Real-Time Video Indexing System for Live Digital Broadcast TV Programs

Ja-Cheon Yoon[1], Hyeokman Kim[2], Seong Soo Chun[1], Jung-Rim Kim[1], and Sanghoon Sull[1]*

[1]Department of Electronics and Computer Engineering, Korea University, Seoul, Korea
{jcyoon,sschun,jrkim,sull}@mpeg.korea.ac.kr
[2]Department of Computer Science, Kookmin University, Seoul, Korea
hmkim@kookmin.ac.kr

Abstract. In this paper, we introduce a real-time metadata service system that is implemented for live digital broadcast TV programs. The system is composed of three parts: an indexing host which indexes broadcast programs in real-time, a broadcaster where the segmentation metadata delivered from the indexing host is multiplexed into the broadcast stream and transferred to clients, and a client PVR that receives the metadata and locates a segment of interest from the recorded stream according to the time description of the delivered metadata. We propose to utilize broadcasting time for a time description of the segmentation metadata, so as to be free from the media localization problems in broadcast environment. In addition, we utilize a spatiotemporal visual pattern of a video for a verification tool of real-time indexing, such that we can reduce the false alarms of video segmentation caused by lack of an efficient tool for verifying video segment. As a result, we show the real experiments that are performed without requiring a return channel and demonstrate the feasibility of the proposed system.

1 Introduction

Recently, digital set-top boxes (STBs) with local storage known as a personal video recorder (PVR) begin to penetrate TV households. With this new consumer device, television viewers can record broadcast programs into the local storage of their PVR for viewing later. Due to the nature of digitally recorded video, viewers now have the capability of directly accessing to a certain point of recorded programs in addition to the traditional controls such as fast forward and rewind. Furthermore, if a segmentation metadata for a recorded program is available, the viewers can browse the program by selecting some of predefined video segments within the recorded program and play highlights as well as summary of the recorded program.

The metadata can be described in proprietary formats or in international open standard specifications such as MPEG-7 [1] or TV-Anytime [2]. The media location used in typical metadata such as TV-Anytime format are usually described by using either byte offset specifying the number of bytes to be skipped from the beginning of

* Corresponding author

P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp. 261–269, 2004.

the file or media time specifying a relative time point from the beginning of the file. However, it might be ambiguous to describe a specific position of a broadcast stream using media time or byte offset, since it is hard to clearly identify when or where a program starts within the broadcast stream in which a number of programs or commercials are multiplexed and that is continuously being streamed without a program boundary marker through the broadcast network.

One possibility for random access to a specific position of broadcast streams is to use MPEG-2 DSM-CC Normal Play Time (NPT) [3] that provides a known time reference to a piece of media. For applications of TV-Anytime metadata in DVB-MHP broadcast environment, it was proposed that the NPT should be used for the purpose of time description [4, 5]. In the proposed implementation, however, it is required that both indexing system and client PVRs can handle NPT properly, thus resulting in highly complex controls on time.

Another possibility is to use the MPEG-2 Presentation Time Stamp (PTS) which indicates the time that a presentation unit is presented in the system target decoder. However, it requires parsing of packetized elementary stream (PES) layers, and thus it is computationally more expensive. Further, if a broadcast stream is scrambled, the descrambling process is needed to access to the PTS. Moreover, most of digital broadcast streams are scrambled, thus an indexing system cannot access the stream without an authorized descrambler if the stream is scrambled.

From a practical point of view, we propose to use broadcasting time as reference time, which is the simplest and most cost effective way of describing time index within a broadcast stream comparing to the above methods that require the complexity of implementation of DSM-CC NPT in DVB-MHP and computational cost and descrambling problems of PTS. Broadcasting time is carried on the broadcast stream in the form of system time table (STT) of ATSC [6] or time date table (TDT) of DVB [7]. Using broadcasting time as reference time does not require for an indexing system and client PVRs to be connected for synchronization through an interactive communication channel such as Internet. Also, it provides an efficient method to locate same position of the broadcast stream in both side of indexing system and client PVRs since the STTs or TDTs are contained in its temporal position of the broadcast stream according to the broadcasting time. For example, STT of ATSC is repeatedly broadcast once every second.

Fig. 1 shows the overall structure of proposed system composed of an indexing host (real-time indexing system: RTIS), a broadcaster, and a client PVR. A segmentation metadata for a live broadcast program is generated at the indexing host and delivered to the client PVR through the broadcasting network. The detailed descriptions will be shown in the following sections: the section 2 shows the detailed description of methods used in RTIS for the media localization and real-time segmentation, the section 3 presents the implementation of the test-bed and the experimental results, and the section 4 concludes the paper.

