Multimodal Approach for Summarizing and Indexing News Video

Jae-Gon Kim, Hyun Sung Chang, Young-tae Kim, Kyeongok Kang, Munchurl Kim, Jinwoong Kim, and Hyung-Myung Kim

A video summary abstracts the gist from an entire video and also enables efficient access to the desired content. In this paper, we propose a novel method for summarizing news video based on multimodal analysis of the content. The proposed method exploits the closed caption data to locate semantically meaningful highlights in a news video and speech signals in an audio stream to align the closed caption data with the video in a time-line. Then, the detected highlights are described using MPEG-7 Summarization Description Scheme, which allows efficient browsing of the content through such functionalities as multi-level abstracts and navigation guidance. Multimodal search and retrieval are also within the proposed framework. By indexing synchronized closed caption data, the video clips are searchable by inputting a text query. Intensive experiments with prototypical systems are presented to demonstrate the validity and reliability of the proposed method in real applications.

Manuscript received June 27, 2001; revised Oct. 24, 2001.

Jae-Gon Kim (phone: +82 42 860 4980, e-mail: jgkim@etri.re.kr), Hyun Sung Chang (e-mail: chs@etri.re.kr), Young-tae Kim (e-mail: ytkim@etri.re.kr), Kyeongok Kang (e-mail: kokang@etri.re.kr), Jinwoong Kim (e-mail: jwkim@etri.re.kr) are with Broadcasting Media Technology Department, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea.

Munchurl Kim (e-mail: mkim@icu.ac.kr) is with the School of Engineering, Information and Communications University (ICU), Daejeon, Korea.

Hyung-Myung Kim (e-mail: hmkim@csplab.kaist.ac.kr) is with the Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea.

I. INTRODUCTION

With the increasing rate of growth of digital video data, more and more interest is being focused on efficient access to desired contents. For the past few years, there have been a lot of R&D activities on related issues such as video analysis, representation, and browsing on the basis of the content [1]-[7]. Video summaries, in which the entirety of a video is abstracted by a gist, result in very compact representation of the video without losing essential contents. A video summary enables efficient browsing as well as a fast overview of the original contents by reducing the costs spent on such tedious operations as fast-forward and rewind. In this aspect, video summarization is popularly regarded as a good approach to the content-based representation of videos.

In general, the existing methods for summarizing video can be classified into one of the following classes according to their styles: *static summary* [1]-[3] and *dynamic summary* [4]-[6]. In a static summary (e.g., simple or more complicated presentation organized with key frames, shot mosaic), a small number of images are arranged to represent the video. On the other hand, a dynamic summary is a video synopsis of greatly reduced duration, which means a dynamic summary is made up of key segments or clips rather than key frames. Most approaches for video summarization that have appeared in the literature are based on a static style. However, a dynamic style summary is thought to be more comprehensive to users in the sense that it still keeps the audio and motion dynamics.

A critical aspect of summarizing a video in a dynamic style is semantic analysis, which is the key to choosing highlight segments to be contained in the summarized video. Generally, implementing a system of tailoring a dynamic summary in a

purely automated way is regarded as a very difficult task because there is a large gap between low-level features and high-level semantics even with the current state-of-the-art technology. Therefore, most of the existing studies resort to their own strategies applicable to some specific domains of video data: news, documentaries, sports, and so on.

In this paper, we propose a novel method for automatic summarization of TV broadcast news videos, and show its applicability in practice. When dealing with news video, first of all, a summary should be able to concisely deliver the reporting of each news item to provide fast overview of the given news video. To do this, it is highly desirable to abstract the news program in a well-structured form by the unit of each news item while preserving semantic information in it. In this sense, the closed caption (CC) provided with a news video can serve as useful media for summarizing the video since it is rather easy to capture semantics automatically from textual information given by CC data. In the proposed approach, we use CC data as a key source for semantic information and speech signal in audio track for synchronization between the CC text and video. Additionally, shot boundaries are detected for visual indexing of the video.

In order to store and deliver the information on the tailored summaries, an effective and interoperable scheme should be used. This paper also presents the description of the summaries based on the MPEG-7 Summarization Description Scheme (DS) [8], [9], which was refined partially by our earlier works on MPEG-7. Although we focus on summarization of news videos and the associated applications of browsing and navigation, we also utilize CC data for the text-based indexing and retrieval within the same framework. More specifically, the system to be presented indexes the CC data for providing multimodal search where relevant video clips are retrieved by queries in textual format.

In Section II, we briefly overview our approach to news video summarization and indexing. Then, we present the details of the system architecture and implementation of the proposed method in Section III. An explanation is also made on how the generated summary is described on the basis of MPEG-7 Summarization DS. In Section IV, we show the feasibility of the proposed system through experiments. Finally, conclusions are presented in Section V.

II. NEWS VIDEO SUMMARIZATION

In dynamic summarization, semantic analysis is essential to extract highlight segments that involve semantically meaningful portions of the given video as well as video structure extraction. While many works have been done in the context of content-based indexing to provide ways of extracting news video structures, mainly by detecting anchor person shots [10], [11], extraction of high-level semantics is still a highly difficult task.

In summarizing a news video, the following two characteristics are noteworthy. A news video is presented in a very structured way to be easily understood by many viewers. Moreover, in a news video, the most important pieces of information, which are semantic in essence, are delivered in voices rather than in images. This means that we need to focus some of our effort on speech understanding for news video summarization. Some papers have proposed to automatically generate the transcript of a video by speech recognition so that some keywords could be located to get into a video skim afterward [4]. However, the transcript generation only with an audio channel is a rather difficult task. Moreover, its reliability is also doubtful for the time being due to the limited performance. In this sense, CC data conveying speech information in textual forms are very utilizable. In nearly all news videos currently being broadcasted in Korea, CC data is also provided with the video itself in an undisclosed way. Some researchers have utilized CC data in news video indexing under the assumption that the available CC data is synchronized with the video [12], [13]. However, the assumption is not true in many cases. Therefore, it is necessary to make temporal alignment between both media.

