Journal of Universal Computer Science, vol. 8, no. 10 (2002), 965-984
submitted: 24/4/02, accepted: 15/10/02, appeared: 28/10/02 © *J.UCS*

XML and MPEG-7 for Interactive Annotation and Retrieval using Semantic Meta-data

Mathias Lux, Werner Klieber, Jutta Becker, Klaus Tochtermann
(Know-Center Graz , Austria
[mlux|wklieber|jbecker|ktochter]@know-center.at)

Harald Mayer, Helmut Neuschmied, Werner Haas
(JOANNEUM RESEARCH, Austria
[harald.mayer|helmut.neuschmied|werner.haas]@joanneum.at)

Abstract: The evolution of the Web is not only accompanied by an increasing diversity of multimedia but by new requirements towards intelligent research capabilities, user specific assistance, intuitive user interfaces and platform independent information presentation. To reach these and further upcoming requirements new standardized Web technologies and XML based description languages are used. The Web Information Space has transformed into a Knowledge marketplace where worldwide located participants take part into the creation, annotation and consumption of knowledge. This paper points out the design of semantic retrieval frameworks and a prototype implementation for audio and video annotation, storage and retrieval using the MPEG-7 standard and semantic web reference implementations. MPEG-7 plays an important role towards the standardized enrichment of multimedia with semantics on higher abstraction levels and a related improvement of query results.

Keywords: MPEG-7, content-based Multimedia Retrieval, Hypermedia systems, Web-based services, XML, Semantic Web, Multimedia
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

1 Introduction

Web technologies play an increasingly important role within a variety of environments to fulfil the demands for platform- and location-independent access to all kinds of multimedia information. And the Web Information Space has transformed into a Knowledge marketplace where worldwide located participants take part into the creation, annotation and consumption of knowledge to fulfil the demands of different application areas for education, communication, e-business or amusement (e.g. WebTV, iTV, eLearning, Infotainment, etc.). Keeping this information universe up-to-date is only possible with efficient knowledge distribution concepts.

However the continuously growing amount of worldwide accessible information resources causes an increasing complexity concerning the location of relevant information. Furthermore, multimedia information includes richer content than text-based information so that the concept of “relevance” becomes much more difficult to model and capture. Conventional search engines can not be extended towards an integrated specialized architecture for content- and feature-based information extraction, information filtering and the integration of heterogeneous information

resources at once. To prepare web-based retrieval systems towards a Semantic Web the underlying information space has to take into account heterogeneous media formats and rich mark up descriptions (like XML) along with meta-data schemes. With this add-on the annotation of content-specific information becomes possible and the content becomes not only machine-readable but also machine-processable. According to the fact, that semantics are not available at once, existing multimedia content has to be (semi-) automatically analyzed or interactively annotated to create the semantic meta-data.

Beside the distribution of multimedia content the complexity of (semi-)automatic content analysis and annotation requires the integration of distributed human and technological expertise. According to these requirements intelligent retrieval frameworks need an uniform communication language, meta object repositories and innovative and efficient query specification and annotation user interfaces.

The XML schema based standard MPEG-7 is used as uniform communication and semantic meta-data description language. The presented semantic retrieval framework consists of prototypical implemented components for (semi-)automatic analysis, interactive annotation and query specification, speaking "MPEG-7". Semantic annotation as well as query specification at higher levels of abstraction are based on MPEG-7 meta object catalogue with semantic agents, places, time and relations.

The specification of semantic queries requires the usage of application dependent semantic structures. In the field of news and documentation, a retrieval needs capabilities to search for speakers, spoken content, topic areas like sport or politics. E-learning applications in contrast might require query specifications that take into account machinery or interface handling of machines, production specific places and so on. According to the diversity of application areas an intelligent retrieval framework has to be adaptable concerning kind of agents, objects, places, time and relations e.g. "person A is throwing a ball into the window".

In section 2 MPEG-7 existing annotation and retrieval system have been regarded towards their capabilities to describe audiovisual content on the base of MPEG-7 meta data and the integration of higher levels of semantics. In section 3 the interplay between MPEG-7 semantic meta data and interface capabilities is regarded towards the improvement of retrieval quality and usability. Section 4 presents the realised Retrieval Framework Prototype with a special focus on each of the three main components and the underlying semantic Meta data catalogue. Conclusion will point out the role of MPEG-7 today following with Future Work.

2 Existing Video Annotation and Retrieval Systems

A variety of projects have designed and implemented multimedia retrieval systems. The focus is on covering multimedia databases, meta-data annotation, specialized multimedia analysis methods and web-based front-ends. A special focus had been laid on projects and systems already using MPEG-7 or providing extended retrieval features. In addition to the usage of MPEG-7 it was important to analyse the level of semantic, that can be described and used.

2.1 MPEG-7 Annotation Tool

IBM released on 10th of July 2002 the first implementation of their MPEG-7 Annotation Tool 1.4 "VideoAnnEx" [ibm02]. This tool allows to interactively describe static scenes of a video using free-text annotations.

The first step of annotation is an automatic shot detection tool that recognizes dissolves and fades to detect scene cuts. A couple of key frames for each shot is used to represent the content of each shot. Content description in form of meta-data can be added to each shot by selecting entries from the tree view. The entries are described in MPEG-7 and can be loaded from a separate file to use customized lexicons. Each shot can interactively be annotated with object descriptions, event descriptions, other lexicon sets and own keywords. Finally the annotated video description is saved as MPEG-7 XML file. A lexicon is an MPEG-7 based definition of application dependent description components, that has no standardised format.

Every description consists of free-text annotations without any given semantic structure. no differences between objects, places or video structure (e.g. shot) descriptions are made. Listed free-text annotations do not allow any relation specification to construct high-level semantic graphs.

