Multimedia information description and search : technology and perspective

Jinwoong Kim, Jae-Gon Kim, Hankyu Lee, and Jae-Woo Yang
Electronics and Telecommunications Research Institute
161 Gajung-dong Yousung-ku, Daejon 305-350, Korea
E-mail: jwkim@video.etri.re.kr

Abstract

As digital audio video data compression and transmission techniques are matured, huge amount of digital multimedia material is produced and delivered via broadcasting, digital storage media and world-wide web(WWW). Thus it became very important to provide a standardized way of multimedia data content description, so that efficient and effective access and reuse of valuable multimedia information can be possible. In this paper, enabling core technologies and our research directions on this are presented with brief introduction on the scope of the multimedia content description interface, called MPEG-7, in terms of objective, application and requirements.

1. Introduction

As the digital audio video data compression and transmission technologies become matured, the amount of digital multimedia data available grows so fast that it became an urgent topic of research and development how to represent those multimedia data in a content-based way. MPEG has already started to work towards that direction under the name of "MPEG-7 : Multimedia Content Description Interface". Development of this technology would make it much easier to access and reuse valuable information on the WWW, from increasing broadcasting channels and future multimedia digital library. While current technology supports generation, data compression, transmission and storage of a large number of digital images, video and audio, existing practice of content-based indexing, searching and retrieval of visual data is in its infancy.

In this paper, the requirements and applications of the new content-based multimedia description technology are briefly presented from the MPEG-7 standardization perspective. Then core techniques related to the standardization activity are described with some details on video data description modeling, feature extraction on the compressed domain and cross-modal processing. Then the scope and strategy of ETRI's research on the MPEG-7-related technology are presented followed by conclusions.

2. MPEG-7

2.1 MPEG Standardization

MPEG has started a new standardization activity about content-based multimedia information description and coding, which is called MPEG-7, which will follow the current MPEG-4 standardization about object-based multimedia coding. Contrary to the past, new trend of international standardization efforts, including MPEG, is not to require giving up intellectual properties on the techniques adopted to the standard. This promotes many companies and institutes having relevant technology to actively participate in the procedure, resulting a good new technology development as well as expedited standardization process.

Especially, MPEG-2 standard has been adopted for digital TV broadcasting, digital versatile disk, and digital camcoder, which would generate a huge amount of royalty for the use of basic MPEG-2 coding algorithm. Thus it becomes more and more important to lead and contribute each party's proprietary algorithms or techniques to new standards which has large potential market. We, in Korea, have contributed little to the core technology of MPEG-2 standard, but has built up its implementation technology such as video encoding/decoding chip set, HDTV CODEC and DVD player. For MPEG-4, we have been adopted to the Committee Draft of video and system part of the standard with automatic segmentation, computational graceful degradation, and scalable bitstream architecture and synchronization techniques in Text-to-Speech(TTS). Based on this experience, we are expecting to contribute a lot more core ideas and algorithms on video and cross-modal processing and feature description part of MPEG-7.

2.2. MPEG-7 technology

- Scope

As more and more audiovisual information is available in digital form, it becomes more difficult but necessary to find potentially interesting material to users. Currently, many text-based search engines help people find documents and textual information on the web. However, it is not possible yet to search for audiovisual information based on its contents or characteristics. MPEG-7 has been started to provide a solution to this problem. MPEG-7, which is a standard representation of audio-visual information satisfying particular requirements like the other members of MPEG family. Figure 1 shows a highly abstract block diagram of a possible MPEG-7 processing chain and the scope of the standard[1]. Though efficient data base structure and the search engine is crucial to retrieval applications, it will not be in the standardization scope since it does not affect to the interoperability. Automatic or semi-automatic feature extraction methods are certain to have big effect to the fast spread of MPEG-7 standard, but it will not be included in the standard either for the same reason. MPEG-7 focuses only on the description of features and contents of multimedia material. Entities of the standard will include multimedia data itself, regardless of its storage, coding, display, transmission, or medium characteristics, description which consists of descriptors and description scheme, and coded description which is a compressed description tailored for easy indexing, efficient storage and transmission.

