Content-based Video Retrieval and Summarization using MPEG-7

Werner Bailer*[a], Harald Mayer[a], Helmut Neuschmied[a], Werner Haas[a],
Mathias Lux[b], Werner Klieber[b]
[a]JOANNEUM RESEARCH, Steyrergasse 17, A-8010 Graz, Austria
[b]Know-Center, Inffeldgasse 16c, A-8010 Graz, Austria

ABSTRACT

Retrieval in current multimedia databases is usually limited to browsing and searching based on low-level visual features and explicit textual descriptors. Semantic aspects of visual information are mainly described in full text attributes or mapped onto specialized, application specific description schemes. Result lists of queries are commonly represented by textual descriptions and single key frames. This approach is valid for text documents and images, but is often insufficient to represent video content in a meaningful way. In this paper we present a multimedia retrieval framework focusing on video objects, which fully relies on the MPEG-7 standard as information base. It provides a content-based retrieval interface which uses hierarchical content-based video summaries to allow for quick viewing and browsing through search results even on bandwidth limited Web applications. Additionally semantic meaning about video content can be annotated based on domain specific ontologies, enabling a more targeted search for content. Our experiences and results with these techniques will be discussed in this paper.

Keywords: video summarization, content structuring, content-based retrieval, MPEG-7, semantic annotation

1. INTRODUCTION

Due to the constant evolution of available bandwidth on Internet connections, Web technologies play an increasingly important role to fulfill the demands for platform and location independent access to multimedia information. In contrast to text documents, the concept of "relevance" and "summarization" in retrieval applications is much more difficult to model and capture, because of the richness of multimedia information, in particular of audiovisual content.

Today multimedia retrieval is in many cases limited to browsing and searching based on low-level visual features and explicit textual descriptors. While low-level features can be extracted automatically, more high-level semantic information has to be described either in full text attributes or mapped onto specialized, application specific description schemes. This approach results in an expensive annotation task, which is one of the main challenges of multimedia archives.

There are many research and commercial multimedia content management tools which already offer automatic feature extraction methods (cf. section 2). Most of them are only able to extract low-level features like color, texture and shape. Up to now the automatic extraction of high-level semantic descriptions is restricted to specific application characteristics. For example, speech in the recording of news broadcast can be recognized because only a few and clearly speaking persons can be heard. In general high-level information can not be automatically extracted in a reliable manner.

Even if there is a full textual description of the content available, it often can not be found directly using conventional text queries. Also today's text-based Web search engines have to deal with this problem. In recent research the technical answer for this problem is the usage of ontologies, which has also been proposed for multimedia applications [14]. Modeling of semantic information with ontologies for managing and querying data enables a targeted search for content. The additional annotation effort is one of the main reasons why ontologies are not used on a regular basis in text

* Werner.Bailer@joanneum.at, phone +43 316 876 1218; fax +43 316 876 1191; www.joanneum.at/iis

Copyright 2004 SPIE and IS&T.
This paper was published in S. Santini and R. Schettini (eds.), *Proc. Internet Imaging V*, San Jose, CA, USA, Jan. 2004, pp. 1-12 and is made available as an electronic reprint with permission of SPIE and IS&T. One print or electronic copy may be made for personal use. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

databases. However, as manual annotation for multimedia objects usually can not be avoided, ontologies should be used for multimedia content descriptions.

But even with a full set of ontology based content descriptions, which enables high relevance rates of query results, there still remains the issue of presenting query results in a way, that users can anticipate the complete content of a multimedia object at a glance. Where text retrieval systems usually try to display a brief text summary of the found text document, multimedia objects need another kind of visual, which takes their abundance into account. Key frames are an efficient and intuitive way to represent visual summaries. They do not need much bandwidth capacity for transmission and can be grasped easily by the viewer. Research work is necessary in the optimal selection of meaningful key frames and to minimize the amount of key frames necessary. Additionally also the description of key frames of a multimedia object should follow agreed standards to enable interoperability between different retrieval applications.

In this paper we present a multimedia retrieval framework, which contains a standardized approach to represent semantic objects and relations in addition to the more traditional low-level features. Agents, events, objects, locations and their relations can be specified in a semantic object graph editor. The framework itself is composed of several independent components, which fully rely on the MPEG-7 standard as information base. This framework provides a rich set of automatic feature extraction modules and a Web based retrieval interface which will be described briefly in a later section.

The issue of describing semantic information, i.e. objects and relations, and the generation of video summaries is the main part of this paper.

2. RELATED WORK

Multimedia retrieval systems have been designed and implemented within a variety of projects. The focus covers multimedia databases, metadata annotation, specialized multimedia analysis methods, Web-based front-ends and presentation of search results.

