An innovative system for formulating complex, combined content-based and keyword-based queries

Herwig Rehatschek*, Peter Schallauer, Werner Bailer, Werner Haas, Alfred Wertner
JOANNEUM RESEARCH Forschungsgesellschaft mbH, Institute of Information Systems & Information Management, Steyrergasse 17, A-8010 Graz, Austria, Europe

ABSTRACT

We propose a search & retrieval (S&R) tool, which supports the combination of a text search with content-based search for video and image content. This S&R system allows the formulation of complex queries allowing the arbitrary combination of content-based and text-based query elements with logical operators. The system will be implemented as a client/server system. The entire S&R system is designed in such a way that the client system can be either a web application accessing the server over the Internet or a native client with local access to the server. The S&R tool is embedded into a system called "MECiTV – Media Collaboration for iTV". Within MECiTV a complete authoring environment for iTV content will be developed. The proposed S&R tool will enable iTV authors and content producers to efficiently search for already existing material in order to reduce costs for iTV productions.

Keywords: content-based, image, video, retrieval, MPEG-7

1. INTRODUCTION

The growth of information is nowadays enormous and has reached a level which has never been reached before. We currently produce in one year more information than it was produced in the entire history of humans so far. In Europe, the holdings of the broadcast archives alone are estimated at about 10 million hours of film, 20 million hours of video and 20 million hours of audio content [1]. For example the holdings of the Dutch national archive "Beeld en Geluid", yearly grow by 5,000 hours of video and 15,000 hours of audio [2].

In the context of production for interactive TV (iTV) the use of archive material is of special relevance. To provide interactivity, several parallel streams must be produced, which significantly increases production costs. The use of archive material could help to keep these costs lower. The work presented in this paper is part of the project "MECiTV" (Media collaboration for iTV), which aims at providing tools for the entire iTV production workflow.

The main requirements for the re-use of archive material are to have a digital multimedia archive and to be able to access its content easily and efficiently. Recently, broadcast archives have started to make their content accessible, in some cases to the general public via web interfaces (cf. the research projects AMICITIA [3], BIRTH [4] and PRESTO [5]). Efficient searching in huge amounts of video and image data contained in digital archives is a scientific challenge. Especially the extension of the search by content-based methods is currently an ongoing research topic all over the world. Many efforts are invested in so called "low level feature extraction". Typical low level features are colour histogram extraction (and similarity search based upon it), key-frame extraction and shot detection. The derive of semantic information – the next logical step after having low level features – is currently investigated by research teams all over the world. The reason for this is: only when having also semantic information on the digital content a typical human search process can be supported efficiently. This is because humans always have a special context and pre-existing knowledge when searching for content, which the machine does not have a priori. There are already a number of prototype systems going into the direction of offering semantic search (see section 2), however, commercial systems in this sector are still not available or very limited in their functionality.

We propose a search & retrieval (S&R) tool, which supports the combination of a text-based search with content-based search for video and image content. By "text-based search" we mean searching for annotations added manually or

* Herwig.Rehatschek@joanneum.at; phone: +43 316 876 1194; fax: +43 316 876 1191; www.joanneum.at/iis

This paper was published in S. Santini and R. Schettini (eds.), *Proc. Internet Imaging V*, San Jose, CA, USA, Jan. 2004, pp. 160-169 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.

extracted from external sources (such as existing metadata descriptions or production information). In the area of content-based search our search and retrieval system supports similarity search by global color and texture features and camera motion (dominant motion), as well as by color, texture and shape descriptions of moving objects. Furthermore the S&R tool allow the formulation of complex hierarchical queries allowing the arbitrary combination of content-based and text-based query elements with logical operators. Special emphasis was put on visualization of video content, since this is a very important aspect in content based search & retrieval applications. It can be time-consuming to view a number of search results in order to judge their relevance. A summarized presentation of a media item allows to speed up this judgement.

The search and retrieval system is implemented as a client/server system. The metadata extracted by automatic content analysis modules as well as that annotated manually is fully MPEG-7 compliant and is the basis for the content-based search.

2. RELATED WORK

This chapter contains an overview on previous work on search & retrieval and visualization tools for multimedia content. Special emphasis was put on systems providing content-based retrieval capabilities and their user interfaces for query formulation and result presentation.

The keyframe based GUIs suggested by [6] are already implemented in a video database system for more than two years. This approach introduces some new concepts in the direction which keyframes should be displayed in connection with a summary (e.g. clustering of shots, giving importance scores for shots), besides only displaying keyframes.