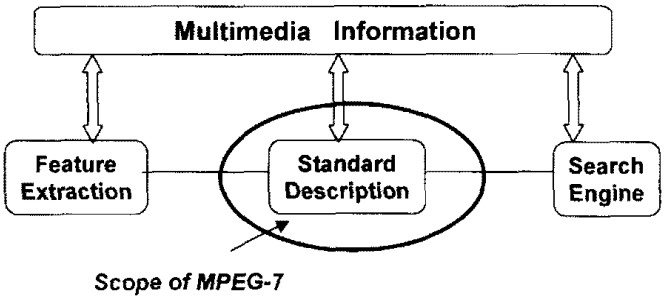

Figure 1. MPEG-7 processing chain

- Applications

MPEG-7 has lots of application areas[3]. The primary application areas will be visual and auditory retrieval applications such as television program and film archives, tele-shopping, bio-medical atlases, Karaoke and music sales, sound effect libraries, and so on. Different applications pose different requirements on the features and functions of multimedia content

descriptions. For example, while video archives need language independent full-text descriptions, medical applications benefits from image-driven queries and cross-modal search. User interface for Karaoke would be more friendlier if we can select a music by query-by-humming. In addition to basic search and retrieval applications, there are also other important application areas: user agent driven media selection and filtering, intelligent multimedia presentation, educational applications, information access facilities for people with special needs, and surveillance systems. Especially, provision for multi-modality for multimedia information presentation can greatly improve the information accessibility of a disabled person in one form or another, thus reducing the communications gap between normal and disabled people.

- Requirements

In order that all these application areas be benefited from the standard, different aspect of diverse requirements should be taken care of[2]. In general, they should include support of various description classes like free text, N-dimensional spatio-temporal structure, statistical information, objective and subjective attributes, and so on. Support for cross-modal search, content-based retrieval, similarity-based retrieval, feature hierarchy and scalability, distributed multimedia databases and robustness to information errors and loss should also be considered. In order to efficiently describe multimedia material, a reference description model for each media should be developed. Figure 2 shows a video description model we proposed at a MPEG meeting[5]. It has hierarchical information abstraction levels: video, shot, key frame, object, feature and representation, to efficiently support users' need at different semantic levels. It also supports associated relations between components of a representation scheme.

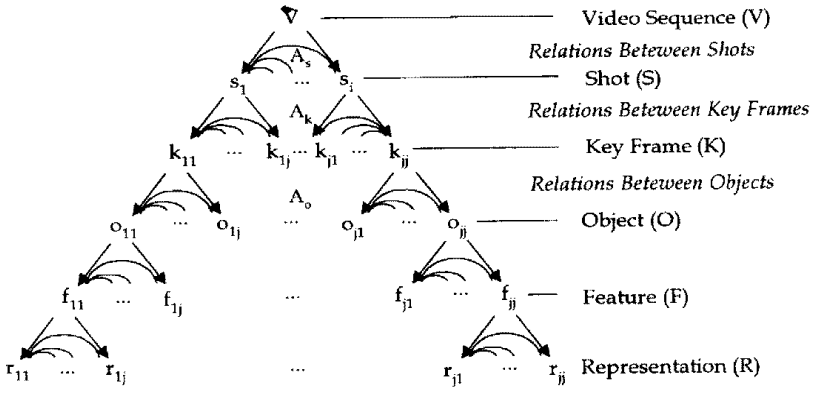

Figure 2. A video description model

In addition to the general requirements, there are modality-dependent requirements on the visual and audio material as well as application-dependent requirements. For visual information, it should support query-related description classes, scalable visualization scheme, various data formats and data classes. Description classes range from the basic features like color, texture, shape to more conceptual features like sketch, object composition. Visualization has means for presenting from sketch to full resolution video. Likewise, it should support for audio information description classes, means for scalable sonification, various auditory data types and classes. Audio description classes include basic features like frequency contour, timbre, harmony and more abstract features like sonic approximations or prototypical sound. Auditory data classes include soundtrack, music, speech and synthetic audio, et. cetra. Extracted features and associated information shall be coded into metadata, which may or may not be wrapped together with content. SMPTE and EBU are also working together for

standardization of metadata and wrappers along with compression issues, transfer protocols and physical medium in order to achieve interoperability on bitstreamed program material exchange[7].

3. Core technologies for multimedia information processing

From the signal processing and communications side, data compression technology has been successfully developed and are widely used now. From the computer side, multimedia data storage and search technology has been extensively developing, but their proprietary schemes are not interoperable. As more and more multimedia data are available in compressed digital form, it becomes necessary to have a bridge between these two ends. In other words, standardized content description technology should be applied to both compressed and uncompressed form of digital multimedia information for efficient search and reuse. MPEG-7 is a natural outcome of the demand for this information description technology. Among the enabling technology for this are automatic segmentation, analysis and extraction of features, methods for description of the features, tagging and indexing of source material, and search engines. In Figure 3, a technology tree for multimedia information coding technology is shown.