The Informedia Digital Video Library project [9] is a research initiative at Carnegie Mellon University funded by the NSF, DARPA, NASA and others that studies how multimedia digital libraries can be established. The project uniquely combines speech recognition, image understanding and natural language processing technology to automatically transcribe, segment and index linear video. Different methods for content-based summarization of video content, such as video skims [29], i.e. temporally condensed videos, and video collages [18] have been proposed.

The VideoQ system [5] provides a number of content-based search and retrieval features and allows query by example and by a user drawn sketch. A Web interface is provided for both text and content based search. The search results are visualized as a matrix showing one key frame per video.

The goals of the MARS [12] project is to design and develop an integrated multimedia information retrieval and database management infrastructure, entitled Multimedia Analysis and Retrieval System (MARS), that supports multimedia information as first-class objects suited for storage and retrieval based on their content. Specifically, research in the MARS project is categorized into the following four sub-areas: multimedia content representation, multimedia information retrieval, multimedia feature indexing, and multimedia database management. An approach for content-based structuring of video content has been developed in this project [23].

The annotation framework CREAM (Creating Relational, Annotation-based Metadata) allows constructing relational metadata, i.e. metadata that comprises class instances and relationship instances [8]. The CUYPERS research prototype system, developed to experiment with the generation of Web-based presentations, is used as an interface to semi-structured multimedia databases [19].

In the Video Browsing and Retrieval system (VIRE) [20] a number of low-level visual and audio features are extracted and stored using MPEG-7. The presentation of search results is a key frame matrix, a similar interface is used for content-based browsing.

Automatic shot and scene segmentation for news content is provided by the MediaMill system [21]. Additionally extraction of low-level features and semantic classification of scenes is supported.

The Físchlár Digital Video System [16] allows content-based search using a number of visual and audio low- and mid-level features. The media descriptions are MPEG-7 compliant. A Web interface for search & retrieval is provided. Results are presented using HTML and SVG with key frames as visualizations. Current research concentrates on modules such as personalization and programme recommendation, automatic recording, SMS/WAP/PDA alerting, searching, summarizing and content based navigation. Key frame based summarization of news stories is supported in the Físchlár system [25].

IBM's CueVideo system [1] automatically extracts a number of low- and mid-level visual and audio features. Visually similar shots are clustered using color correlograms. The annotation tool *VideoAnnex* [28], which creates MPEG-7 compliant descriptions, can be used to manually assign semantic labels to scenes and annotate keywords and main objects. The search result presentation and browsing interface is key frame based. A Web based search interface is provided.

The Q-Video system [11] allows to extract a number of low-level visual and audio features and to classify shots by semantic concepts. The system is capable of searching by color similarity or assigned semantic label. The results are presented as list using key frames as visualizations.

pViReSMo [10] provides a personalized, semantic selection and filtering of multi-media information based on a client/server infrastructure. The generated content description is MPEG-7 compliant. The system supports text queries only and provides a Web interface. For manual annotation, Ricoh's *MovieTool* [22] is used, which can do shot boundary detection and allows to add some text annotations. As the annotations are done in a GUI that depicts the MPEG-7 XML document, it is not very intuitive and can be confusing for non-technical users.

In [7] an approach for key frame based summaries is proposed. The frames of a video are hierarchically clustered by visual similarity. The presentation of the selected representative key frames is discussed in [26], where a comic book style visualization is used.

Although MPEG-7 is becoming more commonly used, many projects still use proprietary representations for many tasks. In contrast, we strictly use the MPEG-7 standard in our project. We also realized a component based architecture using standard Web technology, while many systems provide Web access only as an add-on. Finally our systems uses a summarized presentation of search results, while in other systems simple lists with a single key frame per video item are predominant.

3. INTELLIGENT MULTIMEDIA LIBRARY

In the next sections we present an intelligent retrieval and annotation framework which is composed of exchangeable, extensible and adaptable components. Although the system had been designed for any kind of multimedia content, the focus of our *IMB* (German abbreviation for intelligent multimedia library) is laid on semantic video and audio retrieval.

3.1. Architecture

The general system architecture of the intelligent multimedia library is depicted in Figure 1. There are three main components: the annotation and content analysis component, the multimedia database, and the retrieval component.

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

The multimedia objects together with the according MPEG-7 document are transferred to the multimedia database. Two different kinds of metadata are contained in the MPEG-7 documents. These are the general metadata (e.g. text annotation, shot duration), which can be searched for by conventional database queries, and content-based metadata (such as the. color histogram of an image or a graph describing semantic relations) where specific indexing and comparison algorithms are necessary. Therefore the multimedia database has to manage the real multimedia data (the essence), the MPEG-7 documents and some specific low-level feature 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.

The multimedia retrieval is completely separated from the multimedia database, according to the objective of modularity. However, we concentrated in our prototype on a Web interface to specify queries and to display the search results. The Web server forwards the queries of each Web client (user) to a broker module. This broker communicates with the multimedia database, groups received search results, and caches binary data. The usage of a 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.

