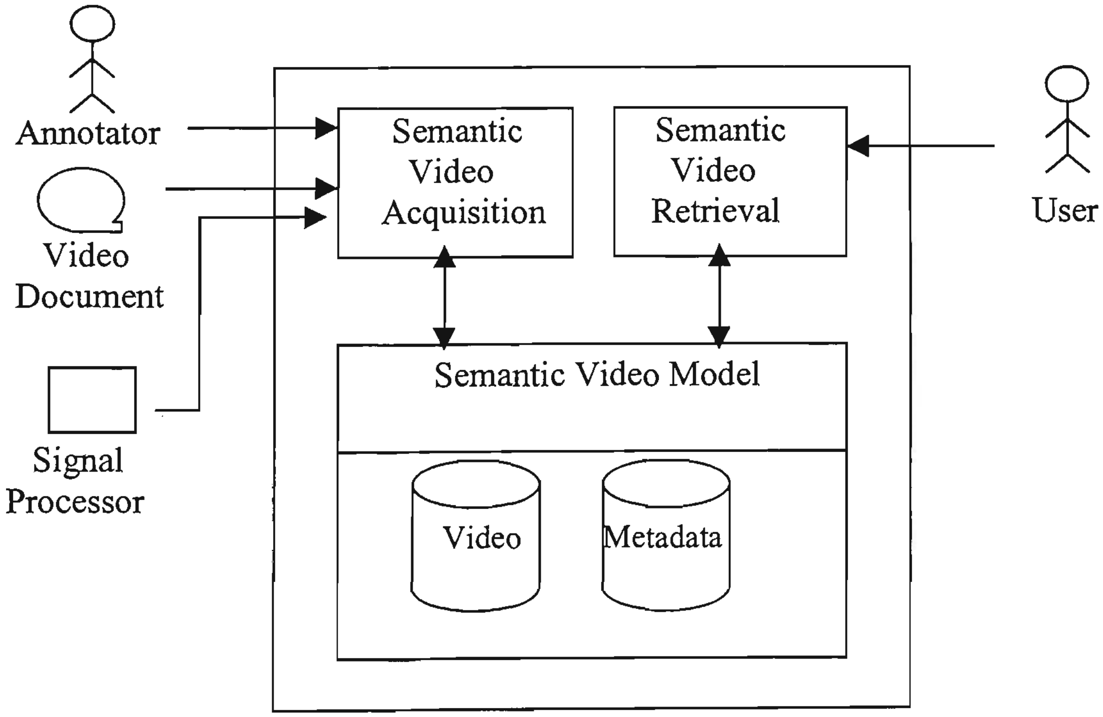


Figure 1. Semantic content-based video retrieval system main components

3.2 Semantic Video Acquisition

A semantic video acquisition serves as a tool for accepting a video document, extracting its content, creating semantic indexes and assigning semantic descriptions to its video clips. The semantic video acquisition proposed in this thesis is a semi-automatic process. It is based on the human analysis approach of video documents and it aims at utilizing works that have been carried out so far in signal processing. This should maximize the use of procedures that can be automatically conducted to serve the acquisition process. The semantic video acquisition will also provide an interface for viewing signal processor's output, entering semantic information and storing the information in the meta-database. Semantic video acquisition is described in detail in chapter 5.

3.3 Semantic Video Retrieval

This thesis proposes semantic content-based video retrieval component that adopts the *query* approach for video retrieval rather than the browse approach. A formal query language is presented to query the video database based on their semantic content. Interfaces and operations for creating and executing queries are implemented. These operations in the semantic video retrieval component are mainly concerned how to accept the user's query, map with the existing stored video model and retrieve similar results. Semantic video retrieval will be discussed in detail in chapter 6.

3.4 Semantic Video Model

The semantic video model describes the semantic structure of a video document, the elements stored in video repositories and the relations between data elements. In order to describe a video document that represents a real world, a formalism is needed to describe reality. So far, the conceptual model is well-developed and has provided satisfactory results in describing reality. The *conceptual model* of a document represents content in an abstract way that conforms to real world representation. It gives a description close to the way users perceive it in terms of real world objects, relationships and attributes. In semantic video modeling, the conceptual model bridges the gap between the physical video media and the user view of the video content. Until now, the conceptual model has been used mainly to describe static reality. This thesis decided to use the conceptual model in describing video content and to extend it to address dynamic reality and more sophisticated problems. The conceptual model in this work aims at being rich in its semantic capabilities and at providing a representation with various levels of detail addressing elementary as well as complex video content. This thesis proposes an extension to the traditional conceptual model applied to the video domain. The semantic video model will be discussed in more detail in chapter 4.

3.5 Semantic Video Repositories

Semantic content-based video retrieval provides an interface to two main repositories: video database and meta-database. The *video database* stores physical video documents in compression formats. The metadata in the *meta-database* stores information about video documents available in the video database. It maintains information concerning semantic video content to facilitate the semantic content-based query. The metadata allows the examination of the content of the video database without retrieving the actual data. This actual data retrieval usually results in an expensive computation and semantically insufficient results. During retrieval, queries are posed against the meta-database rather than against the video database. The structure of the meta-database will be defined in chapter 4.

3.6 Summary

This chapter presents an outline of the main components required for developing the content-based video retrieval system proposed in this thesis. This chapter is considered an introduction to the rest of this thesis. A formal approach for conceptual model, has been adopted for video content representation.

4
SEMANTIC VIDEO MODEL

The first step toward developing a semantic content-based video retrieval system is the development of a formal semantic modeling of video content description. This chapter aims at providing an elaborate semantic model to describe the semantic content of a video. This model addresses the semantic structure, the high-level semantics composition, and the video content indexing and storage.

4.1 User View of a Video Document

> *"Indexing is an idiosyncratic affair: One person's indexes are not another's. Humans would construct different indexes because what they pay attention to and what they have experienced are different, not because the indexing schemes differ in principle.*
>
> *Yet, we are standard enough "*
>
> *(Schan 1990)*

When considering the human nature in describing video content, a major task is to investigate how a user creates his/her own view of a video document. A *user view* of a video document is the perception of the content of the proposed video. Understanding the user view of a video document helps in deciding what aspects of the video document should be considered and stored. This will enable the model proposed to depict the user's various perspectives of a video document, which will help building a system capable of answering the user's heterogeneous queries.

The approach for generating user view

A user view can be generated through exposing a number of meaningful entities or *semantic units*. The user would reveal different perspectives depending on units and descriptions of their interest. A number of users could be watching the same video clip but are interested in different semantic units. For instance, one would be interested in the man walking and another in the moving car. Subsequently, users would refer to the video clip based on their units of interest. In other words, two different queries would be submitted: 'Find a video clip of a man walking' and 'Find a video clip of a moving car'. Moreover, users may describe a semantic unit in different ways, for instance, 'a man in the red shirt' or 'a man in the blue jeans'.

Semantic units in a video document are related to each other in the video space. The user may refer to a semantic unit based on its relationship with another, such as a video clip of 'the son-of John Kennedy' or a video clip of 'an accident under a bridge'. In a video retrieval system that does not consider relationships, when posing different queries, such as 'accident behind a bridge', 'accident under a bridge' and 'accident above a bridge', the same video clip, which contains both semantic units (accident and bridge), regardless of the relationship will be retrieved.

Therefore, the author argues that *relationships* are extremely important in semantic video modeling and should be captured between semantic units in the various video spaces. Neglecting relationships leads to inaccurate or even wrong answers as shown in the above examples.

End users often get a fuzzy understanding of their own need. *Fuzzy* needs could be expressed with a number of possible interpretations or representations. This could explain why, most of the time, Internet end users fail to find what they want using Internet search engines. In semantic content-based video retrieval, end users are unaware of the video structure and annotations stored. For instance, if an object was stored in the database and annotated as `student`, the system would fail to retrieve it when the end user asks for a `person`, while it is semantically correct. End users employ various types of abstraction to construct their own view. Therefore *abstraction* is an important mechanism for imitating the user view of video content. It associates a physical element with a real world concept. In addition, abstraction is important for generating high-level semantic units as will be elaborated later in this chapter.

The proposed semantic video model is based on the human perspective in order to have a system that could retrieve clips capable of answering human query. Hence, the semantic model based on the user view constitutes:

- Semantic units
- Associations among semantic units
- Descriptions of semantic units and associations
- Abstraction mechanisms over semantic units, descriptions and associations.

4.2 Semantic Units

A significant issue is the identification of the *meaningful* units (semantic units) in a video. The choice of a semantic unit determines the expressiveness, completeness and flexibility of the model. The objective of this section is to provide an informal description of the logical structure of video documents.

At the semantic level, a video document is an *unstructured* media type. It has no underlying semantic structure. From the physical point of view, a video is a sequence of frames (visual and aural). An *aural frame* is a set of audio parameters of an interval of 10-30 ms (Peacocke & Graf; 1990). A fundamental task in the semantic video modeling is to identify a semantic logical structure of a video document known as *video structuring*.

A video needs to have a meaningful (semantic) structure. The physical structure (frame, pixels and frequencies) and the screenplay-based logical structure (episode, scenes and shots) do not capture the underlying semantic structure of a video perceived by the end users. End users will not refer to a video in terms of pixels or scene cuts but in terms of the semantics represented by the visual and aural objects. As mentioned earlier in chapter 2 (Current Works in Content-Based Video Retrieval), most of the effort in the content-based video retrieval has gone into physical and screen-play logical structure and very little has gone into structuring the *semantic* of a video document delivering a simple structure with limited query capabilities. Most works have mainly been discussing the formal abstract structure of a video (the syntax or grammar) while not giving much attention to the actual semantic content. Hence, the model propsoed in this thesis aims at going beyond the physical or screenplay structure of a video and move towards a sophisticated video indexing based on the semantic content.

Current trends in semantic video modeling aim at addressing frame-based semantic units where the only type of semantic unit captured from a frame is *object.* Object-based semantic models are so simple and cannot express complex aspects in semantic video content. The author believes and will prove later in this section, that semantic video content is more complex than objects and that the different types of semantic units and relationships need to be distinguished in order to construct a very detailed semantic video content. The advantage of a sophisticated video model is that it captures various human perspectives and consequently answers a variety of queries at various levels of detail.

To the viewer, a video as a whole is organized in a way that tells a story. Hence, the user's semantic comprehension of a video is based on the *story-line* structure. In order for a human to understand a story, he/she must first break it down into the conceptual actions underlying the events (Schank; 1990). A *story* is a recorded sequence of *events* (Mandler & Johnson; 1977). These events involve real world *objects* and *activities* performed by them. This entails the choice of *objects* and *activities* as *elementary semantic units*.

4.3 Elementary Semantic Units

A *physical object* is an instance of a salient object captured in a video's physical space and represented visually, aurally or textually. A *semantic object* is a physical object identified by the viewer as it belongs to real world objects, such as a `car` or a `person`.

An *activity* is the interpretation of continual changes in the values of an object's observable attributes over a sequence of frames (interval of time). A *semantic activity* is an activity identified by the viewer as belonging to real world activities, such as class `walk` or `run`. An *actor* is the object performing the activity. Activity and actor are associated in a 1:1 *performed-by* relationship. A fact that an object *O* performs an activity A is represented by $A(O)$. For instance, `run(car)` and `write(author)`. Actors could be elementary objects, such as `author` as well as composite object like `team` or `group`. Composite objects will be elaborated on later in this chapter. Activities are most of the time performed by actors on objects, which we refer to as the *object* of the activity. An activity A performed by actor O on object B is represented by $A(O,B)$. For instance, `write(author, book)` and `eat(man, cake)`. The object of an activity could be elementary, such as `book` or composite, such as `food`.

4.4 Observation Slots

Each semantic unit may appear a number of times in a video or in multiple videos. Therefore, a semantic unit is associated with a video identifier (*VID*), and a pair of two numbers (t_s, t_e) representing the time in which they are valid and identified by a frame number or time in milliseconds. The triple $[VID,\ t_s,\ t_e]$ is called the *observation slot* of x and denoted by $T(x)$. The observation slot links an abstract concept of a semantic unit with a physical chunk of video document.

4.5 Associations

To represent the various interactions among semantic units within a video space, the chapter introduces the concept of *association*. A key characteristic of the video is the various relationships embedded in, and connecting, semantic units. Each semantic unit identified in a video has an entry in the real world knowledge base. For instance, `Pyramid` semantic object is associated in the real world to the city of `Cairo` in `Egypt` and to the age of `Pharaoh`.

Semantic units are interrelated in context, semantic structure, space and time. This indicates four types of semantic associations: contextual, structural, spatial and temporal. Like semantic units, associations are attached with observation slots.

1. *Contextual association* is an n-ary relationship between *n* semantic units in context. For instance, a contextual connection, such as in 'X *father-of* Y' may exist between two semantic units of class `person`. Contextual association is denoted by $R(A_1, \ldots, A_n)$ where A_i is a semantic units and R is an association name such as `father-of` and `friend-of`.

2. *Structural association* is a binary association between instances of semantic units in composition structure. For instance in 'conference speech', the order of semantic units indicates an implicit structural relationship between a `conference` event and its component `speech`. Structural association is denoted by $R(A, B)$ where A and B are semantic unit and R is the association name, such as `component-of` and `part-of`. Composition structure will be elaborated later in this chapter.

3. *Spatial association* is a binary association between two semantic units indicating relationship in space, expressed qualitatively based on the order of units in space, and denoted by $R(A_1, A_2)$, where $R \in$ {*above, left, in front, between, overlap*} and their inverse. The choice of spatial associations comes from (Sistla, Yu and Haddad; 1994). For instance, 'book *above* table' is a spatial association between two objects.

4. *Temporal association* is a binary association between two semantic units interpreted in time, expressed qualitatively based on the order of units in time, and denoted by $R(A_1, A_2)$, where $R \in$ {*before, meet, during, overlap, starts, ends, equal*} and their inverse. The choice of temporal associations comes from (Allen; 1993). Allen introduces the interval-based temporal logic to represent the knowledge and interference concerned with time. For instance, 'man runs *after* a clerk has been attacked' is a temporal relationship between two activities in time `run(man)` and `attack(man, clerk)`.

4.6 High-level Semantic Units

An *event* is defined as a partially ordered set of transitions of activities and objects, where a transition is indicated by the changes of values of observable attributes. Partial order is denoted by $\angle$. Suppose E is an event and A is a set of semantic units (activities or objects), then $a_i \angle E$ where $a_i \in A$ iff $T(a_i) \subseteq T(E)$. Events are denoted by *E(A, S)*, where *S* is a set of contextual associations such that for every semantic unit there exists another semantic unit and an association to relate them. Events are formally defined as:

$$\forall\ a_i \in A, \exists\ a_j \in A \wedge \exists\ s \in S \text{ where } s(a_i, a_j) \wedge i \neq j.$$

A *semantic event* is an instance of an event that belongs to real world events, such as `conference`. Consider for example a sequence of frames representing the `leaving` event depicted in Figure 2. There are two objects o_1 of class `person` and o_2 of class `door`. Changes in the spatial parameters of the two objects over a sequence of frames are captured as activities: a_1 of `walk` performed by o_1, a_2 of class `open` performed by o_1 on o_2, and a_3 of class `swing` performed by o_2.

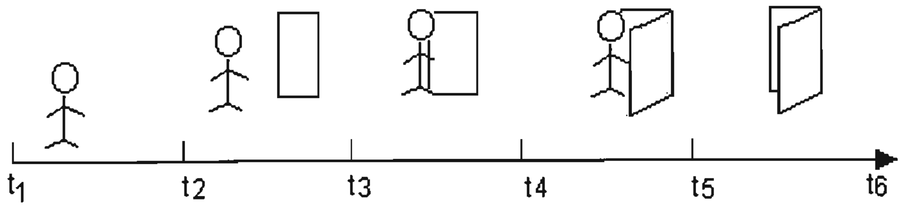


Figure 2. A sequence of frames constructing a `leaving` event

A *story* is a collection of partially ordered events, denoted by $e_i \angle S$ where E is a set of events, $e_i \in E$ and S is a story. The sequence of events is important in defining the story.

4.7 Composite Semantic Units

In a story, a number of objects of class `person` could be related to each other, such as in `team`. One man running after another leads to the concept of `chase` and a number of people talking to each other structures a `conversation`. All this leads to the concept of composite semantic units, which allows the construction of new semantic units from existing ones.

A *composite semantic unit* is a structure built of instances of elementary and possibly other composite semantic units, which could be of heterogeneous type, with a semantic interrelationship to express a complex fact. For instance, a group of objects of class `man` with a *collaboration* interrelationship express the complex object `team`. A man runs *after* another man expresses the `chase`.

4.8 Description

Descriptions are important features in real world modeling. In the model proposed in this thesis, an optional open set of content attributes is tightly related to each semantic unit and association. Modeling associations by simple semantic constraints is insufficient to express real-world relationships. Associations need to be described as well as semantic units for a more precise result. For instance, *5 milliseconds* `before` and *good* `friend`.

The *description* of semantic units or associations is an open set of attributes and values representing features of interest to the end user. Descriptions could be perceptual (media-dependent), such as `color` or semantic (media-independent), such as `name`. Semantic units or associations may appear in a video or in multiple videos a number of times, leading to two categories of content attributes:

- *Static* attributes that have fixed values, such as `name` and `date of birth`.
- *Dynamic* attributes that change their values over time, such as the spatial position.

Semantic units and associations may have dynamic properties that can change in various frames or according to the context. This may lead to the concept of the states of both semantic units and associations. *State* transition is determined by a change in the value of an observable dynamic attribute. Each state is associated with an observation slot and a set of dynamic attributes representing the description of the state.

4.9 Abstractions

Classification, generalization and aggregation abstractions are the common abstraction mechanisms available for grouping instances of semantic unit, description or association within classes, building class hierarchies and constructing complex semantic units. As elaborated in chapter 2, few works have considered abstraction and only of objects. Abstraction is essential for modeling real world features and associations as well as semantic units. Some may argue that perceptual content are *specific*, where they hold one constant interpretation and no abstraction needed. Perceptual content in the semantic layer could have multiple interpretations, for instance, `square` and `rectangle` could be referred to as `quaternary`, and `reddish brown` as `red`. Therefore, in semantic video model, abstraction should be considered for content attributes and attribute values.

1. *Classification* abstraction allows for the definition of the classes of semantic units. For instance, class of object `person`, class of activity `run`, class of events `conference`, color description of class `red`, and association class `sponsor-of`.

2. *Generalization* abstraction allows for defining the hierarchies of the classes of semantic units, as for instance `postgraduate-student` class is a subset of `student` class which is itself a subset of `person` class. Class activity `run` is

a subclass of class `move`. Let C be a set of homogeneous classes of semantic units, descriptions or associations. Generalization abstraction G is defined as a subset of C X C. Generalized concepts are organized into a hierarchy of *IS-A* relationship, where sub-classes inherit all properties of super-classes. The hierarchy leafs corresponds to specific concepts, such as `postgraduate`, and higher nodes corresponds to more unspecified concepts such as `person`.

3. *Aggregation* abstraction is a class structuring mechanism for assembling complex semantic units, descriptions and associations from elementary or composite ones with a *component-of* relationship. For instance the object people is an aggregation of more elementary objects of classes `person` and `car` is an aggregation of `engine`, `wheels`, `etc`. `Address` content attribute is assembled of one or more attributes, such as `street name`, `state`, `country`, etc. Semantic unit aggregation is a special case of structuring composite unit.

4.10 Definition of Semantic Units and Associations

This section shows how semantic information is stored in databases. A semantic unit or an association is a quadruple (uid, F, V, ∂), where uid is the identifier, F is a set of content attributes, and V is a set of attributes' values $V = \cup_{f \in F} domain(f)$. Then ∂ maps attributes into their values $\partial : F \rightarrow V$ such that $\partial\,(f) \in domain(f)$.

Suppose for instance an object person with a quadruple (123, F, V, ∂) is given where:

F = { name, date-of-birth, shirtcolor, class, …} is a set of content attributes.

V = { Ali, 2-5-1970, red, person, ...} is a set of attributes' values.

∂(name) = Ali, ∂ (date-of-birth) = 2-5-1970, ...

4.11 Definition of a State

The states of semantic units and associations are each recorded in a 7-tuple (*S, uid, T, F, V,* ϑ, λ), where S is a set of state identifiers, *uid* is the semantic unit or association identifier in which states belong, T is a set of observation slots triple [VID, t_s, t_e], F is a set of dynamic attributes, V is the set of their values, ϑ maps states into a set of attributes and values such that ϑ: $S \rightarrow$ P(∂) and $\vartheta(s) \in$ { ∂_1, ∂_2, ...} where $\partial \in \partial$,and λ maps states into observation slots such that λ: $S \rightarrow T$ then λ(s) $\in$ t

Suppose for instance the semantic object person with a 7-tuple (*S, 123, T, F, V,* ϑ, λ) is given where:

S = { s_1, s_2, …} set of states of a unit.

T = { [222, 20, 45], [333, 120, 127], …} set of observation slots where object appears.

F = { shirtcolor, X, Y, …} set of dynamic attributes.

$V = \{$ red, white, 20, 30, … $\}$ set of attributes' values.

∂_1(shirtcolor)= red, ∂_2(X) = 30, ∂_3(Y)= 20, ...

$\vartheta(S)$ maps states into attributes and attributes' values as follows:

$\vartheta(s_1) = \{ \partial_1, \partial_2 \}, \vartheta(s_2) = \{ \partial_3, \partial_4 \}, \ldots$

$\lambda(S)$ maps states into observation slots as follows:

$\lambda(s_1) = [222, 20, 45], \lambda(s_2) = [222, 70, 95], \ldots$

4.12 Video Logical Layers

In this proposed model, a semantic logical layer is built on top of the physical layer of a video to provide a semantic structure to the video document and a semantic abstract view of the video content. Semantic content-based video retrieval does not work with the physical layer directly but with the semantic layer. Video layers are shown in Figure 3 and are decsribed as follows:

1. *Physical layer* is the raw data stream, which contains frame-based objects and objects motion over a sequence of frames.

2. *Semantic layer* is an abstract layer where the physical layer contents are linked into the real world. This layer provides the semantic structure to the video document. Two levels of semantic layer are distinguished, Intermediate and High-level.

2.1 *Intermediate level* semantics are directly extracted from the physical layer. These are the elementary objects and activities, perceptual features, and spatial and temporal associations. Signal processors can automatically capture intermediate level semantics.

2.2 *High-level* semantics are composed of intermediate level content. In the proposed semantic video model, events, story, composite units, high-level descriptions, and structural and contextual associations are considered high-level semantics. Knowledge representation and inference rules are needed to detect high-level semantics.

3. *User View* represents the user's perspective of a video clip. It is constructed from units from the semantic layers, descriptions and relationships.