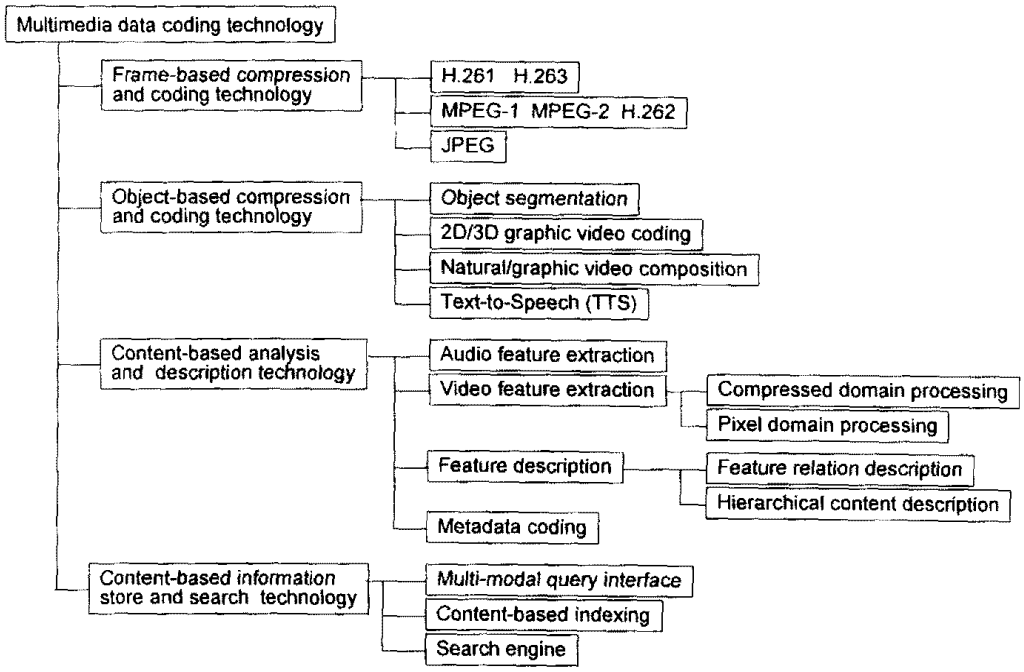

Figure 3. Technology Tree for multimedia data coding

One of prerequisite to video indexing is to have a video program segmented into scenes according to its semantic meaning. This requires various techniques for scene change detection for scene-cut by edit, gradual change by special effects like dissolve, fade-in, fade-out and wipe[10]. Feature analysis and editing in compressed domain save lots of processing power and thus is a key factor for real-time operation[8]. Compressed video data according to MPEG-2 or MPEG-4 standard have already many information about features of the source contents, like luminance level, motion vectors, macroblock types, shape of an object, and many others. So, definition and description of features which can be extracted or analyzed in compressed domain is one of important research areas. Furthermore, it is expected that typical current encoding algorithms may be re-evaluated and modified in a way that some compression ratio are traded for more content representing ability of the coded bitstream. Cross-modal search is a potentially capable technique for efficient multimedia data handling. Speech recognition and natural language processing techniques are used for informedia news-on-demand system[11]. Among some prototype

systems integrating many of these techniques are, WebSEEk and VisualSEEk systems in Columbia university, QBIC of IBM, Informedia digital library of Carnegie-Mellon University.

4. Research Directions and Strategy

Our MPEG-7 technology development project will have mainly threefold targets: (1) multimedia data feature extraction technology, (2) Content-based feature description technology, (3) Metadata coding technology, which enables hierarchical content representation. While the focus will be given on the core technology development and having results of them adopted in the standard, a test-bed system will be developed at the same time to experiment and demonstrate the effectiveness of proposed schemes and algorithms. The test-bed system will have following features: medium will be broadcasting news and web information, applications targets will be real-time as well as non real-time, compressed domain data processing, multi-modal query processing and search. Block diagram of the test-bed system is shown in Figure 4.