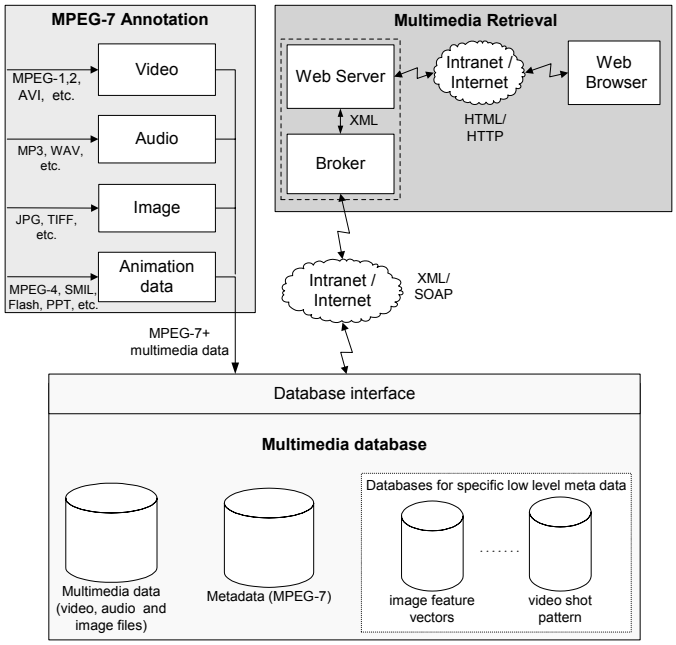


Figure 1: Software architecture.

The search results, which are displayed by the Web user interface, contain extracts of the annotations, a visual summary of the content 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.

3.2. Content Analysis and Annotation

Describing multimedia objects is generally done as part of the ingestion process. An annotation component (see Figure 2) has been implemented for this purpose, which allows to enter parts of the description manually and also to use automated content analysis modules as plug-ins. These modules just need to write their descriptions into the corresponding parts of the MPEG-7 document.

The description of multimedia data comprises the areas described below. These attributes are available for retrieval applications later on.

- *Description of the storage media*: file and coding formats, image size, image rate, audio quality, etc. Depending on the encoding format, these attributes can be detected automatically.
- *Creation and production information*: creation date and location, title, genre, etc. This information typically has to be entered manually, unless it is available in an already existing electronic catalogue.

- *Content semantic description*: content summary, events, objects, etc. A manual description task, which is supported by a graphical editor (described in detail in section 4.1).
- *Content structural description*: shot, scenes, key frames and summaries, etc. Extracted automatically by modules. (described in detail in section 4.2).
- *Low-level content features:* color, texture, shape, motion etc. descriptions to enable content-based retrieval. Extracted automatically by modules.
- *Meta information related to the usage of the content*: right holders, access rights, etc. This information typically has to be entered manually, unless it is available in an already existing electronic catalogue.
- *Meta information about the description*: author, version, creation date, etc. This information can be determined by the annotation application automatically.

All these types of information can be represented using MPEG-7 description schemes.

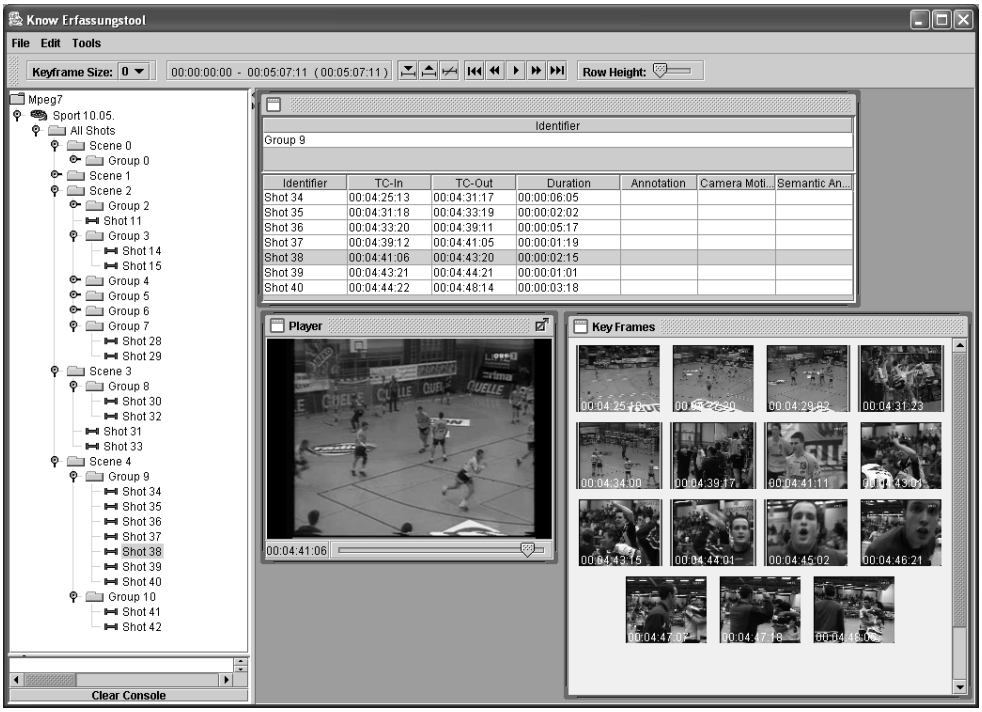


Figure 2: MPEG-7 annotation tool.

