Summary Description Schemes for Efficient Video Navigation and Browsing

Jae-Gon Kim[a], Hyun Sung Chang[a], Munchurl Kim[a], Jinwoong Kim[a], and Hyung-Myung Kim[b]

[a]Broadcasting Media Technology Department, ETRI
[b]Dept. of Electrical Engineering, Korea Advanced Institute of Science and Technology

ABSTRACT

The Summary Description Scheme (DS) in MPEG-7 aims at providing a summary of an audio-visual program that offers the effective mechanism for efficient access to the program by abstracting the contents. In this paper, we present, in details, the Summary DS proposed to MPEG-7 that allows for efficient navigation and browsing to the contents of interest as well as overview of the overall content in an incorporated way. This efficiency is achieved by a unified description framework that combines static summary based on key frames and key sounds with dynamic summary based on a series of highlights. The proposed DS also allows efficient description for the event-based summarization by specifying summary criteria. In this paper, we also show the usefulness of the Summary DS in real applications largely based on the results of the Validation and Core Experiments we performed in MPEG-7 activities. We also describe a methodology for the automated generation of a dynamic summary.

Keywords: MPEG-7, summary description scheme, video abstraction, highlight, video navigation and browsing

1. INTRODUCTION

Recently, the capability of efficient access to the desired video content is of growing importance because more digital video data are available at an increasing rate. Many researchers have focused on the issue including video content analysis, representation, and browsing, which are the bases for accessing video content for the past few years [2]-[7]. The solution of the issue is largely dependent on the representation of the video content. In other words, to enable efficient access to the desired content, an efficient and interoperable content-based representation of the video content, which is one of the goals of the MPEG-7 standard [8], is required. The MPEG-7 standard aims to satisfy the need by standardizing a framework for description of audio-visual content. Video content analysis deals with the extraction of the features to be contained in the descriptions and video browsing directly supports users' access to the video content using the description.

Video summaries abstract the entirety of video with the gist without losing the essential content of the original video result in effective representation of whole content. A video summary can be used to enable the efficient browsing as well as the quick overview of the original contents, by saving the costs spent on the tedious operations such as 'fast-forward' or 'rewind.' In this aspect, video summarization is regarded as a basic and powerful method for content-based representation of the video content in lots of research.

In general, the existing techniques for the video summarization fall into either of the following categories according to their styles: *static summary* [1] and *dynamic summary* [2], [3]. In static summary such as storyboard, a fixed number of representative images are arranged, which gives direct access to different parts of the original video. A dynamic summary is an audio-visual sequence of reduced duration containing a series of highlight segments.

This paper presents the summary DSs proposed to MPEG-7, which describe static and dynamic summary in a unified framework for efficient video browsing. In order to organize the audiovisual information in a structural way, a generic audiovisual description scheme is currently being designed as the Multimedia Description Scheme (MDS) in MPEG-7 [26]. The MDS also provides the Summarization DS in which the summarized information of the whole content is conveyed. The Summarization DS provides a set of summaries of an audio-visual material. Each of the summaries, which are represented by Summary DS, is an audio-visual abstract of the entire contents. It enables the efficient browsing as well as the fast skimming of the contents.

However, the primitive version of the Summary DS specified in [24] was faced with some severe limitations as follows: 1) it does not allow users to further navigate from the contents of the highlight segments into the original content; 2) it lacks

In *Visual Communications and Image Processing 2000*, King N. Ngan, Thomas Sikora, Ming-Ting Sun, Editors, Proceedings of SPIE Vol. 4067 (2000) • 0277-786X/00/$15.00

the abstraction capability of audio information which is more important than visual information in some cases; 3) it can not support efficient description for event-based summarization in terms of description size and search complexity.

In order to overcome such weaknesses, we have proposed a modified description scheme that extends the existing Summary DS with aspects of improved access mechanism, incorporated audio information, and efficient description for the event-based summaries. Especially, a more flexible description structure is proposed to allow browsing relevant data in an efficient way. After overview of the highlight summary video, users more efficiently navigate and/or browse desired content based on the overview. In this sense, the proposed Summary DS can be thought of providing a unified framework for summary description that combines static summary based on key frames and key sounds, as well as dynamic summary based on a series of highlights of audiovisual segments into a unified description structure. The details of the proposal are described in the subsequent sections.

In this paper, we also present a methodology for the automated generation of a dynamic summary. Most video summarization approaches that have appeared in the literature rely on static summary. However, dynamic summary is considered more comprehensive to the users in the sense that it still keeps the audio and the temporal dynamics. Dynamic summary can speed up the browsing of video through its fast sequential playback. A compacted short video of dynamic summary can be used as an electronic program guide or a video-browsing tool in the personal storage devices. Although we focus on the dynamic summary approach, strong aspect of static summary is included in the description of the dynamic summary through the unified framework.

In the next section, we describe the overall scheme for generation of dynamic summaries that are described by the proposed summary DS. In Section 3, we present the details of the proposed Summary DS with the inspection of the HierarchicalSummary DS. The experiments show the feasibility and the efficiency of the proposed DS in Section 4. Finally, conclusions are presented in Section 5.

2. DYNAMIC SUMMARY GENERATION

1. Overall Scheme

The critical aspect of summarizing a dynamic video is context understanding, which is the key to choosing highlight that should be included in the final summarized video. The dynamic video summarization technique has attracted the attentions since the Informedia project [2], executed by Carnegie Mellon University. Although the existing methods report quite good performances, their results may not necessarily be semantically meaningful since they adopt feature based approaches [2], [3].