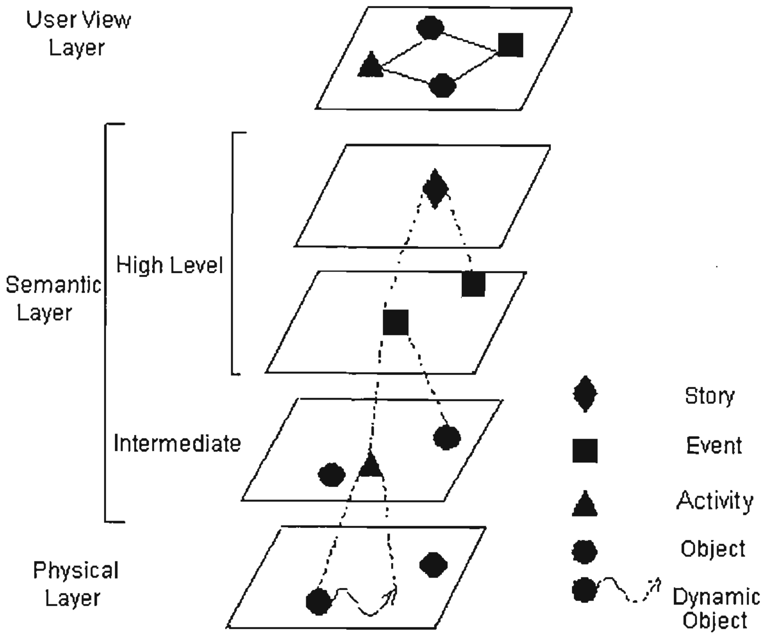


Figure 3. Video layers

4.13 Video Structure

So far, the author has been dealing with the components of the semantic layer of the video document. This thesis aim at structuring the video document based on its story-line structure. This section aims at presenting the approach adopted in structuring the semantic layer. As elaborated in section 2.5, a number of works that deal with the problem of structuring the logical representation of a video stream. Among these approaches are *segmentation* and *stratification.*

- *Segmentation* in Figure 4.a, partitions the video into chunks and assigns a set of keywords to each chunk (Little et al.; 1993).

- *Stratification* in Figure 4.b, associates a keyword with distinct pieces of video (Davenport, Smith and Pincever; 1991).

In this proposed semantic video model, the same semantic unit can be extracted at different levels of abstraction. These are levels of object, activity, event or story. For instance, `Ali` can be extracted from the object layer, the `running` activity performed by `Ali` can be extracted from the activity layer, and `Ali` involved in a chase can be extracted from the event level. To support the various levels of abstraction and share of annotations, a video stream is not physically segmented. The stratification enables the assignment of several annotations to a time interval. Hence, the author decides to extend the stratification approach with the proposed semantic structure rather than the free annotation.

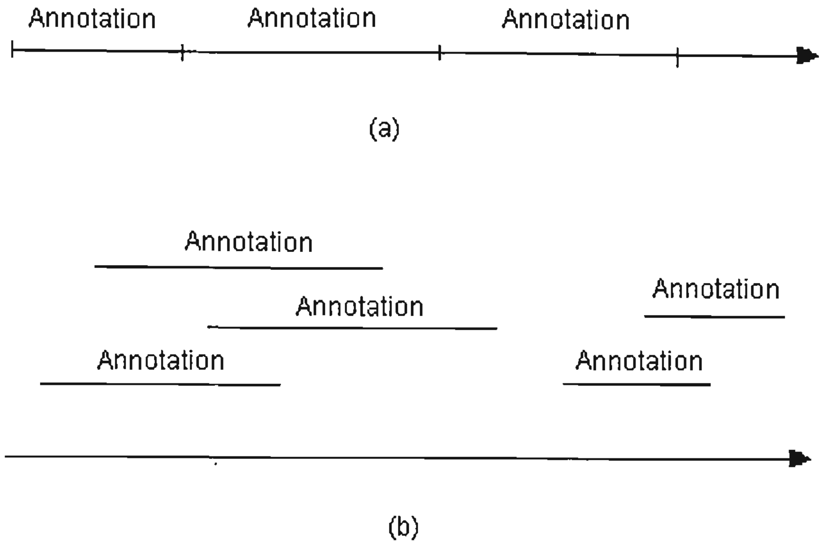


Figure 4. (a) Segmentation and (b) Stratification

4.14 Graphical Conceptual Model for Video Content

A salient characteristic feature of this proposed semantic video model is its ability to compose semantic units and associations to structure a new complex fact. Components of the model could be static units or dynamic units with state transitions, including a set of associations. The assemblage of various associations within the video space is considered a remarkable characteristic of this proposed semantic model. The graphical conceptual model has thus far proved its usefulness and efficiency in representing the connection between concepts and relationships, and also a reliable technique to improve the understanding mode. This section is aimed to introduce a graphical notation for the proposed model, described earlier in this chapter, as well as to represent the interplay among semantic units constituting a composite unit and map to an abstract model into video.

Why graphical representation of video content?

The methodology presented in this section can manifest itself in conceptualizing heterogeneous views of a video as perceived by individual user. This can be a major step toward an easy-to-grasp graphical user interface, where, for each input video stream, semantic units and relationships are captured and encoded on the proposed graphical notation. Later, the proposed graphical notation can be used as a searching mechanism that model information at various levels of granularity and in various video spaces. It is believed that the proposed unified framework may help users to express their heterogeneous queries and utilizing the system to process those queries.

Current graphical representation

Temporal relationships among objects have been modeled using Petri Nets, time-line, time intervals, time flow graph, and others (Chang & Chang; 1996). Little has been accomplished to express both spatial and temporal relationships in the same model. VSDG (Day, Dagtas, Iino, Khokhar and Ghafoor; 1995) is a graphical model that captures both spatial and temporal objects in a video. Spatial relationships are described graphically as a set of attributes associated with each circular node in the graph, which, in the author's opinion, is not a true *graphical* representation of spatial relationships.

The Object Composition Petri Net (OCPN) model (Li, Goralwalla, Özsu and Szafron; 1996) is a directed graph with transitions and places. It is an extension of the augmented Petri Nets (Coolahan & Roussopoulos; 1983) and is suitable for representing concurrence and synchronization between entities. As shown in Figure 5, OCPN uses the following notations:

- *Circles* are places representing interval of media object;
- *Duration* is assigned to each place representing the time interval in which a place is active;
- *Vertical bar* represents a transition or point of synchronization, when components synchronize their presentation, and project the temporal order of components;
- *Token* specifies active places where a transition fires when each of its input places contains a token for each of its output places.

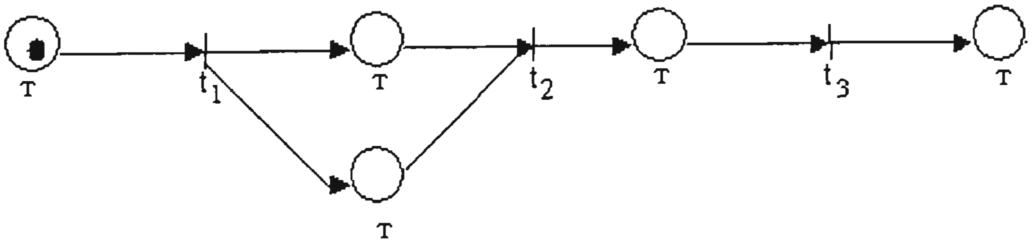


Figure 5. The Object Composition Petri Net (OCPN) model

Applying current graphical representations to the proposed semantic model

In this proposed semantic video model, an event is constructed of one or more synchronized *related* states of objects and activities, which impose synchronization and relationships in presentation. Several issues need to be considered during the graphical representation of an event: different states, state duration, state transitions, the establishment and termination of association between various states, and the synchronization between states and associations.

OCPN is reported to suffer with certain limitations. One of the limitations of OCPN is its inability to express all semantic relationships between components, but only temporal relationships. Semantic nets (O'Docherty & Daskalakis; 1991) represent objects and their interrelationship. The ER model (Storey & Goldstein; 1988), is a very efficient graphical conceptual model for representing the relationship between the entities. Both Semantic nets and ER fail to express synchronization and composite entities.

The proposed graphical conceptual model

The proposed graphical notation may be defined as a directed graph, extended from OCPN by adding a temporal unidirectional *lightning arrow* to describe a relationship between two components in space and context. Arrows are labeled while representing explicit relationships, such as `father` and `above`. This lightning arrow requires the presence of both components. In other words, the removal of one or a change in state will lead to the termination of the relationship. The reason for introducing relationships in space and context spaces lies in allowing the representation of high-level semantic units, which are otherwise rather cumbersome while expressing graphically.

In the proposed graphical representation, notations are redefined as follows:

- *Circles* represent a state of a semantic unit while state modification is associated with the change in video presentation time;

- *Duration* is assigned to each state representing the time interval in which a state is active;

- *Vertical bar* represents a transition or point of synchronization, which in video is represented by time, and also indicates the creation or termination of a new state or association;

- *Lightning arrow* describes a relationship between two components and assigns the lightning arrow a use for representing spatial and contextual association.

A simple example is cited to clarify this proposed graphical conceptual model. For instance, we have a video clip where a `book` is placed *above* a `table`. OCPN representation is shown in Figure 6.a, where it captures the relationship between the two objects (book and table) in time but fails to express their spatial relationship (above). Figure 6.b demonstrates the graphical representation of the same frame based on this proposed graphical model, which captures both spatial and temporal association between both components.

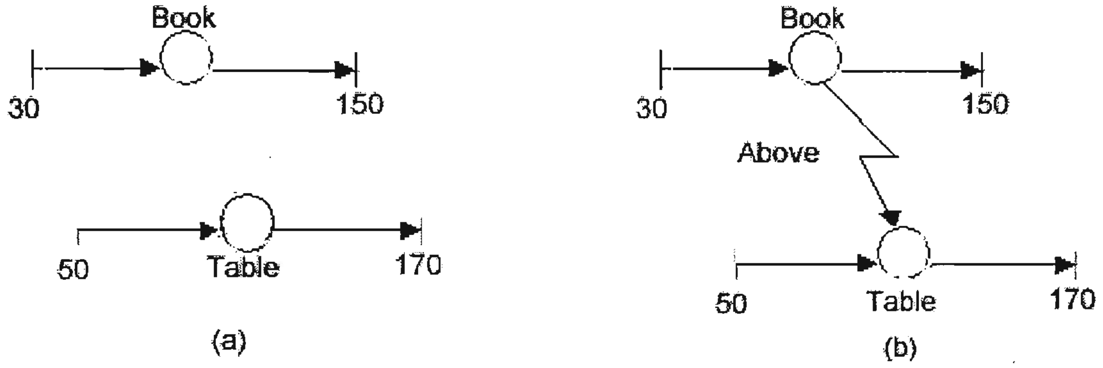


Figure 6. Graphical representation of a video clip content

An example

To illustrate the idea proposed in this section, consider the sequence of frames representing the `leaving` event depicted in Figure 2. The graphical representation of the event is given below in Figure 7. A number of objects and activities are involved. Objects o_1 and o_2 represent `person` and `door` respectively. Activities a_1, a_2, and a_3 represent `walk`, `open` and `swing` activities respectively. Object o_1 appears at moment t_1 performing activity a_1, object o_2 appears at t_2. At t_3, activity a_2 performed by o_1 on o_2 $a_2(o_1, o_2)$, and activity a_3 performed by o_2 denoted by $a_3(o_2)$

appears at t_4 and associated with activity a_2 in *cause-by* relationship; o_1 disappears at t_5.

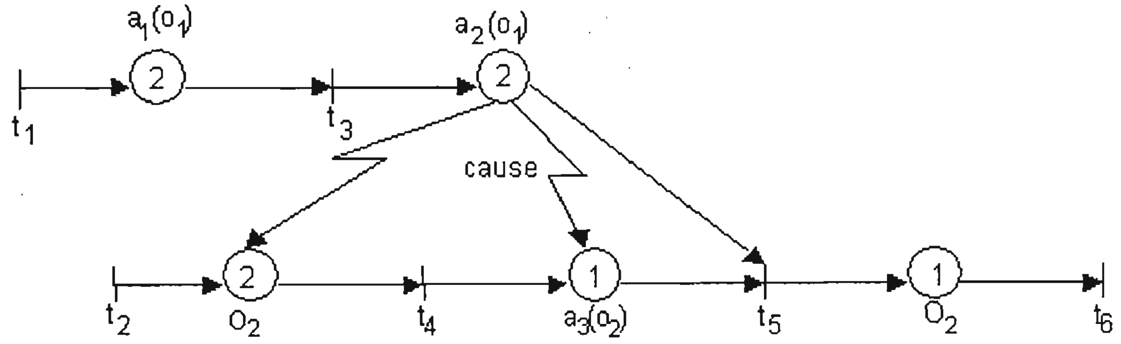


Figure 7. Graphical representation of a `leaving` event

The proposed graphical representation allows encapsulating part of a diagram and identifying it as a separate semantic unit intended to be used in other diagrams. Current conceptual models, such as in ER, do not support this view. The encapsulation feature provides flexibility to the graphical notation and supports extendibility in order to build composite units. For instance, a composite unit, such as `evacuation` is composed of a number of sub-events of type `leaving`. Hence, the diagram in Figure 7 above can be encapsulated, represented by circle, tagged as e_i and reused in the graphical representation of `evacuation`.

4.15 Summary

This chapter, constituting a core of this thesis, attempts to imitate a human understanding of the semantic content of a video and consequently develop a formal semantic model for video content. It explains in detail the proposed semantic video model and the way this model is stored in databases. A graphical conceptual model is proposed for representing video content interrelationship.

5
SEMANTIC VIDEO ACQUISITION

One of the chief components of semantic content-based video retrieval systems is the *semantic video acquisition,* which undertakes an analysis of a video and extracts its semantic content. This chapter suggests an approach to analyze video documents which are believed to behave in a manner resembling human analysis of a video document and create their own semantic model capable of describing the video content. It then transforms the approach into a tool for semantic video acquisition. The proposed semantic video acquisition possesses the potential of utilizing current state-of-the-art signal processors to maximize procedures that can be automatically conducted to speed up the acquisition process.

5.1 Human Approach to Semantic Video Acquisition

The main purpose of this section is to explain the mechanism of human analysis of a video which is proposed to be used as the basis of evolving formal methodology suggested for the semantic video acquisition. By taking into consideration the human approach to analysis, the author intends to develop a semantic video acquisition system. This system meets the human nature in video analysis and subsequently makes it user-friendly and increases the possibility of answering as many of the users' queries.

Definitions

Video analysis is a process of understanding a video document, enabling to extract contents of the latter and organizing the extracted information to be comprehended by the users. As video documents contain two categories of content: perceptual and semantic, consequently two kinds of analysis are required: perceptual analysis and semantic analysis.

- The *perceptual video* analysis refers to the process of extracting and addressing the perceptual features of a video, such as color, texture and frequency;

- The *semantic video* analysis refers to the process of extracting and addressing the semantic content of a video in order to comprehend.

In psychology, it is claimed that this is the type of representation of stories that end users employ to guide to comprehension during encoding and in retrieval (Mandler & Johnson; 1977). In this work, *video analysis* will refer to both perceptual and semantic analysis unless specified otherwise.

Unpredictable behaviors of humans

> *"When asking the question: what index might have labeled a given story? One thing we must continue to bear in mind while asking such a question is that no right answer exists, only possible answers"*
>
> *(Schank, R. 1990)*

Human behavior in the semantic video analysis is complicated although not very predictable. In the way a human analyzes a video, a number of factors may play a role, such as the human nature in being guided by expectations, jumping into conclusions, concentrating on elements of their interest and neglecting much detail in a video. These behaviors and more lead to unpredictable results. Humans do not think in the same way, but, on the other hand, contents have a basic structure in common (Schank, R.; 1990) and humans are expected to share these components and structures. For instance, all humans represent a `chase` by a number of objects, two or more, running one after the other.

In this work, the author concentrates on the expected behaviors of humans in semantic video analysis which leads to a common model and predictable results.

Story comprehension and structure

As elaborated in chapter 4 (Semantic Video Model), from the human perspective, semantic video content can be described as a relationship between components constructing a story. Components are events, activities and objects. Usually, semantic units along with associations are grouped and encapsulated in context (single location and time duration). The unit containing a collection of partially ordered semantic units appearing within the same context is referred to as ***scene***. For instance, a scene may contain `leaving` and `conversation` events. Scene is not a semantic unit, as it is considered in many semantic models, because it represents no meaningful unit and cannot be a variable factor in the end user's query.

The way a story is comprehended depends on the order of its content. In video, frames are in a certain sequence. Data exposed in each frame are most probably necessary for understanding the content of later frames. This could explain why viewers cannot understand a video when watching it in backward or in a random order. Most stories export some information, especially in earlier scenes, to serve as basic information throughout the story. Hence, a scene viewer is not dependent on the current scene only. Most of the time, viewers determine current information with the aid of additional information, mostly that provided by preceding scenes or by the general knowledge of the world. In a conclusion, two kinds of information constitute a semantic unit at a point of time. These are:

1. *Time-dependent* information which is directly based on a point of time and is represented by the dynamic attributes defined in section 4.8 (Description);

2. *Time-independent* information which is not based on a point of time, but drawn and assigned by the end user's perception based on the information presented earlier or the real world knowledge, and is represented by the static attribute defined in section 4.8 (Description).

This leads to the concept that the whole video document should be treated as one entity and the content of each segment should not be considered as an independent entity, as advocated in most current works in video modeling.

Human process of semantic video analysis

By investigating the phenomena to be modeled, which is the human approach in analyzing video documents, understanding content, extracting semantic units, and building his/her own semantic model describing video content, the author summarizes the process as follows:

- A human observes and classifies a salient object that appears in a frame;

- The human assigns information to this object by linking it to that observed in the preceding scenes, the real-world knowledge base and information extracted from the current frame;

- Within the scene, the viewer tracks the object until it disappears and observes its changes. A human interprets these changes into an activity performed by the specified tracked object;

- The recognized activity could be linked to a previous occurrence or to the real world knowledge to draw time-independent information. New time-dependent information is assigned to the extracted activity;

- The process of capturing salient objects and activities operates in a cycle for all objects in the same context (scene);

- Within the same scene, the viewer captures associations between extracted semantic units (objects and activities) in the various spaces in a video (time, space, context and structure);

- The human-annotator applies his/her own knowledge and constructs high-level semantics by assembling captured objects, activities and associations;

- The constructed semantic unit could be linked to previous occurrences and to the real-world knowledge to draw time-independent information. Time-dependent features are extracted from the current point of time;

- The process of assembling and identifying high-level semantics is repeated in the same context until no more high-level semantics need to be constructed;

- The viewer will move to the next context (scene) and will repeat the same process until the end of video stream;

- The sequence of scenes captured and analyzed is assembled into a story.

The output of this process is a semantic model describing video content.

In conclusion, the semantic video analysis starts by breaking a video document into scenes and observing their components. This process is repeated several times to construct the semantic model describing the content of the whole story. Within a scene, the semantic video analysis process is best described in a bottom-up approach starting by observing components of the bottom level (objects), and combining them to build higher level of semantics.

5.2 Algorithm for Semantic Model Construction

With the human approach in semantic model construction in mind, this thesis suggests the following semantic model construction algorithm as illustrated in Figure 8 below. It should be noted that humans may think in different orders, but they are still expected to think within the same format (Schank, R.; 1990).

(a) A video document is segmented into a sequence of scenes;

(b) Object extraction algorithms are applied to extract objects from a scene;

(c) By tracking dynamic objects over a scene using motion detection algorithms, activities are detected;

Processes (b) and (c) are repeated for all objects in the scene.

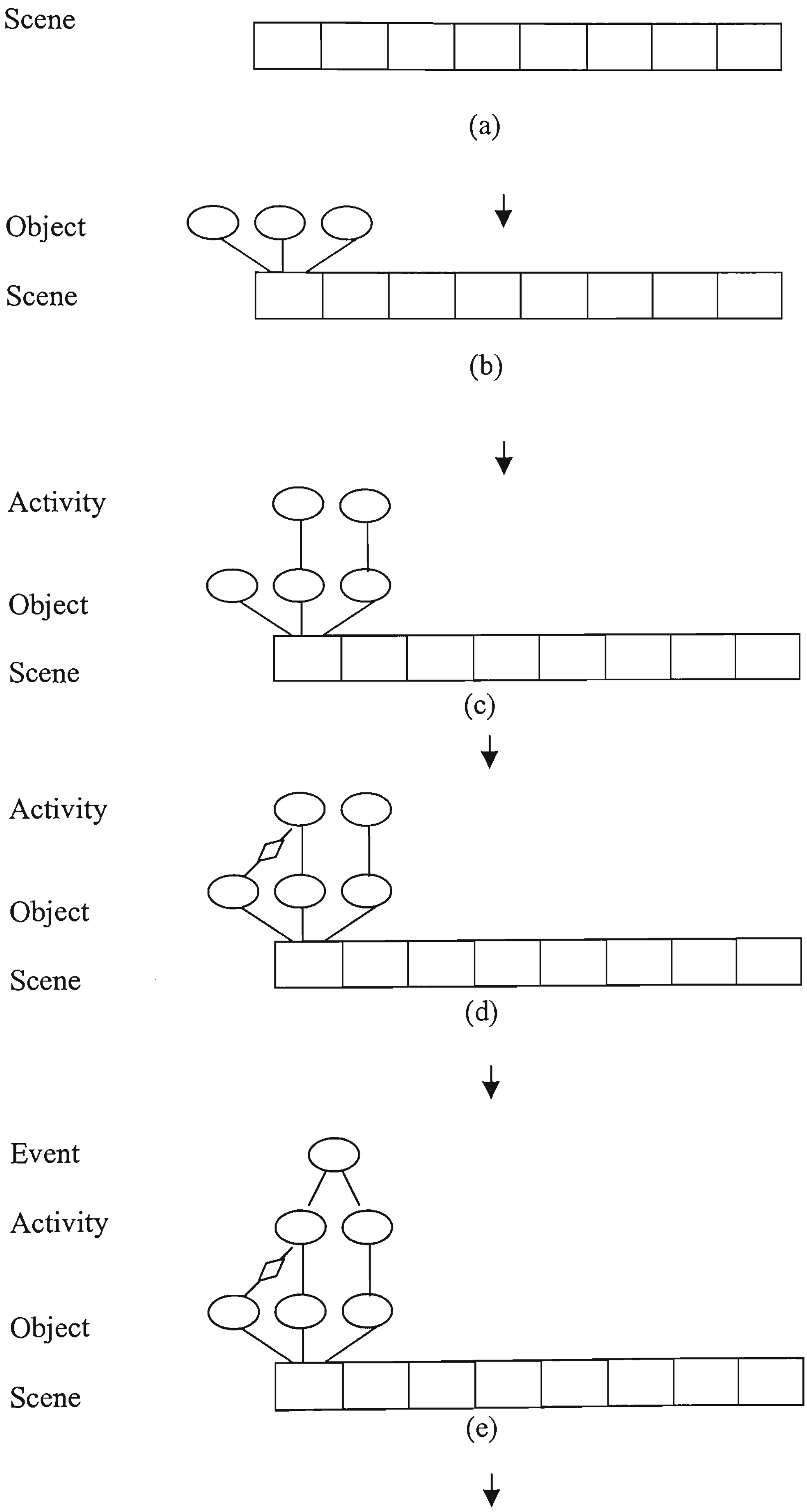
(d) The various relationships are captured between extracted objects and activities;

(e) Captured objects, activities and associations are assembled to construct events;

The event construction process is repeated for all events in a scene;

(f) Processes (b) to (e) are repeated for all scenes;

(g) Assemble the sequence of scenes into a story.



