[RFSt00] *Ivica Rimac, Stephan Fischer, Ralf Steinmetz;* **The ICOR Framework: A Top-down Approach to Media Indexing and Retrieval;** IEEE International Conference on Multimedia and Expo (ICME), July 2000, pages 149-152.

The ICOR Framework: A Top-down Approach to Media Indexing and Retrieval

Ivica Rimac[1], Stephan Fischer[1] and Ralf Steinmetz[1,2]

[1]Industrial Process and System Communications (KOM)
Dept. of Electrical Engineering and Information Technology
Darmstadt University of Technology
Merckstr. 25 • D-64283 Darmstadt • Germany

[2]GMD IPSI
German National Research Center
for Information Technology
Dolivostr. 15 • D-64293 Darmstadt • Germany

E-mail: {Ivica.Rimac, Stephan.Fischer, Ralf.Steinmetz}@kom.tu-darmstadt.de

Abstract

In this paper we present the ICOR framework which enables the user to easily integrate content description schemes and content analysis algorithms. We show how the generic framework can be adapted to provide appropriate metadata for applications from different areas, such as music clips or sports games.

Keywords: audio/video analysis, audio/video indexing, audio/video retrieval, content description, metadata.

1. Introduction

The rapid increase of digital audio and video data available nowadays has brought about the need for tools and systems which enable the user to search for particular content in an efficient way — especially when he has to pay for a service. As a consequence computer supported media processing as well as indexing and retrieval of audio-visual data became major research issues in the last decade. Early works dealt with the automatic recognition of video cuts [18], and with the automatic recognition of film genres [4] [17]. Since that a great variety of publications have been dedicated to different approaches to the automatic analysis of audio and video. These can be categorized as follows:

- Many disciplines — e.g. signal and image processing, computer vision, artificial intelligence — are involved in fundamental content-based research, which addresses problems ranging from low-end algorithms (e.g. image similarity) to complex high-end applications (e.g. face detection and recognition in movies). As a support, tools like Vista [13], Dali [11], and MoCA [8] have been described.
- Starting with QBIC [5], general indexing and retrieval systems such as Informedia [15], Virage [16], and VideoQ [2] which provide methods for retrieving images and videos by using visual examples and/or sketches for querying, and matching statical and dynamical visual cues, such as color, texture, shape, motion, and spatio-temporal composition attained the attention of the public.
- Specialized indexing and retrieval systems focus on the implementation and solution of problems which are narrow in a sense that the solution cannot be adapted to other applications. An example is the image retrieval for petroleum applications described in [7].
- Multimedia content description has become a critical issue in a sense that there are several initiatives and frameworks such as MPEG-7 [10] or RDF [14], to name a few, who address standardized metadata for multimedia content [6] [12].
- Surveys of important advancements and open issues in the area are provided for example by [1] and [3].

Many of the systems which have been described follow a bottom-up approach, as the algorithms which have to be used to implement the specific parts of the system are developed and the system is then assembled from those. However, it can be assumed that only those systems will experience a wide acceptance which provide a straightforward look and feel and which hide the details of complex algorithms for content analysis from the end-user. It is the goal of the ICOR (Intelligent COntent and Retrieval) project of Darmstadt University of Technology to examine ways to bridge the gap between the needs of the end-user and the core of algorithms for content analysis of digital audio and video.

In this paper we present the ICOR framework which is a generic and extensible framework applicable to a wide range of application areas of content retrieval. The main goal of the framework is the integration and mapping of data models to algorithms of the core, which can be state-of- the-art or new developments.

The paper is structured as follows. In Section 2 we discuss the system architecture of ICOR while section 3 describes its functionality. Section 4 illustrates the use of our framework by a case study and Section 5 concludes with a summary and future research directions.

2. System Architecture of ICOR

To be able to map the needs of the end-user to the core of algorithms for content analysis of digital audio and video we propose an architecture consisting of 4 layers:

- the user interface level
- the meta level, where the metadata models are stored,
- the system level, which comprises the analysis modules and the control unit,
- the data level, which includes the raw audio-visual data (input) and the description data (output).