Corridini, Del Bimbo et al propose in [7] a new conceptual model for describing films together with a hypermedia navigation system to search and browse in the digital movies. The system is based on a feature extraction engine which offers shot detection, a simple scene analysis and a spatial temporal logic extraction. The resulting browsing interface is quite complicated and only meant for people who have deep background in image processing.

VICAR is a project which was finalized in 1999 [8]. VICAR aimed at the development of a system for the documentation and analysis of digital video material. The project result was a tool named *VIN–VideoNavigator* [9] for video browsing, text annotation and content analysis tools restricted to MPEG-1 format.

VIZARD introduced a new intuitive concept to deal with video content [10]. The software supports users to create a sketch or a story they want to tell. The system is targeted for typical SOHO users (small office and home users) and thus very restricted in its functionality towards professional video producers. One of VIZARD's components, the *VExplorer*, provides simple methods for searching, but does not support content-based queries, and the *VPublisher* offers a simplified media summary view.

pViReSMo [11] provides a personalized, semantic selection and filtration of multi-media information based on a client/server infrastructure. The generated content description is MPEG-7 compliant. For manual annotation, Ricoh's *MovieTool* [12] is used. The system supports text queries only.

Caliph and Emir are two tools for the management of digital photo archives [13]. Caliph is an annotation tool which allows creating high-level MPEG-7 compliant semantic descriptions. Emir supports retrieval in file system based photo repositories and supports queries for text annotations, global color similarity and similar semantic descriptions. The tool is currently restricted to smaller photo collections.

The Físchlár system [14] is a content analysis, search and retrieval framework for digital video. Both automatically and manually annotated content is stored using MPEG-7 format. The system provides a web-based search and retrieval interface and supports content-based navigation and keyframe based summarization.

The Informedia project aims at combining speech recognition, image understanding and natural language processing technology to automatically transcribe, segment and index linear video. Within Informedia-II new paradigms for video information access and understanding were introduced, enabling summarization and visualization that provides responses to queries in a useful broader context (with historic or geographic perspectives) [15]. The prototype database developed in Informedia-II allows for rapid retrieval of individual video paragraphs which satisfy an arbitrary spoken or typed subject area query based on the words in the soundtrack, closed-captioning or text overlaid on the screen. One

emphasis within Informedia-II was laid on the generation of summaries for each story segment: headlines, filmstrip story-boards and video-skims.

3. THE CONTEXT OF THE SEARCH & RETRIEVAL TOOL

This section describes the system in which the S&R tool is used. The MECiTV project aims at developing tools to support the entire iTV production workflow as it is depicted in Fig. 1. The search and retrieval system is at the beginning of the production workflow, supporting authors of iTV content at finding archive material for use in a new production. The material can then be used in the iTV authoring tool, either directly or after additional editing in an off-the shelf non-linear editing system.

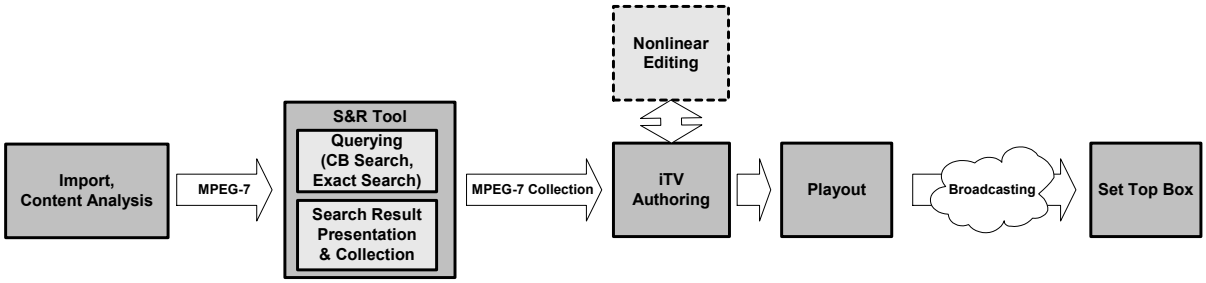

Fig. 1: Typical iTV production workflow.

The components of the search and retrieval system are shown in Fig. 2 and briefly described in the following.