(a)

(b)

(c)

(d)

(e)

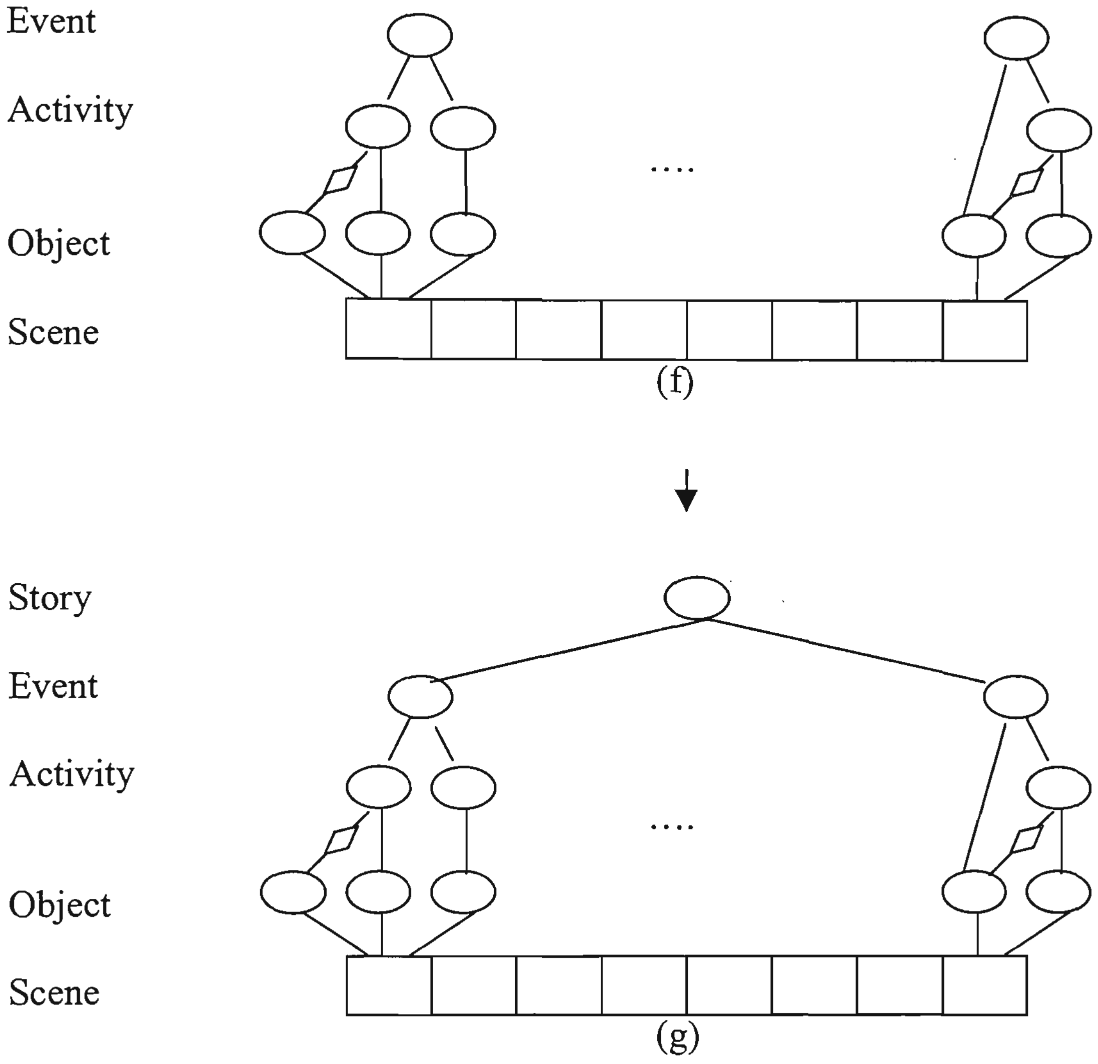


Figure 8. Semantic model construction

Handling repeated processes

Many events, activities and objects may appear in a video document for a number of times. With a view to reduce the mental process where a human-annotator does not have to extract and construct an already captured semantic unit, this section introduces the concept of scripts. A *script* is a sequence of patterns in a program to describe the structure of semantic units. When similar structure appears, the system automatically retrieves the already predefined script. This should save the time and effort of the human-annotator, and serve the consistency of the annotations. For instance, the `leaving` script is defined in the following sequence, `walk(man)`, `contain(door, man)` and `disappear(man)`.

5.3 The Architecture for Semantic Video Acquisition

The author attempts to use the algorithm inferred from the human approach to semantic video acquisition into building a semantic video acquisition tool designed to be used on the video's physical stream. On the outset, it is necessary to accurately identify tasks which are possible for the human observer and others for the machine. On the basis of this identification, the author will suggest an implementation protocol of the acquisition system and illustrate in detail the suggested structure of the video semantic acquisition.

Human versus machine analysis

Pure manual annotation is perceptually and analytically difficult, tedious, expensive, inconsistent and time-consuming. Yet, it may be admitted that perceptual content analyzers (signal processors) do not offer a satisfactory solution to the semantic analysis.

A number of work programs in vision, audio and text processing have been reviewed and listed in the later part of this chapter. A signal processor called IMAQ Vision (National Instrument; June 1997) has been tested in the laboratory (see Appendix B for product review). With current signal processing techniques, it is possible to achieve the following:

- To extract salient objects from the video stream;

- To identify a group of pixels or a sequence of frequency as selected predefined objects, such as a `person` and a `car`;

- To track the motion of an object and identify primitive activities, such as `walk` and `remove`;

- To address particular properties existing in the perceptual level, such as color, width, area of vision object, and amplitude and frequency for audio object; and

- To capture relative spatial and temporal positions for visual salient objects.

However, the current signal processing techniques suffer with limitations which may be stated as under:

- To recognize content properties beyond the perceptual level, such as role and name;
- To recognize contextual and structural associations; and
- To detect composite semantic units and high-level semantic units, namely, events and story.

The observations support the conclusion that, in practice, intermediate semantics can be extracted automatically but an automatic detection of high-level semantics is hard and sometimes impossible.

Semantic video acquisition process

The entire video analysis and acquisition process are performed off-line. The semantic video acquisition is implemented through two steps: data acquisition and data analysis. The *data acquisition* acquires data that reveal no information. The *data analysis* extracts information from received data. Signal processing techniques achieved success in the data acquisition. But, at this stage with available machine capabilities, we cannot escape from the need for using human intelligence in the data analysis. For instance, a machine can recognize an object and classify it as a `person` or, in more intelligent systems, may classify it as `student`, but machines fail to distinguish `postgraduate` student from `undergraduate`.

The author suggests an integration of both human and machine analysis in a computer-aided digital video analyzer. In the suggested architecture, video documents are analyzed manually with the assistance of the state-of-the-art processing techniques.

Computer- aided analyzer functions

With a computer-aided analyzer for video semantic, a human-annotator can perform the following functions:

- To view automatic processing technique outputs;
- To cancel, add or modify extracted data;
- To assign semantics to extracted data;
- To store information in the meta-database;
- To build contextual and structural relationships between stored semantic units;
- To construct high-level semantic units; and
- To assign semantics to constructed units.

Computer-aided analyzer components

To perform semantic video acquisition, the following components are needed:

- *Video encoder* to digitize the raw video and audio signal, compress it and collect some metadata associated with the compressed stream;

- *Video server* to store and manage video documents and meta-database;

- *Semantic video-aided analyzer* to provide web-based user interfaces for semantic video acquisition through which the operator can view and monitor automatic extracted units and features and cancel, modify or add new ones. In addition, it provides an interface from which the human-annotator can assign semantics, contextual and structural associations between predefined units, and, lastly, construct high-level semantic units;

- *Scene detector* searches for scene boundaries, extracts and returns one scene at a time to the human-annotator for analysis;

- *Vision, audio and text analyzers* automatically extract perceptual features;

- *Motion detector* accepts a visual salient object as an input and returns a set of frames representing the trajectory of an object;

- *Video database* stores physical documents; and

- *Meta-database* stores information describing videos in the video database.

The architecture of a computer-aided analyzer

Figure 9 shows the computer-aided analyzer architecture. The acquisition of video semantic goes through seven steps which are explained in detail.

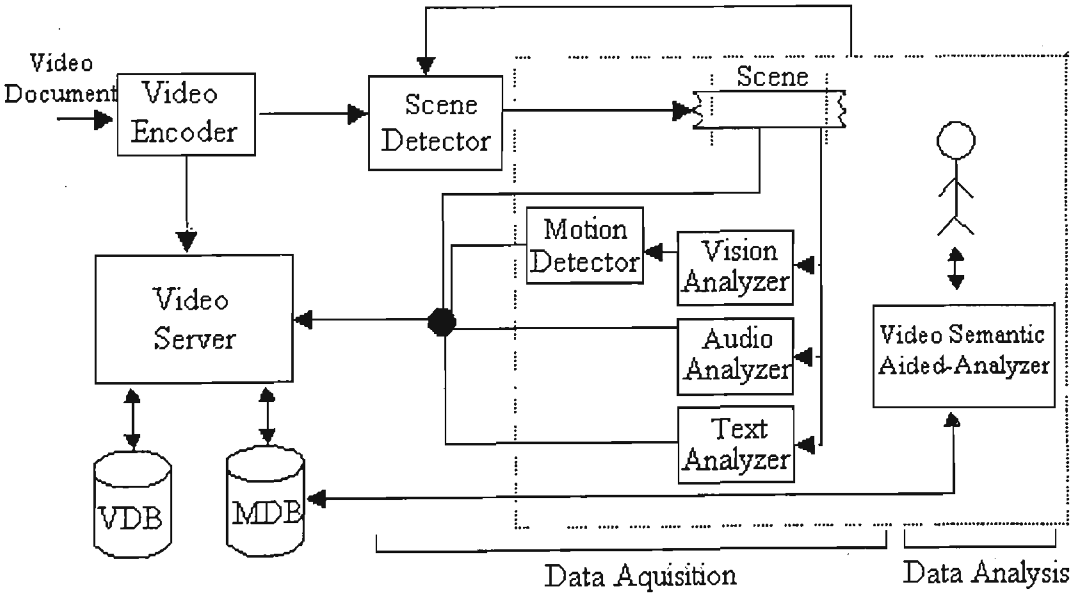


Figure 9. A general structure for semantic video acquisition

1. Capturing live video and encoding

Input: Live video

Output: Compressed format of video + metadata (format, length, encoding date and time)

Digital video documents are created from actual videos through a third-party encoding software or hardware called a *codec*. The way in which an encoder compresses video frames facilitates to occupy less space and is called a compression format. Compression formats which the video server can stream include: MPEG (Motion Pictures Experts Group), Iterated Systems ClearVideo, Radius CinePak, Intel Indeo and motion JPEG (Joint Photographic Experts Group)

2. Scene Detection

Input: Compressed video

Output: Scene + metadata (start and end frame, key frame)

Scene detection has been the focus of many researches where maximum attention is paid to detect scenes based on the visual changes (fade, dissolve, color histogram, ...) (Arman, Hsu and Chiu; 1993, Smoliar & Zhang; 1994, Aoki, Shimotsuji and Hori; 1996, Meng & Chang; 1996, Meng, Juan and Chang; 1995, Nagasaka & Tanaka; 1991, Patel & Sethi; 1997, Vinod & Murase; 1997, Yeo & Liu;. 1995, Zhang; 1993, and Yeung, Yeo and Liu; 1996); others are based on audio changes (speaker, musical interludes, silence, ...). Both visual and audio changes at scene boundaries constitute an accurate transition. VISION (Gauch, Gauch, Bouix and

Zhu; 1999) automatically partitions a video into short scenes using video, audio and closed-caption information.

The goals of video segmentation into scenes are summarized as under:

(1) The location of the start and end points of each scene;

(2) The extraction of a key frame to represent the scene; and

(3) The extraction of an area to search for semantic units and associations appearing within the same context.

The human-annotator should run the scene detector on the video stream to extract a scene and return the scene metadata (start and end frame, and keyframe). As shown in Figure 10, frames from 1362 to 1588 are returned as a result of a scene detector applied on news video stream, with a keyframe representing the scene content. A number of work programs have been conducted on keyframe extraction (Teodosio & Bender; 1993, Aoki, Shimotsuji and Hori; 1996, and Irani & Anandan; 1998). The human-annotator is provided with the screen shown in Figure 11 to view extracted information. The human-annotator should review, modify or acknowledge extracted information, then subsequently store it. Scenes are stored in the meta-database as shown in Appendix D.

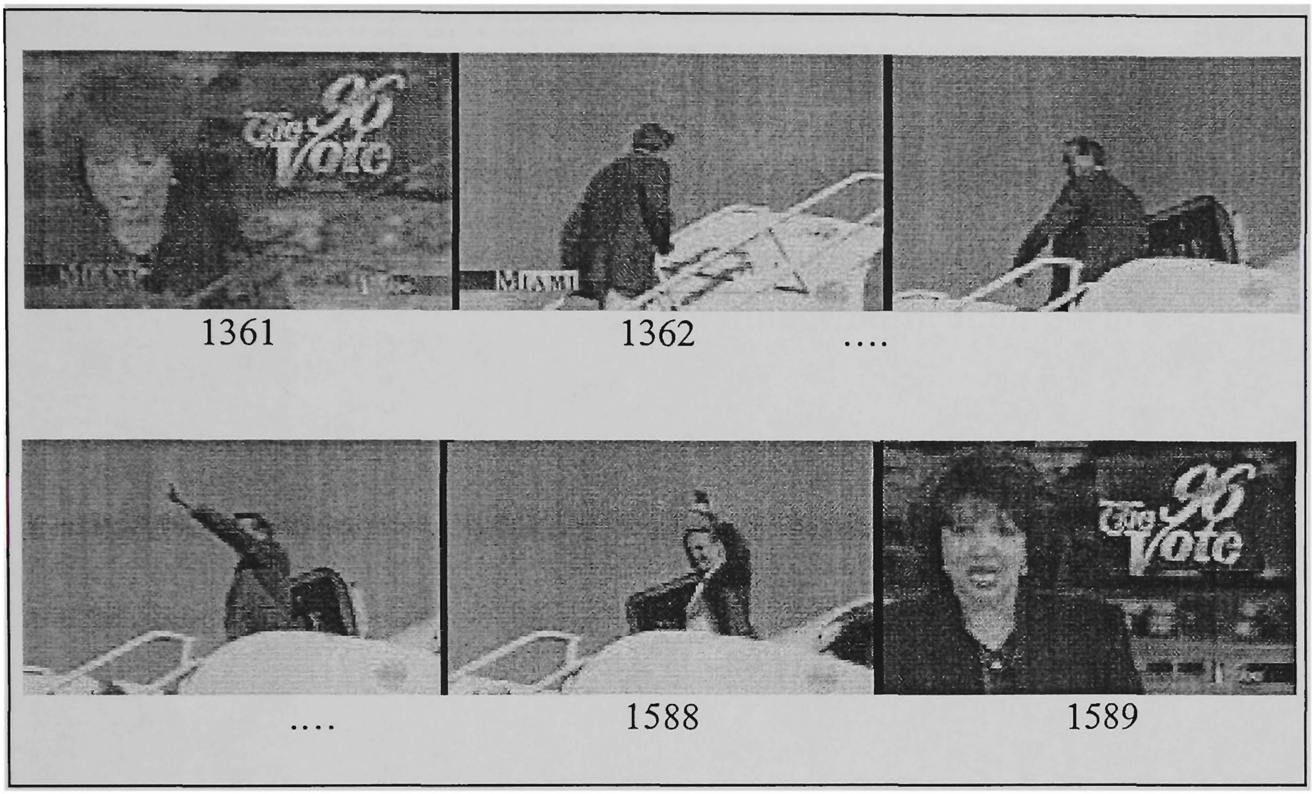


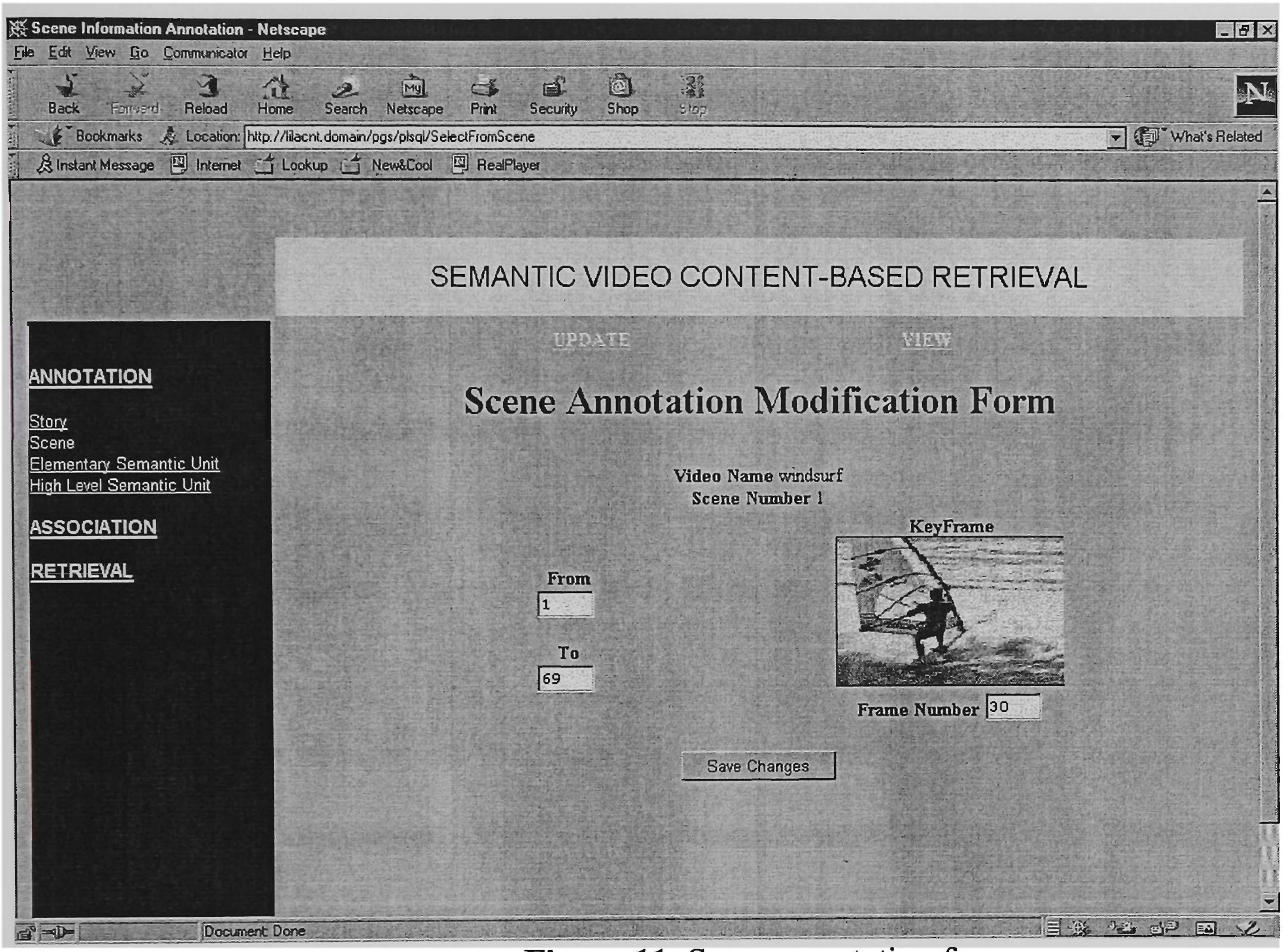
Figure 10. Sequence of frames constructing a scene

3. Signal Processing

Input: Scene + metadata

Output: Salient object + perceptual features

A detected scene is submitted for processing. Digital signal processors split out the scene into media streams (visual, speech, non-speech and textual). Several suitable processing techniques are applied to each media stream to extract salient objects and their perceptual features.


Scene Information Annotation - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://lilacnt.domain/pgs/plsql/SelectFromScene
What's Related
Instant Message Internet Lookup New&Cool RealPlayer
ANNOTATION
Story
Scene
Elementary Semantic Unit
High Level Semantic Unit
ASSOCIATION
RETRIEVAL
UPDATE
VIEW
Save Changes
Document: Done

SEMANTIC VIDEO CONTENT-BASED RETRIEVAL

Scene Annotation Modification Form

Video Name windsurf
Scene Number 1

From 1

To 69

KeyFrame

Frame Number 30

Figure 11. Scene annotation form

The aims of the signal processing include the following:

(1) Extracting perceptual features and salient objects from the visual stream (Flickner et al.1995, Meng & Chang; 1996, Smith & Chang; 1996, Smoliar & Zhang; 1994, Srihari; 1995, and Wu, Ang, Lam, Moorthy and Narasimhalu; 1993) (Figure 12 shows a blob detected and extracted by a vision processor);

(2) Identifying faces (Wu et al.; 1993);

(3) Capturing embedded captions (Lienhart; 1996);

(4) Obtaining spoken words contained in a speech stream (Brown, Foote, Jones, Jones and Young; (1995, 1996), and Peacocke & Graf; 1990); and

(5) Capturing salient units in a non-speech stream and classifying by using parameters such as smoothness and bandwidth (Blum, Keislar, Wheaten and Wold; 1995), MuscleFish technique for audio searching (http://www.musclefish.com), and (Wold,Blum, Keislar and Wheaten; 1996).

The resulting data of a signal processor tested (National Instrument; June 1997) consists of a data structure of salient objects found, related features (color, texture, histogram, frequency, amplitude, ...), spatial data for visual objects (x, y, mass size, width, length), and temporal information (start and end frame). The resulting data are presented to the human-annotator for review, modification and addition. Figure 13 shows the review screen for one of the elementary semantic units. Later, the human-annotator adds semantics to extracted features and objects. Signal analyzers are capable for checking the database for similar predefined objects. Extracted elementary semantic units and their features are stored in meta-database as shown in Appendix D.