2 Media Localization and Real-Time Segmentation

We encounter two problems in implementing the proposed real-time metadata service scheme. One is how to localize the broadcast stream with broadcasting time in both

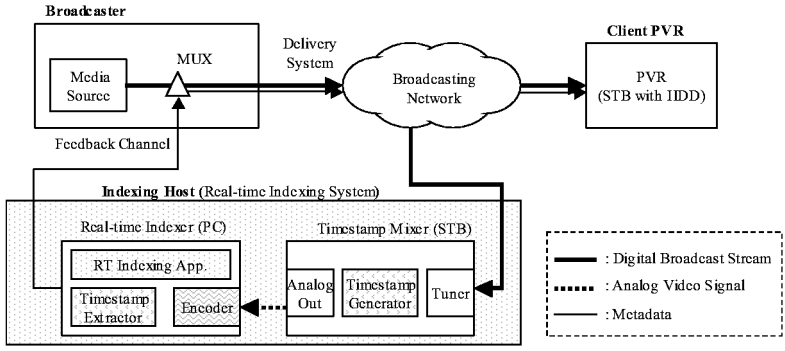


Fig. 1. Overall structure of test-bed for the real-time metadata service scheme

sides of the indexing system and the client PVRs. Another is how to index a live broadcast program in real-time, that is, how to detect shot boundaries (or scene changes) and group the shots into the segments of interest and how to easily verify the detected shot boundaries in real-time. The other problem is how to deliver the segmentation metadata to user's PVR in broadcast stream.

2.1 Media Localization Using Broadcasting Time

To solve the media localization problem in broadcast environments, we use the broadcasting time carried on STT or TDT of the broadcast stream in both sides of the indexing system and the client PVRs due to the convenient features of it as described in above section.

In the indexing system RTIS of Fig. 1, the timestamp mixer is introduced to index a digital broadcast stream with broadcasting time regardless of whether the stream is scrambled or not. The timestamp mixer superimposes the visual timestamp, such as a structured color-code [8], showing the current broadcasting time onto each frame of broadcast stream received through the tuner. The visually time-stamped analog output signals of the timestamp mixer are then encoded in low bit-rate at the real-time indexer. Using the stream encoded in low bit-rate, we can avoid a possible problem of directly accessing scrambled broadcast stream as well as a burden of indexing very high bit-rate stream such as HDTV broadcast stream.

In order to superimpose the timestamp for the current broadcasting time, the timestamp mixer examines broadcasting time carried on the STT or TDT of the received broadcast stream via its tuner.

In case of ATSC, it is recommended that I-frames shall be sent at least once every 0.5 second in order to have acceptable channel-change performance. Further, there exists a delay between the arrival time of a frame and its presentation time due to the VBV delay with maximum delay time of 0.5 second and decoding time delay. Fig. 2 shows an example of indexing and accessing the start position of a segment specified by the broadcasting time *BT* based on the above properties of ATSC.

The broadcasting time *BT* carried on the STT or TDT is represented with a discrete second unit. Thus the frames presented on screen during a discrete second have the

same broadcasting time with which they are time-stamped with the same broadcasting time

When the real-time indexer indexes the re-encoded video resulting from timestamp mixer, it extracts the broadcasting time for each video frame from the timestamp superimposed onto the frame. The extracted broadcasting time represents the current broadcasting time of the frame, at which the frame is presented on screen. For example, in case of frame A in Fig. 2(a), the broadcasting time of frame A is BT_n when the frame is displayed on screen on which the broadcasting time BT_n is time-stamped. Whereas in case of frame B, the broadcasting time of frame B is BT_{n+1} although the frame is arrived at previous time of BT_n, because the frame B is displayed on screen at BT_{n+1} with which the indexing system indexes the frame B.

Let $PTS(\alpha)$ and $PTS(\mathrm{I}_n^1)$ denote the PTS value for the first frame α of a segment S_α presented at the broadcasting time BT_n and for the first I-frame since BT_n, respectively. Then, the time difference $TD(S_\alpha)$ is defined as:

$$TD(S_\alpha) = PTS(\alpha) - PTS(\mathrm{I}_n^1). \tag{1}$$

In Fig. 2(a), the time difference $TD(\mathrm{S_A})$ for the segment $\mathrm{S_A}$ displayed at BT_n has a positive value because the PVR will display the video starting from I_n^1 including the segment $\mathrm{S_A}$ as shown in Fig. 2(c). However, the time difference $TD(\mathrm{S_B})$ for the segment $\mathrm{S_B}$ displayed at BT_{n+1} has a negative value because the client PVR will display the video starting from the first I-frame I_{n+1}^1 since BT_{n+1} which results in missing frame B that is desired to be presented as the first frame of the segment $\mathrm{S_B}$.