2.2 Ricoh MovieTool

MovieTool is a tool for interactively creating video content descriptions conforming to MPEG-7 syntax [Ricoh, 02]. Ricoh intends its use for researchers and designers of MPEG-7 applications. The software interactively generates MPEG-7 descriptions based on the structure of a video already during loading. In-built editing functions allow the modification of the MPEG-7 description. Visual clues assist the user during the interactive structure editing in combination with candidate tags, to choose appropriate MPEG-7 tags. Relations between video structure and MPEG-7 description are visualised. Every MPEG-7 description is validated in accordance with MPEG-7 schema. Meta-data defined in MPEG-7 can be used to enrich videos with free-text annotation. Future MPEG-7 changes and extensions can be reflected. No semantic description with meta-data of higher level of abstraction is possible. This tool might be combined with MPEG-7 based retrieval tools, but no experiences are mentioned.

2.3 Informedia digital video library project

The aim of Informedia project of the Carnegie Mellon University School of Computer Science [Informedia, 02] is to achieve machine understanding of video and film media including all aspects of search, retrieval, summarization and visualization. Informedia uses speech recognition, content based image retrieval and natural language understanding mechanisms to automatically segment, transcribe and segment video data. The summarization and visualization of the data happens through a web-based interface. The project was extended to support cross-media retrieval, including visualization through document abstraction for each media.

This is a very powerful retrieval framework, but retrieval is only performed on feature-based information as well as implicit semantics within speech and textual descriptions. A combination with other annotation and retrieval systems might be

possible, but no standardised format like MPEG-7 for a uniform communication is mentioned.

2.4 GNU image finding tool (GIFT, MRML)

GIFT is the result of the VIPER project (Viper Project – Visual Information Processing for Enhanced Retrieval [VIPER02]). The GIFT (the GNU Image-Finding Tool [GIFT02]) is a Content Based Image Retrieval System (CBIR) that enables users to search for images "Query By Example". It uses relevance feedback to improve the query results. The communication is done with a XML-based communication protocol named MRML (Multimedia Retrieval Markup Language) [MRML, 02].

GIFT allows indexing a whole directory tree containing images. Separate client programs are used to query and browse these image collections. The MRML protocol is a client-server protocol with focus on simplicity and extensibility. In MRML a basic query consists three components.

1. A list of images referred by their URLs.
2. A relevance level for each image.
3. A list of algorithms that has to be used on the server for image matching.

The default VIPER algorithms uses colour histograms and colour features for searches. The retrieval process is only based on low-level features but may be a powerful combination with semantic retrieval frameworks. The XML based MRML is designed to be extendable. A MRML description can be combined with MPEG-7 but no experiences have been made.

3 Semantic based Retrieval using Meta data

A complete retrieval system requires a pre-processing with data extraction and storage components, that makes implicit information explicit storable within the information space. Stored data will be (re)used with focus on search ability and fast data retrieval. Finally a retrieval framework is needed that enables the user to specify a query, matches the query against the data stored in the database and presents the results to the user. The concept of the presented retrieval framework based on standardized meta data description based on MPEG-7.

Before a user is able to use a retrieval system the database has to be filled with suitable MPEG-7 meta data. To enable a semantic based retrieval the level of semantics of the meta data descriptions have to be increased. The interplay between semantic meta data and interface capabilities is regarded towards the improvement of retrieval quality and usability.

3.1 Creation of Meta data

Data can be extracted from the multimedia data automatically or manually (figure 1). Automatic extraction algorithms work well for low level features such as colours or file size. For higher semantic features such as shape or object recognition the algorithms have too high error rates to get usable results. Here manual correction or annotation is required.

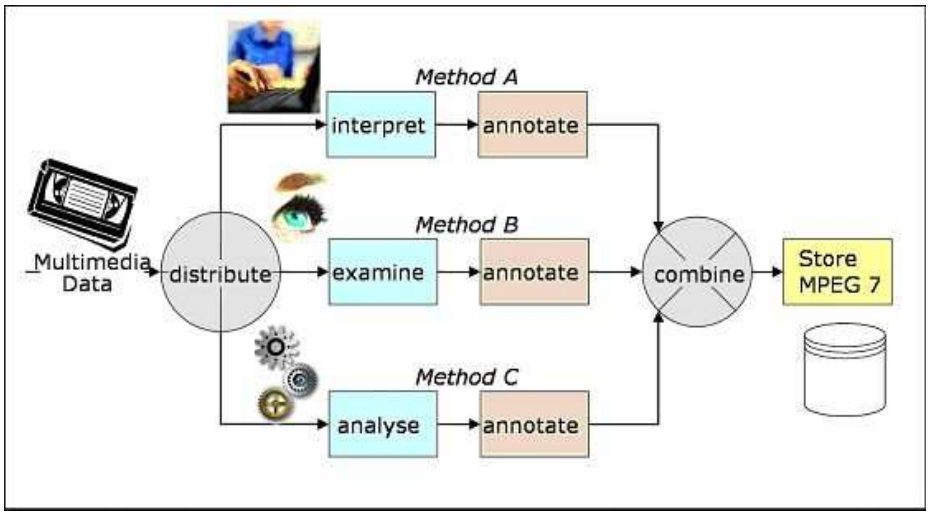

Figure 1: Different kinds of annotation methods – a) human based interpretation using interactive annotation, b) human based examination using "query by example", c) (semi-)automatic analysis of multimedia

The resulting meta-data contains various information:

- different levels of abstracted meta-data: colours, faces, names of objects
- information about the original media: size, length
- logical structure of objects and their relation
- rules and intelligence how to interpret the meta-data
- Integration of human knowledge

To enable a unified communication between annotation components and retrieval frameworks every kind of meta data is stored as MPEG-7 description.