In this paper, we attempt to utilize the correlation between physical appearances (features) and the underlying semantics to detect highlights. First, a set of events is defined according to the kind of a video sequence. Then, knowledge-based rules are applied to detect the events and the episode boundaries around them. Finally, we select highlights to be included in the dynamic summary according to the skimming rule where the users' needs and perception capabilities are considered.

Figure 1 shows the overall procedure of our approach for generating dynamic summaries. The proposed approach consists of three consecutive processing steps as follows:

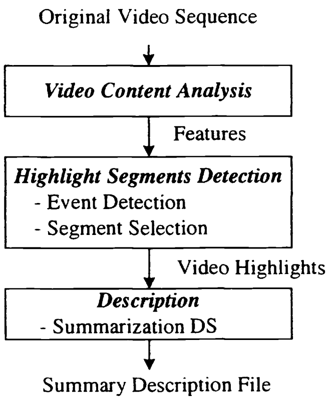

Figure 1. Overall procedure for our approach for dynamic summary generation

- *Video content analysis* – extracts the structural information (shot) and visual features such as camera motion, embedded caption, etc. from the video.
- *Highlight segments detection* – It consists of event detection and segment selection steps. In the first step, the event detection module detects main events and estimates the episode interval for each event, based on the information provided by the video content analysis module. In the second step, the segment selection module determines the video intervals to be included in the summary.
- *Description* – describe a set of highlight segments composing an audio-visual dynamic summary based on the structure specified by the Summarization DS. Details are described in section 3.

2. Video Content Analysis

In terms of content-based representation, the goal of video content analysis is to extract the visual features and the structural information from video in a way that allows meaningful and efficient classification, indexing, and retrieval. In general, the very first step in the video content analysis is shot boundary detection to recover the basic structure. Once the shots have been detected, the resulting shots are analyzed by further steps such as key-frame extraction, camera motion characterization, and significant object detection, etc.

Although meaningful features for summarization are well known, different features are required by the different approach and the different kinds of video to be summarized. In our approach, we assume that the predefined event is accompanied by several features. In a soccer video, for example, it is very probable that a goal is followed by captions bearing the shooter's name and/or the replays. We specify the features to be extracted in the video content analysis step to detect the event.

For the soccer, we specify shoots and goals as the events to detect. To do this, we performs the extraction of the following features for automatic summarization:

- Degree of camera views – from 'distant' to 'close,' based on the object-to-screen ratio.
- Shot categories – Rough categorization for each shot. In the soccer example, most of the shots can be categorized into the ground shot (G), bench shot (B), or cheering audience shot (A).
- Shot boundary information - cut, dissolve, wipe.
- Video playing speed – normal, slow
- Camera motion characterization – zoom-in, zoom-out, pan, etc.
- Embedded caption

Video content analysis for the automatic extraction of features is important for the automated generation of the dynamic summary. The tool set of the feature extraction is being developed. To support camera motion characterization, we have developed techniques to detect camera motion in the MPEG compressed domain [10]. The affine motion parameters that fit motion vector field are qualitatively interpreted to classify camera motion types of zoom, rotation, pan, tilt, object motion, and static. We also have developed the tool of closed caption analysis to extract key word and key-sentence in the case of new video summarization. We can detect a highlight segment associated the key-word/key-sentence incorporating the speech analysis techniques applied to sound tracks [11].

3. Highlight Segments Detection

As described before, we attempt to utilize the correlation between visual features and the underlying semantics to detect highlights. It is often the case that an event (esp. semantic event) is accompanied by several of other events or features. These cause-and-effect relationships can be utilized to detect events even if they are not deterministic. In order to detect the events in a video sequence, an inference framework based on the Bayes rule will be considered. For an event E and its associated feature F, Bayes rule says

$$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F)}. \tag{1}$$

If we let $\frac{P(F|E)}{P(F)}$ be denoted by $G_F(E)$, (1) is rewritten as

$$P(E \mid F) = G_F(E)P(E). \tag{2}$$

Note, in (2), that $G_F(E)$ can be interpreted as the *gain* to the occurrence probability of E, obtained from the observation of F. It means that the feature F satisfying

$$G_F(E) = \frac{P(F \mid E)}{P(F)} > 1 \tag{3}$$

can serve as a hint of the occurrence of E in a probabilistic sense. In other words, it helps detect an event to detect the feature which occurs more frequently along with the event than used to. For the multiple features $F = \{F_1, F_2, \cdots, F_N\}$,

$$P(E \mid F) = \prod_{k=1}^{N} \frac{P(F_k \mid F_1, \cdots, F_{k-1}, E)}{P(F_k \mid F_1, \cdots, F_{k-1})} \cdot P(E). \tag{4}$$

Similarly to (3), the condition

$$P(F_i \mid J, E) > P(F_i \mid J), \ \forall J \subset F - \{F_i\} \tag{5}$$

forces all of the (conditional) gain terms in (4) to be larger than 1, which means all of the features contribute to the overall composite gain. The feature set F satisfying (5) constitutes a diagnostic support set for the event E. On detecting any subset $J \subset F$ whose composite gain $G_J(E)$ is sufficiently large $\left(> \frac{1}{2P(E)}\right)$ we are able to make a decision that the event E occurred. The diagnostic support set for each event may have to be found empirically from the observations.

In the followings, some examples are given for soccer videos although the approach (framework) is also applicable to other types of videos as well. For the soccer, we specify shoots and goals as the events to detect. To do this, we assume the extraction of the following features: degree of camera views, shot categories, shot boundary information, replay, camera motion characterization, and embedded caption. The empirical results on the gain of each feature to the occurrence possibility of the events are shown in Table 1.