3.3. 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 objects), 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, Tamino, Xindice, etc.). However, it still is a challenge to store and manage such complex and large XML documents as MPEG-7 descriptions of audiovisual media typically are. We decided to use an Oracle 9i DBMS for our prototype. The multimedia data (essence) are saved in a file pool and only the file references are managed by the DBMS.

The multimedia database provides a set of Web services as interface to retrieval applications, i.e. the broker. Web services have the advantage that they usually do not interfere with possible firewall mechanisms. All search queries and results are specified in XML format. The retrieval process is implemented in two steps. First the MPEG-7 descriptions of the documents matching the request are retrieved. They also include the structural description of the content-based media summary. All textual information of the result can be displayed immediately. If there are references to derived essence such as key frames, they are requested in a second step and added to the previously received result, depending on the retrieval application. The corresponding multimedia object can be downloaded or streamed by a separate request.

3.4. Multimedia Retrieval

The retrieval tools use a WASP SOAP Client to communicate with the multimedia database's WASP server. The interface between the Web server and the broker is designed as a lightweight protocol for exchanging parameters. The user interface sends a request with the according data to the broker that responds with a result MPEG-7 XML record. Requests are enveloped in a XML structure to be as independent of the communication protocol as possible.

The user interface for search and retrieval is described in detail in section 5.

4. ANNOTATION

In this section two important annotation modules are described in more detail. The first one covers the description of high-level semantics based on agents, events, objects, locations and their relations. The second one automatically generates a visual summary by intelligently selecting key frames from a given video sequence.

4.1. Semantic Annotation

With the current state-of-the art technologies the possibilities for automatic extraction of semantic content are limited. A significant contribution to the annotation of video and audio data is gained by speech recognition [2][3][15] and by recognition of text [17][24]. Text recognition is only applicable in domains with many text inserts, such as news. Our experiments with speech recognition systems have shown that the performance is poor, if many different speakers are involved and the vocabulary cannot be restricted to a small domain. Another possibility to get semantic information is the semantic labeling of scenes (e.g. outdoor, indoor) and detection of objects (faces, cars, etc.) by processing of low-level feature descriptors (e.g. [1], [9], [11], [21]). This requires a small predefined set of scene classes.

Manual annotation therefore remains the most reliable way of adding semantic descriptions to multimedia content, which also is the basis for good retrieval results. The necessary time for this task can be reduced by providing better annotation tools.

While free text annotations are useful when presented to humans, they provide limited search capabilities, as the meaning and context are not understood by the system performing the search. Therefore structured semantic annotations are used in IMB. The semantic descriptor is used to describe the semantic meaning of multimedia content. It can describe real-life situations and concepts in terms of objects and their relation to each other. MPEG-7 distinguishes between several semantic entities: objects, agent objects, events, concepts, states, places, times, and narrative worlds.

MPEG-7 does not define any entities itself. It specifies just the structure for a semantic description of objects. Instances of objects have to be created by applications themselves, because these objects are very context specific. Within IMB we took soccer as an example to evaluate the semantic description approach. Descriptions of soccer players and soccer events stored in a separate semantic object repository. These entities are referenced to multimedia content. A directed graph expresses the relation of the entities. A relation consists of a name, a source entity and target entity. A basic set of 45 relations already comes with MPEG-7; new relation sets can be defined if needed.

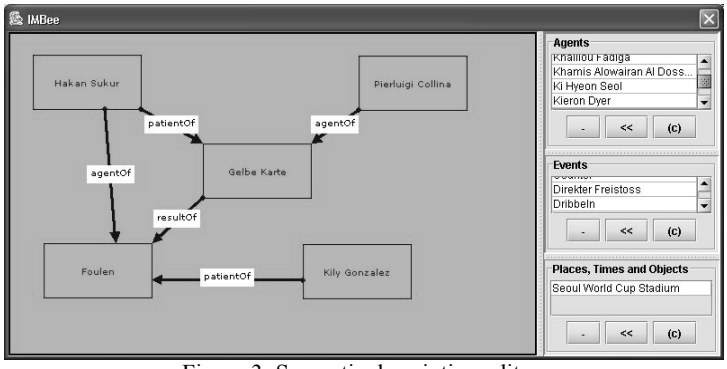


Figure 3: Semantic description editor.