3.2 Classification of semantic

According to the model of J.P. Eakins meta data or meta data attributes are classified according three levels (figure 2). The integration of semantic meta data requires the definition of abstract attributes and objects.

Extraction methods of the first level retrieves primitive features such as colour, texture, shape or the spatial location of image elements. At this level, retrieval features are directly extracted from the source data itself without involving any external knowledge database.

The second level extracts derived or logical features by involving some logical inference with objects in the image. To extract such logical features usually external knowledge is needed. For example, to answer queries such as "find pictures with the Eiffel Tower" the extraction tool needs the knowledge that a certain building has been named "Eiffel Tower". These criteria are reasonably objective.

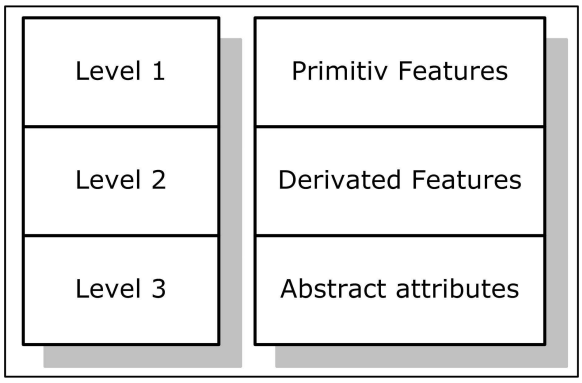

Figure 2 : Three level model of J. P. Eakins

The third level is used for a retrieval by abstract attributes, involving a significant amount of high-level properties such as happiness, laughing or events (goal, explosion, ...). These retrieval methods are highly subjective and topic specific. The second and third level together are often called the 'semantic retrieval'.

MPEG-7 provides the tools for describing the semantics of multimedia content by defining a schema for constructing semantic descriptions [MPEG, 01]. These tools can be used to describe real life concepts or narratives which includes objects, agent objects, events, concepts, places times and narrative worlds which are depicted by or related to multimedia content. These tools can also be used to describe semantic attributes and semantic relations. According to these capabilities of MPEG-7 semantic meta data can be specified during the meta data creation process using human knowledge and for higher level semantic query specification. To enable the reusability of semantic objects (for example persons, places or time describing objects) and semantic relations meta data catalogues will be introduced.

3.3 Usability and retrieval quality

The model of Eakins classifies on the analysis point of view. For the retrieval side this model has to be extended to further points of view to cover the requirements for search and matching algorithms. Extracted data has to be stored in a way to enable a semantic retrieval, e.g. to allow a rule based system to process the data or to support a semantic object catalogue. The semantics fall in two categories: 1) user semantics – what does the user mean and expects – and 2) computer semantics – semantic from a mechanism (algorithm) that is stored in meta-data. Open questions are:

- What does the user mean?
- What information is in the data?
- What meaning has some special data in the context of the entire data?
- How is data with different semantic meaning combined?
- What knowledge has the user about the system?
- How user-friendly is the query specification?

- What knowledge does the user need about the internal structure of the system?

In a typical retrieval process the user specifies a query and gets the results. The effort of the user is to define the query and to pick up the interesting data sets from the result. A more intelligent retrieval mechanism can filter and rank the results more exact to present the user the optimal results. The query of a user is the start point of the retrieval process. The better the system understands and interprets the query of the user, the better the results will be and the effort of the user will be reduced to a minimum (figure 3).

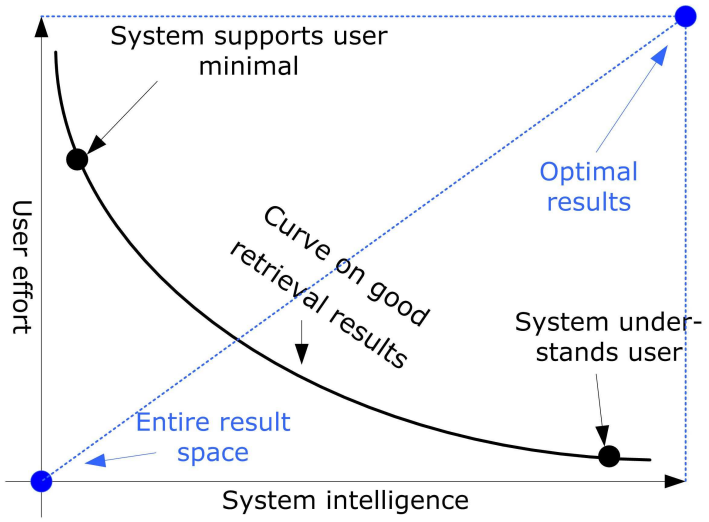

Figure 3: Interface intelligence versus user effort

A user specifies a query to filter out the relevant information from the whole result space. No filtrating process means either to retrieve all results from the result space or to retrieve no results. With the use of complex query languages and intelligent decision algorithms (AI), the user interface can retrieve (more) optimal results. Figure 3 illustrates the user effort depending on the intelligence of the system.

For optimal results an intelligent user interface is needed and the user has tp exactly specify his or her query. If the system has enough semantic information about the user and user demands, it understands the user and good results can be retrieved with few user effort.

The quality of retrieval frameworks will be increased using high semantic and intelligent interface and system components to reduce the user effort for query specification.

The semantic retrieval level can be increased in three ways:

1. provision of an intelligent user interface
2. optimizing of the retrieval mechanism
3. improving the quality of the meta data

The user interface semantic level is not directly combined with the semantic level of the stored meta-data. The combination is done on the retrieval framework.