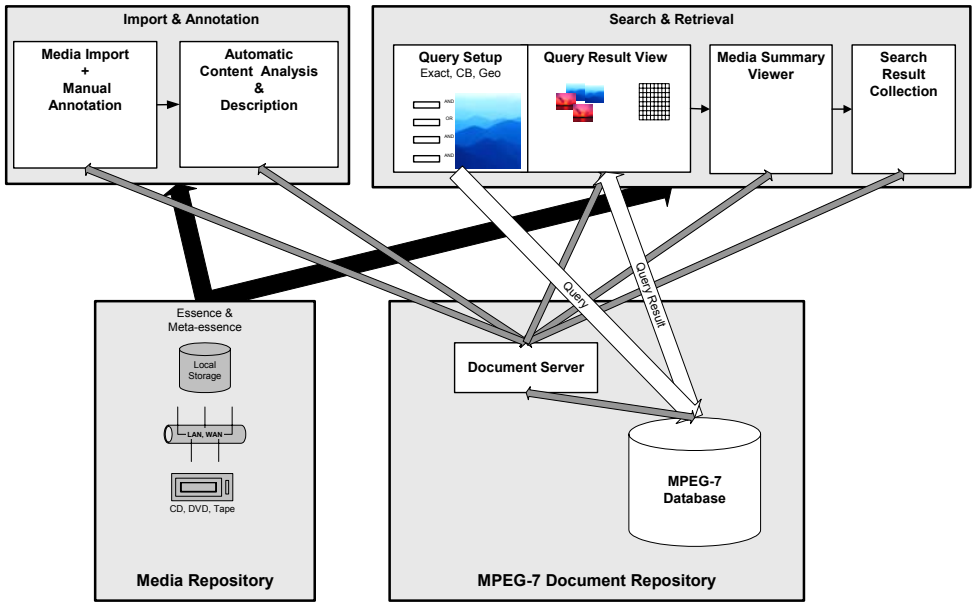

Fig. 2: Architecture of the search & retrieval system.

3.1. Data Repositories

The **Media Repository** holds all the imported media data (essence), typically in both master and preview quality. It also holds media data derived from the essence during the content analysis phase, which we call meta-essence, such as key frames or stripe images (temporal overview images generated from the center column of every n-th frame, cf. Fig. 7). The media can be stored on local or network disk storage and on accessible video/audio tape devices.

The **MPEG-7 Document Repository** holds the metadata descriptions of the media items. The MPEG-7 documents are stored in a relational database, which provides extensions for storing and querying XML documents. The document server component allows accessing and manipulating whole MPEG-7 documents as well as of parts of documents. Both text based and content-based queries are processed directly by the database.

3.2. Import and Annotation

3.2.1. Import Tool

The **Import Tool** is the GUI for adding new media items to the repository, controlling status and parameters of the automatic content analysis modules and for manual annotation (e.g. production information). To view the results of automatic content analysis, the Media Summary View (cf. Section 0) can be invoked from the Import Tool. The GUI of the import tool is shown in Fig. 3.

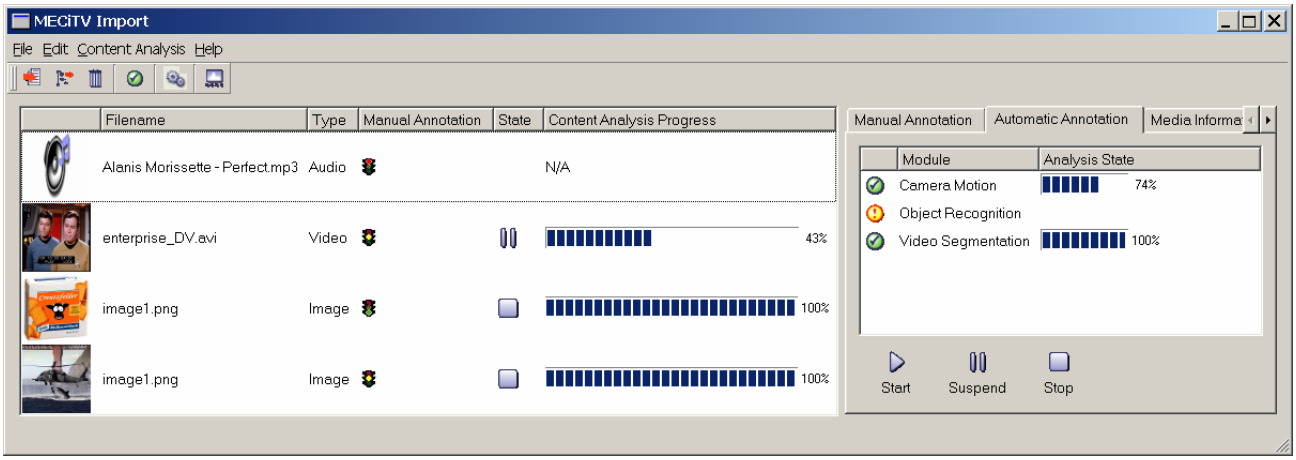

Fig. 3: The import tool of the S&R system.

3.2.2. Automatic Content Analysis Modules

Video content analysis is performed by the analysis modules described below. The implementation of the modules does not depend on any specific video format or encoding. The output of the analysis modules is a description in MPEG-7 format.