In our method, CC data is utilized in speech recognition to reduce computational complexity while considerably enhancing the reliability. Speech recognition is adopted just for aligning the words in the CC text with those in the real audio channel. Once we have audio-synchronized CC data, audio-based summarization of news video becomes equivalent to text-based one. Then, in the next step, CC text is analyzed for understanding the context and extracting the news structure. Generally, a news program consists of several news items, called *events*, and each event consists of several *scenes*. Here, the term "a scene" means a semantic unit identified by the change of speakers. Three kinds of scenes–*anchor*, *reporter*, and *interview*–are found in news videos. In a typical case, a news item begins with an anchor scene providing a short description of the event, and continues to more detailed reporting with some comments, followed by reporter and interview scenes. There may be anchor scenes again to give a summary or conclusion. This pattern is common in most news programs. In the proposed approach, we can easily extract the above hierarchical structure by utilizing additional tags contained in the CC data. The structure extraction is the most basic step and the extracted structure itself is the most fundamental semantic attribute in our summarization approach in that it represents the whole sequence by a few types of units that have their own semantic roles.

In the next step, we detect highlights involving the core of the contents through text analysis combined with the semantic attributes of the extracted structure. As mentioned before, each

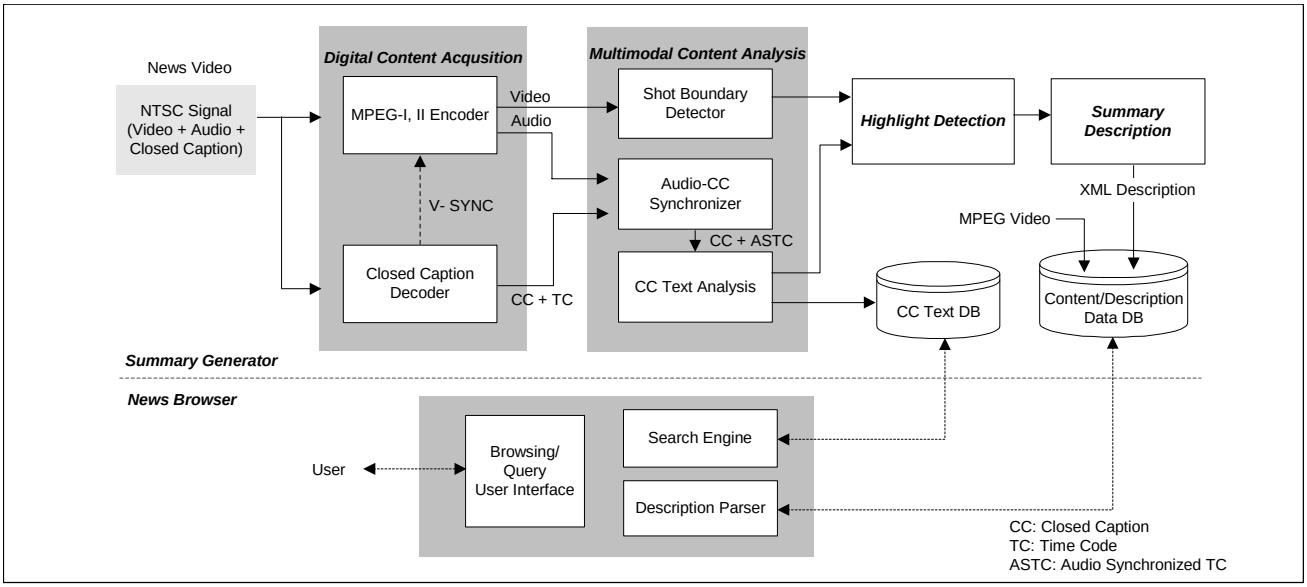

Fig. 1. Overall system architecture of the news video summarization and indexing.

scene has a role that can be interpreted as a semantic attribute in a news item. For example, an anchor scene appears at the beginning to give a short description for the event, and reporter and interview scenes follow the anchor scene to describe the details of the event. In this sense, it is far simpler to compose a summary that only consists of anchor shots in a program [13]. There are two drawbacks to this approach; we need a more compact summary instead of including a full portion of each anchor shot and there is no other visual information except the anchor scene.

To take into account these factors, we generate a summary as follows. At first, we split each event into two parts. One is the *title* part whose scene type features the anchor, and the other is the *article* part that features reporter and interview scenes. Then, through the language analysis work on the closed caption, we extract a few key sentences to be included in a summary from the two parts in each event, respectively. In our approach, we can basically generate two-levels of summary with different time durations. A coarse-level summary consists of segments, to which the key sentences extracted from title parts correspond, while the fine-level summary has additionally added segments, to which the key-sentences extracted from article parts correspond to the coarse-level ones. We can also generate various summaries that have different time durations by selecting the number of key sentences to be detected in the language analysis stage.

III. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The overall architecture of the proposed system for news video indexing and summarization is shown in Fig. 1. The system is comprised of two parts: a news summary generator and the associated browser. Given a news video, the summary generator automatically generates a summary based on a multimodal approach. The generated summary, described with MPEG-7 Summarization DS, is delivered to the video browser. Then, the user can skim and navigate the news video on the browser that consumes the summary description, validating the functionalities offered by the summary generator such as browsing, navigation, and retrieval.