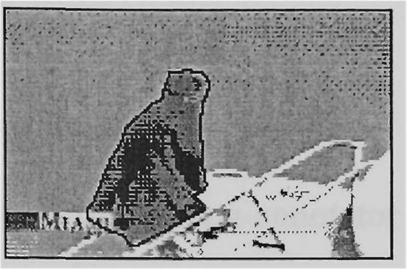
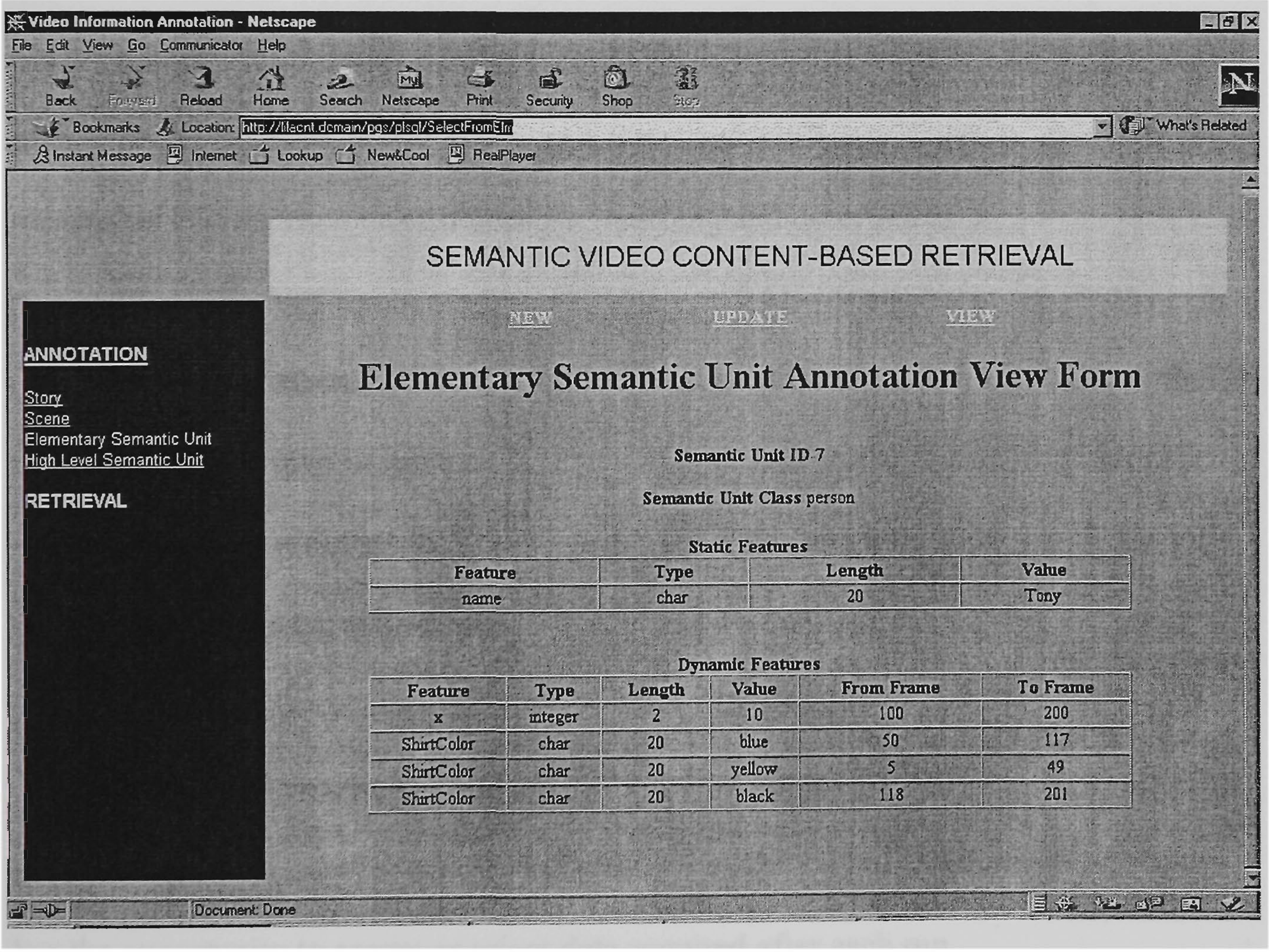


Figure 12. Automatic blob extraction by a vision processor


Video Information Annotation - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://litecnt.dcmain/pgs/plsql/SelectFromElm
What's Related
Instant Message Internet Lookup New&Cool RealPlayer
SEMANTIC VIDEO CONTENT-BASED RETRIEVAL
NEW UPDATE VIEW
ANNOTATION
Story
Scene
Elementary Semantic Unit
High Level Semantic Unit
RETRIEVAL
Document: Done

Elementary Semantic Unit Annotation View Form

Semantic Unit ID 7

Semantic Unit Class person

Static Features

Feature	Type	Length	Value
name	char	20	Tony

Dynamic Features

Feature	Type	Length	Value	From Frame	To Frame
x	integer	2	10	100	200
ShirtColor	char	20	blue	50	117
ShirtColor	char	20	yellow	5	49
ShirtColor	char	20	black	118	201

Figure 13 . Elementary semantic unit annotation form

4. Motion Detection

Input: Scene + blob

Output: Set of frames, graph, metadata

For each extracted visual object, the human-annotator highlights the target object and submits the object to motion detection software for tracking the motion of an object within the scene. Considerable amount of research has been done in the area of object motion detection. Primitive motions may be identified and classified automatically (Lee & Koa; 1993, Courtney; 1996, and Aggarwal & Cai; 1999). The output is the position in 2D space and time, and attributes, such as direction of movement and appearance attributes for matching and description.

For each object in motion, the motion detector returns a graph representing the motion path as shown in Figure 14, which is sent back to the human-annotator for semantic analysis. Motion detectors enable the human-annotator to search for all motions which have a similar path.

As for object detectors, extracted information by motion detectors is presented to the human-annotator for acknowledgment or modification. The human-annotator has the opportunity to accept or reject data acquired after each run.

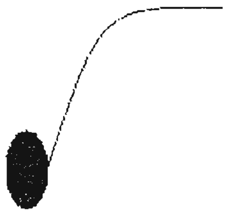


Figure 14. Returned graph representing motion path of an object

5. Assigning Contextual and Structural Associations

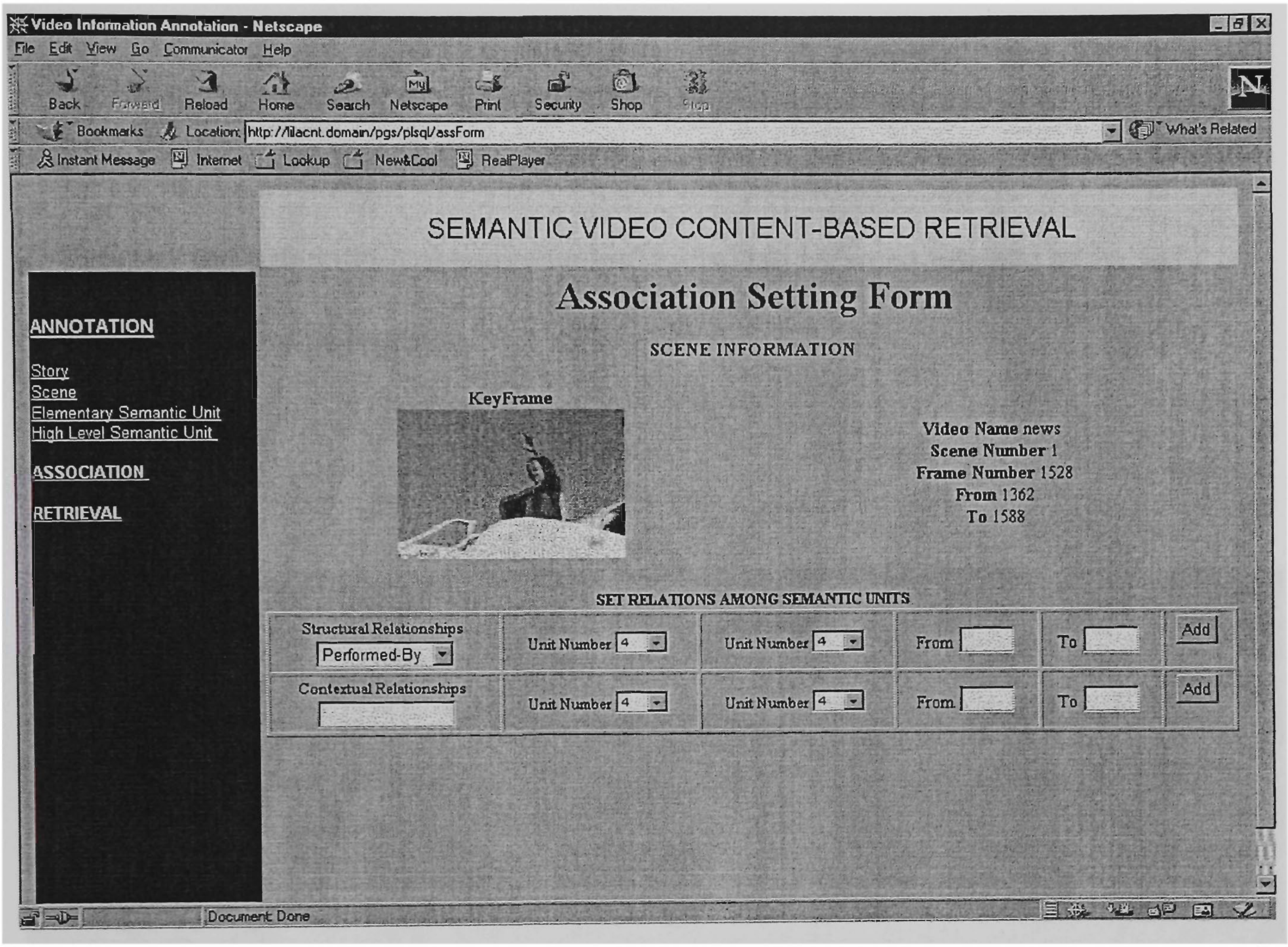
Input: Scene + meta-database

Output: Contextual associations + structural associations

Spatial and temporal associations are automatically inferred from captured spatial and temporal attributes. Table 1 lists the interpretation of spatial and temporal association functions calculated from the captured spatial and temporal attributes of a semantic unit. The human-annotator selects predefined semantic units and manually assigns contextual and structural associations as shown in Figure 15. Associations are stored in meta-database as shown in Appendix D.

Table 1 Association Predicates Interpretation

Predicate	**Interpretation**
A before B	$A.t_e < B.t_s$
A meet B	$A.t_e = B.t_s$
A during B	$A.t_s \geq B.t_s$ and $A.t_e \leq B.t_e$
A overlap B	$A.t_s \leq B.t_s$ and $A.t_e \leq B.t_e$ or B overlap A
A starts B	$A.t_s = B.t_s$
A ends B	$A.t_e = B.t_e$
A equal B	$A.t_s = B.t_s$ and $A.t_e = B.t_e$
A left B	$A.x < B.x$
A above B	$A.y > B.y$
A between B	$A.x \geq B.x$ and $A.x+width \leq B.x+width$ and $A.y \geq B.y$ And $A.y+height \leq B.y+height$
A overlap B	$A.x \leq B.x$ and $A.x+width \leq B.x+width$ and $A.y \leq B.y$ and $A.y+height \leq B.y+height$ or B overlap A


Video Information Annotation - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://ilacnt.domain/pgs/plsql/assForm
What's Related
Instant Message Internet Lookup New&Cool RealPlayer
ANNOTATION
Story
Scene
Elementary Semantic Unit
High Level Semantic Unit
ASSOCIATION
RETRIEVAL
Document: Done

SEMANTIC VIDEO CONTENT-BASED RETRIEVAL

Association Setting Form

SCENE INFORMATION

KeyFrame

Video Name news
Scene Number 1
Frame Number 1528
From 1362
To 1588

SET RELATIONS AMONG SEMANTIC UNITS

Structural Relationships Performed-By	Unit Number 4	Unit Number 4	From	To	Add
Contextual Relationships	Unit Number 4	Unit Number 4	From	To	Add

Figure 15. Contextual & structural association assignment form

6. Constructing a High-level Semantic Unit

Input: Scene + meta-database

Output: High-level semantics units

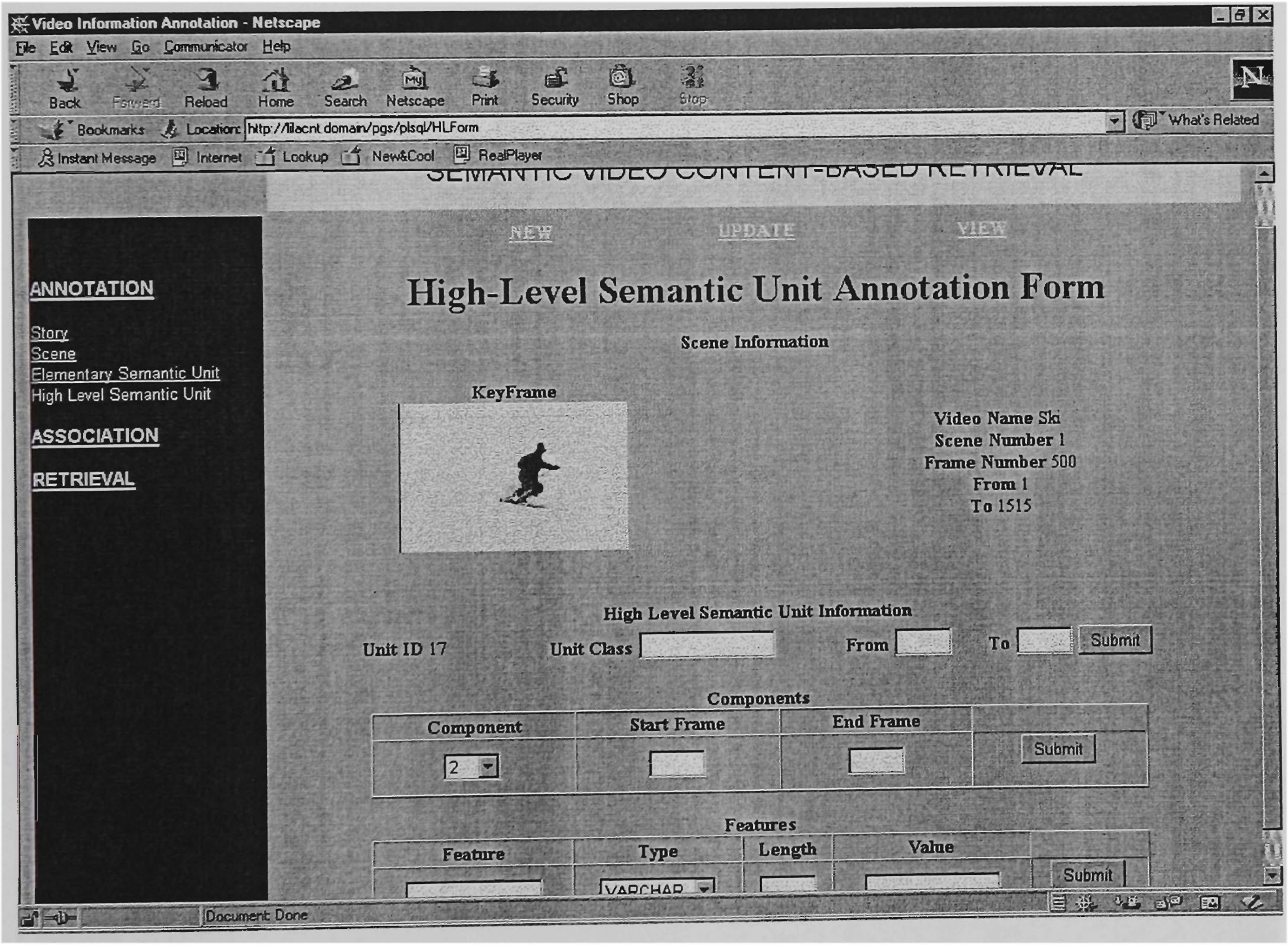
The process of constructing a high-level semantic unit is completely manual. Figure 16 shows the high-level semantic unit construction screen.

The process of constructing a high-level semantic unit involves the following:

1. Selecting constructing units within the observation slot;

2. Defining the structural and contextual relationships between selected units; and

3. Assigning semantic features.



High-Level Semantic Unit Annotation Form

Scene Information

KeyFrame

Video Name Ski
Scene Number 1
Frame Number 500
From 1
To 1515

High Level Semantic Unit Information

Unit ID 17 Unit Class From To Submit

Components

Component	Start Frame	End Frame	
2			Submit

Features

Feature	Type	Length	Value	
	VARCHAR			Submit

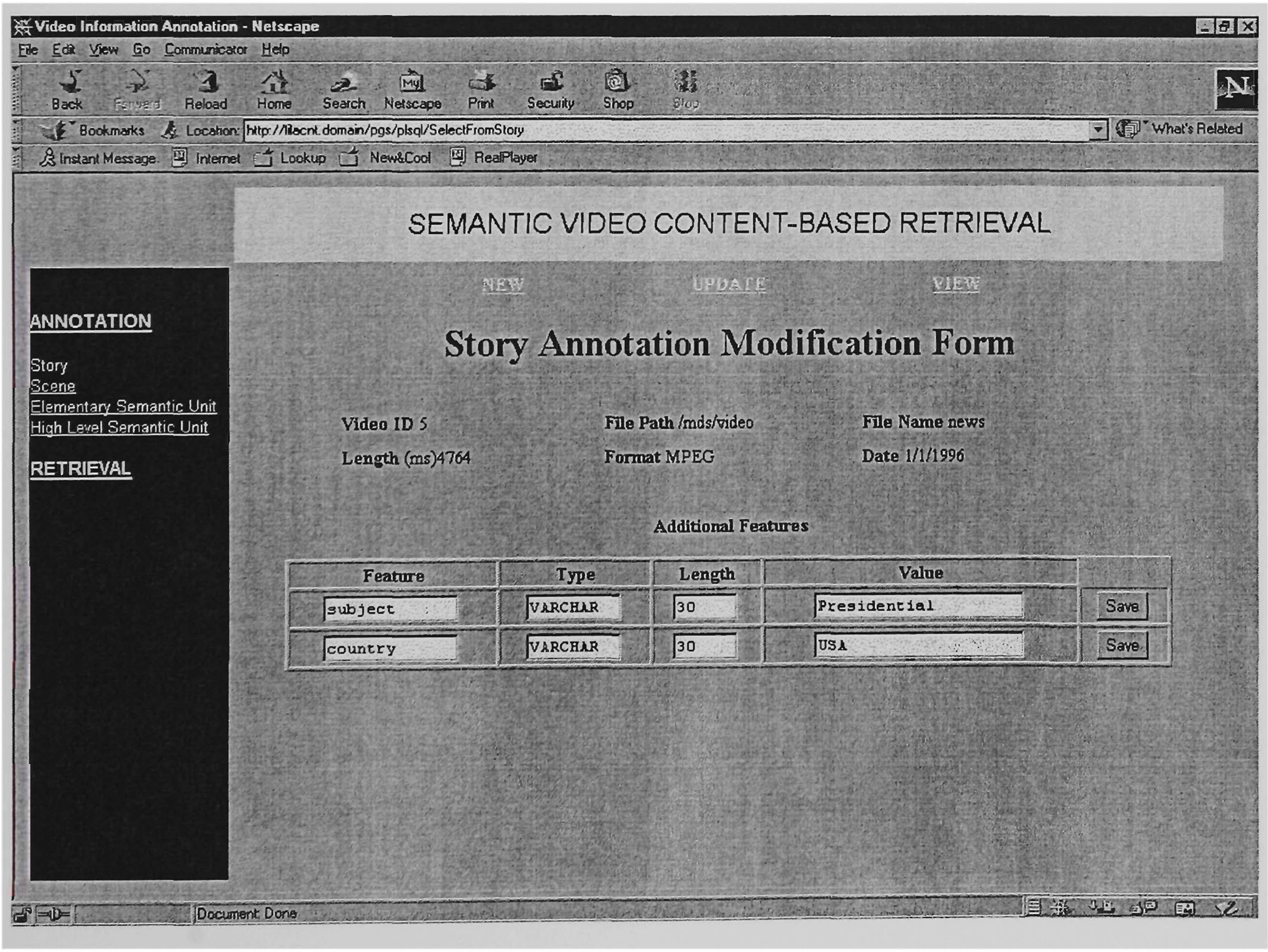
Figure 16. The high-level semantic unit construction screen

7. Annotating Story

Input: compressed video + meta-database

Output: Semantic metadata

In addition to the automatic information returned by the codec while digitizing the video document (path, video name, length, format, …), the human-annotator should assign semantics to describe the overall story as shown in Figure 17.


Video Information Annotation - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://ilecnt.domain/pgs/plsql/SelectFromStory
What's Related
Instant Message Internet Lookup New&Cool RealPlayer
ANNOTATION
Story
Scene
Elementary Semantic Unit
High Level Semantic Unit
RETRIEVAL
NEW UPDATE VIEW
Document: Done

SEMANTIC VIDEO CONTENT-BASED RETRIEVAL

Story Annotation Modification Form

Video ID 5
Length (ms)4764
File Path /mds/video
Format MPEG
File Name news
Date 1/1/1996

Additional Features

Feature	Type	Length	Value	
subject	VARCHAR	30	Presidential	Save
country	VARCHAR	30	USA	Save

Figure 17. Story annotation form

5.4 Summary

This chapter describes a semantic video acquisition component for analyzing video documents, extracting content and organizing in the way they serve the semantic-based video retrieval. The developed system is based on the human approach at semantic video analysis. Hence, the human approach at understanding a video document and building a semantic model is investigated. Also, human and machine capabilities in video content extraction have been studied. As a result, a computer-aided analyzer has been proposed based on the semantic video model presented in chapter 3 (Content-Based Video Retrieval System Architecture) in order to overcome the limitations of both human and machine, and thus provide a tool for acquiring the semantic content of a given video document.

6 SEMANTIC VIDEO RETRIEVAL

This chapter discusses the last component in semantic content-based video retrieval systems namely the *semantic video retrieval.* A video query language is proposed in this chapter which is based on the first ordered logic for querying video information. This video query language provides operations for utilizing compositional data, description, and contextual, spatial and temporal relationships in end user queries. In addition, effort is made to describes the overall architecture of the semantic video retrieval suggesting a model for accepting end user's query.

6.1 Query Language

A number of query languages have been evolved and presented for retrieving multimedia documents such as VideoText (Jiang, Montesi and Elmagarmid; 1997), Orenstein & Manola; 1988, Roussopoulos, Faloutsos and Sellis; 1988, VideoSQL (Oomoto & Tanaka; 1993), MMSQL (Amato, Mainetto and Savino; 1998) and VQL (Hee, IK and Kim; 1999). Query langauges are built on top of the database, hence, the query langauge supports the media type of the database. For example:

- VideoText query langauge (Jiang, Montesi and Elmagarmid; 1997)

  VideoText query lanauge was developed based on the VideoText data model. Therefore, the query langauge supports keywords search connected with temporal relations and logical operators.

- PSQL (Pictorial SQL) (Roussopoulos, Faloutsos and Sellis; 1988)

  PSQL retrieves data from pictorial-alphanumeric databases. Therefore, PSQL is an extension of the standard SQL to support abstract data types that are used for defining pictorial and alphanumeric domains.

- VideoSQL (Oomoto & Tanaka; 1993)

  A query langauge based on the OVID for retrievig video objects. VideoSQL is a SELECT-FROM-WHERE query formulated in a fill-in-the-blank manner. The SELECT paragraph is quite different from the ordinary SQL. It specifies only the category of the resulting object, that is, continuous (single frame), incontinuous (sequence of frames), and anyobject (independent objects). FROM used to specify the video name. WHERE specify the condition, consisting of attribute/value pair and comparison operators. As OVID does not support relations between video object or complex objects, VideoSQL provides no relational or boolean operators. OVID is based on keyword annotations, therefore, it supports keyword-match.

- VQL (Video Query Langauge) (Hee, IK and Kim; 1999)

  VQL is an SQL SELECT-like statement, in the form of FIND-FROM-WHERE. In the FIND paragraph, user can specify what he/she wants to retrieve (video document, sequence, scnene or an object). FROM paragraph defines the search field of query and a WHERE paragpraph defines the retrieval condition. The retrieval condition is defined on the basis of attributes, color and spatial-temporal relations on a scene or an object for similarity retrieval. The query language includes operations relevant to their suggested data model. For instance, formulas for returning similarity degree on a scene or an object, and color queries which are not supported by the model proposed in this thesis.