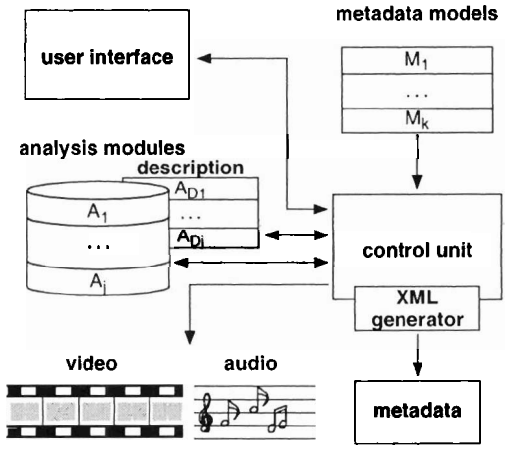

Figure 1. System architecture of ICOR.

Using a subset of these layers it is possible to derive a set of modular algorithms necessary to implement a specific application. The functionality of the meta and the system level are described in the next paragraphs.

2.1. Metadata Level

The idea behind the ICOR-approach is based on the assumption that a content processing application can be described by unique features and characteristics to a great extent, as an application belongs in most cases to one or only a few domains. Each domain can be represented by one or more metadata models. The characteristics of a metadata model can be categorized into bibliographical, structural, and content descriptive metadata. When searching for particular content usually only a subset of these characteristics will be needed which have to be applied in a combined way. An example could be the query "Search that part of the music clip of a particular singer in which the chorus is performed".

To be able to integrate metadata schemes like the Resource Description Framework (RDF) and MPEG-7 once they will become standards, we use the syntax of XML for the description of content. An example, which has been described in more detail in [9], is provided in section 4.

2.2. System Level

The indexing model can be used to derive analysis modules automatically without the involvement of the end-user. Hence both modules represent the "knowledge" of a media indexing system. ICOR uses modules to encapsulate analysis algorithms, as well as description schemes. As both the needs of the user and research results in core content processing can change rapidly, it is essential to be able to insert, modify, or replace these modules when necessary. The only constraint for the integration of new metadata modules is that they apply the syntax of XML.

Algorithms for core content analysis may process media, extract content, transform media, and also integrate neural networks. In most cases combinations of these algorithms yield better results, requiring means to combine algorithms via logical rules. As a consequence the data types to be exchanged between the modular algorithms should be as universal as possible since a restriction to a specific type would create incompatibilities between modules [8]. In ICOR we hide the functionality of the modular algorithms in JavaBeans and increase the computational speed by using native implementations connected via JNI. If an algorithm has to be exchanged we remap the corresponding interface to the new algorithm if necessary. To provide support for the tasks described above, each analysis module is enriched by meta-information according to both a machine processible and a human readable XML-description, as presented in Figure 2.

```
<AModule ID="..." name="...">
   <input_type>...</input_type>
   <output_type>...</output_type>
   <usage>
      <metamodel ID="...">[Name of Module]
      <status>1|2|3</status>
      <input_from ID="...">[Name]</input_from>
         ...
      <output_to ID="...">[Name]</output_to>
         ...
      </metamodel>
      <metamodel ID="..." ...>
      ...
   </usage>
</AModule>
```

Figure 2. Description scheme for analysis modules.

These descriptions include the input and output data type of a module, the usage in specific metadata modules, the status of the output — description object, input for other modules, or both — in each specific module, and the other analysis modules to which the modular algorithm is connected.

To guarantee a frictionless processing the control unit performs the interaction with the user and communicates with other system components. In cases when two or more modules need the same input, it is obvious that calling the

same preprocessing module for each processing chain leads to computational overhead. But since each analysis module is provided with a description, the control unit utilizes this information to calculate the most effective sequential diagram and to check for incompatibilities. Furthermore it is the task of the control unit to guarantee the existence of the output of a module in the memory as long as it is necessary, but to erase it as soon as possible to keep the memory usage low.

3. Functionality

We support two usage modes in the ICOR system: the expert mode, in which the system is maintained, and the user mode, which provides the functionality to index audio-visual data based on a selected metadata model.