Figure 3 shows the editor which allows creating a semantic description graphically. The boxes represent entities, the arrows their relations. On the right side, the defined entities are listed, from where they can be dragged into the graph. The semantic descriptions are associated with a segment in the video. The same editor is used to create query examples for semantic annotations. When searching for semantic descriptions in the retrieval tool, video segments having similar descriptions are returned.

4.2. Structural Content Description and Summarization

The result of a search in a multimedia database is typically a large amount of audiovisual data. In order to enable the user to deal with this amount of content, it has to be presented in a form which facilitates the comprehension of the content and allows to quickly judging the relevance of search results. Media summarization techniques are strategies that aim at selecting parts of multimedia content which are expected to be relevant for the user. The main task is therefore to decide which parts of a video are important and to select a visual representation for these parts. The result of a media summary can be organized sequentially (a list of key frames or a video skim [29]) or hierarchically (e.g. [7]). The hierarchical presentation has the advantage of guiding the viewer from a rough overview to a relevant segment.

The summarization strategy used in IMB works in three steps: First shot boundaries are detected and key frames are extracted. Then a hierarchical structure of the video is generated by clustering visually similar shots into shot groups and scenes. Finally, representative media items (in IMB key frames are used) for each of the shots, groups, and scenes are selected.

Shot Boundary Detection and Key Frame Extraction

Shot boundary detection is a preprocessing step which is required for many tasks in video content analysis. The algorithm used in IMB is an extension of that proposed in [30] and supports detection of hard cuts, dissolves and fades. The approach for hard cuts is based on color histograms. Following our goal to consequently use MPEG-7, we employed the scalable color descriptor for that purpose. For transition detection, the edge energy is a suitable feature. We use the MPEG-7 edge histogram descriptor, which has two advantages over the simple edge energy measure: It describes the edge direction, so that changes between different shots with similar amount of edges can be detected more reliably, and it contains a rough, block based spatial information, which makes the transition detection more robust against local changes (e.g. text inserts). The use of these MPEG-7 descriptors for shot boundary detection has been independently proposed in [13].

Key frames are extracted in a motion adaptive way, i.e. the number of extracted key frames is proportional to the amount of visual change in the image sequence. For each key frame, a number of MPEG-7 color and texture descriptors are extracted, which are later used for similarity search.

Hierarchical Structuring of Videos

The hierarchical structure of the video is based on the shot structure, i.e. assuming that the boundaries between higher-level blocks coincide with shot boundaries. We follow the method proposed in [23] to create *groups* of shots, which are

defined by similarity of visual features, and *scenes* defined as sequences of groups. In contrast to groups, scenes are required to be temporally continuous. Therefore temporally overlapping groups will be put into the same scene.

Like in [23], a similarity measure between any shot pair in the video is calculated. In our algorithm, we use the following features:

- Average visual activity between the shots, calculated from frame pair differences in RGB color space.
- The color similarity of two shots is determined as the maximum of the pair-wise color similarities of the first and last key frames of the shots. In [23] the first and last frame of a shot are used for efficiency reasons, but the key frames are in most cases more representative and as they are extracted for later visualization using them causes no additional computational burden. The color similarity between two key frames is calculated as one minus the distance between the MPEG-7 scalable color descriptors of the two frames.
- A temporal weighting measure is used to avoid that visually similar, but temporally distant shots are put into the same group. The temporal weight can be used as a parameter to determine the average length of a group.

The formation of scenes is a much harder problem than that of visually similar groups, especially when groups are temporally overlapping. This means that there is an alternating pattern of visually similar shots. In some cases, e.g. in the case of a dialog scene, where the dialog partners are alternately shown, it makes sense to put these groups into the same scene. This can for example be a problem in a news broadcast, where very short news stories cause shots showing the anchorperson before and afterwards to be put into the same scene because of their similarity.

To overcome this problem, we introduced *scene boundary indicators*, i.e. a set of frames that typically introduce a new scene. If the first key frame of a shot is visually similar in terms of color (MPEG-7 scalable color descriptor) to one of the scene boundary indicators, the likelihood of a scene boundary at this position is increased.

The structure of the video is described in MPEG-7 using a hierarchy of video segments and temporal decompositions. The resulting structure is also visualized as a tree in the annotation tool (cf. Figure 2).

Representation of Video Segments

Key frames are commonly used as visual representation of videos. Of course this is a limited representation because of the lack of motion and the selection of distinct time points. Their advantage is that a user can view many of them simultaneously, while he can only view one video at a time. Key frames are therefore a tool to provide a summarized overview, provided that they have been selected in order to be a meaningful visualization for the video segment they represent.