In this thesis, query language aims at showing that the semantic model represented earlier in chapter 4 (Semantic Video Model) facilitates the execution of different types of queries capable of answering human heterogeneous needs and expressing complex concepts with relationships. The author adopts and extends the formal query language based on the first ordered logic notation to build queries to video database (Maier; 1983). Commercial query langauges are more "English-like" and based on some aspects of the formal query language. Any commercial langauges based on the formal query language can be extended to serve the retrieval of video documents presented in this thesis.

The formal query language

The formal query language builds the query language using $\neg$ (not), $\wedge$ (and), $\vee$ (or), $\forall$ (for all), $\exists$ (there exist), | (such that), set of predicates, functions, constants (e.g. 123, red, Ali, etc.), and variables representing semantic units and the values of the content attributes.

A number of predicates are defined in this proposed query language. These are class, association, description, and semantic structure. Some association and description predicates may be created automatically from existing identified semantic units and associations stored in database. In other words, the identification of new class, association or description automatically has its impact on the query language by obtaining a new predicate.

1. *Class* predicates written in upper case letters identify the class to which a semantic unit or an association belongs. For instance, STUDENT(x) and SPONSOR-OF(x, y).

2. *Association* predicates are driven automatically from registered associations. Some *temporal association* predicates (*before, meet, during, overlap, starts, ends, equal*) and *spatial association* predicates (*above, left, between, overlap*) are predefined.

3. *Semantic structure* predicates *actor(x, y), object(x, y)* and *component-of(x, y). Actor(x, y)* determines if the dynamic object x performs the activity y. *Object(x, y)* determines if the activity x is performed on object y. *Component-of(x, y)* predicate determines if x is a sub-component of a composite unit y (not necessarily a direct component). For instance, a `lecturing` activity is a component of a `speech` event, and that is a component of `conference`. However, *component-of*(`lecturing, speech`) and *component-of*(`lecturing, conference`) all return TRUE.

4. *Description* predicates associate a semantic unit with the values of the attribute representing an attribute name in the form *att(x, θ, v)*, where *att* is an attribute's name, *x* is a variable name identifying a semantic unit or an association, *v* is the attribute's value and $\theta \in \{ =, <, >, \neq, \ \ , \ \ , \ \ \}$. Exact match is denoted by = and similar match by . For instance, color(x, , red) indicates that required description values belong to class red.

6.2 Types of Queries

This section presents some examples of queries that may be submitted by end user and answered by the semantic content-based video retrieval system proposed in this thesis. The section shows how these queries can be formally expressed using this proposed query language. Based on the semantic video model described earlier in chapter 4 (Semantic Video Model), user's common needs are expressed in one of the following types of queries: elementary object, elementary activity, elementary event, relational, and compound queries. The five types of queries are defined on top of the proposed semantic model and described furnishing examples in each case. Other types of queries can be submitted by users, as we will explian in section 6.3 (Limitation of the Proposed Query Language). User query is unpredictable. No system can fully capture all user queries but the five types of queries listed below are most common queries user may pose to semantic retrieval system. Other types of queries can be defined with the extension of the semantic model.

1. *Elementary object* query

Form: 'Retrieve a video clip of an *object*'.

Example: 'Retrieve a video clip of a red car'

Formal expression using query language:

$$\{\ x \mid \mathrm{CAR}(x) \wedge \mathrm{color}(x, =, \mathrm{red})\ \}$$

In this example, the end user searches for an object that belongs to class CAR and described by having an exact red color. Color(x,   , red) returns colors similar to red, which may include reddish brown or orange.

2. *Elementary activity* query

Form: 'Retrieve video clip of an *activity*'

'*activity* performed by an *actor*' or

'*activity* performed by an *actor* on *object*'.

Example: 'Retrieve a video clip of a walking man'

Formal expression using query language:

$$\{ x \mid \exists y \ (MAN(y) \wedge WALK(x) \wedge actor(y, x)) \}$$

In this example, the end user is interested in the activity WALK, performed by an object of class MAN. If no actor is specified, the activity is returned regardless of the performer.

3. *Elementary event* query

Form: 'Retrieve video clip of an *event*' or

'Retrieve video clip of an *event* with *component*'.

Example 1: 'Retrieve a video clip of a SIGMOD conference'

Formal expression using query language:

$$\{ x \mid CONFERENCE(x) \wedge name(x, =, SIGMOD) \}$$

The event CONFERENCE in question is described by having the exact name SIGMOD.

Example 2: 'Retrieve a video clip of a conference with editorial presented by Ali'

Formal expression using query language:

$$\{\ x \mid \exists\ y,\ j,\ k$$
$$(\ \text{CONFERENCE}(x) \wedge \text{component-of}(y, x) \wedge$$
$$(\ \text{EDITORIAL}(y) \wedge \text{component-of}(j, y) \wedge$$
$$(\ \text{PRESENT}(j) \wedge\ \text{actor}(k, j) \wedge$$
$$(\ \text{PERSON}(k) \wedge\ \text{name}(k,\ =, \text{Ali})\ )\ )))\}$$

A complex event CONFERENCE is queried in this example. The event is composed of an EDITORIAL sub-event. This sub-event is constructed of a PRESENT activity performed by an object of class PERSON and described by having the name Ali.

4. *Relational* query

Form: 'Retrieve video clip of a *relationship* between semantic-unit$_1$ and semantic-unit$_2$'

Example 1: 'Retrieve a video clip of a man approaches a car from *left*'

Formal expression using query language:

$$\{x\ \ y \mid \exists j\ \ (\ \text{MAN}(j) \wedge\ \text{WALK}(x) \wedge\ \ \text{actor}(j, x)$$
$$\wedge\ \text{CAR}(y)\ \wedge\ \text{left}(x, y)\ )\ \}$$

The query involves a spatial relationship *left* between the activity RUN performed by an object of class MAN and a CAR.

Example 2. 'Retrieve a video clip of the Son of John Kennedy'

Formal expression using query language:

$$\{\ x \ \mid \exists\, y \ \ (\text{PERSON}(x) \ \wedge \text{PERSON}(y)$$

$$\text{name}(y, =, \text{'John Kennedy'}) \wedge \text{son-of}(x, y)\ )\ \}$$

The object in question is referred to through its contextual relationship with another object described by having the name John Kennedy.

Example 3: 'A video clip of a conference with editorial presented by Ali followed by a multimedia lecture'

Formal expression using query language:

$$\{\ x \mid \exists\ y,\ j,\ k,\ z$$

$$(\ \text{CONFERENCE}(x) \wedge \text{component-of}(y, x) \wedge \text{component-of}(z, x) \wedge$$

$$\text{before}(y, z) \wedge$$

$$(\ \text{EDITORIAL}(y) \wedge \text{component-of}(j, y) \wedge$$

$$(\ \text{PRESENT}(j) \wedge\ \ \text{actor}(k, j) \wedge$$

$$(\ \text{PERSON}(k) \wedge\ \ \text{name}(k,\ =, \text{Ali})\ )\ ))\ \wedge$$

$$(\ \text{LECTURE}(z) \wedge \text{subject}(z, =, \text{multimedia})\ )\ )\}$$

This is a query of an event CONFERENCE composed of an EDITORIAL and a LECTURE sub-event. The EDITORIAL event is constituted of a PRESENT activity performed by an object of a class PERSON described as having the name Ali. The LECTURE event is described as being on a multimedia subject.

5. *Compound* query

Involves a number of semantic units and relationships connected by logical operators (AND, OR, NOT).

Example: 'Retrieve a video clip of a book above a table *and* a man walking'

Formal expression using query language:

$$\{\ x \wedge y \mid \exists\ j, z$$

$$(\ \mathrm{TABLE}(x) \wedge \mathrm{BOOK}(z) \wedge \mathrm{above}(z, x) \wedge$$

$$\mathrm{MAN}(j) \wedge \mathrm{WALK}(y) \wedge \mathrm{role}(j, y)\ )\ \}$$

This query is a compound of two queries, the former is a spatial relational query between two objects of class BOOK and TABLE, and the latter is an elementary activity of class WALK performed by an object of class MAN.

Compound queries are executed by decomposing the query into elementary ones. Processing an elementary query produces a set of results in the observation slot form [VID, t_s, t_e]. Based on the connecting logical operators, suitable operations are used to compute the final results. The intersection operation computes the compound of two observation slots connected by AND operator. It takes two observation slots $T(A_1)$ and $T(A_2)$ as input, and returns $T(A_1 \cap A_2)$ as output. The union operation computes the compound of two observation slots connected by OR operator. It takes two observation slots $T(A_1)$ and $T(A_2)$ as input, and returns $T(A_1 \cup A_2)$ as output. The complement operation is performed when the end user asks for all video clips that do not contain a specific semantic unit by preceding it with NOT operand. The query processor returns the result by first finding the observation slot $T(A)$ in which the given semantic unit A appears in, then it returns the complement of A.

6.3 Limitations of the Proposed Query Language

- Performs only semantic similarity and ignores the media-instance that is just like the current instance type of queries such as 'retrieve video clip similar to this sound or picture' (currently viewing). As mentioned earlier in chapter 2 (Current Works in Content-Based Video Retrieval), many perceptual-based retrieval system has been implemented and can be integrated with the system proposed in this thesis.

- Does not support video-based query, which ask for semantic units or relationships that appears in a specific video document. For instance, 'Retrieve semantic units appeared in The Sound of Music'.

- Does not support frame-based query, those asking for semantic units or relationships that appears in a specific sequence of frames. For instance, 'Retrieve all semantic units appeared from frame 1430 to frame 2140'.

- Does not support features-based queries. Some users may not be interested in retrieving video clips but features associated with a given semantic unit. For instance 'Retrieve features associated with a Plane'.

- Does not support relation-based queries. Queries in the form 'Retrieve all relationships associated with a Plane'.

- Query results are video clips of predefined sizes. Unlike VideoSQL, users are not allowed to specify how many frames he/she would like to see in a presentation. The whole video clip is returned for presentation. The work proposed in this thesis concentrate on semantic contents and does not allow formulating conditions which involves operations on frames or time intervals.

The proposed system concentrates on semantic retrieval. Movies, frames, scenes or shots are not considered semantic units. Hence, they are not variables in the presented query language. Also, the result of the proposed query language is video clips. However, the query language can be extended easily to support the previous operations. May be it does not fully express all user's queries, but it supports the objectives of this thesis in its current status.

6.4 The Query Mechanism

The mechanism of query processing is summarized in the following steps:

- End user issues a query;
- Application re-writes the query;
- Application processes the query;
- Application presents the hit list;
- End user selects from the hit list;
- Application streams selected video clip.

6.5 Semantic Video Retrieval Architecture

The architecture for the semantic video retrieval component is built on the top of the semantic video model defined in chapter 4 (Semantic Video Model), which is defined on top the data repositories: video database and meta-database defined in chapter 3 (Content-Based Video Retrieval Architecture), and thesauri. The *thesauri* define relationships between concepts and phrases to support the concept of abstraction defined in chapter 4 (Semantic Video Model) and the concept-based match defined in section 2.6 (Concept-Based versus Keyword-Based Match). Abstraction occurs when a similarity exists between query and retrieved data, but both are not identical. Hence, end users can retrieve documents that contain relevant concepts by expanding queries to include similar or related terms as defined in a thesaurus.

Figure 18 shows the main components of semantic video retrieval system. These are: the video client, video server and repositories. The video server contains the query processor, database management system and video streamer. The following steps illustrate how these components interact to process a query:

1. The end user sends a query from the *video client* user interface to the *video server;*

2. The *query processor* in the video server receives the end user's query, analyzes, parses the syntax, extends using the thesauri and creates a formal query with the same semantic;

3. The created query is sent to the *database management system (DBMS)* for processing;

4. The DBMS poses the query to the meta-database searching for relevant results;

5. The resolved query returns a list of video clip identifiers in the observation slot form [VID, t_s, t_e] for those clips that satisfy the query's constraints;

6. A list of returned clip identifiers is sent to the client and displayed for end user;

7. The end user picks a clip identifier or submits a new query;

8. The client carries the end user's selection to the video server for streaming;

9. The *video streamer* in the video server locates the video clip from the video database and streams back to the client; and

10. The client receives the video stream display, and provides end user's control over the stream playback (stop, play, rewind, forward and pause).

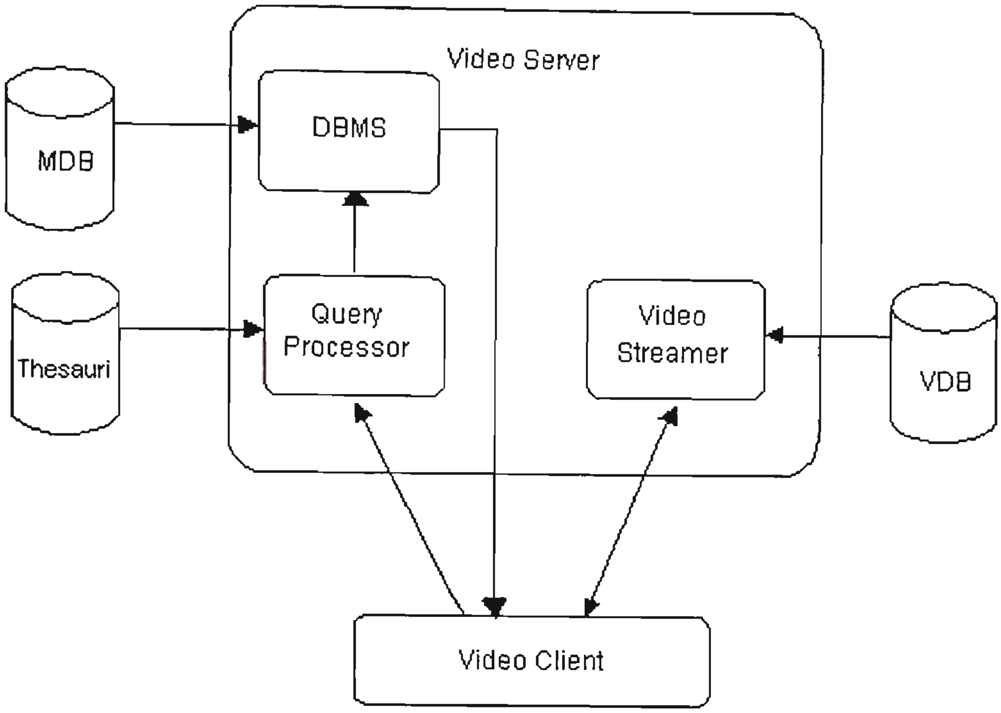


Figure 18. Structure for semantic video retrieval component

6.6 An Interactive User Query

The semantic video model proposed in this thesis has an open set of content attributes, which may cause a schemaless description conflict. With *schemaless description conflict*, end users cannot predict stored content attributes and it is hard to distinguish descriptions from main concepts by automatically parsing the end user's query. Although one could argue that, based on the arrangement of query, it is possible to turn into internal representation and access database. For instance, <description> of <semantic unit> implies that a content attribute precedes the semantic unit. However, it is hard if a mixture of description is derived. For instance, "USA 1990 SIGMOD conference" query consists of an event of class conference, with a mixture of country, year, and name content attributes where no predefined description order could be predicted.

A two-phase interactive user's query obtain process is suggested to resolve the conflict:

Phase 1. End users are asked to think in terms of concepts and relationships interpreted later into semantic units and associations, respectively;

Phase 2. Each semantic unit and association has a particular schema that is tightly attached with a set of content attributes in database. Hence, end users narrow their search to restrict the set of retrieved video clips by selecting from the list of attached content attributes for determination.

Example: 'SIGMOD conference that contains multimedia presentation' query is expressed in two phases.

Phase 1. Identify concepts and relationships: conference *contains* presentation.

Phase 2. Determine description of interest: name(conference, =, SIGMOD) and subject(presentation, =, multimedia).

The two-phase technique should increase precision by taking advantage of the form-based technique and at the same time it narrows down the number of content attributes to be displayed.

6.7 Implementation

A simulator of the semantic video acquisition user interface described in chapter 5 (Semantic Video Acquisition) has been implemented. The human-annotator is provided with the screen to review, modify or acknowledge information extracted from signal processors. Extracted information then are stored in the meta-database described in Appendix D. The screen shown in Figure 11 process scenes. The screen in Figure 13 process elementary semantic units. The human-annotator selects predefined semantic units and manually assigns contextual and structural associations as shown in Figure 15. Figure 17 allows assigning semantics to describe the overall story.

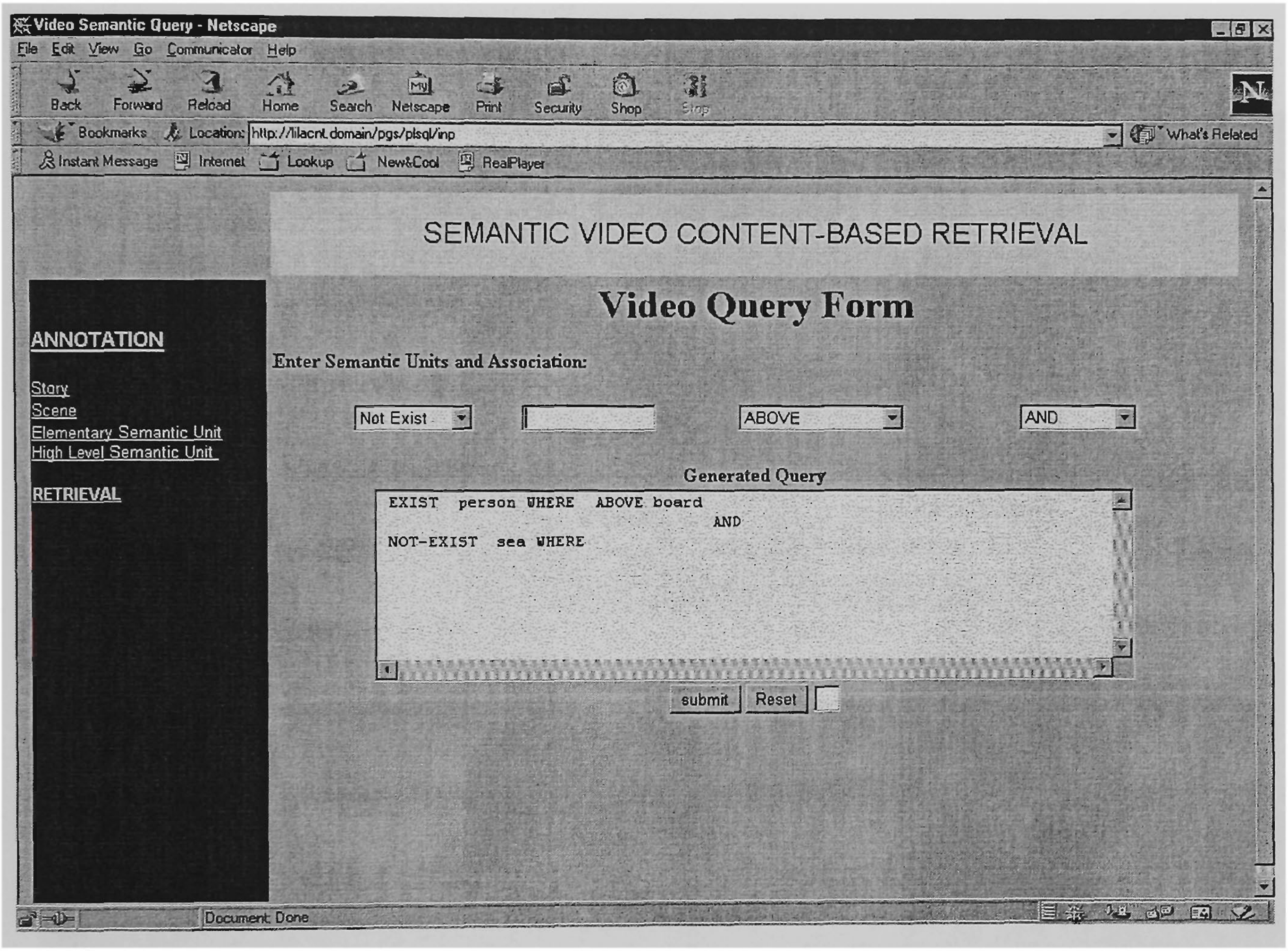
In this chapter, a simulator of the semantic video retrieval user interface is implemented and the query algorithm is described. The simulator is written in PL/SQL and runs on top of Windows NT 4.0 and Oracle Enterprise 8.0.5, Oracle Web Application Server 4.0, Oracle Video Server 3.0.4.2, and Oracle Video Client 3.0.4.2.

The simulator aims at implementing the data repositories, user's interface, the two-phase interactive user's query, the list of query results and video clip streaming.

Among data repositories, video database and meta-databases have been defined and implemented. Appendix D shows the relational database representing the semantic video model. As thesauri have not been implemented, this version of the simulator performs a keyword-based match rather than concept-based.

When a user invokes the semantic video retrieval, a screen comes up as shown in Figure 19. The user enters the query in terms of semantic units and association (phase 1 of the interactive user's query). In this version of the simulator, only association and semantic structure predicates are defined. The entry is executed when the user presses the **submit** button. Figure 19 shows a specific query requesting all video clips where a `person` *above* a `board` and no `sea` exist. This is a compound type of query where a relational and an object query are connected with AND and NOT operators. The result of the query is a screen with all content attributes attached to each semantic unit specified in phase 1 for determination (phase 2 in the interactive user's query). Figure 20 shows a list of attributes attached to semantic units appeared in phase 1 (`person,` `board` and `sea`). User selects attributes and assigns values. Figure 20 shows `name` and `role` as attributes attached with `person`. Attributes display automatically based on meta-data stored for each semantic unit. User set the person's name into Tony. The query is executed by pressing the **submit** button. A list of the identifiers of the matching video clips is returned to the end user. Figure 21 list identifiers of matching the query 'person *above* a board and no sea exist. The person name is Tony'. The user may select any of the returned identifiers and click the **play** button for streaming the selected video clip. Figure 22 shows the streaming of the selected video clip and the control provided to the end user over the stream playback.

The implementation suggests that the proposed semantic model and the theoretical algorithms work in practice.