Again, it has to be distinguished between user semantics and system semantics. User semantics is the knowledge the user must have about the system to specify a query and the manual effort of the user to specify a query. System semantics is the knowledge that is stored in the meta-data and knowledge to adapt the retrieval process to the users' needs. The intelligence of the user interface is directly linked with the amount of knowledge the user needs about the internal structure of the information space, about complex query languages or logical combinations of search terms to obtain the optimal results. If the user query is semantically interpreted by the system, the need of user knowledge about system and information space semantic is reduced: "The system talks the users language", "The system understands what the user wants" (c.f., figure 3).

Generally, users are more interested in the content of an image than in its features. An intelligent semantic retrieval requires the possibility to describe content with semantic descriptions. Furthermore human beings and computer systems need a uniform communication language to "talk about" these semantic descriptions. According to this view the philosophy of the presented semantic retrieval framework is:

The level of semantics in any retrieval framework is defined by user knowledge, interface and system intelligence and available semantic metadata.

With MPEG-7 descriptions, distinguish between agents, places, time and relations, the user is able to specify, what he "knows" during the meta-data creation process and he can "say" what he is looking for during the retrieval process. The machine readable semantic MPEG-7 description allows to improve the intelligence of system components using standardized display, exchange and matching methods.

4 The Retrieval Framework

Our main objective is to design and implement a multimedia retrieval framework, which is modular, highly scaleable and fully based on MPEG-7. The modular approach and the usage of MPEG-7 allow the separation between all system components and the exchange of information with other systems. The general system architecture is depicted in figure 4 below. There are three main components: the MPEG-7 annotation, the multimedia database, and the retrieval component.

To provide well performing retrieval methods the availability of high quality content descriptions is essential. This is achieved by content analysis modules, which are integrated into an annotation tool. All information on different multimedia data is stored as MPEG-7 conforming descriptions. Already existing meta-data, e.g. within MPEG-4 streams, is inherited to the MPEG-7 document without human interaction.

The multimedia data together with the according MPEG-7 document are transferred to the multimedia database. Three different kinds of meta data are contained in the MPEG-7 documents. The general meta data (e.g. text annotation, shot duration) can be searched for by conventional database queries. Low-level meta data (e.g. colour histogram of an image) are necessary for specific compare and search algorithms. Semantic descriptions like agents, events and relations are stored within the semantic meta object catalogue. Therefore the multimedia database has to manage the real multimedia data (the essence), the MPEG-7 documents and some specific low-level data, which can be extracted from the MPEG-7 documents. The requirements on the database are manifold: managing XML schema data (MPEG-7 documents) and indexing and querying multidimensional feature vectors for specific low level data.

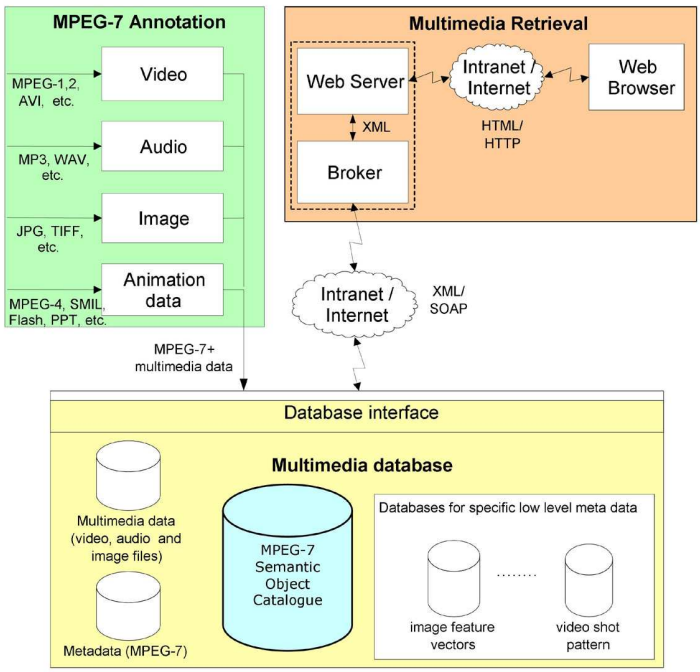

Figure 4: The three components of the Retrieval Framework

The multimedia retrieval is completely separated from the multimedia database, according to the objective of modularity. In our prototype a Web interface is used to specify queries and to display the search results. The Web server forwards the queries of each user to a broker module. This broker communicates with the multimedia database, groups received search results, and caches binary data. The usage of broker architecture makes it possible to integrate other (e.g. text) databases or search engines by simply adding the brokers of these data sources to the Web server. Vice versa also

other search engines can have access to the multimedia database by using the broker from this system.

The search results, which are displayed by the Web user interface, contain extracts of the annotations, parts of the content (e.g. key frames of a video) and references for downloading the multimedia data and the appropriate MPEG-7 document. A streaming server is used to display only the interesting part of the multimedia data.

All data interchange between the annotation tool, the multimedia database and the retrieval interface is based on standardized data formats (MPEG-7, XML, HTML) and protocols (SOAP, HTTP) to achieve a maximum of openness to other systems.

4.1 Content Analysis and Annotation

Describing multimedia data is a very time consuming process. Therefore the automatic extraction of any information on the content is highly desirable. These information can be directly used for the description or can facilitate the manual annotation. For this purpose an annotation application (c.f., figure 5) has been implemented, which stores all information as MPEG-7 documents. Content analysis modules can be integrated as plug-ins, however also manual annotation is possible.

The description of multimedia data, which is used for content retrieval, comprises the following areas:

- *Description of the storage media:* file and coding formats, image size, image rate, audio quality, etc.
- *Creation and production information:* creation date and location, title, genre, etc.
- *Content semantic description:* content summary, events, objects, etc.
- *Content structural description:* shot and key frames with colour, texture and motion features, etc.
- *Meta-data about the description:* author, version, creation date, etc.