Therefore, when we display a segment whose start time is BT_n, we propose that I_{n-1}^1, which precedes I_n^1 with a broadcasting time unit ($BT_n - BT_{n-1}$: one second in case of STT) from BT_n, should be used to avoid missing the first frame of the segment we use.

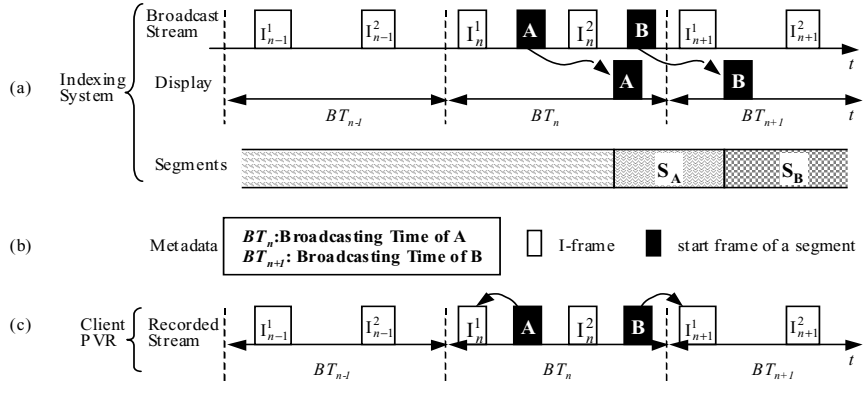


Fig. 2. An example of the media localization: (a) The indexing system indexes broadcast stream with the broadcasting time *BT*. (b) The generated metadata is described the broadcasting time. (c) The client PVR locates the start position of the segment by the broadcasting time.

2.2 Real-Time Segmentation Using Spatiotemporal Visual Pattern

Several approaches [9-11] have recently been proposed for an automatic video indexing by analyzing video, audio and closed caption. However, with the current state of art technology on image understanding and speech recognition, it is still hard to accurately detect highlights and generate a meaningful metadata in real-time.

In order to index a broadcast program in real-time, an operator might have to watch carefully the current broadcast program and manually determine the start and/or end times of events before a broadcast program ends. The event is usually composed of a shot or a set of subsequent shots many of which might be automatically detected by a suitable algorithm with false alarms and missing shots due to editing effects such as zooming in/out, fading, dissolve, and wipe. To get the exact time information of the events, the operator might have to verify the result of automatic algorithm by playing back suspicious segments repeatedly, which will take lots of time. Thus, in order to overcome such problems and quickly index the live broadcast program, we need a new tool for easily verifying shot boundaries.

A spatiotemporal visual pattern called Visual Rhythm [12] also known as spatio-temporal slice [13] provides an efficient way of verifying video segments, which is a two-dimensional abstraction of the entire three-dimensional content of the video.

The most distinguished feature of the visual rhythm is that different video effects including edits and others such as cuts, wipes, zooms and camera motions manifest themselves as different visual patterns on the visual rhythm, as shown in Fig. 3. Due to the features, an operator can find out missing shot boundaries, for example, the wipe in $shot_n$ in the right side of Fig. 3, which might not be detected by the automatic scene change detection. The operator divides manually the $shot_n$ into two shots, $shot_{n1}$ and $shot_{n2}$ so as to determine the segment boundary of $segment_m$ and $segment_{m+1}$.

Therefore, inclusion of the visual rhythm in user interface of the real-time indexing application aids an operator to easily and quickly identify segment boundaries as well as visual rhythm itself might be used as a primitive material for automatic shot detection.

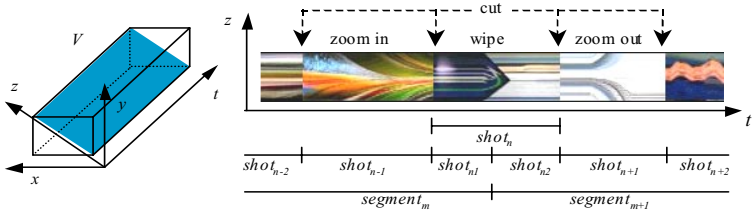


Fig. 3. (a) *VR* extraction from the video *V*. (b) Editing effects presented in *VR*.