Table 1. Gain of Each Feature to the Event Occurrence Probability (*empirical*)

Event	Goal/Shoot					Replay	
Feature	Zoom-in	Caption	Far view	Ground shot	Replay	Slow motion	Gradual transition
Gain	2.236	9.396	2.577	1.076	7.934	10.02	4.956

An event belongs to more global context built up around it rather than forms a story unit by itself. The global contexts are referred to as *episodes*, *scenes*, or sometimes *story units*. As an example for a soccer video, in general, several types of associated shots follow the primary goal shot – close-up shots for the shooting hero with or without his name overlaid, the applauding audience shots, replayed event shot from different views, etc. The shots constituting the episode should not be considered separately. Thus, it is desirable to select highlight segments for each episode once after the episode boundaries are identified.

The episode boundary identification process is based on the domain-dependent rule obtained from the prior observations. From the experiments, we found out that the episode for a goal or shoot starts from the event shot and continues until the next play that is captured as a 'ground shot in distant or middle view' at the normal speed in many cases, as in Figure 2. After the identification of episode boundaries, the priorities may be assigned for each episode according to the kind of the event, the length of the playback time or audiovisual activity.

Then, we finally generate the summary video according to the skimming rule where the users' needs are reflected. Each user may adjust the length of the summary video, etc. by the following steps of inputs through the user interface. Figure 2 illustrates the example of how the skimming rule is applied to real soccer videos.

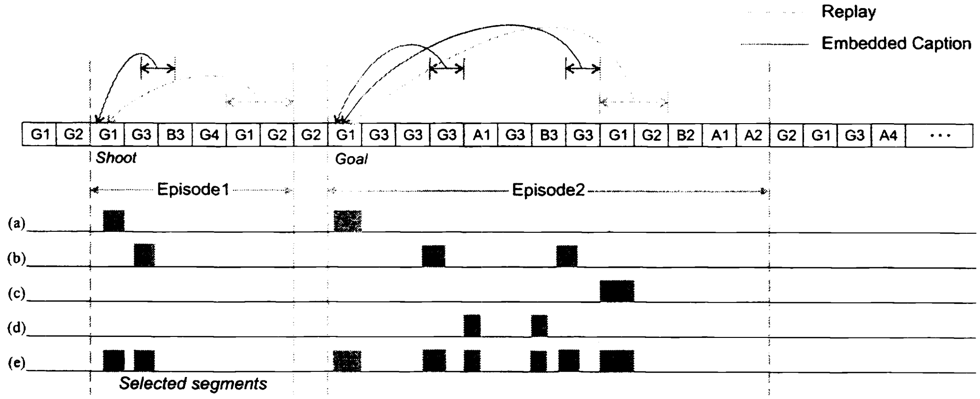

Figure 2. The event detection and episode boundary identification applied to a soccer sequence. Each shot is represented by a block that carries the information of shot category and degree of camera views (from 'distant' (1) to 'close' (4)).

We finally extract highlight segments in Figure 2(e) by the steps as follows.

- the event shot (goal or shoot)
- embedded caption
- replayed shot
- auxiliary contents such as bench and audience shot

The length of each segment depends on the priority of the corresponding episode. The selection of each segment interval is made on the basis of the camera motion characterization. For example, the zoom-in interval is chosen for the event shot since it is very likely that the camera zooms in the ball kicked into the goal-net. To reduce possible nuisances, the length of each segment is forced to exceed the minimum threshold (1sec).

3. SUMMARY DESCRIPTION

The Summarization DS in MPEG-7 provides a set of summaries of an audio-visual material. Each of the summaries, which are represented by Summary DS, is an audio-visual abstract of the entire contents. As illustrated in Figure 3 where UML notations are used, in real instances, either HierarchicalSummary DS or SequentialSummary DS is derived from the Summary DS, and used instead of it.

1. Basic Architecture of the HierarchicalSummary DS

The primitive version of HierarchicalSummary DS [24], which is to convey the essential information about the content of an audio-visual material, incorporates several proposals submitted in responses to the MPEG-7 Call for Proposals [23]: synopsis [12], the highlight, event, and close-up view [13], multi-level digest [14], and so on.

In its simplest case, the primitive HierarchicalSummary DS just concatenates the important portions of a video to be displayed. Then, users can understand the outline of the video by the summary without watching the full version of the video. When users navigate a multimedia server system such as a Video-on-Demand and a Broadcast system to select a video or audio program, they might want to access the summarized version, in advance, rather than the entire stream in order to know the outline of the program.

Sometimes, it is necessary to provide multi-level summaries for the multimedia data. For example, users may want to see five-, ten-, or twenty-minute summaries of a two-hour video stream according to their needs. As shown in Figure 4, the primitive version of HierarchicalSummary DS organizes a summary into a succession of levels, each represented by a HighlightLevel DS. In general, levels closer on the top of the hierarchy provide a coarse summary and levels closer to the bottom correspond to finer levels of detail. The HierachicalSummary DS distinguishes two different types of hierarchies, based on the parent-child relationships, specified by the hierarchyType attribute, in two consecutive levels: *independent* and *dependent*. In an independent hierarchical summary, the information in the children levels is self-complete, without reference to the information in its parent element. On the other hand, in a dependent hierarchy, the children levels add to, or refine, the information in its parent element. Without the parent level's information, the child levels cannot be interpreted.