Temporal Video Segmentation and Keyframe Extraction

This module performs temporal segmentation of the video. This includes the detection of shot boundaries and the classification of transitions into abrupt transitions (cuts) or gradual transitions (e.g. dissolves). For gradual transitions the duration of the transition is estimated. For each of the shots that were identified during the temporal segmentation, one or more key frames are extracted, depending on the visual activity within the shot. For each of the keyframes, MPEG-7 compliant colour and texture descriptors are extracted.

Camera motion analysis module

The module estimates the background motion in the video sequence. In most cases this will equal the camera motion. For search and retrieval purposes an exact parametric description of the background motion is neither necessary nor reasonable. We therefore generate a simplified description restricted to nine types of camera motion (fixed, pan left/right, tilt up/down, roll left/right, zoom in/out).

Moving object extraction

The task of this module extracts moving semantic video objects, i.e. video regions which correspond to real-world objects. A semantic video object is described by shape, colour and texture. The extraction of semantic video objects is

based on colour segmentation, which usually leads to over-segmentation of the video, i.e. too many small regions will be found. This problem can be overcome by additionally using motion information, so that regions, which move similarly over time, are grouped to yield a semantic video object. As motion information is required to create the segmentation, only moving objects will be extracted.

The result of the automatic segmentation will be a set of semantic video objects. The objects' descriptions are stored the in MPEG-7 document describing the video from which they were extracted. Each object is described by its contour and by its colour and texture features. This allows a generic description of any video object, which is not restricted to a set of predefined objects. To search within the set of extracted objects, the same contour, colour and shape descriptors are extracted from a query sample and the most similar objects in terms of these descriptors are returned.

3.3. Search and Retrieval

The **S&R Tool** is the main issue of this paper and is described in detail in Section 4.

The communication between the components of the search and retrieval system is done using CORBA [16]. The components may therefore be remotely distributed over a network. The use of web services for the interfaces between the components has been considered, and would especially be beneficial for the communication with a planned web-based search & retrieval client. Because of the fact that asynchronous communication is required between the components, which is only available for web services in a technology dependent way (i.e. there are specific asynchronous call implementations for Java and MS .net web services), we have decided to use CORBA for the first version of the system.

4. THE SEARCH & RETRIEVAL TOOL WITH COMPLEX QUERY FORMULATION

The search and retrieval tool is the GUI for query formulation and result presentation. It provides an intuitive interface for the formulation of hierarchically structured combined text- and content-based queries. In addition to a simple result list view, a media summary view can be used to quickly judge the relevance of a search result.

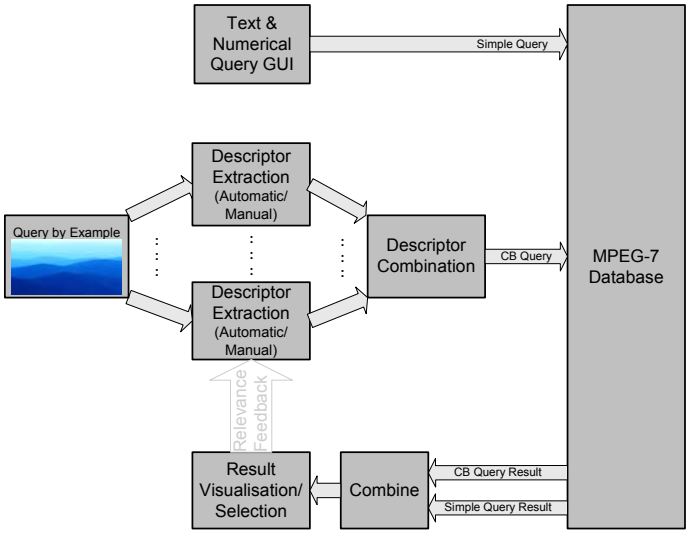

Fig. 4: Formulation of text and content-based queries. Optional relevance feedback is possible.

The search and retrieval system supports the following types of query elements (cf. Fig. 4):

Text based elements: This class of query elements is used for searching metadata entries. The result will be a set of media items, which have a metadata description that exactly matches the query (by equality, greater or less). The parameters for these search criteria are entered by the user as text or numerical values.

Content based elements: A media item (image, visual, audiovisual) is used as query example. The result set will be consist of media items which are similar in terms of the selected search criteria. The query example is either

automatically analyzed by the content analysis modules described in Section 3.2.2 or the parameters are entered manually by the user, for example by selecting a type of camera motion or by drawing a shape to define a region being used as query example.

4.1. Search & Retrieval User Interface

The S&R tool enables the researcher/author to effectively search in the digital multimedia archive by formulating complex queries containing both text-based and content-based elements.