2.3 Metadata Delivery

One way to describe segmentation metadata is by utilizing international standards on metadata specification such as MPEG-7 or TV-Anytime. The MPEG-7 or TV-Anytime metadata can be multiplexed into MPEG-2 transport stream that is broadcast to clients through broadcasting network. There might be several solutions of

delivering the standard metadata to clients through broadcast stream: defining a new MPEG-2 private section or descriptor, using the DSM-CC sections, or specifying new type of MPEG-2 PES.

These approaches have two inherent problems. First, the segmentation metadata generated based on the metadata standards are often large in size and thus occupies non-negligible amount of bandwidth for data broadcasting that the current DTV service providers want to minimize. Second, it will take much time for the approaches to be realized because they will require many changes or adoption of existing or new software and hardware components in existing broadcasting environment.

Therefore, a new technique is needed to deliver the segmentation metadata that is smaller in size compared to segmentation metadata based on MPEG-7 and TV-Anytime, through the existing broadcasting environment.

In the proposed system, instead of defining new field for the segmentation metadata, we adopt the existing EPG (Electronic Program Guide) as a carrier of the segmentation metadata because it could be used without any modification of broadcast equipments. That is, we utilize the field for detailed description (synopsis) of a program in EPG data structure. Since the detailed description of a program is presented in the viewer's screen, we have designed new compact metadata format to be legible and informative for viewers who do not have metadata browsing modules only ported on our test-bed client PVR. In table 1, the syntax of the segmentation metadata is represented according to BNF (Bacchus Naur Form) grammar, and one example used in our test-bed is given. The size of the example metadata in table 1 is only 239 bytes whereas the TV-Anytime format for the metadata requires more than 5K bytes for same segmentation information. Due to the small size, we can carry it on the detailed description of a program in EPG which is practically restricted in size of 250 bytes in our test-bed.

Table 1. BNF grammar for the our segmentation metadata format and the example.

BNF grammar for metadata format	Example
<segment_info> ::= <title> <segments> <title> ::= <string> LF <segments> ::= | <segments> ::= <segment_locator> SP [<segment_title>] LF <media_locator> ::= <2digit> ':' <2digit> ':' <2digit> <segment_title> ::= <hierachical_sequence> SP <string> <hierachical_sequence> ::= '<' <sequence_number> '>' <sequence_number> ::= DIGIT | DIGIT '.' <sequence_number> <string> ::= CHAR | CHAR <string> <2digit > ::= DIGIT DIGIT	Survival English SEP 4 06:20:41 <1> Introduction 06:22:49 <2> Today's Dialog 06:23:19 <3> Dialog Part I 06:27:04 <4> Dialog Part II 06:31:00 <5> Dialog Part III 06:33:08 <6> More Expressions 06:34:25 <7> Review Dialog 06:35:48 <8> Help Me

3 Implementation and Experimental Results

Real experiments with ATSC terrestrial HDTV programs are performed by porting our software into a commercially available PVR. The scenario we have implemented is as follows. Firstly, we index a broadcast program in real-time and immediately send the resulting metadata to a broadcast station.

Secondly, the delivered metadata of the program is inserted into the field for synopsis of the program in EPG that is transmitted to client PVRs through the broadcasting network.

Finally, the client PVR detects the EPG update and retrieves the metadata of the program in the delivered EPG. The client PVR then locates a segment of interest from the recorded stream according to the broadcasting time described in the delivered metadata. Thus, the client PVR user can browse the recorded program through functionalities such as segment play/replay and random access to the segment of interest.

Fig. 4. The timestamp and the real-time indexer using spatiotemporal visual pattern.

The RTIS is composed of a real-time indexer (personal computer) equipped with an encoder for low bit-rate encoding, and a timestamp mixer that is a STB including timestamp generator. We implemented the timestamp mixer by programming the timestamp generator module and then porting it onto the commercially available PVR. Fig. 4 shows the example of the timestamp represented with structured color-code [8] superimposed onto the frame, and the screen shot of indexing application that indexes broadcast program in real-time using the visual timeline called the visual rhythm shown in the top of the application.