The summary generator consists of four main functional modules: digital content acquisition, multimodal content analysis, highlight detection, and summary description. In the first step, analogue TV news signals are captured, and filed in some standardized digital formats (audio-visual data in MPEG and CC data in ASCII), at the digital content acquisition module. The multimodal content acquisition module extracts some features to be used as clues in the highlight detection step from the acquired digital contents. The news database for text-based retrieval is also populated in this step. Next, the system applies a predefined rule to detect highlights of the video, which is detailed in Section III.3. Finally, the detected highlights are described and stored into the database with the original content.

The news video browser provides efficient mechanisms for accessing news videos in two ways. The one is search and retrieval of desired news clips by a textual query, and the other is browsing and navigation of the news video when the desired content is already searched. The news video browser, which integrates a search engine, a description parser, and others on a common graphical user interface (GUI), consumes a summary description data and accesses CC database as shown in Fig. 1.

1. Digital Content Acquisition

In the digital content acquisition module, MPEG systems' streams and CC data are acquired by an MPEG encoder and a CC decoder from an incoming NTSC TV signal, respectively.

Generally, no time information is attached to the words in the CC data. Furthermore, the time that the CC appears in the video signal does not coincide with that of the corresponding speech in the audio track, because the CC data is manually generated by typing the broadcasted speech on-line and is inserted into the video signal with some temporal delay. Therefore, in order to use CC data as a source media to be processed in the proposed summarization and indexing system, it is necessary to align the words in the CC text and those in the audio track of the video. This kind of time alignment between the CC data and video stream is accomplished by the following two-step approach. First, the CC data is extracted together with a time code indicating the time of its appearance in the video signal in the CC decoding. Then, in the next step of speech recognition, we update the time code attached to the CC data at the value that gives the exact synchronization with the video by compensating for the delay mentioned above.

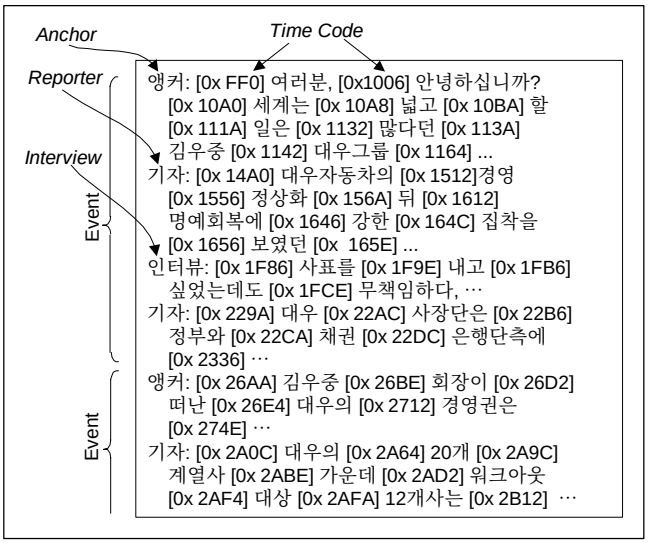

Fig. 2. An example of CC data decoded from a news video by the closed caption decoder.

CC data is carried by two ASCII characters on every other video field. We developed a CC decoder that extracts characters together with the number of the associated field in which the characters are encoded. Whenever a white space is detected in the CC data, a time code whose value is the counting number of fields from the start of decoding to the moment of detection is attached. This way, the time code gives temporal information that indicates the relative time of the words appearing in the CC data. Figure 2 shows an example of an output of the CC decoder that includes uncompressed caption data with the extracted time code.

As a result, we have a link between the CC data and video in a time-line. However, synchronization between both media is not achieved yet. In the subsequent section, we will explain how to do this.

2. Multimodal Content Analysis

A. Audio-CC Synchronization

As mentioned before, the synchronization of the CC data and video is achieved by speech recognition that works on the audio track in a video signal. From an acquired MPEG systems stream, we extract an audio stream by demultiplexing. The MPEG audio is converted to a PCM wave file and downsampled to 16 kHz before the recognition process.

As feature parameters for speech recognition, the 12th order mel-frequency cepstral coefficients and signal energy measured every ten milliseconds were used [14]. A context dependent semi-continuous hidden Markov model (HMM) is used for each phoneme-like unit HMM. The recognition network is constructed by concatenating words that include the current word and about ten words before and after it in temporal order of the text stream. Note that we can easily set the search area for the word in question which needs to be recognized by referring to the time code attached to the associated word in the input CC data. The underlying word is detected from the speech signal using speech recognition in which a probabilistic similarity measure is used in the search area of the audio track. We update all of the time codes attached to each word in the input CC data as those of the audio-synchronized value obtained by the above procedure.

In our approach, we observed through experiments in which real broadcast videos including CC data were used that CC data with a time code considerably enhances the reliability of the speech recognition in the audio track. CC data might contain a few word errors in the broadcast news such as spelling errors, insertions and omissions. In terms of the word detection task, we achieved a 97.34% detection rate in the experiment using real CC data that contained some word errors. One type of error was caused by the spoken word not given in the CC, and others were caused by severe background noise in the reporter's voice.

B. Shot Boundary Detection

A shot is the most basic unit in structuring a video sequence in terms of content-based indexing. In our approach, higher level structuring for summarization is solely based on CC data.

Although shots are not directly used in the summarization, they are included in the summary description to enable efficient navigation, in which the key frames of shots have the role of navigation guidance in combination with the key frames of highlights in a hierarchical manner. This way, we can directly access more relevant shots around a given highlight segment in a finer granular unit of shot. We applied the existing method [15] that is applicable to compressed video to the detection of the shot boundary. It gives fairly good performance in the case of abrupt change but needs manual modification for accurate results in the case of gradual transition. For the key frames, simply the first frame in each shot was used.