In contrast to other approaches (e.g. [7]), which start from the whole video, we use the previously determined content-based structure as the starting point. In IMB, we do not have any prior knowledge about the user or the context of the search, so the main criterion for the relevance of a specific key frame is the duration which it represents. Another general criterion which has been proposed [26] is the rarity of a frame. Of course it is useful in some cases to see that a certain visual content appears only in one shot and nowhere else. But especially if the number of key frames used in the summary is low (which makes it easier to comprehend), it is more important to choose a key frame representing as much as possible of the video segment.

The key frame selection algorithm we propose works as follows:

- Within a shot, the key frames are selected by the duration they represent. The visual similarity within a shot is not considered, as key frames have been extracted based on changes of the visual content.
- All key frames of a group (also those which have not been found to be representative for a shot) are clustered by color similarity using the MPEG-7 scalable color descriptor. For each cluster, the duration it represents is calculated from the duration for which its member key frames are valid. The clusters representing the longest duration are chosen.
- A scene is visualized by the representative key frames of its group. These key frames cannot be visually similar, as this was the criterion for forming the groups.

- The key frames to visualize the whole video are chosen from the representative key frames of all groups, which are clustered by visual similarity, again using the scalable color descriptor.

The generated summaries are stored using the MPEG-7 hierarchical summary description scheme. Shots, groups and scenes are represented using summary segment groups. For each group, the representative key frames are described as summary segments, using the order attribute to indicate the order of their relevance.

5. RETRIEVAL AND RESULT PRESENTATION

5.1. Query User Interface

A query interface has to enable an easy and intuitive query specification for the user. IMB offers a classic Web browser based search interface with text edit boxes and selection lists. Such an interface is necessary, as it does not require anything else than a Web browser but it does not provide means for formulating content-based queries. Therefore a multimedia query specification interface has been developed with realizes components for the following search facilities:

- Keyword search.
- Query by example with using images.
- Query by sketch based on MPEG-7 metadata using camera movement and color distributions.
- Generate MPEG-7 data matching parameters.
- Support of semantic queries based on the MPEG-7 semantic descriptor.

Main points in the query interface are components dealing with MPEG-7 descriptions. The components have a graphical interface that visualize the data and interacts with the user. These MPEG-7 components are used for search specification, search result visualization and annotation. Search results can be re-used as input for the MPEG-7 components and the user can easily inspect whether the results are similar to the query or not.

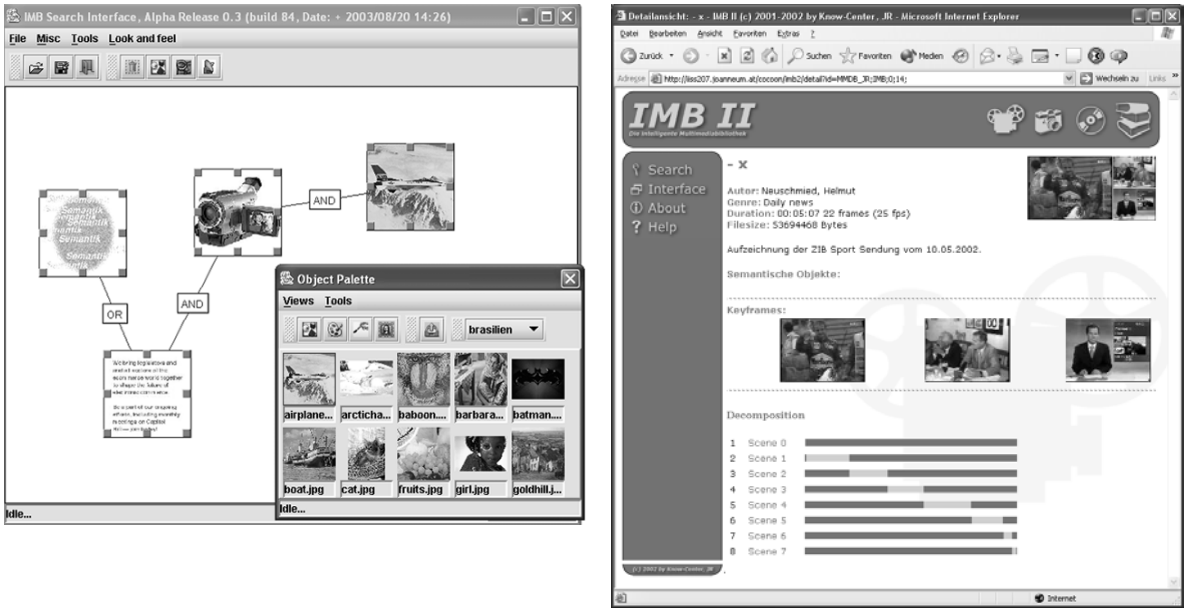
The query user interface uses various MPEG-7 components to visualize and edit specific multimedia data like the color distribution or semantic graphs. To provide an uniform multimedia query interface, the various multimedia terms, their data and their combinations need to be presented in a consistent way. However, each multimedia term has to be handled and presented according to its specific behavior. In the user interface, each multimedia term has its own panel that covers its specific needs. The panel opens when the user selects a multimedia term for inspecting. The search interface includes two main panels. An object catalogue is used as repository containing selectable objects and a drawing area where the user can group the query terms, set their relations to each other and open panels to edit specific multimedia terms.