All these types of information can be handled by MPEG-7 description schemes. Obviously some of these descriptions can only be inserted manually, like creation and production information, the meta-data about the usage and about the description itself. This is supported by specialized user interfaces within the annotation tool.

Other content descriptions can partly be extracted automatically. Basic information about the storage media can directly be read from the raw data. Other information has to be extracted by content analysis methods. In the annotation tool the following content analysis methods are integrated.

4.1.1 Shot Detection and Keyframe Extraction

By the shot detection method the video can be segmented automatically into shots. A shot is a contiguous sequence of video frames recorded from a single camera operation. The method is based on the detection of shot transitions (hart cuts, dissolves, and fades).

From the shots one or more keyframes are extracted in dependence of the dynamic of the visual content.

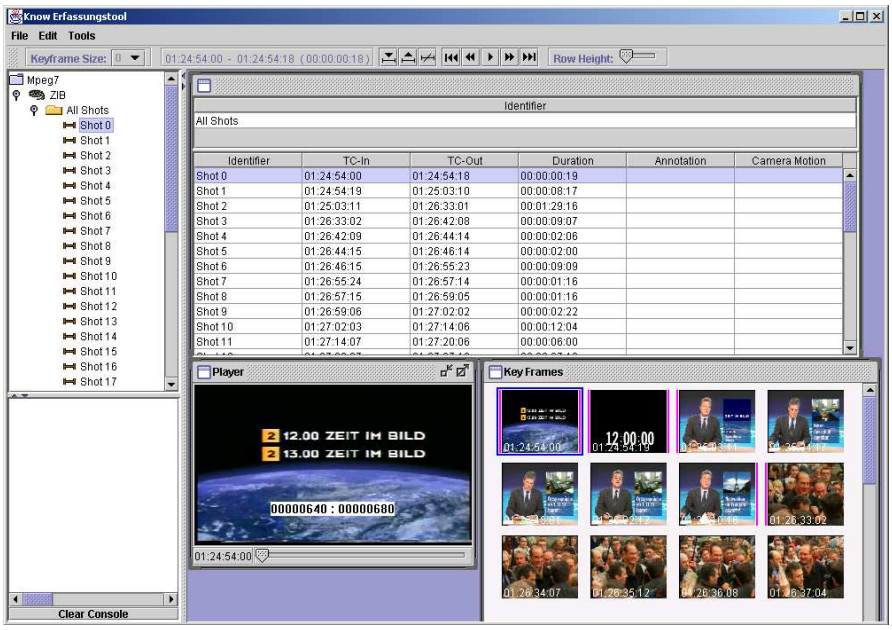

Figure 5: MPEG-7 annotation tool.

4.1.2 Scene Segmentation

Related Shots are grouped into high-level units termed scenes. The used algorithm evaluates the similarity of keyframes and the amount of the dynamic of the visual content in the shots.

Often scenes are start with specific sequences where for instance an anchor person can be seen. By the detection of these sequences the begin of scenes are recognized.

4.1.3 Summery Generation

The calculation of a rating of keyframes, shots and scenes enables the retrieval of a hierarchical video structure which can be used for a summery description. The rating is based on

- the duration of a shot or scene,
- image contrast of keyframes,
- characteristic colour distribution in keyframes,
- camera movements (zoom may indicate an import event), and on the
- audio level.

4.1.4 Visual Keyframe Annotation

From the extracted keyframes color and texture features are computed. The features are saved by using MPEG-7 visual descriptors. For video content retrieval these features can be used for queries based on example images.

4.1.5 Camera Movement Detection

There is also an analysis tool which determines the camera movements in a video. Camera movements like pan, tilt, zoom and rotation are detected. E.g. searching for scenes in a soccer video, where a zoom-in can be seen, delivers retrieval results of important events (fouls, goals, etc.).

4.2 Intelligent Multimedia Database

The multimedia database consists of two major parts, the front-end system, which processes incoming requests and a database backend system, which stores the content itself. Three types of data, binary (multimedia data), XML (MPEG-7) and multidimensional data (low-level meta-data), have to be managed.

XML documents can be divided into document centric and data centric documents. Relational databases are well suited for data centric documents. The MPEG-7 schema is very complex and heterogeneously structured and belongs to document centric documents. Most database management systems (DBMS) started to support such document centric XML data (e.g. Oracle 9i [Oracle, 02], Tamino [Tamino, 02], Xindice - former dbXML [Xindice, 02], etc.)

The multimedia data (essence) are saved in a file pool and only the file references are managed by a DBMS. The Oracle 9i DBMS have specific data types for the management of these file references. When importing new objects to the database key frames are extracted automatically from video data. These key frames are stored in a compressed format (JPEG).

The management of multidimensional data is very limited within conventional DBMS. Specific index structures (e.g. hybrid tree [Chakrabarti, 99]) have to be implemented to enable a fast access to such data.

A web service is used for the interface to the retrieval system (to the broker), with emphasis on that the communication does not interfere with possible firewall mechanisms. All search queries and results are specified in XML format.

Two different kinds of retrieval functions were created as web service of the multimedia database:

- Retrieval of meta-data
- Retrieval of essence

The retrieval process is implemented in two steps. At first the MPEG-7 description is retrieved. All textual information of the result can be displayed immediately. If there are references to navigational information like key frames, they are requested in a second step and added to the previously received result. The corresponding multimedia file can be downloaded or streamed by a separate request.

4.3 Retrieval

Through the modularized architecture of the retrieval program we can support multiple data brokers, multiple databases and multiple user interfaces. The broker is the module of the retrieval framework which does the actual retrieval. It uses SOAP as a standardized interface to connect to one or more multimedia databases. It can also use available third party Web-Services to boost the retrieval quality, like phoneme based similarity matching or retrieval through other software agents. The search interface and the visualization engine are main components of the human-computer interface. They can be replaced by other components to support further environments (e.g. mobile computing, speech driven retrieval, …).