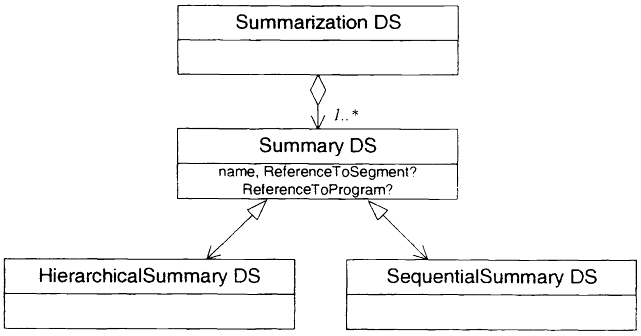

Figure 3. The structure of the Summarization DS (in UML)

Each HierarchicalSummary DS is associated with a summaryType attribute, which specifies what this summary represents. The summaryType can take one out of five possible values – keyVideoClips, keyAudioClips, keyFrames, keyEvents, and mixed type. For example, if the summaryType is keyEvents, the summary might correspond to instances of a particular type of semantic event, and the clips of these events can be accessed and visualized (e.g. all the slam ducks in a basketball game). Or if the summaryType is keyFrames, the summary is reduced to the same as the key-frame hierarchy as in [7], [9]. The HighlightLevel DS specifies one particular level of a highlight summary. The HighlightLevel DS contains an arbitrary number of segments (visual or audio), located by a SegmentLocator DS. A particular level may correspond to, for example, a highlight with particular time duration or a particular set of events.

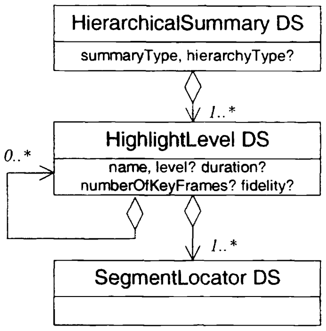

Figure 4. Basic architecture of the HierarchicalSummary DS (in UML)

2. Extensions to the HierarchicalSummary DS

The primitive HierarchicalSummary DS in Figure 4 is a simple structure which effectively meets the assigned functionality – delivery of the essence of a video content. However, as mentioned before, the primitive version is face with some severe limitations. First, it does not allow users to further navigate from the contents of the highlight segments into the original content. Notice that, after the quick overview of the summarized videos, users are very often eager to navigate or browse the scenes that they are particularly interested in. Second, it lacks the abstraction capability of audio information which is more important than visual information in some cases. Moreover, the primitive version of HierarchicalSummary DS turned out to require too much cost (i.e. description size and search complexity) to support the event-based summarization [16]-[18].

In order to overcome such weaknesses, some extensions were made to the HierarchicalSummary DS. As illustrated in Figure 5, the HighlightSegment DS was put under the HighlightLevel DS with an attribute segmentType, and four sub-DSs – VideoSegmentLocator DS, AudioSegmentLocator DS, ImageLocator DS, and SoundLocator DS, respectively. The SummaryCriteria DS was also added to the HierarchicalSummary DS. The details of the extensions will be explained in subsequent sections.

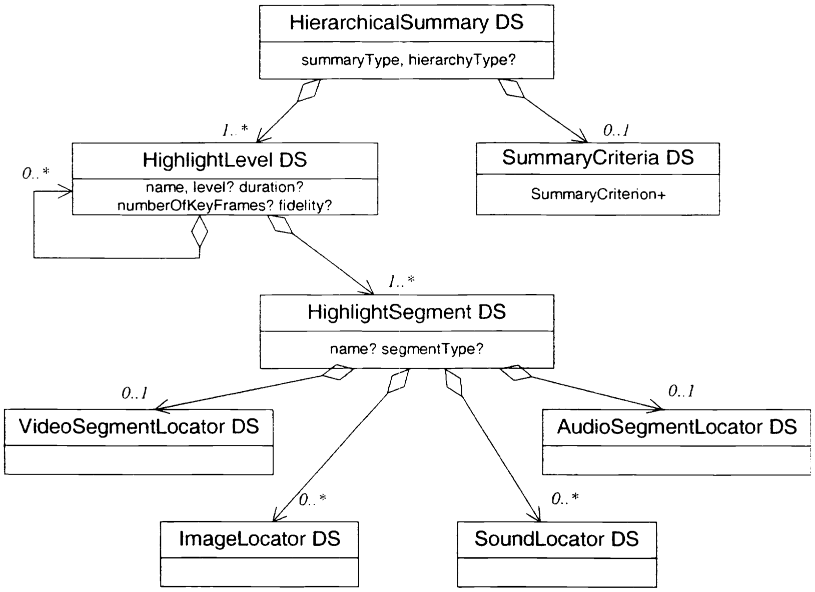

Figure 5. Extensions to the basic architecture of HierarchicalSummary DS (in UML)

2.1 Improved Access Mechanism

Often people want to take a look at the dynamic summary as form of audio-visual summary such as a film trailer at the first glance in order to capture the underlying story of a long program. The objective of the primitive HierarchicalSummary DS in Figure 4 is mainly focused on such functionality. However, sometimes the overview is not sufficient to appreciate the underlying story of the content. So in the next step, with what they feel interested in after the overview, they want to navigate the content of the program in order to reach subjects of interest, starting from the auido-visual dynamic summary.