The S&R GUI consists of two parts: a query space (shown in Fig. 5) and a result presentation space (shown in Fig. 6).

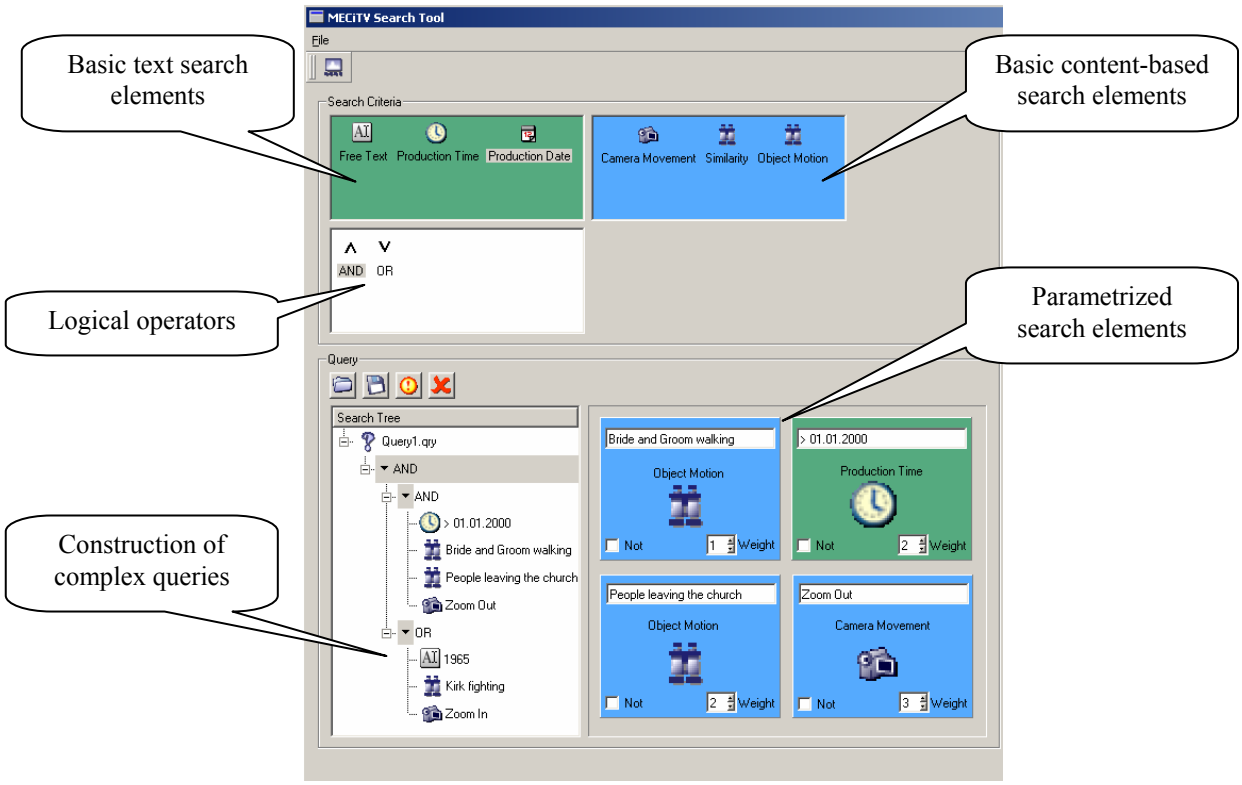

Fig. 5: Graphical user interface for formulating complex combined text and content based queries.

The query interface enables a user to formulate a complex hierarchical query in an intuitive and graphical way. A query is constructed from basic query elements (search criteria) which can be combined with logical operators. These elements of a query can be selected from tool boxes on top and dragged into the search tree in the lower left corner. The search criteria comprise text-based query elements (such as title, production date, director, etc.) and content based elements. The search criteria can be easily extended to support new types of queries. To formulate a query the user builds up a so called query tree, which is visualized in the bottom left area of Fig. 5. The leafs of the tree are search criteria, all other nodes are logical operators. Within one level of the tree one logical operator (AND, OR) can be applied, additionally for each query element a relative weight and the logical NOT operator can be applied.

Right of the search tree, the search criteria of the selected branch of the tree are shown in detail. Here the parameters for the search criteria can be set. For text based and some content-based search criteria (e.g. camera motion), this can be done by entering text or selecting values from a list. For content-based search criteria, this is done by dragging an example (e.g. a video segment or an image) onto the query element. The example can originate from the file system or from the results of a previous query.

From the query tree defined by the user, a query formulated in SQL/XML [17] is generated and sent to the MPEG-7 database. The result is a list of references to the metadata descriptions of the matching media items.