The entire concept of retrieval of multimedia data was split into three main parts (see figure 6): definition of search parameters, finding the desired entities and visualizing the results. Search results are MPEG-7 documents, which describe one or more multimedia entities. Each MPEG-7 document has to match criteria's specified in the search parameters. So we are searching for the meta-data, which is described in MPEG-7 and stored in a multimedia-library as textual information.

For our concept of content-based retrieval it does not matter how the MPEG-7 description is created, it can be done by hand, by automatic information extraction like speech recognition, OCR or face recognition or a combination of this technologies.

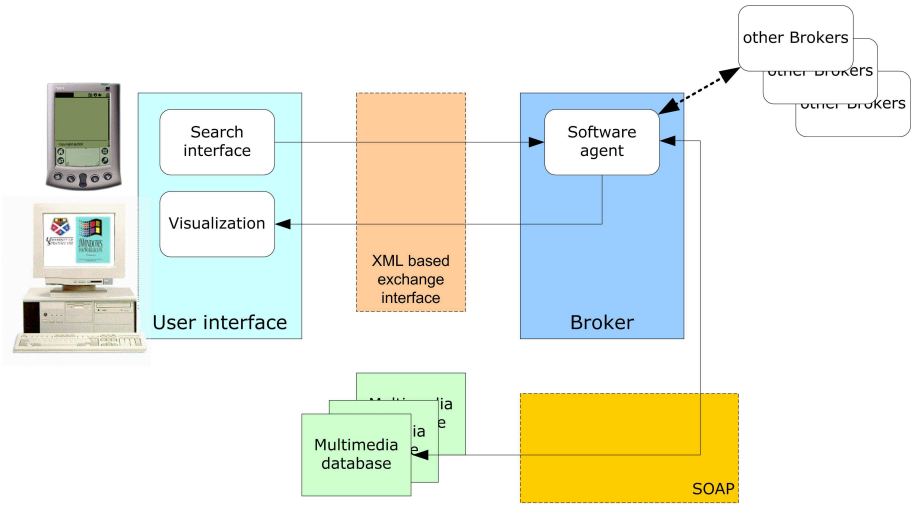

Figure 6: Retrieval framework

4.3.1 Search MPEG-7

The multimedia retrieval starts with the definition of a set of search parameters. Within the first prototype the search is defined as set of text-based parameters. The query dialog allows the user to specify different kinds and sets of meta-data constraints:

- The matching constraints define, which meta-data has to match the given text (e.g. title, summary, genre, description of a scene, name of a person featured in the multimedia data, everywhere)
- According to the combination constraints the given text parameter will be handled differently (search for phrase, single words, sounds like etc.)
- Priority constraints allow the relevance definition of several results (sorting , not more than 2 weeks old, size etc.)

The search parameter definition is a combination of interactive parameter definition and matching strategies.
Within the second prototype the query definition has been extended to support higher level semantic query specifications using a semantic object catalogue (see section 4.4).

```
<IMBResult>

   <ResultId>result_id_1</ResultId>
      <RecordSet>
      <mpeg7/>
      </RecordSet>
      <DatensatzAnzahl>1</DatensatzAnzahl>
      <StartPosition>1</StartPosition>
      <Format>IMBMpeg7</Format>
      <Detailierungsgrad>
           Vollansicht
      </Detailierungsgrad>
      <SoundsLike>
            <term>Herde</term>
            <term>Heirat</term>
      </SoundsLike>
</IMBResult>
```

Figure 7: IMBResult – the resultset of a search request

4.3.2 Find MPEG-7

The previous specified query parameters are send to an agent to gather the information. The user has no influence on where the agent tries to gather information. The agent executes an order of the user automatically and returns the results of its work. When the user initiates a search request, the broker converts the search parameters to an IMB-query and sends it to the multimedia database.

4.3.3 Visualizing MPEG-7

Once the agent has finished its work the results has to be visualized. The result is obtained from the broker using the format IMBResult:

4.4 MPEG-7 Meta Object Catalogue

As a base for describing semantics the descriptor of type "objectType" is used. All other descriptor types are heirs of "objectType". Using instances of these semantic descriptors, which we will call "semantic objects" from now on, we can build a set of objects (nodes or vertices) which can be used to build a graph using semantic relations as edges of the graph, connecting the nodes (see also figure 8).

An example for a semantic description is:

Alex is shaking hands with Ana in New York on the 9th of September. The event is showing "comradeship".

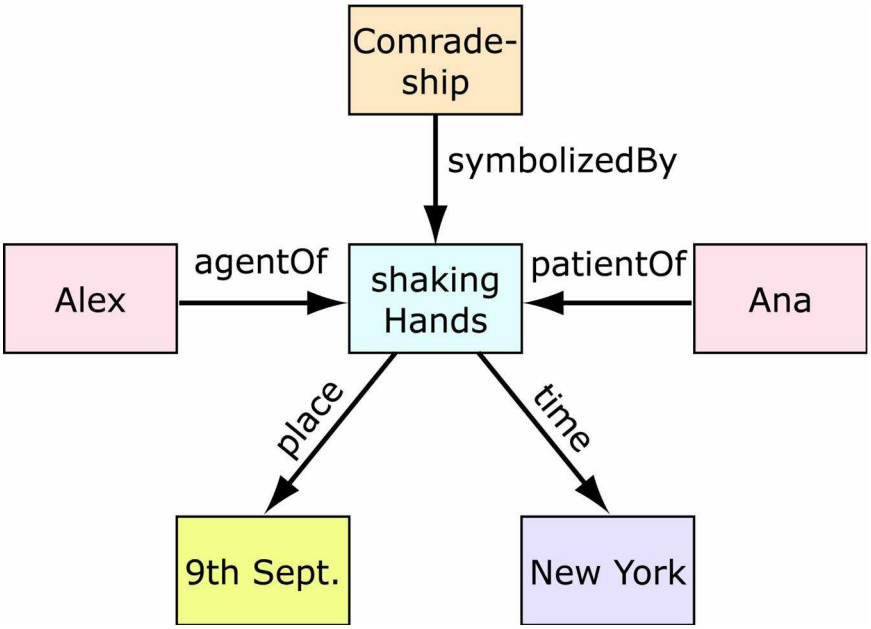

Figure 8: Graph based on semantic description above

The above graph shows the following objects:

Label	Type	Relation	Target
Alex	AgentObjectType	agentOf	shaking Hands
Ana	AgentObjectType	patientOf	shaking Hands
Comradeship	ConceptType	symbolizedBy	shaking Hands
shaking Hands	EventType	place, time	9th Sept., New York
9th Sept.	SemanticTimeType		
New York	SemanticPlaceType		

Table 1: Semantic objects and their relations

4.4.1 Building a catalogue

The above example shows reusable and computer readable semantic descriptions based on the tools defined in the MPEG-7 standard. Building the above graph is a time consuming task, but the objects are reusable, so they can be taken from a description and inserted into a database for later retrieval and usage. We will call this database "semantic catalogue" since it is meant for browsing the semantic objects and for building description graphs.

This is of great use to describe content in a specific context, where events and objects, states, times and places are predefined. In our project we indexed media from FIFA World Cup™ 2002. The soccer ontology, the FIFA website and common soccer knowledge built the base for the catalogue for adding events, states, places and objects. So the catalogue was not built from the scratch it was merely transformed from other formats to MPEG-7, DAML in case of the ontology and HTML in case of the website.

4.4.2 Extending MPEG-7

Although the given MPEG-7 tools will prove sufficient for most contexts, we extend MPEG-7 in defining SoccerPlayerType, SoccerRefereeType and VenueType to meet our requirements.

The "Referee" shows "Soccer Player #1" the red card (figure 9).

Figure 9: Graph based on description above

Label	Type	Relation	Target
Soccer Player #1	SoccerPlayerType	patientOf	Event red card
Referee	SoccerRefereeType	agentOf	Event red card
Event red card	EventType		

Table 2: Example of semantic objects in context "soccer"

The procedure of extending MPEG-7 is well described according to the suggestion of how to integrate Dublin Core with MPEG-7 [Hunter, 00]. Each type we introduced extends the "objectType" type, which is defined in MPEG-7.

4.4.3 Using the catalogue

The pre-built catalogue is the base for constructing semantic descriptions using MPEG-7 tools. In our tool selected parts of the catalogue are displayed in tables on the right hand side of the builder. They can be used to build the descriptions using the common "drag and drop" mechanisms. Objects that are not present can be created or, in case they were created before, imported.

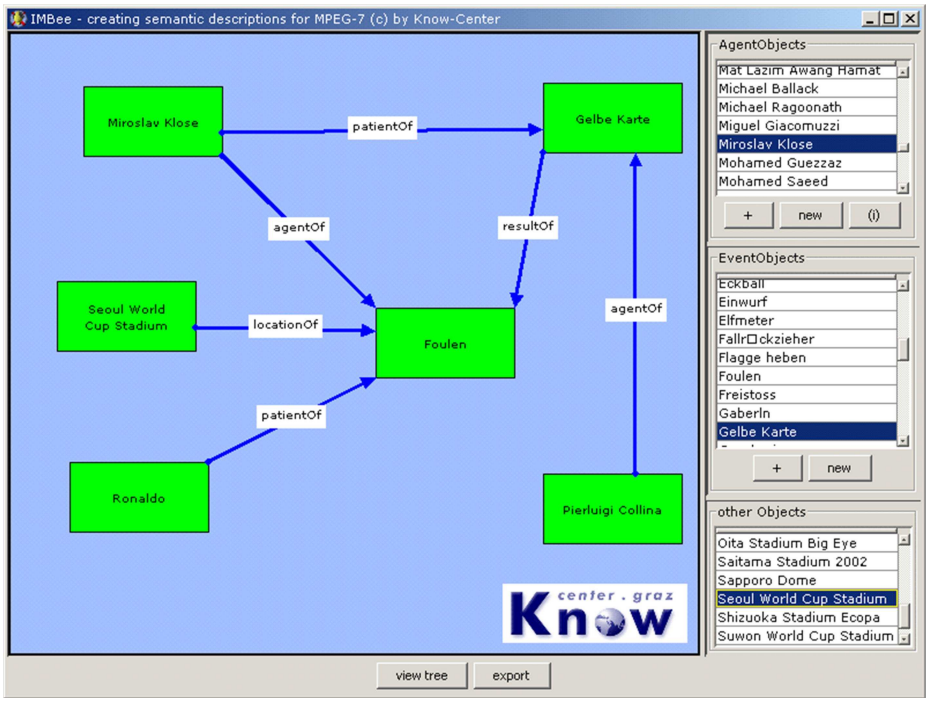

Figure 10: Screenshot of semantic description builder

4.4.4 Benefits of the catalogue

Annotation

The benefits for annotations are obvious. Using an intuitive and user-supporting environment the annotation task can be accelerated. Also the objects can be kept unique as instances of or references to objects stored in the catalogue.

Retrieval

The retrieval offers much more opportunities thanks to the semantic descriptions. Users can define a graph with wildcard objects or relations to find a matching structure. Queries such as “Show me all scenes in which Ronaldo receives a pass” and “Show me all scenes in which Ronaldo and Ronaldinho are related in some way” are possible and computable.