Meanwhile, a static summary gives direct access to different parts of the original video. However, common drawbacks of static video summaries are that they do not preserve the time-evolving dynamic nature of video content. Furthermore, in general, there are too many key frames to fine appropriate for browsing. It is reported, for examples, that the movie *Terminator – the Judgement Day* contains three hundred shots in a fifteen-minute video clip [6]. The movie lasts about a hundred and forty minutes, which means approximately three thousand shots. So it is a very tedious job to rummage a large number of key-frames in order to find subjects of interest in the whole description of the content. It also requires large amounts of time to directly navigate the whole description of the program content.

In our approach, we combine the strong point of static summary's facility of direct access and dynamic summary's capacity of quick skimming. In other words, to navigate and browse the contents, we utilize the key-frames of the highlight segments composing dynamic summary to reach and find the subjects of interest. So, the key-frames of the audio-visual dynamic summary play stepping-stones between the summary description and the content of the original program. This is achieved by introducing the HighlightSegment, where VideoSegmentLocator DS, ImageLocator DS, and SoundLocator are located together as shown in Figure 5. This unified framework enables coarse-to-fine navigation to traverse from the summary to the original program or more relevant information in the generic description. Meantime, the extended HierarchicalSummary DS can be a multi-level of dynamic summary result in providing key-frames in a hierarchical structure so that efficient navigation and quick access are possible to reach the relevant information in hierarchical manner.

Figure 6 shows a timeline layout of audio-visual segments and navigation path from audio-visual highlight segments to relevant audio-visual information. If, after reviewing highlight video, a user is interested in the first key-frame or key-sound in the first highlight audio-visual segment, he/she can navigate the original content based on the key-frame or key-sound.

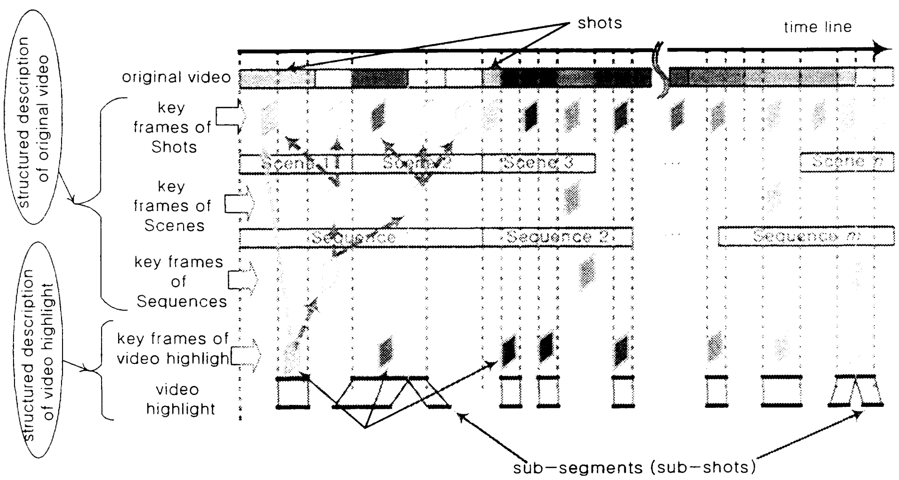

Figure 6. Timeline layout of a video segments and navigation from highlight to original audio-visual content

In general, when abstracting the original video based on key-frames as shown in Figure 6, the visual abstraction results in many key-frames as aforementioned. In contrary, since the summarized video is composed of highlight segments, it can be abstracted with a much smaller number of key-frames. A key-frame in a highlight segment guides a navigating user to its related key-frame(s) which is (are) at the lowest level in the description hierarchy of the whole description of a long program. As shown in Figure 6 where the navigation path is indicated by bold dash lines, the focus of attention on desired subjects to be found narrows down progressively in more details into deeper levels in the hierarchy through the navigation path.

2.2 Key-Sound as a Counterpart of Key-Frame

In many videos, audio tracks are associated with their visual counterparts. Since the audio information usually conveys semantic information, it is treated more importantly in limited computation or transmission environments.

In the extended version of HierarchicalSummary DS, as analogy to the key-frames for video segments, the key-sounds (specified by SoundLocator DS) represents aural abstract of their associated highlight segments. Each highlight segment can contain highlight audio segments (indicated by AudioSegmentLocator DS) which can be abstracted with key-sounds. The key-sounds include keywords in speech, sound effects represented within some time interval, emotional sounds, exploding sounds, loud sounds, specific instrument sounds, turning point words in dialogues, names, etc. As with the key-frames in video, the key-sound is also efficient abstraction of underlying semantics in audio tracks. For example, in a soccer game, important events can be realized with several key-sounds such as "goal-in," "throwing," "corner-kick," "player's name in sound," "foul," "tackle," "spectators' outcry," etc. If these key-sounds are abstracted with the key-frames for the soccer game, the key-frames can better be understood by listening the semantic key-sounds. So it is easier for a user to narrow down focus of attention to more relevant subjects of interest.

2.3 Event-Based Summary

Event-based summary is one of the most useful functionalities provided by the HierarchicalSummary DS. A drama, for example, can be summarized well in terms of its Person-Action-Locale (PAL) constituents [16], [21]. However, the primitive version of HierarchicalSummary DS entails a serious problem that it makes redundant descriptions for the highlight segments related to more than one event [16]-[18]. Note that in the primitive version a highlight segment involved in multiple events should be instantiated as many times as the number of events. Further, it may lead to the increased complexity when a user wants to browse some particular subjects of interest in events, since it requires exhaustive search of all event categories and complicated operations.