C. CC Text Analysis

Figure 3 shows the outline of the procedure for the CC text analysis. Since each piece of CC data contains all the news reports in a daily unit, the first necessary task is to partition it by the unit of an event like the example in Fig. 2.

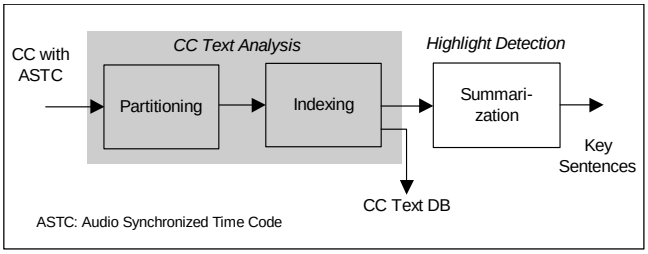

Fig. 3. A procedural flow for CC text analysis.

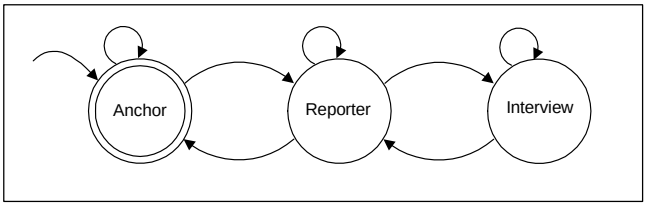

Fig. 4. Finite state automata modeling the transitions between speakers.

The CC data provided via terrestrial broadcast in Korea, as illustrated in Fig. 2, bear additional tags (*anchor*, *reporter* or *interview*) indicating who the current speaker is. Each event boundary closely relates to the transition between speakers. The relationship of the transition between speakers can be modeled by finite state automata (Fig. 4). In the figure, the boundaries between events are determined by an anchor scene. This way, we can easily extract the structural information of the given news program. In other words, we segment the given news sequence in the unit of event and then partition each event into title and article parts utilizing the tags.

At the indexing step, each sentence in the CC data is analyzed lexically so that meaningful words like nouns can be extracted to represent the sentence. The words are called *indexing terms* or more often *keywords*. This means an arbitrary sentence, say S, can be decomposed into

$$S = \{k_1, k_2, \cdots, k_n\}, \tag{1}$$

where k_i denotes the keywords extracted.

Once keywords are extracted, inverse document frequency (IDF), which is defined as a reciprocal to the ratio of the number of documents (equivalent to events in our case) containing a keyword, is additionally extracted for each keyword. The IDF measures the discriminating capability of a specific keyword in characterizing the document it belongs to. Finally, all of the keywords extracted from the given CC text, the associated IDF and the structural information are stored in the CC text database.

In the summarization stage, we select a set of sentences to be included in highlights by measuring the degree of importance combining with the structural information. The details of the summarization are described in the subsequent sections.

3. Highlight Detection

As mentioned before, we extract a few key sentences as semantically meaningful parts, namely highlights, from the structured CC data through the CC text analysis. The highlights, which are video segments temporally aligned to each extracted key-sentence, are concatenated to compose a two-level summary. In order to extract a key-sentence, the weight, or degree of importance, is assessed on the basis of the term frequency (TF) and IDF of the keywords in each sentence. Specifically, in our experiments, the weight of S in (1) is calculated by

$$W^S = \frac{1}{|S|} \sum_{k_i \in S} \left(mtf^S(k_i) \cdot \frac{N}{df(k_i)} \right), \tag{2}$$

where

$mtf^S(k_i)$ - modified term frequency of k_i in S,

N - total number of documents (or events),

$df(k_i)$ - the number of documents in which a keyword k_i appears,

$|S|$ - the number of keywords in S.

In our case, we use a modified term frequency (MTF) instead of TF to take into account the characteristic of news content in which important context is usually placed in the beginning. Therefore, the closer the location of the sentence to the

front of the document, the more weight is assigned to the words in the sentence in calculating the MTF, as illustrated in the example of Table 1. The value of *Sentence ID* numbers the sequential order. In this example, the MTF is determined as the value that is the total TF of the current and subsequent sentences plus one.

As mentioned before, we extract key-sentences as highlights from the title and article parts of each news event to generate a two-level summary in which the coarse level is composed of key sentences extracted from the title part of all events, and the fine level includes all key sentences from both parts. Therefore, once the weight of each sentence is calculated, we select key sentences that have higher weights than others from two parts of each event. According to the user's preference, we can generate various summaries with different durations by adjusting the number of key-sentences to be extracted.

Table 1. An example of calculating MTF for a keyword.

Sentence ID	TF within sentence	MTF
1	1	5
2	0	0
3	2	4
4	1	2
5	0	0

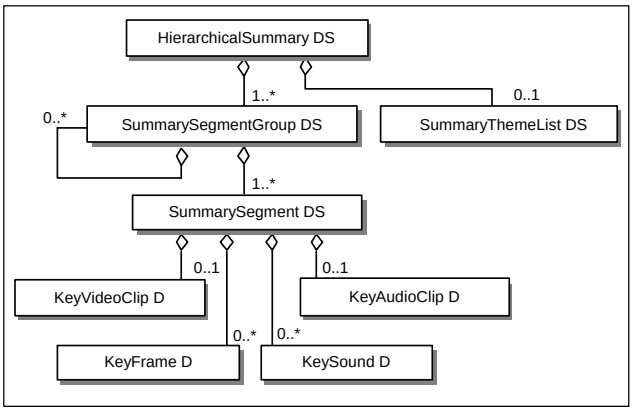

Fig. 5. UML representation of the HierarchicalSummary DS.