Video Semantic Query - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://lilacnt.domain/pgs/plsql/inp
What's Related
Instant Message Internet Lookup New&Cool RealPlayer
ANNOTATION
Story
Scene
Elementary Semantic Unit
High Level Semantic Unit
RETRIEVAL
Document: Done

SEMANTIC VIDEO CONTENT-BASED RETRIEVAL

Video Query Form

Enter Semantic Units and Association:

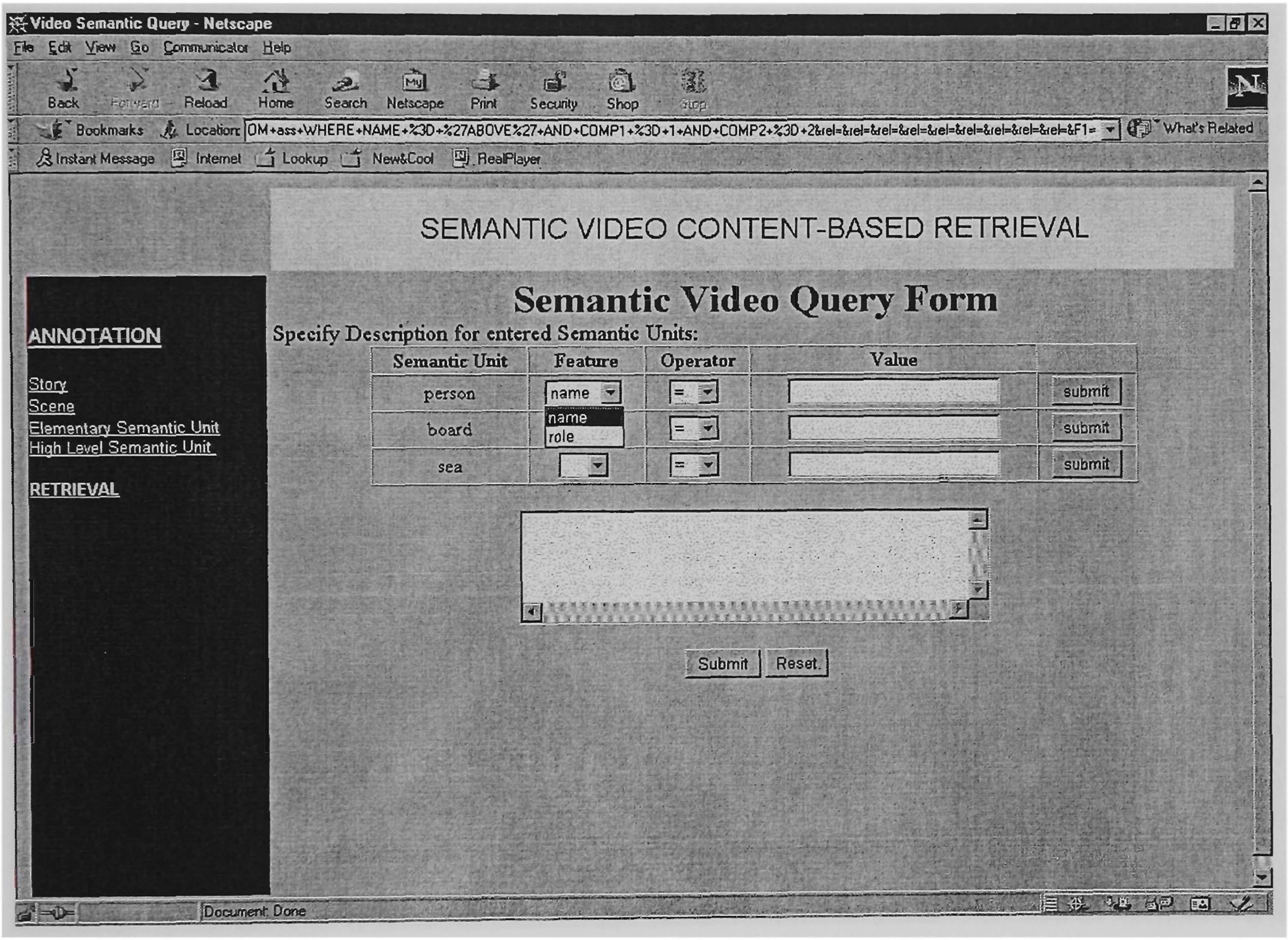
Not Exist ▾ | | ABOVE ▾ | AND ▾

Generated Query

```
EXIST  person WHERE  ABOVE board
                                  AND
NOT-EXIST  sea WHERE
```

submit Reset

Figure 19. Form for accepting a user query in terms of concepts & relationships


Video Semantic Query - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location:
What's Related
Instant Message Internet Lookup New&Cool RealPlayer
Document: Done

Location: OM+ass+WHERE+NAME+%3D+%27ABOVE%27+AND+COMP1+%3D+1+AND+COMP2+%3D+2&rel=&rel=&rel=&rel=&rel=&rel=&rel=&rel=&rel=&F1=

SEMANTIC VIDEO CONTENT-BASED RETRIEVAL

ANNOTATION

Story
Scene
Elementary Semantic Unit
High Level Semantic Unit

RETRIEVAL

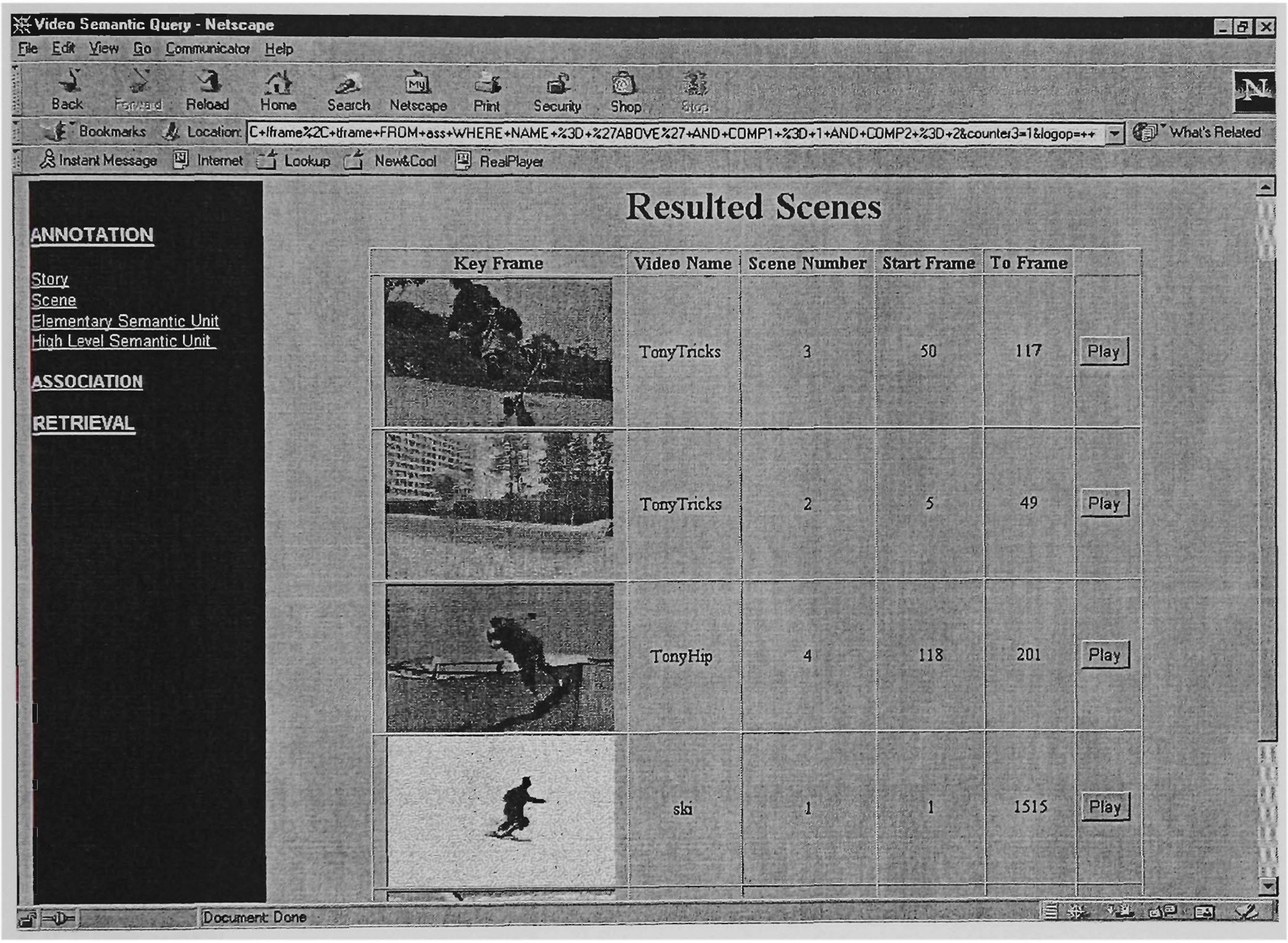
Semantic Video Query Form

Specify Description for entered Semantic Units:

Semantic Unit	Feature	Operator	Value	
person	name name role	=		submit
board		=		submit
sea		=		submit

Submit Reset.

Figure 20. Form to select and assign content attributes

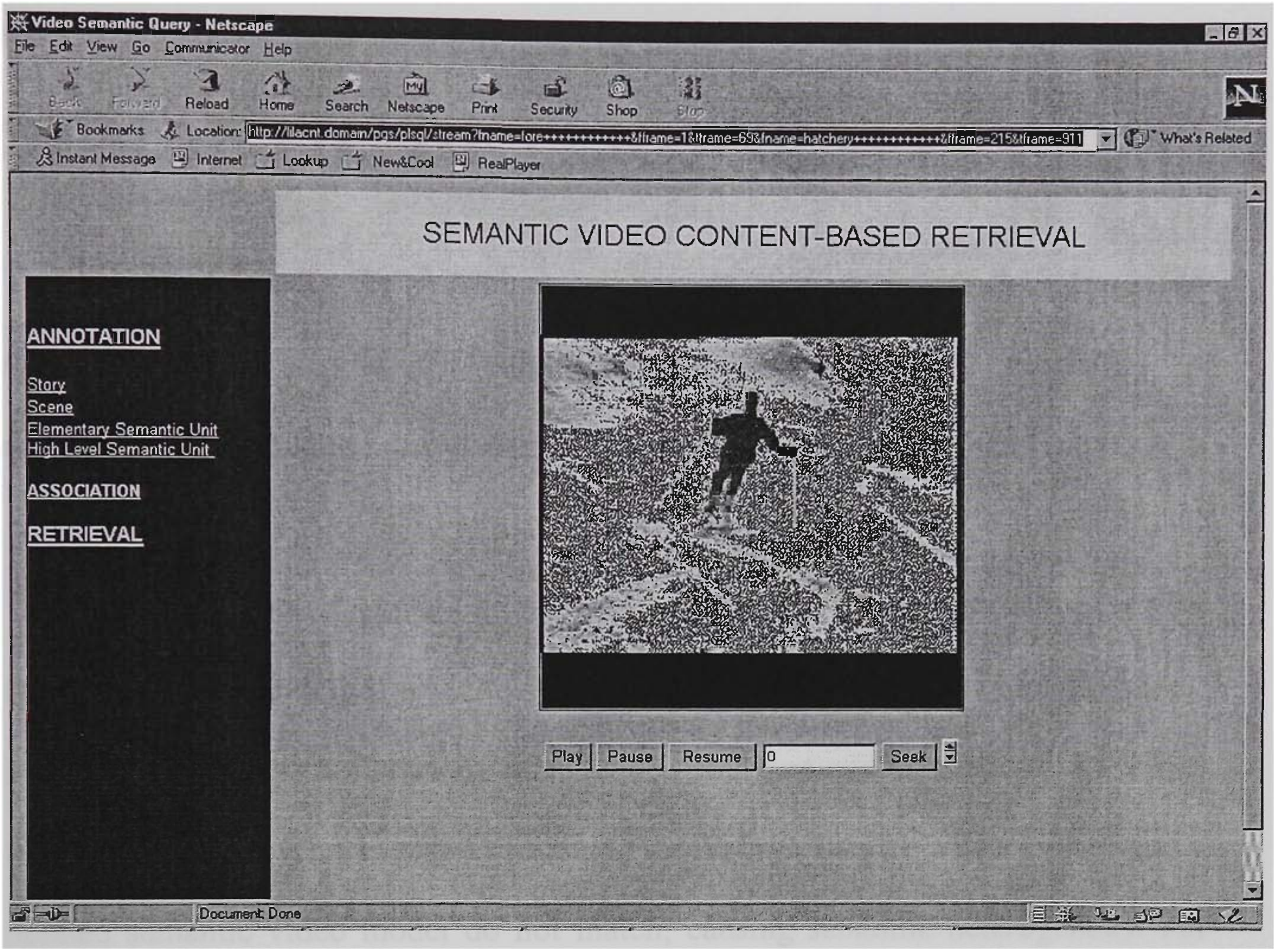


Video Semantic Query - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: C+lframe%2C+tframe+FROM+ass+WHERE+NAME+%3D+%27ABOVE%27+AND+COMP1+%3D+1+AND+COMP2+%3D+2&counter3=1&logop=++
What's Related
Instant Message Internet Lookup New&Cool RealPlayer
ANNOTATION
Story
Scene
Elementary Semantic Unit
High Level Semantic Unit
ASSOCIATION
RETRIEVAL
Document: Done

Resulted Scenes

Key Frame	Video Name	Scene Number	Start Frame	To Frame	
	TonyTricks	3	50	117	Play
	TonyTricks	2	5	49	Play
	TonyHip	4	118	201	Play
	ski	1	1	1515	Play

Figure 21. The list of returned video clips identifiers displayed for the end user


Video Semantic Query - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location:
What's Related
Instant Message Internet Lookup New&Cool RealPlayer
ANNOTATION
Story
Scene
Elementary Semantic Unit
High Level Semantic Unit
ASSOCIATION
RETRIEVAL
Play Pause Resume Seek

http://lilacnt.domain/pgs/plsql/stream?fname=fore++++++++++++&fframe=1&lframe=69&fname=hatchery++++++++++++&fframe=215&lframe=911

SEMANTIC VIDEO CONTENT-BASED RETRIEVAL

0

Document: Done

Figure 22. Streaming the video clip and providing user control over the stream

6.8 Summary

This chapter described a semantic video retrieval module that allows the retrieval and the stream of video clips. A query language has been proposed and developed to allow users to pose their queries in terms of semantic units, associations and descriptions.

7 SEMANTIC HETEROGENEITY

This chapter discusses the possible heterogeneity between a user model and the video content defined in the video database, and studies the possibility of eliminating the semantic heterogeneity between query content and video content.

In order to construct a query, an end user imagines for himself/herself a situation (user's semantic model) of what he/she is going to search using the query language. In general, end users are unaware of the video structure and annotations defined in the video database. This is why a semantic model constructed in a user's mind and a defined semantic video model do not match, causing *semantic heterogeneity.* An obvious consequence is imprecise results.

Why study semantic heterogeneity?

Current semantic video retrievals assume that users are aware by default of the semantic video model and do not address semantic heterogeneity. For instance, in UVRS (Hee, IK and Kin 1999), end users should be familiar with the suggested logical structure in order to query. Such assumption is too restrictive because it forces end users to discover the structure of the video database before any query may be constructed, which is unacceptable for untrained end users.

In the model proposed in this thesis, considerable semantic heterogeneity may occur between the user semantic model and the semantic video model defined in the database because of:

- The semantic language constraints.
- The heterogeneous views of users when it comes to describing a clip.
- The imprecise semi-automatic video analysis with the involvement of human annotators.
- The open semantic video model where no predefined indexes can be predicted for video content as new classes, associations and descriptions are created during semantic video acquisition.

7.1 View Mapping

Semantic heterogeneity is resolved by finding a map between the user semantic model and the semantic video model. *View mapping* is the process of aligning the user semantic model depicted in a query with the semantic video model to make them match. Suppose a query Q is given. $Q \Rightarrow Q'$ is read as Q mapped to Q', where Q' is an intermediate model semantically similar to Q, with a schema equivalent to the video model. View mapping allows a great deal of flexibility over the semantic video model where users query the video database with no need of pre-knowledge of the video structure or the annotations stored. This flexibility also enables reliable match and retrieval. *Reliable* means that end users should be able to retrieve documents that have the most potential of being relevant to their queries.

While query processing, the user and the semantic video models map could be thought of as follows:

- Concepts are detected from the query to search similar semantic units in the video database
- Descriptions and relationships are captured from the query to search the video database for similar values of content attribute and associations, respectively.
- The order of the concepts in the query should specify the order of the arguments of an association and the interclass relationship. For instance 'conference editorial' denotes an `editorial` as a component of a `conference`.

View mapping goes through a number of steps:

- *Conflict analysis,* which is the process of detecting and extracting all possible differences between the user and the semantic video models.

- *Resolution,* which aims at resolving detected conflicts by extending one model to conform to the other.

- *Matching,* which takes the resolved output as an input representing the query, and pose to the meta-database for a traditional map of extended query and for obtaining results.

7.2 Conflict Analysis and Proposed Resolutions

In the proposed model, two types of conflicts are distinguished: naming and structure conflicts.

7.2.1 Naming Conflict

It arises between semantically similar entities with different names. A number of naming conflicts are distinguished: abstraction, primary attributes and spatial-temporal associations.

1. *Abstraction:* An important aspect in the determination of a semantic unit, association and description similarity is abstraction where, for instance, a `man` could be referred to as a `person`, a `father-of` as a `sponsor-of`, and `reddish brown` as `red`. What makes this proposed semantic model different is that abstraction is spread over all names of different concepts and not only over objects as in most current works in semantic modeling. Whenever a name can be used, such as in object, activity, event, description or associations, it may have

different semantics associated with it, for instance, a `client` may refer to a `patient` and a `customer`, and, alternatively, the same semantics may have different names such as `human` and `person`.

Performing a class match, which goes beyond name matching into semantically similar concepts may resolve abstraction conflicts. Two classes could be semantically related if they were mapped to the same taxonomy concept. As elaborated in section 2.4 (Concept-Based vs. Keyword-Based Match), many works have addressed concept-based match. Concept-based match is beyond the scope of this work. Techniques from knowledge representation and natural language processing can make a useful contribution to solving the abstraction conflict.

2. *Primary attributes*: In matching a concept of a query with a semantic unit, the concept could be referred to by its *class name* or one of its *attributes,* such as referring to an object of class `person` as `Ali` (name attribute) or `professor` (specialty attribute). An obvious consequence is that no matching results.

A possible solution is to use the set of primary attributes to be part of the concept search. During video analysis, possible attributes that could be set as key attributes are identified. In the previous example, the extracted concept is matched against the name and specialty attributes as well as the class name. Suppose *A* is a semantic unit and *B* is a query concept. $B \in$ {*class(A)* U *domain*(key-attribute(*A*))} where key-attribute={name, specialty}.

3. *Spatial-Temporal associations*: Some associations are *hyponym* where the same relation gives different meanings in space and time. For instance, *contain, overlap* and *between* are meaningful and exist for any semantic unit that may project its position in either space or time. Whereas, 'fire *overlap* accident', which is semantically correct in both dimensions, gives different meanings.

This conflict is caused by language ambiguity and it is not related to semantic units of specific level of granulation. One solution is that these dimension-dependent associations need identification of dimension name (time or space) in order to resolve the ambiguity of relations. For instance they should be specified as overlap in time dimension, or contain in space dimension. Each association *R* is defined as a pair of *(n, d)*, where *n* is the name, *d* is the dimension associated with *R*, and $d \in$ {time, space}

7.2.2 Structure Conflict

This arises as a result of a different construction of a query which could be different than the structure of the defined video model. Different structures could be semantically similar.

The conflict is resolved by transforming one model into the other in a way it does not change the imbedded semantic. Possible structure conflicts in the proposed semantic model are: virtual-discrete associations, activity-association conflict, layers of granulation and the schemaless descriptions elaborated in section 6.5 (An Interactive User Query).

1. *Virtual-Discrete associations*: Some semantic associations are incorporated into the semantic video model and are not specified as a discrete association. They are *virtual* since they are defined with an interclass association. However, end users may interpret an interclass connection as a discrete association. For instance, a 'man running' with an activity-actor interclass association can be described as 'run *performed-by* a man', as depicted in Figures 23.a and 23.b, respectively. And 'conference speech' with an aggregation interclass connection can be queried as 'speech *part-of* a conference', as depicted also in Figures 23.a and 23.b, respectively.

Assume A is an activity performed by O, then $A(O) \equiv R_X(A, O)$, where $R_X \in \{$ performed-by, done-by, ...$\}$. Assume C is a composite unit composed of c_i, then $component\text{-}of(c_i, C) \equiv R_Y(c_i, C)$, where $R_Y \in \{$ part-of, component-of, ...$\}$.

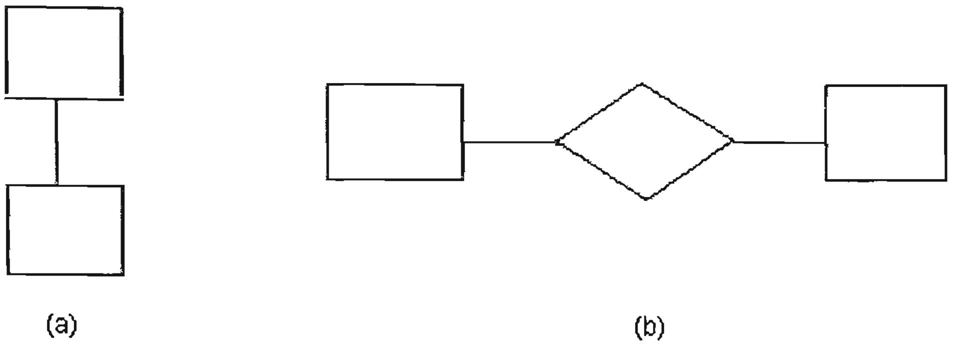


Figure 23. (a) Interclass (virtual) association (b) Discrete association

Auto-inference of activity: Given the query 'run performed-by a man' where the semantic video content is stored as 'man running', an algorithm should be developed to infer an actor-activity interclass association from the unmatched query with the discrete association.

Example:

'Run performed-by a man'

Formal expression using query language:

$$\{\ y \mid \exists\ x\ (\ \mathrm{MAN}(x) \wedge \mathrm{RUN}(y) \wedge \text{performed-by}(y, x)\ )\ \}$$

Formal expression using database tuples:

Based on the entered query, searched tuples are expressed as:

$(s_1, \mathit{uid}, t_1, \{\ \text{class}\ \}, \{\ \text{man}\ \}, \varepsilon, \lambda)$

$\partial_1(\ \text{class}\ ) = \text{man}, \varepsilon(s_1) = \{\ \partial_1\ \},\ \lambda(s_1)=t_1$

$(s_2, \mathit{uid}, t_2, \{ \text{class} \}, \{ \text{walk} \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ walk}, \varepsilon(s_2) = \{ \partial_1 \}, \lambda(s_2) = t_2$

The discrete association *A* is expressed by the following tuple:

$(A, \mathit{uid}, t, \{ \text{class, operand1, operand2} \}, \{ \text{performed-by}, s_1, s_2 \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ performed-by}, \partial_2(\text{ operand1 }) = s_1, \partial_3(\text{ operand2 }) = s_2$

$\varepsilon(A) = \{ \partial_1, \partial_2, \partial_3 \}, \lambda(A) = t$

On the other hand, the interclass association *S* in 'man running' is stored in the meta-database as:

$(S, \mathit{uid}, \text{t}, \{ \text{class, actor} \}, \{ \text{walk}, s_1 \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ walk}, \partial_2(\text{ actor }) = s_1$

$\varepsilon(S) = \{ \partial_1, \partial_2 \}, \lambda(S) = \text{t}$

Method:

When a discrete association returns no match in the meta-database, the following procedure is applied to infer an activity-actor association from a discrete association, if possible.

```
$\forall$ $A$ NOT FOUND
    IF $A$.class $\in$ {performed-by, done-by, …} then
        $\exists$ $S$ : $S$.actor = $A$.operand1 $\wedge$ $S$.class = $A$.operand2
```

Auto-inference of aggregation: Given the query ‘speech part-of a conference’ where the semantic video content is defined as ‘conference speech’, an algorithm should be developed to infer an aggregation structure from the unmatched query with discrete association.