In order to overcome these drawbacks, we proposed, in [15] and [18], some extensions to the HierarchicalSummary DS, which considers the description efficiency and lowers the search complexity. As shown in Figure 5, the modification is made by introducing the SummaryCriteria DS under the HierarchicalSummary DS, and the segmentType attribute in Highlight Segment DS. In this scheme, each highlight segment is associated to possibly multiple key events, which inhibits the redundancies in description and reduces the search complexities.

Specifically, the segmentType is used to describe main event instances in each highlight segment. It may vary according to the kinds of videos. For examples, The segmentType is used to describe main event instances in each highlight segment. It may vary according to the kinds of videos. For example,

- Drama– PAL (Person-Action-Locale) model as in [12]
- Sports– main sports events such as slam dunk, three-point shoot, etc., and
- News– news categories (political, economic, social, etc.)

can be used.

On the other hand, the SummaryCriteria DS is used to enumerate segmentTypes available for a video sequence. In a possible scenario, an application system shows the summaryTypes registered in the SummaryCriteria to the users and lets them select the events of their own preferences. Then, it builds the user-customized summaries dynamically. Both of the above proposed components, SummaryCriteria DS and segmentType attribute, are optional, that is, when the HierarchicalSummary DS instantiates only summaries not based on key events, they are not used.

The summary instances which contains the SummaryCriteria DS, SummaryCriteria D (with all of its attributes), and segmentType describes the event information in a video very efficiently in two aspects.

- Description efficiency — All of the key event information can be described at a time, and the redundancies entailed by the primitive version of HierarchicalSummary DS are inhibited. (For the details, refer to [20])
- Low search complexity — The search of even complicated combinations of the key events is reduced to segmentType checking for each highlight segment which is a relatively simple process.

4. EXPERIMENTS

This section presents an example of the description and browsing experiments to show how the Summarization DS can be applied to the real instances. The example is mainly based on the results of the MPEG-7 DS Core Experiment (whose reference code is ETRI-4 [9], [10]). The example was performed on the mixed types of summaries which can visualize, possibly in multi-levels, the video content on the bases of various key events as well as key video-clips. (Refer to [20] for more details.)

The example to be discussed is the summarization for the video of V8 in MPEG-7 Content Set, which is a situation comedy produced by Portuguese TV [22]. The detailed information for the video is tabulated in Table 2. We summarized the video in two levels (coarse and fine level). Each of them contains four and six highlight segments, respectively. For each highlight segment, its Person-Locale (PL) constituents were extracted manually and defined as key-events. The summary information for the video is illustrated in Table 3, where, for example, the highlight segment which lasts from the frame 10,600 to 10,728 is in the fine level and contains three events – Pedro, Anna, and Living Room. The pictorial representation for the summary of the sitcom is illustrated in Figure 7, where each component is mapped to the Ds, DSs, or attributes defined in the HierarchicalSummary DS.

For the experiment, we first wrote a simple DTD for the Summarization DS on the basis of DDL specification in [28]. Then, we used DOM API to create the description files in XML for the above video material. The well-formedness and validity of the XML file were checked by IBM XML parser as well as Microsoft IE 5.0.

Table 2. Information for the Situation Comedy of MPEG-7 Content Material V8

Test Sequence	Summary Information		
(Duration)	Level	Duration	Key Events
Sitcom (V8, 38:00)	coarse	0:17 (1/136)	Pedro; Anna; Old_man; Pedro_Room; Living_Room
	fine	0:44 (1/52.1)	

Table 3. Detailed Summary Information for the Situation Comedy of MPEG-7 Content Material V8
(Two-Level Hierarchy, Five Key Events Based on Its Person-Locale Constituents)

Segment Id	MediaTime		Level	Person			Locale	
	Start	End		Pedro	Anna	The old	Pedro room	Living room
1	10,600	10,728	fine	o	o			o
2	13,800	13,865	fine	o	o		o	
3	13,866	13,912	coarse	o	o	o		
4	50,579	50,658	fine			o	o	
5	50,659	50,707	coarse			o	o	
6	50,961	51,192	fine			o	o	
7	53,050	53,151	fine	o	o		o	
8	53,705	53,774	fine	o	o		o	
9	54,112	54,343	coarse	o	o		o	
10	55,230	55,319	coarse		o		o	

We have verified the feasibility and the efficiency of the proposed summary DS using the exemplary application system we have implemented, a video browser. The generated description file is fed to the video browser, whose graphical user interface (GUI) is shown in Figure 8. The Video Play Control in the GUI is to display the original or summarized versions of a video. The Video Skimming Control located in right-upper side takes various inputs related to the video summarization from a user. With the Summary Criteria panel deactivated as it was at initial time, the click of Skimming button makes a keyVideoClips summary which is the concatenation of the HighlightSegments in the HighlightLevel specified by the radio button. It should not be missed, here, that the fine-level summary includes all the highlight segments in the coarse-level in case that the hierarchyType of the Summary is set to dependent. The Summary Criteria panel is activated or deactivated by the click of button Event-Based Summary. When activated, this panel shows the segmentTypes registered in SummaryCriteria DS with check-boxes to the user. Now, the user selects the events of his or her interests, and clicks the Skimming button to make a keyEvents summary, which is actually a kind of user-customization. Note that the summary type was changed from keyVideoClips to keyEvents with the activation of the Summary Criteria panel.