4. Summary Description

In order to organize audio-visual information in a structural way, a generic audio-visual description scheme known as the Multimedia Description Scheme (MDS) is devised in MPEG-7. The MDS is comprised of largely six functional groups which are basic elements, content organization, content management, content description, navigation and access, and user interaction. Each functional group contains many description tools: descriptors (Ds) and description schemes (DSs). Summarization DS belongs to the functional group of navigation and access. The Summarization DS provides a set of summaries of audio-visual material. Each of the summaries, which are presented by a Summary DS, is an audio-visual abstract of the entire contents. Since the Summary DS is an abstract DS, in real instances, either HierarchicalSummary DS or SequentialSummary DS is derived from the Summary DS, and used instead of it. In our case, the HierachicalSummary DS is used to describe multi-level hierarchical summaries.

A. HierarchicalSummary DS

Figure 5 represents the HierarchicalSummary DS of MPEG-7 in unified modeling language (UML). It has evolved from the primitive one [6], [16]. At first, the primitive version of the HierarchicalSummary DS faced some severe limitations in its access mechanism and description efficiency for the event-based summaries. In order to overcome such a weakness, we propose a modified description scheme extending the previous one [6], [17]. As a result, the current HierarchicalSummary DS enables browsing and navigation as well as a fast overview in an efficient way through a more flexible description structure. After seeing the overview of the highlight summary video, users can efficiently navigate and/or browse the content based on the overview. In this sense, the HierarchicalSummary DS is thought to provide a unified description framework that combines a static summary based on key frames and key sounds with a dynamic summary based on a series of highlights of the audio-visual segments. Major functionalities of the HierarchicalSummary DS focusing on the news summary description will be further explored subsequently.

B. SummarySegment DS

Viewers sometimes want to first look at a dynamic video summary such as a film trailer in order to get the gist of a longer program. When the summary is not enough to get the gist of a particular portion of the program, they can move to that portion from the dynamic video summary. Meanwhile, the static summary gives direct access to different parts of the original video. However, a common drawback of static video summaries is that they do not preserve the time-evolving dynamic nature of the video content. Furthermore, in general, there are too many key frames to find an appropriate one for browsing. Considering this factor, the HierarchicalSummary DS provides a description scheme that combines the advantage of the static summary's facility of direct access with the dynamic summary's capacity of quick skimming based on a uni-

fied description framework. In other words, to navigate and browse the contents, we utilize key frames of highlight segments composing a dynamic summary to reach and find the subjects of interest. Therefore, the key frames of the audio-visual dynamic summary play the roles of stepping-stones between the summary description and the content of the original program. This is achieved by the SummarySegment DS, where KeyVideoClip D and KeyFrame D are located together as shown in Fig. 5. For each highlight composing a dynamic summary, the information of the video segment and the associated key frames is described in KeyVideoClip D and KeyFrame D, respectively. This unified framework enables coarse-to-fine navigation to traverse from the summary to the more relevant part of the original program. Meanwhile, the HierarchicalSummary DS can be a multi-level summary resulting in key frames in a hierarchical structure so that efficient navigation and quick access are possible to reach relevant information in a hierarchical manner.

C. SummaryThemeList DS for Event-Based Summary

It is very useful to access and consume the given content in units of segments that are related to associated themes. Themes can be associated with certain events, categories, places or other entities. The HierarchicalSummary DS provides this kind of efficient functionality. A news program, for example, can be summarized well in terms of its category: political, economic, social, etc. We also easily define main events such as goal, slam-dunk, and three-point shooting in the case of a basketball game.

As shown in Fig. 5, the event-based summary is described by the SummaryThemeList DS under the HierarchicalSummary DS, and the themeIDs attribute in the SummarySegmentGroup DS and/or the SummarySegment DS. The SummaryThemeList DS is used to enumerate all themes available for the video sequence by unique identifiers. On the other hand, the themeIDs is used to indicate theme instances associated with each summary segment by referring the identifiers defined in the SummaryThemeList DS.

In terms of application, the SummaryThemeList DS enables users to browse a video by several specific themes by listing the enumerated themes. The SummaryThemeList DS provides more efficient description to provide a Table of Content (ToC)–like structure, which makes it much easier to determine the available themes and select the preferred items by using the parentId attribute [17]. Furthermore, the SummarySegmentGroup DS also has the themeIDs when all the segments in a highlight level have the same theme and each highlight segment or highlight level is possibly associated with multiple themes, which avoids redundancies in the description [17].

IV. EXPERIMENTS

To show the validity of the proposed approach in real applications, we implemented prototypical systems on a PC platform, and experimented with a set of TV news programs comprising 10 daily news videos broadcasted in Korea in November 1999, each about forty-five minutes long. This section will present the experimental results in the two aspects of summarization and browsing.

1. Summarization

In our experiments, each daily news video was summarized in a two-level (coarse-and fine-level) hierarchy. As shown in Table 2, each sequence is summarized with a compaction ratio of more than a tenth and less than a tenth for coarse-and fine-level, respectively. These amounts of compaction enable fast overview by using summaries that are up to ten minutes in length.

Figure 6 shows the graphical user interface of the implemented software for news video indexing and summarization, in which three modules are integrated, each for multimodal

Table 2. Information for the test news video.

Test Sequences (min:sec)	Summary Information		
	Level	Duration (min:sec)	Compaction Ratio
News1101 (47:44)	coarse	4:14	1/11
	fine	7:49	1/6
News1102 (47:22)	coarse	3:04	1/15
	fine	5:03	1/9
News1103 (49:07)	coarse	3:37	1/14
	fine	6:34	1/7

Fig. 6. Graphical user interface of the implemented software for news video indexing and summarization.