The objects from the catalogue can be used to define a query like this. If users don’t have access to the catalogue, objects from previous queries can be used to build a new query structure.

Automation

Parallel to the definition of a catalogue of semantic objects a catalogue of computed low-level features, such as colour layout or other content based image retrieval features, can be built. By defining mappings between the semantic objects and the computed visual and audio characteristics of the multimedia content, the annotation tool gains the ability to make suggestions of semantics to the user. Though this is not yet possible it is a preparation for later developments and achievements.

4.4.5 Problems with a catalogue

Creation

There are two ways of creating a new catalogue:

- Find human authors for a new catalogue, these authors should, at least, be supported by an expert for the context in which the catalogue will be used.
- Generate the catalogue from existing data. For a lot of topics existing ontologies can be used and data can be extracted from structured websites or bought from content providers.

The first option is obviously the best choice for very small catalogues, the second method requires in most cases a manual correction of automatically generated data. If a rather big catalogue is built with lots of objects of the same structure in it, an automatic generation is certainly the better choice.

Annotation by different users

Different users often describe semantics in different ways. This is a common problem when untrained people create meta-data. To avoid problems like this a tutorial is provided, which gives users guidelines for annotation. Now the description differences are manageable, the structure of the description graph will show no

insuperable irregularities. Incorrect descriptions can be fixed by the retrieval engine providing imprecise search methods.

5 Conclusions

The current work with MPEG-7 demonstrates that this standard provides an extensive set of attributes to describe multimedia content. MPEG-7 is able to play an important role towards standardized enrichment of multimedia with semantics on higher abstraction levels to improve the quality of query results. However, the complexity of the description schemes makes it sometimes difficult to decide which kind of semantic descriptions have to be used or extended. This may lead to difficulties when interchanging semantic meta-data with other applications. Nevertheless the standardized description language is easy to exchange and filtered with available XML technologies. Additionally the Web-based Tools are available on different platforms and could be extended with further components according to the usage of standardized API´s, Client/Server Technologies and XML based Communication. Furthermore the system architecture allows the broker the communication of any web agent to support user specific retrieval specifications. The different output capabilities could be easily extended for special result representation on mobile devices, first tests had been made with WML.

MPEG-7 or a similar technology will make its way into the "Semantic Web" since it gives computers the possibility to compute semantics in a standardized way. As a result the upcoming software agents are able to interpret, change, index and match the meaning of multimedia to fulfil the needs of the users.

6 Future Work

In near future software agents need to be "educated" to interpret multimedia contents to find semantically corresponding data. Therefore a lot of work has to be done in the area of semantic retrieval and software agents.

The annotation of multimedia content should happen automatically. In case of a video a program could be the viewer and interpreter of the content. Such a development could follow these steps:

1. generation of mappings between low level features and semantics
   evaluation, correction and enhancement of this rule-based system
2. supporting manual annotation by computer generated proposals
3. automatic semantic annotation

Future user interfaces will support the user by "understanding" him or her. Making semantics storable, retrievable and computable will support this trend.

Acknowledgements

The Know-Center is a Competence Center funded within the Austrian Competence Center program K plus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.kplus.at).

References

[Abdel-Mottaleb, 00] M. Abdel-Mottaleb, et al, MPEG-7: A Content Description Standard Beyond Compression, February 2000.

[Achmed, 99] M. Ahmed, A. Karmouch, S. Abu-Hakima, Key Frame Extraction and Indexing for Multimedia Databases, Vision Interface 1999, Trois-Rivières, Canada, 19-21 May

[Chakrabarti, 99] K. Chakrabarti, S. Mehrotra, The Hybrid Tree: An Index Structure for High Dimensional Feature Spaces, In Proc. Int. Conf. on Data Engineering, February 1999, 440-447 http://citeseer.nj.nec.com/chakrabarti99hybrid.html

[Cocoon, 02] Cocoon XML publishing framework, 2002, http://xml.apache.org/cocoon/

[GIFT, 02] The GNU Image-Finding Tool, 2002, http://www.gnu.org/software/gift/

[Hunter, 00] Hunter, Jane, Proposal for the Integration of DublinCore and MPEG-7, October 2000

[IBM, 02] IBM: "VideoAnnEx – The IBM video annotation tool", July 2002, http://www.alphaworks.ibm.com/tech/videoannex

[Informedia, 02] Informedia digital video library, Carnegie Mellon University, July 2002, http://www.informedia.cs.cmu.edu/

[MRML, 02] MRML: "Multimedia retrieval markup language", GNU image finding tool, http://www.mrml.net/

[MPEG, 01] MPEG Consortium, ISO/IEC 15938: Information Technology - Multimedia Content Description Interface, 23.10.2001

[Nack, 99] F. Nack, A. Lindsay, Everything You Want to Know About MPEG-7: Part 1 and 2, IEEE Multimedia, 6(3) and 6(4), Juli-December 1999, 65-77

[Oracle, 02] Oracle 9i Database, 2002, http://www.oracle.com/

[Ricoh, 02] Ricoh: "Ricoh MovieTool Home", June 2002, http://www.ricoh.co.jp/src/multimedia/MovieTool/

[Tamino, 02] Tamino XML database, http://www.softwareag.com/tamino/

[Tomcat, 02] Apache Tomcat, official reference implementation for the Java Servlet and JavaServer Pages technologies, 2002, http://jakarta.apache.org/tomcat/

[VIPER, 02] VIPER server: "Visual Information Processing for Enhanced Retrieval", http://viper.unige.ch/

[W3C, 02] W3C – World Wide Web Consortium; Link: http://www.w3.org/

[Xindice, 02] Xindice XML Database, 2002, http://www.dbxml.org/

[XPath, 99] XML Path Language, Version 1.0, November 1999, http://www.w3.org/TR/xpath