The key-frames are displayed in two rows at the bottom of the GUI. The key-frames in the first row named Key Frame View are the ones for the original video. They are specified within Multimedia DS. On the other hand, the key-frames in the second row, Skim Video Key Frame View, are the ones for the highlight segments, each of which is specified by the ImageLocator in the corresponding HighlightSegment. If the user became interested in a specific scene after the overview, he or she may be able to find the scene by the key-frames in Skim Video Key Frame View panel, and click the key-frame so that the GUI may show the related key-frames of the original video in the Key Frame View panel. The focus of attention is narrowed down by the key-frames or key-sounds of a summary in this way. Finally, the user finds the scenes of particular interests in the Key Frame View panel, and clicks the key-frame so that the portion of the original can be played.

5. CONCLUSIONS

This paper presents a methodology for the dynamic video summarization in content-based manners, which is necessary for the efficient use of vast amount of multimedia data. The dynamic summaries can be described by the MPEG-7 Summarization DS in an efficient and interoperable way to support applications for browsing and navigation.

In this paper, we have inspected the HierarchicalSummary DS of MPEG-7 and have presented the HierarchicalSummary DS proposed in MPEG-7. The proposed HierarchicalSummary DS has evolved from the primitive version which is basically the concatenation of the key video-clips possibly in multiple levels into much more powerful version that 1) it allows for users to further navigate the contents of highlight segments into the original content through the unified framework; 2) it is armed with the audio information; and 3) it reduces the costs (i.e. description size and search complexity) spent to support the event-based summarization. An exemplary application software was implemented, and intensive experiments were performed to show the usefulness of the proposed Summary DS in real life.

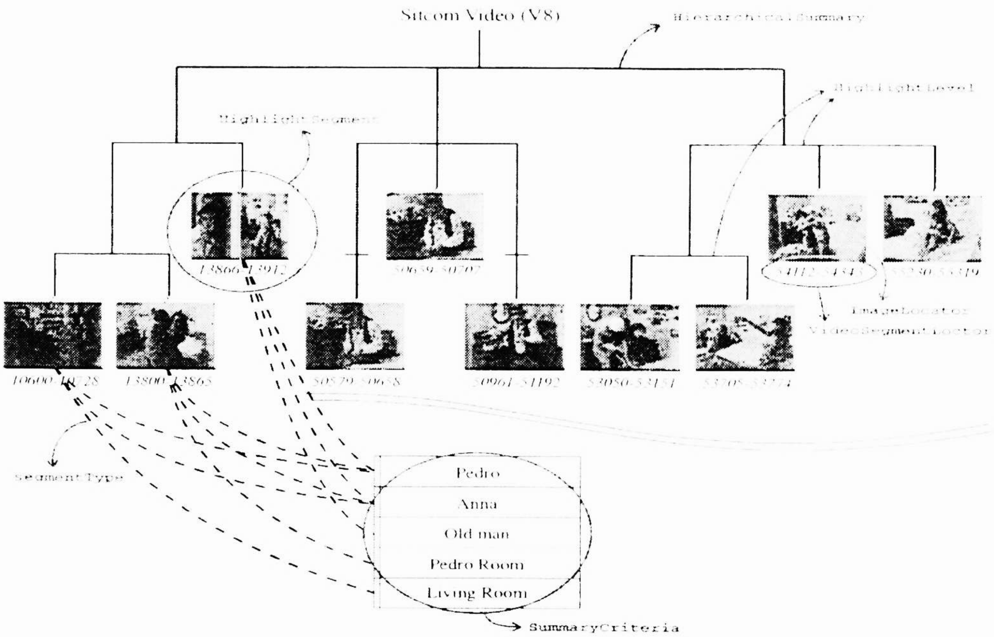

Figure 7. Pictorial representation of the summary of the situation comedy of MPEG-7 content material V8

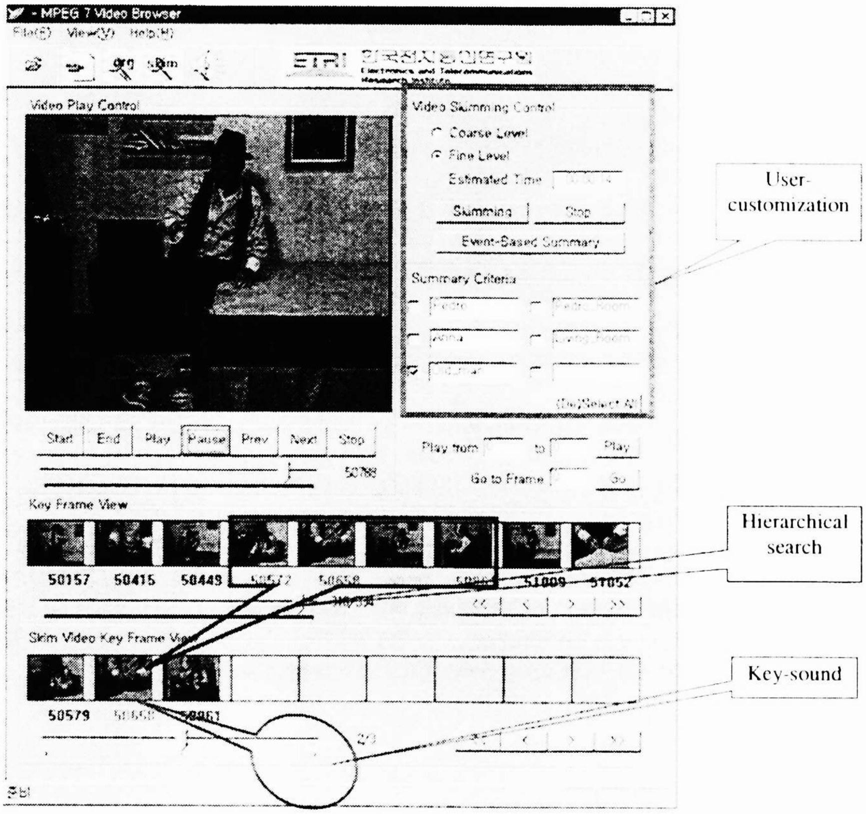

Figure 8. The graphical user interface of the exemplary application system (A Video browser)