As an example, in Fig. 5 a query with two levels of complexity was constructed, which performs the following combined keyword and content search: find all items which are [("produced after 1 January 2000" AND "contain a bride & groom walking" AND "contain people leaving a church" AND "contain a zoom out")] AND [("contain free text 1965" OR ("contains Captain Kirk fighting" OR "a zoom in")]. This is already a rather complex query which contains already an arbitrary combination of content and text based search elements.

Search results are displayed in the right half of the S&R tool GUI (Fig. 6). The results are presented in a list, showing some metadata and a representative key frame. Each item in the list can be viewed in the Media Summary View (cf. Section 0). Apart from a simple list view of the result items users have the possibility to hierarchically organize search results within a freely definable folder structure. This is referred to as the result collection and is a special feature which shall help authors to organize search results according to their principle story line. The result collection is stored as an MPEG-7 Collection and therefore easily interchangeable between MPEG-7 compliant applications.

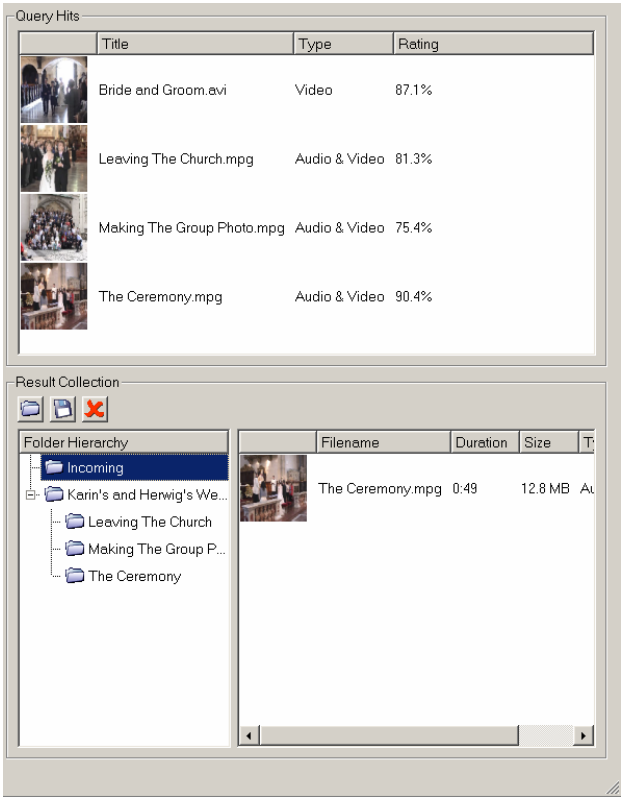

Fig. 6: Query result view of the search & retrieval tool.

4.2 Media Summary Visualization

The result list just gives a very brief overview of the items found in response to a query, but the information that can be presented in this view is often not sufficient to judge the relevance of a multimedia item. The main reason is that the temporal dimension of the video or audio item cannot be sufficiently represented in this view. Playing the media items in the result list is very time consuming and therefore not an option. We therefore provide a summarized media view, which enables the user to quickly get an overview of the content of the media item and judge its relevance.