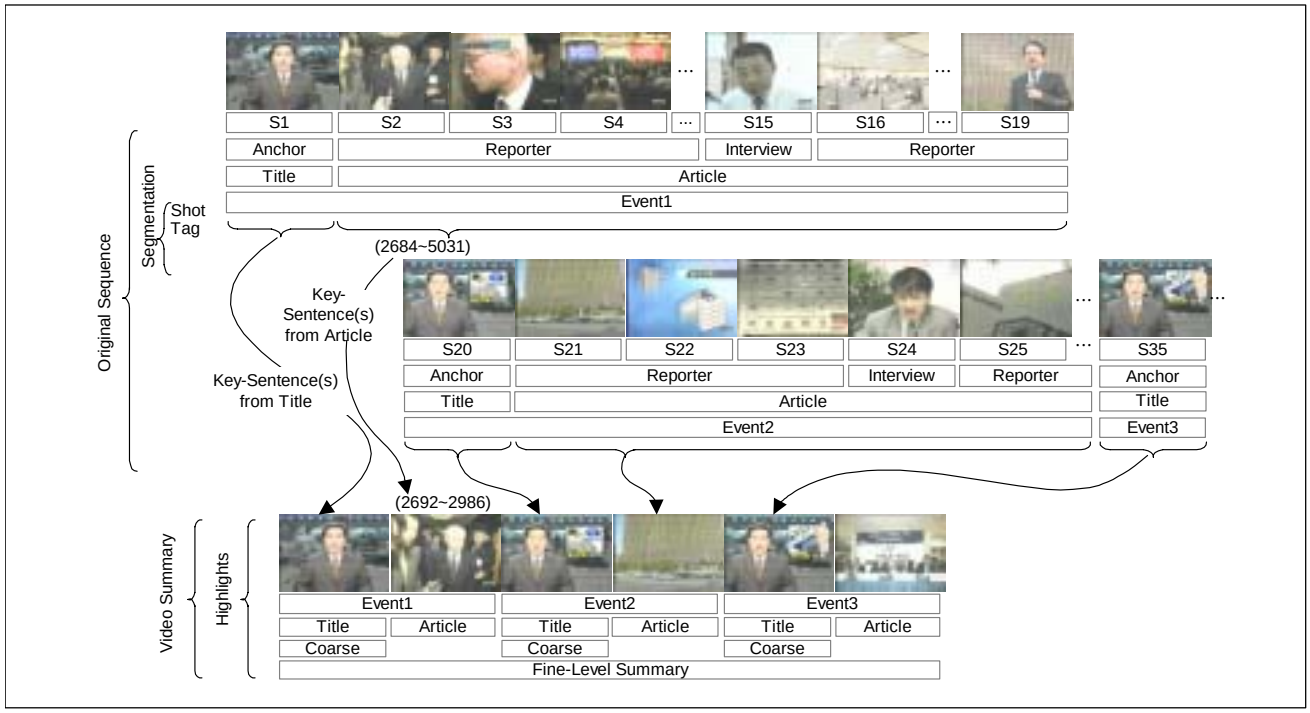

Fig. 7. An example of the procedure of news video summarization.

content. External hardware devices like an MPEG encoder and a CC decoder are also used to compose the digital content acquisition module separately. With the news video inputted, the summary generator analyzes the video multimodally, detects highlight segments, and finally produces a summary description. All this process is done unsupervised. In addition, it gives the functionality of evaluation for the interim results in each processing step. The interface shown in Fig. 6 illustrates the evaluation of the results of the highlight detection step. As shown in the figure, a list of all the extracted highlights are presented to users with key frames attached, so that they can evaluate each item by browsing the associated segments.

In Fig. 7, we illustrate the procedure of summarization for a test sequence, News1101. Dozens of shots are shown from the beginning part of the sequence, represented by their key frames, in the first two rows. Four kinds of segmentations are applied to the original video as indicated. The first one is visual segmentation that partitions the video sequence by the shot, while the others are textual segmentations utilizing tags in CC data. Although, in Fig. 7, it looks as if all the ends shot boundaries are aligned with the others obtained by textual segmentation, this is not true in general. However, we believe it will not make our points misunderstood in describing the relationships among the segments in time line.

It is seen from Fig. 7 that the video sequence is segmented in the unit of scene by tags, and events are detected on the basis of anchor scenes, as described in previous sections. Each event that represents a single news item is then divided into two parts, a title and an article. In this example, a key sentence is extracted from each part of the title and article to form coarse-and fine-level summaries, respectively. As shown in the last row of Fig. 7, the video parts corresponding to the extracted key sentences are marked as highlights for the news video. For example, the second highlight segment lasting from 2,692 to 2,986 in frame units comes from the article part of the first news event that covers from 2,684 to 5,031. The extracted highlights from the beginning two events show that the results reflect the characteristics of the news content in which more significant context is usually located in the preceding part of each news item.

Figure 8 depicts how the extracted summary of the news video is mapped to the Ds, DSs, or attributes of the MPEG-7 HierarchicalSummary DS. The example is an instantiation of a mixed type, including key video-clips, key frames, and key themes summaries. Item categories were manually extracted and specified as themes for each highlight segment. Users may skim the news highlights in two-level alternatives, or view the video content from various points of themes.

2. Browsing

Another application software, named Video Browser, was implemented to verify the validity and effectiveness of the proposed summary generator. The summary description outputted from the summary generator was fed to Video Browser whose

GUI is shown in Fig. 9. By accessing the CC text database, video clips are searchable by inputting text query through *query interface* invoked by *main interface* as shown on the right side in Fig. 9.