The user interface supports two query paradigms:

- In a *Query by Example* automatic metadata extraction algorithms are used to specify the query automatically. The focus is on an automatic generation of the query data without user interaction.
- In a *Query by Sketch*, the user defines a search scenario by hand. The query data is generated semi-automatically. The system generates the query data from the user input.

These two paradigms can be mixed to enable a semi-automatic query specification. The automatic data extraction process can be used to generate some basic metadata used as entry point for a query specification - the system supports the user to define a search sketch.

The query user interface shown in Figure 4a has been implemented using Java Web Start. It can thus be started directly from the Web browser and will install missing components without any further user interaction.



(a) **(b)**

Figure 4: (a) Query user interface, (b) Summary visualization.

5.2. Result Presentation

For presenting the MPEG-7 descriptions of the search results, we use a cross-media publishing framework integrating the capabilities of XML and XSLT: Cocoon 2 [6]. We used Cocoon's XSP (XML Server Pages) generator to generate an XSP document from the response of the multimedia database. The result is then transformed to the desired output format, in the case of the Web interface this is HTML and SVG for visualization.

The MPEG-7 video summary is used to generate an overview visualization of a video item containing the overview key frames for the currently shown video segment (cf. Figure 4b). The three most relevant key frames are displayed. We have also implemented the comic book style key frame visualization proposed in [26], where the size of the key frames depends on their importance. The user can navigate within the summary by clicking on the key frames or by using the time diagram at the bottom of the page.

6. CONCLUSION AND FUTURE WORK

The actual work with MPEG-7 demonstrates that this standard provides an extensive set of attributes to describe multimedia content, which is exactly its purpose. However, the complexity of the description schemes makes it sometimes difficult to decide how a semantically correct description is given. This may lead to difficulties when interchanging data with other applications. Nevertheless the standardized description language is easy to exchange and filter 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 communication of any Web agent with the broker to support user specific retrieval definition. The different output capabilities could be easily extended for special result representation on mobile devices, first tests had been made with WML.

Experiments with scene structuring and summarization have shown, that the approach works well for clearly structured content, such as news, but does not work well for other content. This is also confirmed by other authors [25]. The main issue is that the relevance of a certain part of the content depends on both the type of content and the user.

The challenge of content dependency could be tackled by exploiting more and high-level information. For example, it has been shown, that the use of audio information can increase the accuracy of scene boundary detection [4][27]. The same goal could be reached by exploiting high-level information, such as results of object detection. Furthermore, we plan to use the semantic annotations entered manually for structuring and summarization.

Currently, the summaries generated by IMB are queried independent, although the relevant content may be different in the context of different queries. This could be improved by not creating the summary in advance but at query time, as the relevance criterion is better known then.

ACKNOWLEDGEMENTS

The work described in this paper has been supported by several colleagues within JOANNEUM RESEARCH and the Know-Center Graz, whom the authors would like to thank here. The Know-Center is a Competence Center funded within the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology.