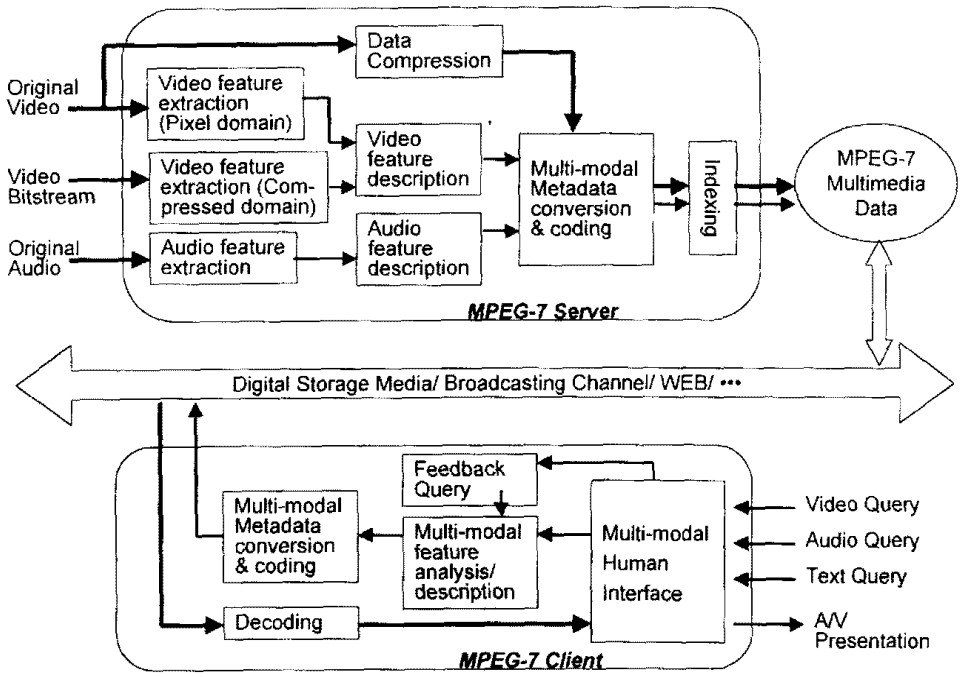

Figure 4. MPEG-7 Test-bed system

The project will consist of two sub-tasks. One is study of media feature extraction and description technology. In this sub-task, requirements on the query and feature description will be defined and feature extraction algorithms for video, audio and speech in both source domain and compressed domain will be studied. Data representation modeling and hierarchical feature description scheme optimized for each level of content abstraction will also be emphasized. The other sub-task will include study of multi-modal metadata coding and human interface technology, and test-bed system development. Efficient query processing techniques will be developed for a textual input from simple key-words to natural language, visual input like sketch or object motion trajectory, as well as aural input. Especially, cross-modal metadata description and search technology and metadata benchmarking method will be intensively studied. The research and development in the multimedia data description and search field requires not only signal processing and media coding experts but also experts on natural

language processing, database architecture, and multi-modal human-computer interface. We are on the discussion now for international cooperation on this project to share each party's accumulated technology in diverse field of relevance.

5. Conclusions

In this paper, we presented background, applications and requirements of an emerging activity on multimedia content description technology development, which is called MPEG-7 standardization, along with our research directions and strategy in terms of relevant core technology and test-bed system development. While we already have lots of useful results on media feature extraction, indexing and search technology, it is expected that fields of compressed-domain processing, cross-modal information handling, feature processing for different level of semantic abstraction, and metadata coding need further research work to provide a viable framework for content-based multimedia information access.

References

[1] Rob Koenen, "MPEG-7: Context and Objective(v.5-Fribourg)", ISO/IEC JTC1/SC29/WG11 N1920, Fribourg, Oct. 1997

[2] ISO/IEC JTC1/SC29/WG11 N1921, "Third Draft of MPEG-7 Requirements", Fribourg, Oct. 1997

[3] ISO/IEC JTC1/SC29/WG11 N1922, "Second Draft of MPEG-7 Appications Documents", Fribourg, Oct. 1997

[4] ISO/IEC JTC1/SC29/WG11 N1923, "MPEG-7 Proposal Package Description(PPD) - Draft V1.0", Fribourg, Oct. 1997

[5] Taehwan Shin, Jae-Gon Kim, Hankyu Lee, and Jinwoong Kim, "Requirements for Video Description Model", ISO/IEC JTC1/SC29/WG11 M2786, Fribourg, Oct. 1997

[6] Yong Rui, Thomas S. Huang, and S. Mehrotra, "MARS and Its Applications to MPEG-7", ISO/IEC JTC1/SC29/WG11 MPEG97/2290, July 1997

[7] EBU and SMPTE, "Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams, First Report: User Requirements", Apr. 1997

[8] J. Meng and S.-F. Chang, "Tools for Compressed-Domain Video Indexing and Editing", *Proceedings of SPIE Conf. on Storage and Retrieval for Image and Video Database*, Feb. 1996

[9] S. W. Smoliar and H. Zhang, "Content-Based Video Indexing and Retrieval", *IEEE Multimedia*, pp. 62-71, Summer 1994

[10] F. Idris and S. Panchanathan, "Review of Image and Video Indexing Techniques", *Journal of Visual Communications and Image Representation*, Vol. 8, No. 2, pp. 146-166, June 1997

[11] A. Hauptmann and M. Smith, "Text, Speech, and Vision for video segmentation: The Informedia Project", *Proceedings of Symposium on Computational Models for Integrating Language and Vision*, Fall 1995