The Video Skimming Control located in the upper right corner of the main interface allows a user to control the video summary play mode. Choosing a level (coarse or fine), one can view the key video-clip summary that is the concatenation of the highlight segments in the level. The Summary Criteria panel shows the themes registered in SummaryThemeList DS in the tree structure like ToC to the user. Now, the user can view the summary relating to particular themes of his/her interests, namely, key theme summary, by checking off boxes that correspond to his/her interests. This is actually a kind of user-customization.

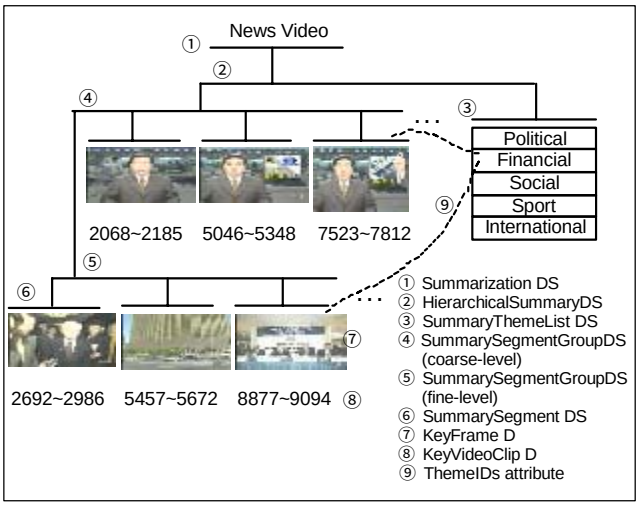

Fig. 8. Pictorial representation for the description of the news video summary.

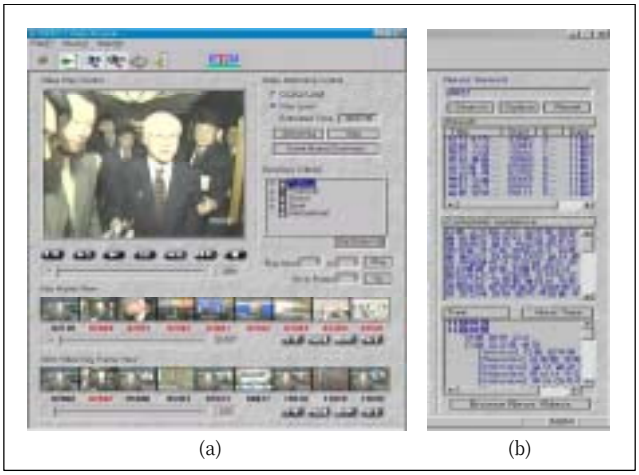

Fig. 9. Graphical user interface of the implemented software for video browsing and text-based retrieval: Video Browser, (a) main interface, (b) text query interface.

The key frames are displayed in two rows at the bottom of the main interface. The key frames in the first row are the ones for the original video, which are specified within the SegmentDS [8]. On the other hand, the key frames in the second row are the ones for the highlights, each of which is specified by KeyFrame D in the corresponding SummarySegments. If a user becomes interested in a particular scene from the overview, he or she may be able to find the scene in the key frames of the highlights, then click the key frame so that the GUI may show the related key frames of the original video. The focus of attention is narrowed down by the key frames of a summary in this way. Finally, the user accesses the original sequence by searching for the key frames most relevant to his/her interests in the first row and clicking it.

In the end, it was shown that efficient access to a video in a content-based manner is achieved by the combination of key video-clip, key frames and key themes summaries in the unified description framework.

V. CONCLUSIONS

This paper addressed the issue of summarizing news videos. It was shown that semantically meaningful highlights were effectively extractable by analyzing multimodal components (*esp*. closed caption) of a news video. Then, the extracted highlights could be described using a Summarization Description Scheme of MPEG-7 in an efficient and interoperable way to support applications for video browsing and navigation. Prototypical systems for summarization and browsing of news videos were implemented and tested with a variety of news sequences, which yielded very promising results in terms of the validity and reliability of the proposed approach. An evaluation framework for video abstraction still remains as a problem to be explored further in the future.

REFERENCES

[1] Y. Rui, T.S. Huang, and S. Mehrotra, "Exploring Video Structure Beyond the Shots," *Proc. IEEE ICMCS'98*, June 1998, pp. 237-240.

[2] D. Zhong, H.J. Zhang, and S.F. Chang, "Clustering Methods for Video Browsing and Annotation," *Proc. IS&T/SPIE Storage and Retrieval for Still Image and Video Database IV*, Feb. 1996, vol. 2670, pp. 239-246.

[3] A. Hanjalic and H.J. Zhang, "An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, Dec. 1999, pp. 1280-1289.

[4] M.A. Smith and T. Kanade, "Video Skimming and Characterization through the Combination of Image and Language Understanding