Example:

‘Speech part-of a conference’

Formal expression using query language:

$$\{\ y \mid \exists\ x\,(\text{SPEECH}(x) \wedge \text{CONFERENCE}(y) \wedge \text{part-of}(\ x, y)\ )\ \}$$

Formal expression using database tuples:

Tuples in question are expressed as:

$(s_1, \mathit{uid}, t_1, \{\text{ class }\}, \{\text{ speech }\}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{speech}, \varepsilon(s_1) = \{\ \partial_1\ \},\ \ \lambda(s_1)=t_1$

$(s_2, uid, t_2, \{\text{ class }\}, \{\text{ conference }\}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{conference}, \varepsilon(s_2) = \{\ \partial_1\ \},\ \ \lambda(s_2)=t_2$

The discrete association *A* is expressed by the following tuples:

$(A, uid, t, \{\text{ class, operand1, operand2 }\}, \{\text{ part-of, } s_1, s_2\}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ part-of}, \partial_2(\text{ operand1 }) = s_1, \partial_3(\text{ operand2 }) = s_2$

$\varepsilon(A) = \{\ \partial_1, \partial_2, \partial_3\ \}, \lambda(A) = t$

The virtual association *S* representing 'conference speech' is expressed in the meta-database by the following tuple:

$(\mathrm{S}, uid, \mathrm{t}, \{\text{ class, component }\}, \{\text{ conference, } s_1\}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{conference}, \partial 2(\text{ component }) = s_1,$

$\varepsilon(\mathrm{S}) = \{\ \partial_1, \partial_2\ \},\ \ \lambda(s_2)=\mathrm{t}$

Method:

When a discrete association returns no match in the meta-database, the following procedure is applied to infer an aggregation structure.

```
            ∀    A    NOT FOUND

    IF A.class ∈ {part-of, component-of, …} then

  ∃ S : S.class = A.operand1 ∧ S.comp = A.operand2
```

2. *Activity-Association conflict*: Some semantic associations can be interpreted as activities, and vice versa. For instance, a 'man write a book' or a 'book written-by a man'. In the former example, `write` is an activity performed by a `man` on a `book` illustrated in Figure 24.a, while in the second example `written-by` is an association between two objects illustrated in Figure 24.b.

As a solution, for unmatched associations, activities should be automatically inferred for resolving the possible activity-association conflict. Suppose o_1 and o_2 are objects, A is an activity, then $A(o_1,o_2) \equiv R_A(o_2, o_1)$, where R_A is an association inferred from A.

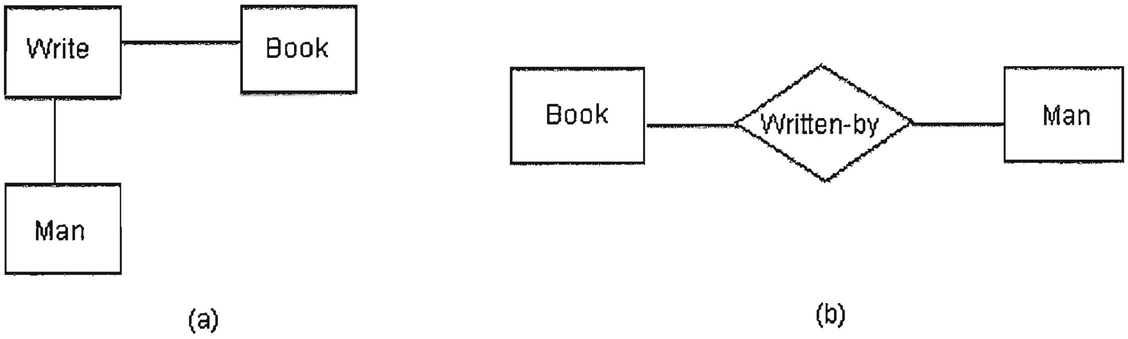


Figure 24. (a) Activity (b) Discrete Association

Auto-inference of an activity: An activity and an association are considered semantically equivalent when they represent the same real-world concept and a map can be established between attributes of both the activity and the association. The

ACTIVITY( ) function is defined to return the activity inferred from an association. For instance, *ACTIVITY(*`written-by`) = `write`. Suppose S is a defined association. β maps observation slots of two semantic units T, content attributes F and their values V such that β: $S \rightarrow A$ is interpreted as ($A.t_1 = S.t_1$, ..., $A.t_n = S.t_n$, $A.f_1 = S.f_1$, ..., $A.f_m = S.f_m$, $A.v_1 = S.v_1$, ..., $A.v_m = S.v_m$) where $t_i \in T$, $f_i \in F$, and $v_i \in V$.

Given the query 'book written-by a man' where the semantic video content is defined as 'man write a book', an algorithm is needed to infer an activity from the discrete association.

Example:

'book written-by a man'.

Formal expression using query language:

$$\{\ x\ \ y \mid (\ \mathrm{BOOK}(x) \wedge \mathrm{MAN}(y) \wedge \text{written-by}(\ x,\ y)\ )\ \}$$

Formal expression using database tuples:

Tuples in question are expressed as:

(s_1, *uid*, t_1, { class }, { book }, ε, λ)

∂_1(class) = book, $\varepsilon(s_1) = \{\ \partial_1\ \}$, $\lambda(s_1)=t_1$

(s_2, *uid*, t_2, { class }, { man }, ε, λ)

∂_1(class) = man, $\varepsilon(s_2) = \{ \partial_1 \}$, $\lambda(s_2) = t_2$

The discrete association *A* is expressed as:

(*A*, *uid*, *t*, { class, operand1, operand2 }, { written-by, s_1, s_2 }, ε, λ)

∂_1(class) = written-by, ∂_2(operand1) = s_1, ∂_3(operand2) = s_2

$\varepsilon(A) = \{ \partial_1, \partial_2, \partial_3 \}$, $\lambda(A) = t$

The activity *S* in the 'man writes a book' is expressed by the tuple:

(S, *uid*, t, { class, actor, object }, { write, $s_{1,}$ s_2 }, ε, λ)

∂_1(class) = write, ∂_2(actor) = s_2, ∂_3(object) = s_1

$\varepsilon(S) = \{ \partial_1, \partial_2, \partial_3\}$, $\lambda(S) = t$

Method:

When a discrete association returns no match in the meta-database, the following procedure is applied to infer activity from discrete associations.

```
IF S  NOT FOUND Then
    ∃ A : β(S) ∈ A ∧
    A.class= ACTIVITY(S.class) ∧
    A.actor = comp2 ∧ A.object = comp1
```

3. *Layers of granulation*: Composite units are built from other units of different granulation. Therefore, different representations of different granulation maybe expected for a composite semantic unit. For instance, in Figure 24, `speech` could be referred to in a lower granulation as `X present` or in a higher granulation as `conference`. This conflict could be resolved by a top-down or a bottom-up refinement process that can find super and subclasses in a hierarchy.

Since associations are maintained between semantic units of specific granularity during video analysis, they no longer exist if units were accessed across different layers of granulation. On the other hand, the same association is semantically true for all its components. For instance, 'editorial *before* speech' implies 'editorial *before* X present'.

The suggested solution is to permit association among components as depicted by Figure 26. Let A be a composite unit and a_i be the *ith* component. Semantically, if A is related to B denoted by $R(A, B)$, then all of its components a_i must be related to B as well as $R(a_i, B)$. Consider for example the situation where we have the query 'editorial *before* X present' and the aggregation structure of the semantic unit `conference` is defined as shown in Figure 25.

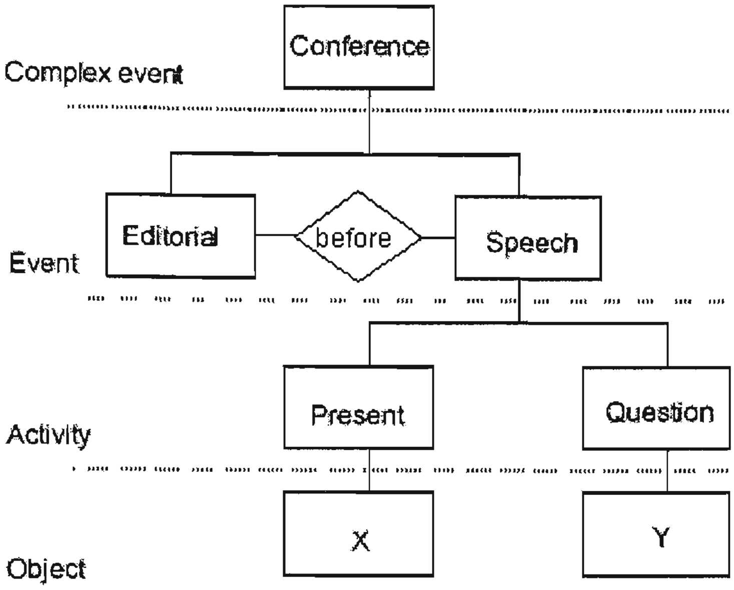


Figure 25. Semantic Unit Layers of Granulation

Example:

'Editorial before Speech'

Formal expression using database tuples:

(s_1, *uid*, t_1, { class, name }, { person, X }, ε, λ)

∂_1(class) = person, ∂_2(name) = X,

$\varepsilon(s_1) = \{ \partial_1, \partial_2 \}$, $\lambda(s_1)=t_1$

$(s_2, \mathit{uid}, t_2, \{ \text{class, actor} \}, \{ \text{present}, s_1 \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ present}, \partial_2(\text{ actor }) = s_1,$

$\varepsilon(s_2) = \{ \partial_1, \partial_2 \}, \lambda(s_2) = t_2$

$(s_3, \mathit{uid}, t_3, \{ \text{class} \}, \{ \text{editorial} \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ editorial}, \varepsilon(s_3) = \{ \partial_1 \}, \lambda(s_3) = t_3$

$(s_4, \mathit{uid}, t_4, \{ \text{class, name} \}, \{ \text{person, Y} \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{person}, \partial_2(\text{ name }) = \text{Y},$

$\varepsilon(s_4) = \{ \partial_1, \partial_2 \}, \lambda(s_4)=t_4$

$(s_5, \mathit{uid}, t_5, \{ \text{class, actor} \}, \{ \text{question}, s_4 \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ question}, \partial_2(\text{ actor }) = s_4,$

$\varepsilon(s_5) = \{ \partial_1, \partial_2 \}, \lambda(s_5) = t_5$

$(s_6, \mathit{uid}, t_6, \{ \text{class, comp1, comp2} \}, \{ \text{speech}, s_2, s_4 \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ editorial}, \partial_2(\text{ comp1 }) = s_2, \partial_3(\text{ comp2 }) = s_4$

$\varepsilon(s_6) = \{ \partial_1, \partial_2, \partial_3 \}, \lambda(s_6) = t_6$

$(A, uid, t, \{ \text{class, operand1, operand2}\}, \{ \text{before}, s_3, s_6 \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ before}, \ \partial_2(\text{ operand1 }) = s_3, \partial_3(\text{ operand2 }) = s_6$

$\varepsilon(A) = \{ \partial_1, \partial_2, \partial_3 \}, \lambda(A) = t$

Example:

'Editorial *before* X present' and 'editorial *before* Y question'

Formal expression using database tuples:

$(A_1, uid, t_1, \{ \text{class, operand1, operand2}\}, \{ \text{before}, s_3, s_2 \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ before}, \ \partial_2(\text{ operand1 }) = s_3, \partial_3(\text{ operand2 }) = s_2$

$\varepsilon(A_1) = \{ \partial_1, \partial_2, \partial_3 \}, \lambda(A_1) = t_1$

$(A_2, uid, t_2, \{ \text{class, operand1, operand2}\}, \{ \text{before}, s_3, s_5 \}, \varepsilon, \lambda)$

$\partial_1(\text{ class }) = \text{ before}, \ \partial_2(\text{ operand1 }) = s_3, \partial_3(\text{ operand2 }) = s_5$

$\varepsilon(A_2) = \{ \partial_1, \partial_2, \partial_3 \}, \lambda(A_2) = t_2$

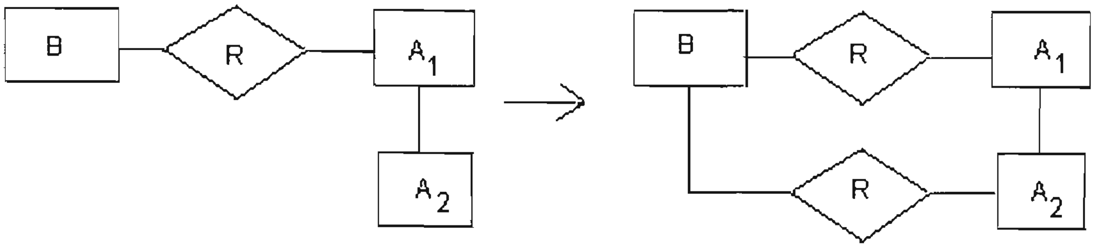


Figure 26. Permitting association among components

Relating components algorithm: The following algorithm permits association among components. We define *component(A)* function to return a list of all *A* components.

```
Algorithm relate-comp(A, B, R): boolean
begin
      SET $R(A, B)$
      IF $component(A)$ = NIL STOP
         ELSE $\forall\ a_i \in component(A)$
            relate-comp($a_i$, B, R)
end
```

The suggestion of an association connecting a composite unit to be enforced on all its components works for spatial and temporal associations involving semantic units

E_1 and E_2 when $T(E_1) \cap T(E_2) = \varnothing$, while it may not always return an accurate result with spatial and temporal associations when $T(E_1) \cap T(E_2)$ $\varnothing$ because:

- The observation slot or bounding volume of the components can be included, but they are not equal. Hence, for instance, $start(A, B) \neq start(a_i, B)$.

- The observation slot of a component can be longer than the observation slot of the composing unit, or the bounding volume of a component can be greater than the bounding volume of the composing unit. Hence, for instance, $simultaneously(A, B) \neq simultaneously(a_i, B)$.

Therefore, during video analysis, both spatial and temporal associations are automatically captured for all semantic units appearing at an instance of components and are not permitted among them.

The search algorithm: The following algorithm searches for the association among different granulation layers. These algorithms do not automatically explore more general terms. For example, if a user asks for conference, the system may get editorial, speech, X present and Y question. But if a user asks for speech, the system gets X present and Y question only.

```
Algorithm match(A, B, R): boolean
begin
        match = false
        IF $\exists$ R(A, B)  return  match = true and STOP
         ELSE IF component(A) = NIL STOP
             ELSE $\forall$ component(A) match(component(A), B, R)
end
```

7.3 Summary

The major contribution of this chapter is to list all possible semantic and schematic conflicts between the user view and the video content. The author proposes approaches for resolving these conflicts through mapping the user view into a semantic video model. This alignment of both user and video models is important for a reliable match and retrieval.

8
CONCLUSION AND FUTURE TRENDS

8.1 Conclusion

The objective of this work is to develop a content-based retrieval system for video documents based on their semantic content. In developing the system, it is essential to define a rich and sophisticated conceptual model powerful enough to describe the semantic content of video documents and to answer users' heterogeneous queries. This thesis develops a semantic video model based on the story-line structure of video, which encompasses objects, activity, events and story. The proposed model extends the plethora of already proposed symbolic modeling tools by the recognition of higher granulation of semantic units and by allowing associations to be defined over them in order to build high-level semantics. Another extension allows for the application of abstraction mechanisms to any type of semantic units, description or association unlike other models which can be applied only to objects.

This thesis has developed a computer-aided analyzer where a human operator, supported by processing techniques, plays a central role in the semantic indexing of video documents. A major step towards implementation is the formal specification of the video conceptual model and the human perception of the content of the video.

8.2 Characteristics of the Proposed Model

The proposed model has a number of characteristics:

- It captures the underlying semantic structure of video documents.
- It provides a representation of high-level semantics for a detailed description of video documents.
- The semantic video structure conforms with the user's perspective and facilitates heterogeneous queries.
- The video content comprises media-independent concepts.
- It provides an open set of annotations. Semantic units and associations are not predefined and have no fixed description schema.
- It shares and reuses semantic units and associations as they may appear in several different video documents.
- The concept of composite semantic units adds extensibility to the video model. This concept is lacking in most current conceptual models.
- The model considers the possible fuzziness in the user's query.
- The concept of abstraction is assigned to semantic units, descriptions and associations.
- There is variable access granularity and representation for a semantic unit.
- It considers the interrelationship between semantic units in various video spaces.
- High-level semantic layers are not totally independent of the physical layer. This will facilitate working on automating semantic acquisition.

8.3 Future Trends

Semantic content-based video retrieval is an active and exciting research area with a wealth of contributing trends in digitizing and processing techniques, knowledge bases, user friendly query languages and much more. It is impossible to cover all these trends in this thesis. Hence, this section highlights some of the future directions.

Video Acquisition and Annotation Interface: One of the major problems of manual annotations is the anticipation of *what* kind of information to record and *how* to record it. It is impossible to tag everything in every way. Having a standard for video annotation and offering intelligent assistance to manual annotation and an easy-to-grasp annotation user interface are essential for annotation precision and consistency. This thesis has presented a graphical model that could be a step towards a graphical user interface.

Automating Semantic video Acquisition: This thesis suggests a semi-automatic video analyzer and content acquisition. As semi-automatic analysis of a video document is tedious, time consuming and imprecise, it is necessary to go further toward automating a high-level semantic analysis of video documents. A major step toward the automation of semantic video acquisition is to have a standard for a semantic video model: this is the aim of this thesis. The integration of a knowledge base of inference rules that describes the construction of events and composite units into the system, along with the consideration of the information exported from all media streams, we believe, will help to identify semantic content that is beyond processing techniques. This will be a step toward minimizing manual processes and automating semantic video acquisition.

User Friendly Video Retrieval User Interface: As consumers do not easily adopt complicated retrieval technologies, designing an efficient user-friendly query interface that is human-oriented is important. In this work, a methodology has been suggested for a gradual interactive query, starting from the query's central concepts and their interrelationship, then giving a description of these selected concepts and relationships in order to overcome the problem of schemaless description. A formal query language has been proposed. Yet an easy-to-grasp user interface needs to be designed for the proposed methodology. Developing the interface could be a research project in itself.

Semantic Units Spatial Relationships: To the best of the author's knowledge, current works have been limiting spatial relationship to frame-based semantic units, i.e. objects (Sistla, Yu and Haddad; 1994, Gudivada & Jung; 1996, Liu & Sun; 1997, Li, Özsu and Szafron; (1996, 1997), Orenstein & Manola; 1988). In reality, a spatial relationship may exist between higher level semantic units, such as in 'accident *under* a bridge', where `accident` is an event over a sequence of frames.

Hence, a research is suggested to investigate how it is possible to compute the spatial relationships between high-level semantic units.

Aural Semantic Units Spatial Attributes: Many image processing works have been conducted to capture and to represent the position of the visual object in a frame (Orenstein & Manola; 1988, Sistla, Yu and Haddad; 1994, Gudivada & Jung; 1996, Liu & Sun; 1997, Li, Özsu and Szafron; (1996, 1997)). Aural semantic objects refer to a physical or virtual object in the video (objects producing sounds). Predicating the position of salient objects presented aurally is needed.

A human recognizes the distance of an aural object by referring to the non-spatial attribute values, which aid in approximating its position. Variation in sound amplitude is what causes a sound to be loud or soft. Distance increases with the decrease in sound amplitude. In other words, a long distance is reflected by a lower sound and vice versa. Hence, the amplitude attribute for the aural semantic object projects its position on the Z axis (distance).

Sounds may appear concurrently with the visual generator, such as the image of a phone with a sound of ring, or may occur independently, irrelevant of a visible source (virtual generator), such as the sound of a clock ringing or of footsteps. In both cases, the sound describes a semantic object, and its position in the video stream. It is well known that a sound becomes audible as loudness increases with an approaching sound source. Hence, the source of a sound is the frame where sound was loudest.

By determining the X and Y axis of an aural semantic unit, and by investigating the way in which human recognizes the direction of the sound generator, the following is found. In real-life, if there is any object in the vicinity, its direction may be determined by natural tendency. For instance, from the sound we can tell the direction of the sound generator (left, above, ...), but that is not true for videos. The existence of some invisible object can be sensed and deducted but are limited in distinguishing the source direction and determining its spatial position. The human ear is what deducts the direction of the sound; however video sounds are delivered from the same source and direction, which is the speakers' direction. That is why sounds suffer from a restricted concurrently directional expressiveness ability. So semantic objects represented aurally have no directional relations.

A specific advantage of the model proposed in this thesis is the ability to combine and formulate requests in a high-level abstraction regardless of the representation media type. This is an open research area for whoever wishes to investigate the possible relations and representations among incompatible media units, such as the spatial relationship between an aural and a visual salient object.

BIBLIOGRAPHY

Abdel-Mottaleb, M., Dimitrova, N., Desia, R. and Martino, J. (1996). CONIVAS. Content-based Image and Video Access System. In proceedings of the Fourth International ACM Multimedia Conference 96.

Adali, S., Candan, K. S., Chen, S., Erol, D. K. and Subrahmanian, V. S. (1996). The Advanced Video Information System, Data structure and query processing. Multimedia Systems, 4.

Aggarwal, J. and Cai, Q. (1999). "Human Motion Analysis: A Review". Computer Vision and Image Understanding, 73(3).

Allen, J. F. (1993) . Maintaining knowledge about temporal intervals. Communications of the ACM 26(11).

Al-Safadi, L. and Getta, J. (1999). Video Semantic Model & Acquisition. IIWAS'99 First International Workshop on Information Integration and Web-based Applications and Services.

Al-Safadi, L. and Getta, J. (2000a). Semantic Modeling for Video Content-Based Retrieval Systems. 23rd Australasian Computer Science Conference ACSC2000

Al-Safadi, L. and Gettta, J. (2000b). Mapping user view and video semantic model. The IEEE 7th technical exchange meeting.

Amato, G., Mainetto, G. and Savino, O. (1998). An Approach to a Content-Based Retrieval of Multimedia Data. Multimedia Tools and Applications.

Ambroziak, J. and Woods, W.(1999, January). Natural Language Technology in Precision Content Retrieval. Sun Microsystems. Retrieved January, 1999 from the World Wide Web http.//www.sun.com/research/techrep/1998/abstract-69.html.

Aoki, H., Shimotsuji, S. and Hori, O. (1996). A shot classification method of selecting effective keyframe for video browsing. Proceedings of ACM International Conference on Multiemdia,

Arman, F., Depommier, R., Hsu, A. and Chiu, M. (1994). Content-based browsing of video sequences. Proceedings of Second ACM International Conference on Multimedia.

Arman, F., Hsu, A. and Chiu, M. (1993). Image processing on compressed data for large video databases. Proceedings of the first ACM International conference multimedia.

Aslandogan, Y., Their, C., Yu, T., Liu, C. and Nair, K. (1997). Using Semantic Content and WordNet in Image Retrieval. Proceeding of ACM SIGIR conference.

Blum, T., Keislar, D., Wheaten, J. and Wold, E. (1995). Audio Databases with Content-based Retrieval. Proceedings of IJCAI workshop on Intelligent Multimedia Information retrieval.

Boykin, S. and Merlino, A. (2000). Machine Learning of Event Segmentation for News on Demand. Communication of the ACM 43(2).

Brown, M., Foote, J., Jones, G., Jones, K. and Young, S. (1995). Automatic Content-Based Retrieval of Broadcast News. In proceedings of ACM Multimedia 95.

Brown, M., Foote, J., Jones, G., Jones, K. and Young, S.(1996). Open-Vocabulary Speech Indexing for Video and Audio Mail Retrieval. Proceedings of the fourth international ACM multimedia conference 96.

Chang, H. and Chang, S. (1996).Temporal Modeling and Inter-Media Synchronization for Presentation of Multiemdia Streams. In Multimedia information Storage and Management.