3.1. Maintaining the System — the Expert Mode

The administrator of the system integrates metadata and analysis modules into the framework. He starts with the creation of a metadata module and the determination of the necessary processing chains for that particular model. The analysis modules can either be obtained from the database of existing modules, utilizing module descriptions for support, or new modules have to be inserted. In the latter case a description is set up for each new module and the entries for I/O types are provided by the administrator. Since the information about the actual metadata module is known and the connected analysis modules are defined by the composed processing chain, the corresponding entries in a module's description (Figure 2) are filled automatically.

The control unit utilizes the interface specifications to perform incompatibility checking to avoid abnormal termination in the indexing mode. The control unit adds all modules to a list according to the composed processing chains, and subsequently sorts the list to avoid multiple calls to the same algorithm. As a result one instance of the sequential control protocol is calculated automatically for each metadata model, stored and linked to the metadata module. In the indexing mode it is invoked when the specific model is selected and the indexing process is controlled by the control unit.

When the administrator starts a modification task of a metadata module, the processing chains are derived according to the instance of the sequential control protocol. Subsequently the administrator can delete a metadata entry, after which the corresponding processing chain is removed and the sequential control protocol is recalculated. If the administrator chooses to exchange an element of a processing chain, the interface specifications are used to check for incompatibility. Since the replaced module is not used in the specific metadata model anymore, the corresponding entries are removed from its description and added to the new module's meta-information. The recalculation of the sequential control finalizes the task. If the analysis module is obsolete and no other meta model entry exists in its description, the administrator may choose to remove it from the database.

In summary the ICOR framework is flexible in a sense that the number of analysis modules which can be integrated into ICOR is theoretically unbound. The integration of new modules requires a minimum of modification overhead

3.2. Indexing of Audio-Visual Data — the User Mode

In the indexing mode the user specifies the raw data file and selects an appropriate metadata model from the model database. According to the linked sequential control protocol the control unit calls the necessary analysis modules and guarantees that all temporal data is kept in memory until all other processes, which need the data as input, are terminated. Finally the output of the top modules of each processing chain is transformed into a content description according to the metadata scheme.

4. ICOR Case Study

In this section we present an example of how ICOR works. First we explain how our audio description scheme and the necessary set of algorithms are implemented by the administrator. We then show how the user plays with the system in order to index a music piece.

In the first step a new model for the indexing of music clips is created by an administrator of ICOR. The model can then be used by common users to index music clips. An example for an audio description is described in Figure 3, details can be found in [9].

```
<audio type="song">
   <title>...</title>
   ...
   <type>live|studio</type>
   <genre>classic|rock</genre>
   <applause_set>
      <applause id="1" start="..." end="..."></a
      ...
   </applause_set>
   <description>
      <bpm extracted_by="matrix|vector|raw">...<
      ...
   </description>
   ...
</audio>
```

Figure 3. Audio description scheme.

In the second step the necessary algorithms have to be derived, for example the recognition of the genre and type of a clip, an applause detection and the extraction of rhythm. In most cases the title of a music clip cannot be recognized automatically. We hence initiate a user interaction request-

ing the necessary input. In [9] we showed that the occurrence of applause usually indicates that the music clip is a live production. The information about beats per minute and rhythm strongly correlates with the genre of the music clip. We hence need two modules to extract the beats per minute and rhythm and to recognize applause, and two additional modules to postprocess the results to be able to identify the genre and type of the music clip. In [9] we explained how applause can be recognized using the frequency distribution of the clip. If the output of the module which calculates the frequency distribution is plugged into the input of the applause detection the chain necessary to recognize occurrences of applause is complete.

In the third step the administrator of the ICOR system creates the descriptions for the modules (if not already present) and finalizes the application with the creation of an instance of the sequential control protocol.

In the fourth step (the indexing mode) the user selects the metadata model for music clips and specifies a file containing the raw audio data. The control unit handles the protocol linked to the model and starts the analysis. After the completion of the analysis a description in XML is created and presented to the user. If necessary the user can then modify the entries.