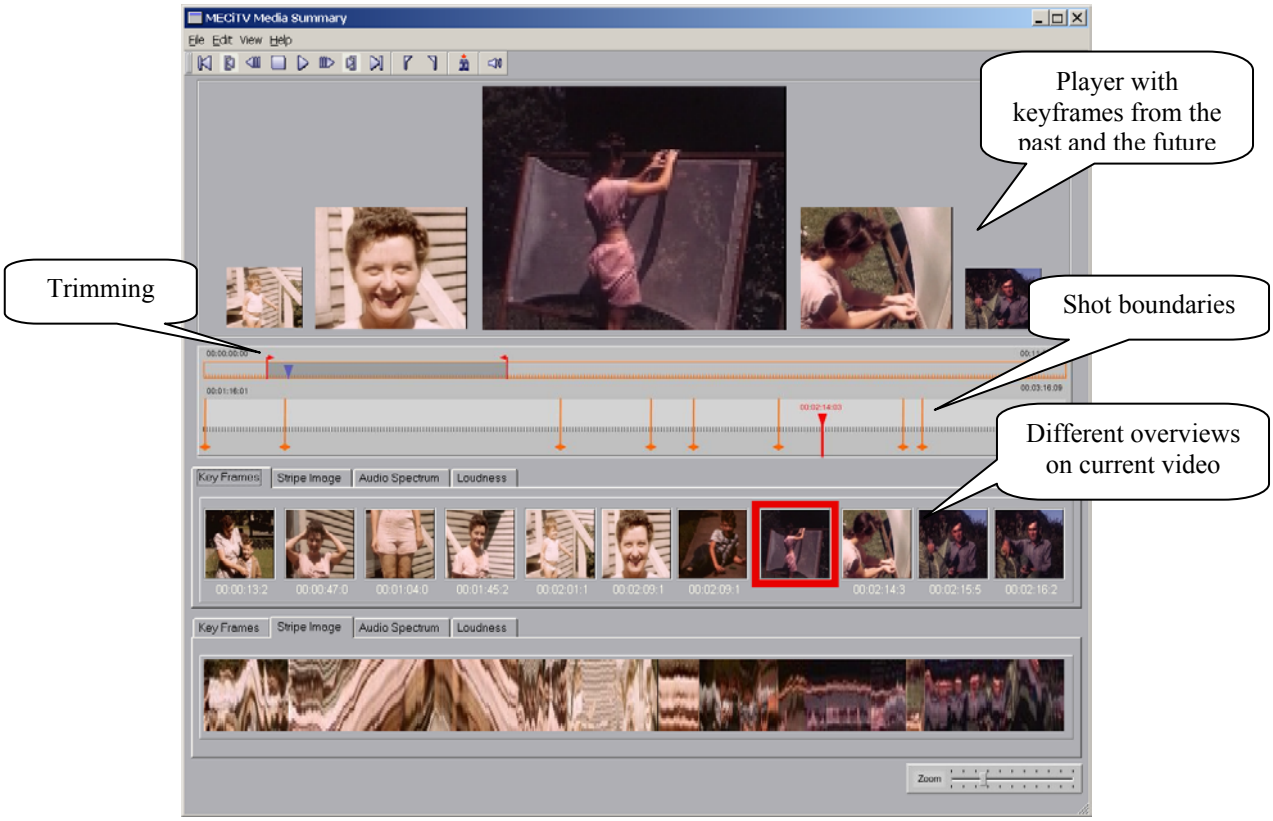

Fig. 7: Media Summary Viewer.

As shown in Fig. 7, this media summary view consists of a player (with presenting relevant key frames of shots from the past and the future) in the upper area of the screen.

The timeline control in the middle visualizes the current position and the segment of the video shown in the current screen. Additionally, the shot boundaries are shown. It also offers a trimming function, which allows to select only a relevant part of a media item to be added to the result collection.

The lower area provides different instruments to get an overview on the entire video, such as relevant key frames of the currently visualized video segment, the stripe image of the segment and audio level and spectrum visualization.

5. SUMMARY AND CONCLUSIONS

We proposed an efficient and intuitive S&R tool for multimedia content which allows the formulation of complex, combined content and text based queries. The proposed tool is part of a search & retrieval system used in an iTV authoring environment (MECiTV system) and enables content producers to efficiently search for already existing material which can be re-used in current production. Reusability is an important topic, especially in connection with iTV productions. This is due the fact, that for real interactivity many parallel streams are necessary, which significantly increases production costs.

The search and retrieval system contains manual and automated annotation components that generate MPEG-7 compliant metadata description which are the basis for both text and content-based search.

The user interface of the S&R tool is the most innovative component of the proposed tool. It enables the researcher/author to effectively search in the digital multimedia archive by a combination of a text and a content-based search. This is fully supported by an intuitive GUI which allows the graphical formulation of complex hierarchical queries. Complex queries can be formulated by defining a search tree using drag&drop, where the query elements (search criteria) in one level are interconnected by a logical operator (AND, OR).

For efficiently representing audiovisual query results the media summary view has been proposed in this paper. This view gives the user an overview of an entire media item within one screen. For this purpose different elements, such as relevant key frames, shot boundary indicators, stripe images, audio spectrum visualization and a player with history/future function are offered.

The modular concept of the search and retrieval framework allows an adaptation to support a web-based search & retrieval client. The inter-component interfaces are already designed to be implemented as web services. The main problem of implementing the search and retrieval tool as web application is the access to media essence and meta-essence which requires a high-bandwidth connection. For that purpose, the media summary visualization has to be reduced to key frames, which may not be sufficient for the professional users which are the target group of this system.

6. FUTURE WORK

Recently the first prototypes for the user interfaces have been finalized, and the implementation phase has started. A first prototype of the S&R system is expected to be available in April 2004, the final system is scheduled for February 2005. A web based-version of the search and retrieval tool is planned. This will also require an adaptation of the media summary view to present a reduced media summary.

As stated in the paper it is a known issue that archive researchers are used to search by text queries rather than by visual queries, such as query by example (cf. [18]). In order to optimally support this we plan to extend our current content based search facilities–which are mainly based on visual input–into the direction of semantic text based queries.