REFERENCES

1. B. Adams et al., "IBM Research TREC-2002 Video Retrieval System", *Proc. Text Retrieval Conference TREC 2002 Video Track*, Gaithersburg, MD, 2002.
2. G. Backfried, J. Riedler, "Multimedia Archiving with Real-time Speech and Language Technologies", *IEEE Conference on Information, Communication & Signal Processing (ICICS)*, Singapore, 2001.
3. W. Brown, S. Srinivasan, A. Coden, D. Pronceleon, J. W. Cooper, A. Amir, "Toward speech as a knowledge resource", *IBM Systems Journal*, **vol 40**, no. 4, 2001.
4. Y. Cao, W. Tavanapong, K. Kim, J.-H. Oh, "Audio-Assisted Scene Segmentation for Story Browsing", *Proc. CIVR 2003*, Urbana-Champaign, IL, 2003.
5. S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ: An Automatic Content-Based Video Search System Using Visual Cues", *Proc. ACM Multimedia '97 Conference*, Seattle, WA, November 1997.
6. *Cocoon*, URL: http://cocoon.apache.org/, 2003.
7. A. Girgensohn, J. Boreczky and L. Wilcox, "Keyframe-based User Interfaces for Digital Video", *IEEE Computer Magazine*, **vol. 34**, nr. 9, pp. 61-67, Sept. 2001.
8. S. Handschuh, S. Staab. A. Maedche, "CREAM — Creating relational metadata with a component-based, ontology-driven annotation framework", *K-CAP'01*, pp. 76-88, Victoria, Canada, Oct. 2001.
9. A. Hauptmann, M. Witbrock, Informedia, "News-on-Demand Multimedia Information Acquisition and Retrieval", In Mark T. Maybury, Ed., *Intelligent Multimedia Information Retrieval*, pp. 213-239, AAAI Press, 1997.
10. O. Hein, "pViReSMo - Personalisiertes Video Retrieval auf der Basis von Metadaten", 4. IuK-Tage Mecklenburg-Vorpommern, Rostock, 2003.
11. X.-S. Hua et al., "MSR-Asia at TREC-11 Video Track", *Proc. Text Retrieval Conference TREC 2002 Video Track*, Gaithersburg, MD, 2002.
12. T. S. Huang, S. Mehrotra, and K. Ramchandran, "Multimedia Analysis and Retrieval System (MARS) project", *Proc. of the 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Retrieval*, University of Illinois at Urbana-Champaign, March, 1996.
13. M. Höynck and C. Mayer, "Usage of MPEG-7 Descriptors for Temporal Video Segmentation", *Proceedings of the 10th Aachen Symposium on Signal Theory (ASST2001)*, pp. 391-396, Aachen, Sept. 2001.
14. J. Hunter, "Enhancing the Semantic Interoperability of Multimedia Through a Core Ontology", *IEEE Trans. on Circuits and Systems for Video Technology*, **vol. 13**, nr. 1, Jan. 2003.
15. F. de Jong, J.L. Gauvain, j. den Hartog, K. Netter, OLIVE, "Speech Based Video Retrieval", *Proc. of the European Workshop on Content-Based Multimedia Indexing (CBMI)*, Toulouse, 1999.
16. H. Lee H, A. Smeaton, C. O'Toole, N. Murphy, S. Marlow and N. O'Connor, "The Físchlár Digital Video Recording, Analysis, and Browsing System". *RIAO 2000 - Content-based Multimedia Information Access*, Paris, France, April 2000.

17. R. Lienhart, “Indexing and Retrieval of Digital Video Sequences Based On Automatic Text Recognition”, *Fourth ACM International Multimedia Conference*, 1996.
18. T. D. Ng, M. G. Christel, A. G. Hauptmann and H. D. Wactlar, “Collages as Dynamic Summaries of Mined Video Content for Intelligent Multimedia Knowledge Management”, *AAAI Spring Symposium Series on Intelligent Multimedia Knowledge Management*, Palo Alto, CA, 2003.
19. J. van Ossenbruggen, J. Geurts, F. Cornelissen, L. Hardman and L. Rutledge, “Towards Second and Third Generation Web-Based Multimedia”, *WWW10*, pp. 479-488, Hong Kong, May 2001.
20. M. Rautianen et al., “TREC 2002 Video Track Experiments at MediaTeam Oulu and VTT”, *Proc. Text Retrieval Conference TREC 2002 Video Track*, Gaithersburg, MD, 2002.
21. S. Raaijmakers, J. den Hartog and J. Baan, “Multimodal Topic Classification of News Video”, *IEEE Int. Conf. on Multimedia and Expo*, Lausanne, 2002.
22. *Ricoh Movie Tool*, URL: http://www.ricoh.co.jp/src/multimedia/MovieTool/, 2003.
23. Y. Rui, T. S. Huang, S. Mehrotra, “Constructing Table-of-Content for Videos”, *ACM Multimedia Systems Journal*, , Sept, 1999.
24. T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, “Video OCR for Digital News Archive”, *IEEE International Workshop on Content-Based Access of Image and Video Databases (CAIVD)*, pp. 52-60, 1998.
25. A. F. Smeaton, “Challenges for Content-Based Navigation of Digital Video in the Físchlár Digital Library”, *Proc. CIVR 2002*, pp. 215-244, London, 2002.
26. S. Uchihashi, J. Foote, A. Girgensohn and J. Boreczky, “Video Manga: Generating Semantically Meaningful Video Summaries”. *Proc. ACM Multimedia*, pp. 383-392, 1999.
27. A. Velivelli, C.-W. Ngo and T. S. Huang, “Detection of Documentary Scene Changes by Audio-Visual Fusion”, *Proc. CIVR 2003*, Urbana-Champaign, IL, 2003.
28. *VideoAnnex* Annotation Tool, URL: http://www.research.ibm.com/VideoAnnEx/, 2003.
29. H. Wactlar, “Search and Summarization in the Video Medium”, *Proc. Imagina 2000*, Monaco, 2000.
30. H. Yu, G. Bozdagi and S. Harrington, “Feature based hierarchical video segmentation,” *Proc. Int. Conf. Image Processing*, pp. 498–501, Washington, DC, USA, Oct. 1997.