REFERENCES

1. D. Yow, B.-L. Yeo, M. M. Yeung, and B. Liu, "Analysis and Presentation of Soccer Highlights from Digital Video," in Proc. 2nd Asian Conf. Computer Vision, vol. 2, Dec. 1995, pp. 499-503.
2. H. Wactlar, T. Kanade, M. Smith, and S. Stevens, "Intelligent Access to Digital Video: The Informedia Project," IEEE Computer, vol. 29, no. 5, pp. 46-52, May 1996.
3. R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video Abstracting," Communications of the ACM, vol. 40, no. 12, pp. 55-62, Dec. 1997.
4. R. Qian, N. Haering, and I. Sezan, "A Computational Approach to Semantic Event Detection," in Proc. IEEE CVPR'99, vol. 1, June 1999, pp. 200-206.
5. A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated High-Level Movie Segmentation for Advanced Video-Retrieval," IEEE Trans. Circuits and Systems for Video Technology, vol. 9, no. 4, pp. 580-588, June 1999.
6. Y. Rui, T. S. Huang, and S. Mehrotra, "Exploring Video Structure Beyond the Shots," in Proc. IEEE ICMCS'98, June 1998, pp.237-240.
7. H. S. Chang, S. Sull, and S. U. Lee, "Efficient Video Indexing Scheme for Content-Based Retrieval," IEEE Trans. Circuits and Systems for Video Technology, vol. 9, no. 8, pp. 1269-1279, Dec. 1999.
8. N2727, "MPEG-7 Requirements Document," ISO/IEC JTC1/SC29/WG11, Mar. 1999.
9. P90, S. Sull, H. S. Chang, and S. U. Lee, "Tree-Structured Index of Video," ISO/IEC JTC1/SC29/WG11, Feb. 1999.
10. J.-G. Kim, H. S. Chang, J. Kim, and H.-M. Kim, "Efficient Camera Motion Characterization for MPEG Video Indexing," to appear in ICME2000, Aug. 2000.
11. J. Son, J. Kim, K. Kang, and K. Bae, "Content-Based Video Segmentation Using Closed Caption and Speech Recognition," to appear in ICME2000, Aug. 2000.
12. P336, T. Walker, H. Matsubara, P. Kuhn, and Y. Shibata, "Video Navigation Description Scheme," ISO/IEC JTC1/SC29/WG11, Feb. 1999.
13. P429, R. J. Qian, P. V. Beek, and M. I. Sezan, "Description Schemes for Consumer Video Applications," ISO/IEC JTC1/SC29/WG11, Feb. 1999.
14. P675, S. B. Jun, H. J. Kim, J. S. Lee, J. M. Song, and H. Y. Lee, "Multi-level Digest Segment Information Scheme," ISO/IEC JTC1/SC29/WG11, Feb. 1999.
15. M5022, M. Kim, J.-G. Kim, H. S. Chang, S.-Y. Kim, and J. Kim, "An Extended Summary DS for Navigation and Browsing," ISO/IEC JTC1/SC29/WG11, Oct. 1999.
16. M5458, K. Yoon, S. B. Jun, and H.-Y. Lee, "Validation Experiment Result Report on Summary DS," ISO/IEC JTC1/SC29/WG11, Dec. 1999.
17. M5492, M. Kim, J.-G. Kim, H. S. Chang, and J. Kim, "Results of the Validation Experiments on Summary DS," ISO/IEC JTC1/SC29/WG11, Dec. 1999.
18. M5493, M. Kim, J.-G. Kim, H. S. Chang, and J. Kim, "An Efficient Hierarchical Summary DS for Event-Based Summarization," ISO/IEC JTC1/SC29/WG11, Dec. 1999.
19. M5790, P. V. Beek, "Report of the CE on Summary Description Schemes," ISO/IEC JTC1/SC29/WG11, Mar. 2000.
20. M5852, H. S. Chang, J.-G. Kim, M. Kim, and J. Kim, "Results of Core Experiments on Summary DS," ISO/IEC JTC1/SC29/WG11, Mar. 2000.
21. J. Saarela and B. Merialdo, "Using Content Models to Build Audio-Video Summaries," in Storage and Retrieval for Image and Video Databases VII, vol. SPIE-3656, Jan. 1999, pp. 338-347.
22. N2467, "Description of MPEG-7 Content Set," ISO/IEC JTC1/SC29/WG11, Oct. 1998.
23. N2469, "MPEG-7 Call for Proposals," ISO/IEC JTC1/SC29/WG11, Oct. 1998.
24. N2844, "MPEG-7 Description Schemes (V0.5)," ISO/IEC JTC1/SC29/WG11, July 1999.
25. N3112, "MPEG-7 Multimedia Description Schemes XM (Version 1.0)," ISO/IEC JTC1/SC29/WG11, Dec. 1999.
26. N3113, "MPEG-7 Multimedia Description Schemes WD (Version 1.0)," ISO/IEC JTC1/SC29/WG11, Dec. 1999.
27. N3128, "Core Experiments on User Preferences," ISO/IEC JTC1/SC29/WG11, Dec. 1999.
28. N3201, "DDL Working Draft 1.0," ISO/IEC JTC1/SC29/WG11, Dec. 1999.