This requires of course the content analysis modules to be further developed and extended. In addition to object detection and feature extraction this requires recognition of faces and objects. The drawback in comparison to our current approach is that the set of objects to be recognized has to be known in advance, the advantage is of course that semantic information is gained which is searchable for the user. In addition, it will then also be possible to recognize static objects, while currently only moving objects can be detected.

AKNOWLEDGEMENTS

The R&D work carried out for the S&R tool is partially funded under the 5th Framework Programme of the European Union within Key Action III of the IST Programme (project "MECiTV" IST-2001-37330). Further information on the entire MECiTV R&D project can be obtained from the public project website: http://www.meci.tv

REFERENCES

1. R. Wright and A. Williams, *Archive Preservation and Exploitation Requirements*, PRESTO-W2-BBC-001218, Jan. 2001. Available: http://presto.joanneum.at/Public/D2.pdf.
2. B. Retsch, A. Mulrenin, *Digitales Zeitalter bedroht kulturelles Gedächtnis*, Noeo Wissenschaftsmagazin, Aug. 2002, pp. 34–37.
3. *AMICITIA - Asset Management Integration of Cultural heritage In The Interexchange between Archives*, IST-1999-20215, R&D project funded by the European Commission, <URL: http://www.amicitia-project.net/>
4. *BIRTH*, R&D project funded by the MEDIA Plus programme of the European Council for pilot projects, <URL: http://www.birth-of-tv.org/>.
5. *PRESTO - Preservation Technology for European Broadcast Archives*, IST-1999-20013, R&D project funded by the European Commission, <URL: http://presto.joanneum.ac.at>
6. A. Girgensohn, J. Boreczky, L. Wilcox, *Keyframe-Based User Interfaces for Digital Video*, IEEE Computer Journal, Vol. 34, No. 9; September 2001, pp. 61-67.
7. J.M. Corridoni, A. Del Bimbo, D. Lucarella, *Navigation and visualization of movies content*, Proceedings of the 11th International IEEE Symposium on Visual Languages, 1995.
8. *VICAR – Video Indexing, Classification, Annotation and Retrieval*, EC R&D project; <URL: http://iis.joanneum.ac.at/vicar/>
9. W. Haas, H. Müller, P. Uray, *Visual Movie Annotation and Analysis in the VICAR Project*, in: Jean-Yves Roger, Brian Stanford-Smith, Paul T. Kidd (Eds.), Business and Work in the Information Society: New Technologies and Applications, pp. 382 - 387, IOS Press, 1999.

10. Herwig Rehatschek, Gert Kienast, Alfred Wertner: "VIZARD-EXPLORER: A tool for visualization, structuring and management of multimedia data", in Proceedings of the Second IASTED International Conference "Visualization, imaging, and image processing", Editor: J.J. Villanueva, pp. 167 - 172, Sep. 2002, Malaga.
11. *pViReSMo (Personalize retrieval on audiovisual data)*, Zentrum für graphische Datenverarbeitung e.V., 2002. <URL: http://www.rostock.zgdv.de/ZGDV/Abteilungen/zr4/Projekte/ViResMo/index_html_en>.
12. *Ricoh MovieTool*, Homepage, 2003. <URL: http://www.ricoh.co.jp/src/multimedia/MovieTool/ >.
13. M. Lux, W. Klieber, J. Becker, K. Tochtermann, H. Mayer, H. Neuschmied, W. Haas, *XML and MPEG-7 for Interactive Annotation and Retrieval Using Semantic Metadata,* Journal of Computer Science (www.jucs.org), Bd 8., Nr. 10, Springer-Verlag, 2002.
14. *Dublin City University* (centre for digital image processing), Fischlar-TV, <URL: http://www.cdvp.dcu.ie>.
15. H. Wactlar, T. Kanade, C. Faloutsos, A. Hauptmann, M. Christel: "Informedia II Digital Video Library: Auto Summarization and Visualization Across Multiple Video Documents and Libraries ", Carnegie Mellon University, 2001. <URL: http://www.informedia.cs.cmu.edu/dli2/index.html >
16. The Object Management Group (OMG): CORBA, *Homepage.* <URL: http://www.omg.org/>
17. ANSI SQL Part 14 - XML related specifications (SQL/XML), ANSI TC NCITS H2 ISO/IEC JTC 1/SC 32/WG 3, working draft, August 2002.
18. P. Enser and C. Sandom, “Towards a Comprehensive Survey of the Semantic Gap in Visual Image Retrieval”, *Proc. CIVR 2003*, Urbana-Champaign, IL, July 2003.