Techniques," *Proc. IEEE CVPR'97*, June 1997, pp. 775-781.
[5] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting Digital Movies Automatically," *J. Vis. Comm. Image Represent.*, vol. 7, no. 4, Dec. 1996, pp. 345-353.
[6] J.G. Kim, H.S. Chang, M. Kim, J. Kim, and H.M. Kim, "Summary Description Schemes for Efficient Video Navigation and Browsing," *Proc. IS&T/SPIE Visual Comm. and Image Processing*, June 2000, vol. 4067, pp. 1397-1408.
[7] Y.M. Ro, M. Kim, H.K. Kang, B.S. Manjunath, and J. Kim, "MPEG-7 Homogeneous Texture Descriptor," *ETRI J.*, vol. 23, no. 2, June 2001, pp. 41-51.
[8] P. Salembier and J.R. Smith, "MPEG-7 Multimedia Description Schemes, " *IEEE Trans. Circuits and Syst. for Video Techno.*, vol. 11, no. 6, June 2001, pp. 748-759.
[9] MPEG MDS Group, "Text of ISO/IEC FDIS 15938-5 Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes," ISO/IEC JTC1/SC29/WG11 N4205, July 2001.
[10] H.J. Zhang, S.Y. Tan, S.W. Smoliar, and G. Yihong, "Automatic Parsing and Indexing of News Video," *ACM Multimedia Systems*, vol. 2, no. 6, Jan. 1995, pp. 256-266.
[11] A. Hanjalic, R.L. Lagendijk, and J. Biemond, "Semi-Automatic News Analysis, Indexing and Classification System Based on Topics Preselection," *Proc. IS&T/SPIE Storage and Retrieval for Image and Video Databases VII*, Jan. 1999, vol. 3656, pp. 86-97.
[12] A. Merlino, D. Morey, and M. Maybury, "Broadcast News Navigation Using Story Segmentation," *Proc. ACM Multimedia'97*, Nov. 1997, pp. 381-391.
[13] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray, "Automated Generation of News Content Hierarchy by Integrating Audio, Video, and Text Information," *Proc. IEEE ICASSP'99*, Mar. 1999, pp. 3025-3028.
[14] J. Son, J. Kim, K. Kang, and K. Bae, "Application of Speech Recognition with Closed Caption for Content-Based Video Segmentation," *Proc. IEEE DSP Workshop*, Oct. 2000.
[15] T. Shin, J.G. Kim, J. Kim, and B.H. Ahn, "A Statistical Approach to Shot Boundary Detection in an MPGE-2 Compressed Video Sequence," *Proc. IS&T/SPIE Visual Comm. and Image Processing*, vol. 4067, June 2000, pp. 143-150.
[16] MPEG MDS Group, "MPEG-7 Description Schemes (v0.5)," ISO/IEC JTC1/SC29/WG11 N2844, July 1999.
[17] H.S. Chang, J.G. Kim, M. Kim, and J. Kim, "Improved Structure of Hierarchical Summary Description Scheme," ISO/IEC JTC1/SC29/WG11 M6057, May 2000.

Jae-Gon Kim received his BS degree in electronics from Kyungpook National University, Daegu, Korea, in 1990 and MS degree from the Department of Electrical and Electronic Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1992. In 1992, he joined the Electronics and Telecommunications Research Institute (ETRI) where he is currently a Senior Member of Engineering Staff in the Broadcast Media Technology Department. His research interests include content-based video representation and multimedia services in the area of digital broadcasting technology.

Hyun Sung Chang received his BS and MS degrees in electrical engineering from Seoul National University, Seoul, Korea, in 1997 and 1999. Since 1999, he has been with the Broadcasting Media Technology Department of ETRI, Daejeon, Korea, as a Member of Research Staff. His current research interests include multimedia contents analysis and representation, metadata processing, and interactive broadcasting technologies.

Young-tae Kim received his BS, MS, and PhD degrees from the Department of Electronics & Communications Engineering from Kwangwoon University, Seoul, Korea, in 1991, 1993, and 1998. He joined Electronics and Telecommunications Research Institute (ETRI) in 1998 as a postdoctoral fellow. Since 1999, he has been with Broadcasting Media Technology Department of ETRI as a Senior Researcher. His current research interests are digital signal processing, content-based video retrieval, and interactive services in the field of digital broadcasting systems.

Kyeongok Kang received BS and MS degrees in physics from Pusan National University in 1985 and 1988. He is now a PhD candidate in electrical engineering at Hankuk Aviation University. He has has been with ETRI since 1991, and he is now a Senior Member of Engineering Staff and the leader of the Interactive Broadcasting Research Team. His major interests are in low-bitrate audio coding, such as sine transform coder, audio signal processing, and MPEG-7 related issues.

Munchurl Kim received his BE degree in Electronics from Kyungpook National University, Korea in 1989, and ME and PhD degrees in Electrical and Computer Engineering from University of Florida, Gainesville, USA, in 1992 and 1996. After his graduation, he joined Electronics and Telecommunications Research Institute (ETRI) where he had worked in the MPEG-4 standardization related research areas. Since 1998, he has been involved in MPEG-7 standardization work. In the course of MPEG standardization, he has contributed more than 30 proposals in the areas of automatic/semi-automatic segmentation of moving objects, MPEG-7 visual descriptors and Multimedia Description Schemes, and served as the Team Leader on evaluation of Video Description Scheme proposals in MPEG-7 in Lancaster, UK, in 1999. In 2001, he joined, as Assistant Professor in the School of Engineering, the Information and Communications University (ICU) in Daejeon, Korea. His research areas of interest include multimedia computing, communications and broadcasting, and multimedia interactive services.

Jinwoong Kim received his BS and MS degrees from Seoul National University, Seoul, Korea, in 1981 and 1983, and his PhD degree from the Department of Electrical Engineering from Texas A&M University, United States, in 1993. Since 1983, he has been a Research Staff in Electronics and Telecommunications Research Institute (ETRI), Korea. He is currently the Director of the Broadcast Media Technology Department. He has been engaged in the development of TDX digital switching system, MPEG-2 video encoder, HDTV encoder system, and MPEG-7 technology. His research interests include digital signal processing in the field of video communications, multimedia systems, and interactive broadcast systems.

Hyung-Myung Kim received his BS degree in electronics engineering from Seoul National University, Seoul, Korea, in 1974, and MS and PhD degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, in 1982 and 1985. He is now a Professor at the Department of Electrical Engineering and Computer Science, the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. His research interests include digital signal/image processing, digital transmission of voice, data and image and multidimensional system theory. Dr. Kim was the Treasurer of the IEEE Daejeon Section in 1992. He has been an editorial board member of *Multidimensional Systems and Signal Processing* since 1990.