Change, C., Lin, K. and Lee, S. (1995). The Characteristics of Digital video and Considerations of Designing Video Databases. Proceedings of the 1995 conference on International conference on information and knowledge management.

Chua, T., Pung, H., Lu, G. and Jong, H. (1994). A Concept-Based Image Retrieval System. Proceedings of the Twenty-Seventh Annual Hawaii International Conference on system Sciences.

Coolahan, J. and Roussopoulos, N. (1983). Timing requirements for time-driven systems using augmented Petri nets. IEEE Transactions Software Engineering. September.

Courtney, J. (1996). Automatic, Object-Based Indexing for Assisted analysis of Video Data. Proceedings of the fourth international ACM multimedia conference 96

Croft, W. and Thompson, R. (1987). IIIR. a new approach to the design of document retrieval system. Journal of ASIS, 38.

Davenport, G., Smith, T. and Pincever, N. (1991). Cinematic Primitives for Multimedia. IEEE Computer Graphics and Applications, July.

Day, Y., Dagtas, S., Iino, M., Khokhar, A. and Ghafoor, A. (1995). Object-Oriented Conceptual Modeling of Video Data. Proceedings of the Eleventh international Conference on Engineering.

Day, Y., Khokhar , A., Dagtas, S. and Ghafoor, A. (1999). A multi-level abstraction and modeling in video databases. Multimedia Systems 7.

Decleir, C., Hacid, M. and Kouloumdjian, J.(1999). A database Approach for Modeling and Querying video Data. The 15th International Conference on Data engineering.

Decleir, C., Hacid, M. and Kouloumdjian, J. (1998). A Generic Data Model for Video Content Based Retrieval. SAC '98 ACM Symposium on Applied Computing.

Dimitrova, N. (1995). The Myth of Semantic Video Retrieval. ACM Computing Surveys, Communication of ACM, 27.

Djeraba, C., Bouet, M. and Briand, H. (1998). Concept-Based Query in Visual Information Systems. Advances in Digital Libraries Conference.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M. and Hafner, J. (1995). Query by Image and Video Content. The QBIC System. IEEE Computer, September.

Gauch, J., Gauch, S., Bouix, S. and Zhu, X. (1999). Real Time Video Scene Detection and Classification. Information Processing & Management, 35(3). Available at http://www.ittc.ukans.edu/~sgauch/DVLS.html

Griffioen, J., Yavatkar, R., and Adams, R. (1996). Automatic and Dynamic Identification of Metadata in Multimedia. First IEEE Metadata Conference.

Gudivada, V. and Jung, G. (1996). An Algorithm for Content-Based retrieval in Multimedia Databases. Proceedings of the International Conference Multimedia computing and Systems.

Hauptmann, A. and Witbrock, M. (1998). Story Segmentation and Detection of Commercials in Broadcast News Video. Proceedings of the IEEE Forum on Reasearch and Technology Advances in Digital Libraries, IEEE ADL '98.

Hee, M., IK, Y. and Kim, K. (1999). Unified Video Retrieval System supporting similarity retrieval. IEEE International conference on Multimedia Computing and systems.

Hibino, S. and Rundensteiner, E.A. (1995). A Visual Query Language for Identifying Temporal Trends in Video Data. International Workshop on Multi-Media Database Management Systems IW-MMDBMS'95.

Hjelsvold, R. and Midtstraum, R.(1994). Modeling and Querying Video Data. Proceedings of the 20th VLDB conference.

Iannizzotto, G., Puliafito, A. and Vita, L. (1997). Design and Implementation of a Content-Based Image Retrieval tool. In proceedings of PDSE'97. the 2nd International Workshop on Engineering for Parallel and Distributed Systems.

Irani, M. and Anandan, P., (1998). Video Indexing Based on Mosaic Representations, Proceedings of the IEEE, 86 (5).

Jiang, H., Montesi, D. and Elmagarmid, A. (1997). VideoText database Systems. Proceedings of the fourth IEEE International Conference of Multimedia Computing and Systems.

Jul, E. (1991).Howard Besser explores the development of image databases. OCLC Newsletter, 190.

Kim, K. W., Kim, K. B. and Kim, H. (1996). VIRON. An annotation-Based Video Information Retrieval System. Proceedings of the 20th International Conference on computer software and Applications COMPSAC'96.

Koh, J., Lee, C. and Chen, A. (1999). Semantic Video Model for Content-Based Retrieval. IEEE International Conference on Multimedia Computing and Systems, ICMCS 1999

Kubala, F., Colbath, S., Liu, D., Srivastavva, A. and Mackhoul, J. (2000). Integrated Technologies for Indexing Spoken Language. Communication of the ACM 43(2).

Lee, S. and Koa, H. (1993). Video Indexing – an approach based on Moving Objects and Track. Storage and Retireval for Image and Video Databases.

Li, J., Goralwalla, I., Özsu, M. and Szafron, D. (1996). Video Modeling and Its Integration in a Temporal Object Model. Technical Report TR 96-02, The University of Alberta.

Li, J., Özsu, M. and Szafron, D. (1996). Modeling of Video Spatial Relationships in an Object Database Management System. International Workshop on Multimedia DBMS.

Li, J., Özsu, M. and Szafron, D. (1997). Modeling of Moving Objects in a Video Database. IEEE International Conference on Multimedia Computing and Systems.

Lienhart, R. (1996). Automatic Text Recognitive for video indexing. Proceedings of the fourth international ACM multimedia conference 96.

Liou, S., Hjelsvold, R., Depommier, R. and Hsu, A. (1999). Efficient and reliable digital media archive for content-based retrieval. Multimedia Systems 7.

Little, T., Ahanger, E., Folz, R., Gibbon, J., Reeve, F., Schelleng, D. and Venkatesh, D. (1993). A digital on-demand video service supporting content-based queries. Proceedings of the First ACM International Conference on Multimedia.

Liu, Z. and Sun, J.(1997). Structured Image Retrieval. Journal of Visual Languages and Computing 8(3).

MAESTRO (2000, August). MAESTRO (Multimedia Annotation and Enhancement via a Synergy of Technologies and Reviweing Operators). Available at http://www.chic.sri.com/projects/Maestro.html

Maier, D.(1983). The Theory of Relational Databases. Pitman Publishing Ltd.

Mandler, J. and Johnson, N. (1977). Remembrance of Things Parsed: Story Structure and Recall. Cognitive Psychology 9.

Maybury, M., Merlino, A. and Rayson, J. (1997). Segmentation, Content Extraction and Visualization of Broadcast News Video using Multistream Analysis. AAAI Spring Symposium.

Meng, J. and Chang, S. (1996). CVEPS – A Compressed Video Editing and Parsing System. ACM Multimedia 96.

Meng, J., Juan, y. and Chang, S. (1995). Scene Change Detection in a MPEG compressed video sequence. IS&T/SPIE Symposium Proceedings, Vol. 2419.

Miller, G. A. (1995). WordNet. A lexical Database for English. Communications of the ACM 38 (11).

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Introduction to WordNet. An on-line Lexical Database. International Journal of Lexicography, 3(4).

Nack, F. and Lindsay, A. (1999). Everything you wanted to know about MPEG-7. Part 1 & 2, IEEE Multimedia 6(3,4).

Nagasaka, A. and Tanaka, K.(1991). Automatic Video Indexing and Full Video search for Object Appearances. In 2nd working conference on visual Database Systems, IFIP.

Nakamura, Y. and Kanade, T. (1997). Semantic analysis for Video Content Extraction – Spotting by Association in News Video. ACM Multimedia 97.

National Instrument (1997, June). IMAQ™ Vision for G. Available at http://www.ni.com/imaq/

O'Docherty, M. and Daskalakis, C. (1991). Multimedia Information Systems – The Management and Semantic Retrieval of all Electronic Data Types. The computer Journal, 34(3).

Oomoto, E. and Tanaka, K. (1993). OVID. Design and Implementation of a Video-Object Database System. IEEE Transactions on Knowledge and Data Engineering.

Orenstein, J. and Manola, F. (1988). PROBE Spatial Data Modeling and Query Processing in an Image Database Application. IEEE Trans. on Software Engineering, 14(5).

Patel, N. and Sethi, I. (1997). Video shot detection and characterization for video databases. Pattern Recognitive, April.

Peacocke, R. and Graf, D. (1990). An Introduction to Speech and Speaker Recognition. Computer 23(8).

Roussopoulos, N., Faloutsos, C. and Sellis, T. (1988). An Artificial Pictorial Database System for PQSL.IEEE Trans. on Software Engineering, 14.

Saur, D., Tan, Y., Kulkarni, S. and Ramadge, P. (1997). Automated Analysis and Annotation of Basketball Video. In Storage and Retrieval for Image and Video Databases.

Schank, R. (1990). Tell me a Story. A real look at Real and Artificial Memory. Maxwell Macmillan International.

Sistla, A., Yu, C. and Haddad, R. (1994). Reasoning About Spatial Relationships in Picture Retrieval Systems. Proceedings of the International Conference of Very Large Databases.

Smith, J. and Chang, S. (1996). Searching for Images and Videos on the World-Wide Web. CRT Technical Report 459-96-25, Columbia University.

Smith, M. and Christel, M. (1995). Automating the creation of a digital video library. Proceedings of ACM Multimedia 95.

Smoliar, S. and Zhang, H. (1994). Content-Based Video Indexing and Retrieval. IEEE Multimedia 1(2).

Srihari, R. (1995). Automatic Indexing and Content-Based Retrieval of Captioned Image. IEEE Computer, September.

Srinivasan, U. and Riessen, G. (1997). A Video Data Model for Content-Based Search. Proceedings of the Eighth International workshop on Database and Expert systems Applications.

Storey, V. and Goldstein, R. (1988). A Methodology for creating User View in Database Design. ACM Transactions on database systems, 13(3).

Sudhir, G., Lee, J. and Jain, A. (1998). Automatic Classification of Tennis video for High-level Content-based Retrieval. IEEE International workshop on content-Based Access of Image and Video Databases CAIVD'98.

Sun Microsystems (2000, January). Conceptual Indexing for Precision Content Retrieval. Retrieved January, 2000 from the World Wide Web http.//www.sun.com/research/knowledge/technology.html

Swabnerg, D., Shu, C. and Jain R. (1992). Structure of a multimedia information system for content-based retrieval. Audio Video Workshop.

Teodosio, L., Bender, W., (1993). Salient Video Stills: Content and Context Preserved. ACM Multimedia '93.

Tong, R., Appelbaum, L. Askman, V. and Cunningham J. (1987). Conceptual information retrieval using RUBRIC. Proceedings of ACM SIGIR Conference.

Vinod, V. and Murase, H. (1997). A feature-based algorithm for detecting and classifying scene breaks. Proceedings of the international conference on multimedia computing and systems.

Wactlar, H., Hauptmann, A., Christel, M., Houghton, R. and Olligschlaeger, A. (2000). Complementary Video and Audio Analysis for Broadcast News Archives. Communication of the ACM 43(2).

Wactlar, H., Kanade, T., Smith, M. and Stevens, S. (1996). Intelligent Access to Digital Video. Informedia Project. Computer, 29.

Wang, X., Chua, T. and Al-Hawamdeh (1992). Probabilistic and semantic based retrieval in hypertext. Proceedings of the South-East Asia Regional Computer Conference.

Wold, E., Blum, T., Keislar, D. and Wheaten, J. (1996). Content-Based Classification, Search, Retrieval of Audio. IEEE Multimedia 3.

Wu, J., Ang, Y., Lam, P., Moorthy, S. and Narasimhalu, A. (1993). Facial Image Retrieval, Identification and Inference system. Proceedings of ACM Multimedia 93.

Yeo, B. and Liu, B.(1995). Rapid scene analysis on compressed video. IEEE Trans. On Circuits and systems for video Technology, 5.

Yeung, M., Yeo, B. and Liu, B (1996). Extracting Story Units from Long Programs for Video Browsing and Navigation. Proceedings of the International Conference on Multimedia Computing and Systems.

Yoshitaka, A., Hosoda, Y., Yoshimitsu, M. and Ichikawa, T. (1996). VIOLONE. Video Retrieval by Motion Example. Journal of Visual Languages and Computing 7(4).

Yoshitaka, A., Kishida, S., Hirakawa, M. and Ichikawa, T. (1994). Knowledge-Assisted Content-Based Retrieval of Multimedia Databases. IEEE Multimedia, winter.

Yow, D., Yeo, B., Yeung, M. and Liu, B. (1995). Analysis and Presentation of Soccer Highlights from Digital video. In Second Asian Conference on Computer Vision ACCV'95.

Yu, H. and Wolf, W. (1997). A Visual Search system for Video and Image Databases. Proceedings of the Eighth International workshop on Database and Expert Systems Applications.

Zhang, H., Tan, S. and Smoliar, S.(1995). Automatic parsing of news video. Multimedia Systems, 1(2).

Zhang, H., Tan, S. and Smoliar, S.(1995). Automatic parsing of news video. Multimedia Systems, 1(2).

Zhuang, Y., Rui, Y., Huang, T. and Mehrotra, S.(1998). Applying Semantic Association to Support Content-Based Video Retrieval. International Workshop on Very Low Bitrate Video Coding VLBV98.

APPENDIX A

PUBLISHED PAPERS

List of published papers related to the content of this thesis.

'Video Semantic Model & Acquisition' presented in IIWAS'99 First International Workshop on Information Integration and Web-based Applications and Service. Yogyagarta, Indonesia.

'Semantic Modeling for Video Content-Based Retrieval Systems' presented in the IEEE 23rd Australasian Computer Science Conference ACSC2000. 31 January – 3 February 2000, Canberra, Australia.

'Mapping user view and video semantic model' presented in the IEEE 7th technical exchange meeting. April 18-19, 2000, KSA.

'Semantic Video Modeling and View Transformation' presented in the 2000 International Resources Management Association International Conference. May 21-24, 2000, Anchorage, Alaska, USA

‘ Semantic Video Content-Based Retrieval for Video Documents’, Design and Management of Multimedia Information Systems: Opportunities and Challenges, 2000.

APPENDIX B
REVIEW OF IMAQ VISION

IMAQ Vision software from National Instruments on Mac OS platform and includes an extensive set of MMX-optimized functions for grayscale, color and binary image display, image processing (statistics, filtering, and geometric transforms), pattern matching, shape matching, blob analysis, gauging and measurement. End users, integrators, and OEMs use IMAQ Vision to accelerate the development of industrial machine vision and scientific imaging applications. IMAQ Vision is used in factory and laboratory automation operations that require extremely reliable, high-speed vision systems. IMAQ accepts images in pic format and video clips in QuickTime format. IMAQ performs frame-based processing for video documents.

IMAQ Vision can manipulate three types of images: gray-level, color and complex images, and performs comparisons between several images and a model. A number of low level features can be automatically and successfully extracted, such as:

- Color. Three planes of color can be extracted RGB (Red, Green and Blue), HSV (Hue, Saturation and value) and HSL (Hue, saturation and lightness). The system equalizes any or all three planes.

- Histogram that indicates the quantitative distribution of pixels per gray-level value, which helps in identifying various components like background, objects and noise.

- Edge detection that reveals texture using *Gradient Filter* and extracts contour of objects and outline details using *Laplacian Filter.* IMAQ vision provides a general description of the appearance of an image and helps identify various components, such as background, objects and noise. The *thresholding* feature provided by IMAQ consists of segmenting an image into two regions: an object region and a background region.

- Quantitative analysis of an image which consists of obtaining densitometry, shape equivalencies and features, and object measurement for object's spatial measurement detection. Spatial calibration consists of correlating the area of a pixel with physical dimensions to extract the X and Y coordination, width, height, area, center of mass(X, Y), upper-left and lower-right corners (min X and Y, max X and Y).

- Pattern matching and high-speed search, but matching does not recognize patterns with significant changes.

The model provides the operator with an opportunity to accept or reject the data acquired after each run, and upon acceptance allows the user to specify features and store in a data structure.

APPENDIX C
REVIEW OF MPEG-7 STANDARD

The purpose of this appendix is to provide a better understanding of the objectives and components of the MPEG-7, "Multimedia Content Description Interface" standard.

MPEG is a working group of the International Organization for Standardization/International Electronics Commission (ISO/IEC), in charge of developing international standards for compression, decompression, processing, and coded representation of movie pictures, audio, and their combination.

MPEG-7 aims to standardize a core set of quantitative measures of audio-visual features, called Description (D), and structure of descriptors and their relationships, called Description Schemes (DS) in MPEG-7 parlance. MPEG-7 will also standardize a language – the Description Definition Language (DDL)- that specifies Description Schemes to ensure flexibility for wide adoption and long life. This allows searching multimedia data (pictures, graphics, audio, speech and video) that has MPEG-7 data associated with it.

MPEG-7 aims to: describe multimedia content, manage data flexibility and globalize data resources.

1. Multimedia content Description

MPEG-7 most important goal is to provide a set of methods and tools for the different classes, which are aspects that a multimedia content may cover, of multimedia content description.

2. Flexibility in data management

MPEG-7 aims to provide a framework that allows references to parts of a document, to a whole document, and to a series of documents. It should be possible to describe multimedia content in such a way as to allow queries based on visual descriptions to retrieve audio data and vice versa.

3. Globalization of data resources

MPEG-7 descriptions maybe physically located with the associated audio-visual material, in the same data stream, or on the same storage system, but the descriptions could also live somewhere else.

The combination of flexibility and globalization of data resources allows humans as well as machines to exchange, retrieve, and reuse relevant materials.

MPEG-7 does not extract description/features automatically. Nor does it specify the search engine that can use the description. Those are outside the scope of the planned standard. Rather, MPEG-7 will concentrate on standardizing a representation that can be used for description. MPEG-7 emphasizes audio and visual content and doesn't aim to create description schemes or descriptors for text.

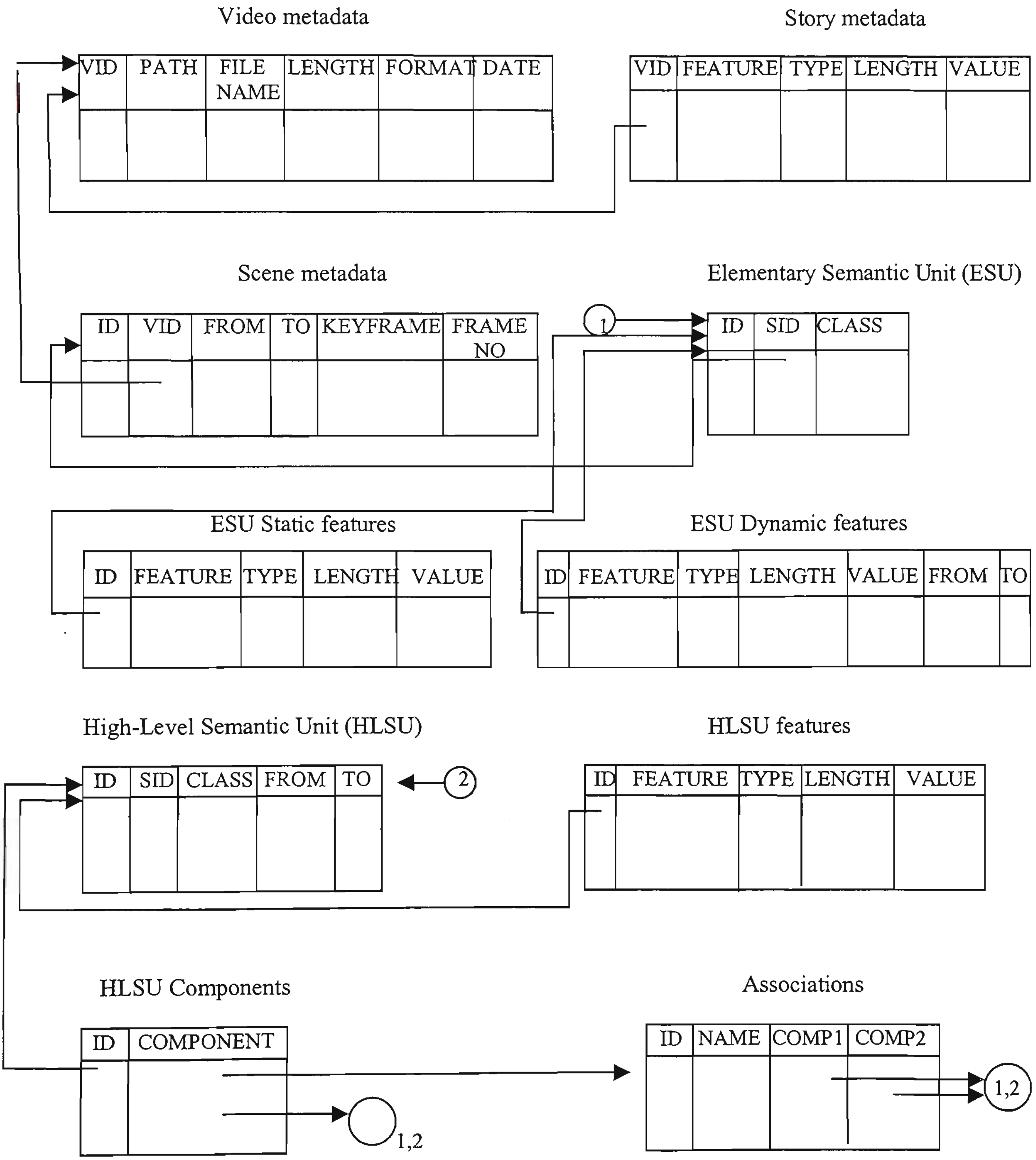
Useful links

- MPEG home page, http://www.cselt.it/mpeg
- Overview of MPEG-7 standard, http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm
- ISO/MPEG N2728, Applications for MPEG-7, MPEG Requirements Group, ISO, Geneva, March 1999, available at http://www.darmstadt.gmd.de/mobile/MPEG7/Documents/N2728.html
- ISO/MPEG N2729, MPEG-7 Context and Objectives, MPEG Requirements Group, ISO, Geneva, March 1999, available at http://www.darmstadt.gmd.de/mobile/MPEG7/Documents/N2729.html
- An overview of the current state of MPEG-7 development can be found in, MPEG-7 Behind the Scenes, September 1999 5(9), available at http://www.dlib.org/dlib/september99/hunter/09hunter

- The current status of MPEG-7 issues, especially as they relate to Music and Audio processing available at meta-labs.com http://www.meta-labs.com/mpeg-7-aud

APPENDIX D

MAPPING SEMANTIC VIDEO MODEL INTO RELATIONAL DATABASE


Video metadata
VID
PATH
FILE NAME
LENGTH
FORMAT
DATE
Story metadata
VID
FEATURE
TYPE
LENGTH
VALUE
Scene metadata
ID
VID
FROM
TO
KEYFRAME
FRAME NO
Elementary Semantic Unit (ESU)
1
ID
SID
CLASS
ESU Static features
ID
FEATURE
TYPE
LENGTH
VALUE
ESU Dynamic features
ID
FEATURE
TYPE
LENGTH
VALUE
FROM
TO
High-Level Semantic Unit (HLSU)
ID
SID
CLASS
FROM
TO
2
HLSU features
ID
FEATURE
TYPE
LENGTH
VALUE
HLSU Components
ID
COMPONENT
1,2
Associations
ID
NAME
COMP1
COMP2
1,2