5. Conclusion and Outlook

In this paper we presented a flexible and adaptable framework for the indexing of audio-visual data. We explained how a top-down approach can be implemented, using the integration of a metadata model from which a set of modular algorithms is derived which in turn create the metadata. Each metadata model is linked to a sequential control protocol which controls the analysis tasks in the indexing mode.

We have implemented the ICOR framework and gained first experience with our approach. We examined, how the indexing of music clips from different genres can be performed in ICOR. The results are very encouraging and show that our modular approach using metadata schemes serves the intended purpose. The application we tested so far is the indexing of music clips from different genres, such as classical music, soul, jazz, and rock music.

Our future work will address the integration of a greater number of metadata models for different domains as well as the creation of a library of content processing routines.

References

[1] R. M. Bolle, B.-L. Yeo, and M. M. Yeung. *Video Query: Beyond the Keywords*. IBM Research Technical Report RC-20586, IBM CyberJournal, October 1999.

[2] S.-F. Chang, W. Chen, J. Meng, H. Sundaram, and D. Zhong. *VideoQ: An Automated Content Based Video Search System Using Visual Cues*. Proc. ACM Multimedia 1997, Seattle, 1997.

[3] S.-F. Chang, Q. Huang, T. Huang, A. Ruri, and B. Shahraray. *Multimedia Search and Retrieval*. Advances in Multimedia: Systems, Standards, and Networks. A. Puri and T. Chen (eds.), Marcel Dekker, New York, 1999.

[4] S. Fischer, R. Lienhart, and W. Effelsberg. *Automatic Genre Recognition*. Proc. ACM Multimedia 1995, San Francisco, 1995.

[5] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and B. Dom. *Query by Image and Video Content: The QBIC System*. IEEE Computer, 28(9), 1995.

[6] M. J. Hu and Y. Jian. *Multimedia Description Framework (MDF) for Content Description of Audio/Video Documents*. Proc. 4th ACM Conf. on Digital Libraries (DL'99), Berkeley, August 1999.

[7] C.-S. Li, R. Smith, V. Castlli, and L. D. Begmann. *Comparing Texture Feature Sets for Retrieving Core Images in Petroleum Applications*. Proc. SPIE Storage and Retrieval for Image and Video Databases VII, San Jose, CA, January 1999.

[8] R. Lienhart, S. Pfeiffer, and W. Effelsberg. *The MoCA Workbench: Support for Creativity in Movie Content Analysis*. Proc. IEEE Conference on Multimedia Computing and Systems, Hiroshima, June 1996.

[9] L. Müller, I. Rimac, S. Fischer, and R. Steinmetz. *Music Retrieval in ICOR*. Proc. SPIE Multimedia Storage and Archiving Systems IV, Boston, September 1999.

[10] F. Nack and A. T. Lindsay. *Everything You Wanted to Know About MPEG-7: Part 1*. IEEE Multimedia, 6(3), 1999.

[11] W.-T. Ooi, B. Smith, S. Mukhopadhyay, H. H. Chan, S. Weiss, and M. Chiu. *Dali: A Multimedia Software Library*. SPIE Multimedia Computing and Networking, San Jose, CA, January 1999.

[12] S. Paek, A. B. Benitez, and S.-F. Chang. *Self-Describing Schemes for Interoperable MPEG-7 Multimedia Content Descriptions*. IEEE/SPIE Visual Communications and Image Processing, San Jose, January 1999.

[13] A. R. Pope and D. G. Lowe. Vista: ASoftware Environment for Computer Vision Research. Proc. CVPR'94, 1994.

[14] Ressource Description Framework. URL: www.w3c.org/RDF

[15] M. A. Smith and M. Christel. Automating the Creation of a Digital Video Library. Proc. ACM Multimedia 1995, San Francisco, 1995.

[16] Virage. URL: www.virage.com

[17] H. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan. Automatic Parsing of News Video. Proc. IEEE Int. Conf. on Multimedia Computing and Systems, 1994.

[18] H. Zhang, A. Kankanhalli, and S. W. Smoliar. *Automatic Partitioning of Full-motion Video*. Multimedia Systems, 1(1), 1993.