For the client PVR in our test-bed, we have utilized a commercially available PVR that is a HDTV STB with a 40G Bytes of HDD, on which we developed our applications. One application is responsible for retrieving the metadata contained in the EPG: checking the EPG update, extracting the metadata of a recorded program from the EPG, and storing the metadata onto the storage. The other application is related with browsing the recorded program with the retrieved metadata: locating a video segment of interest, extracting key frames (thumbnail images) which are used for user interface for browsing window, and managing graphic user interface.

For the experiments, we indexed an educational program that was broadcast at 6:20 AM in Korea. We indexed the program while it was being broadcast using the real-time indexer as shown in the right side of Fig. 4. The indexing process was finished at a minute after ending time of the program. We manually verified the segmentation results, and then generated the metadata such as shown in Table 1. Immediately after generating the segmentation metadata, we sent the metadata through email to an operator who is responsible for updating EPG of a broadcaster. The operator then updated the detailed description (synopsis) of the program using EPG builder with the received metadata. It took some minutes because the operator had to check his email and copy the metadata script and then paste it on the input field of the detailed description of the program manually. Finally, after applying the EPG update in the broadcaster, the metadata was transmitted or broadcast through the broadcast network

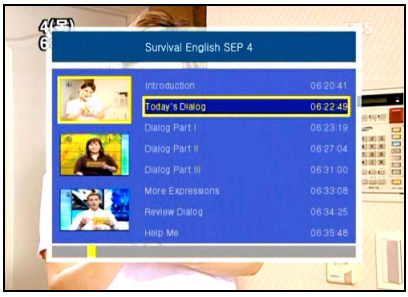
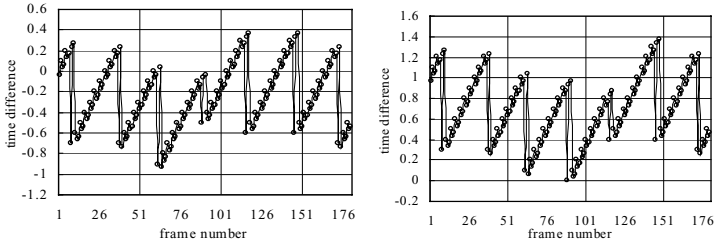


Fig. 5. The graphic user interface of the browser in PVR client.



(a) (b)

Fig. 6. (a) The time difference by (1). (b) The time difference of proposed method.

and finally received by the client PVR that eventually extracted the metadata. In the experiment, it took about 5 minuets from the ending time of the program to the time of receiving the metadata on the client PVR. This time delay is mainly due to the manual works for sending an email and updating EPG. If we have an interactive channel between the EPG builder and our real-time indexer and the update of EPG can be controlled by software, we could reduce most of the time delay. Thus, PVR users can browse a recorded program with corresponding metadata just after the recording is finished.

Fig. 5 shows the resulting TV screen displayed in PVR when we browse the recorded program with the delivered metadata. The key frame shown in the left of the screen is the image extracted from the recorded stream in PVR by using the broadcasting time described in the delivered metadata.

In our experiment, we observed that the first part of a segment was often missed. To see how much time difference was occurred, we measured the time difference (1) with broadcast stream as shown in Fig. 6(a). Negative values of the time differences in Fig. 6(a) indicate that the video was started playing after the absolute time difference from the desired starting time position of the segment. On the other hand, the time difference of the proposed scheme has no negative value as shown in Fig. 6(b) since we subtracted one second (based on ATSC STT) from the broadcasting time described in the metadata to avoid missing frames when we implemented the browsing module onto the PVR.

4 Conclusion

We have introduced a real-time metadata service scheme and implemented a test-bed having an indexing host, a broadcaster, and a client PVR. For the service scheme, we have proposed a novel method of indexing the broadcasting program in real-time, which is to utilize broadcasting time that is carried on the broadcast stream itself. From the experiments, we could show that the method could be applied to the current digital broadcast environments without changing any software and hardware components. Moreover, it was very beneficial demonstration for digital broadcasting, in the point of real-time metadata service for live broadcast program.

Acknowledgments. We would like to appreciate Educational Broadcasting System in Korea and Samsung Electronics Co., LTD for allowing us to use their equipments.

References

1. ISO/IEC 15938-5 Int. Standard Information Technology – Multimedia content description interface – Part 5 Multimedia Description Schemes. ISO/IEC JTC1/SC29/WG11 (2002)
2. TV-Anytime Forum SP003v13 Metadata Specification Version 1.3. TV-Anytime Forum Specification Series: S-3 (Normative). TV-Anytime Forum (2003)
3. ISO/IEC 13818-6 Int. Standard Information Technology – Generic coding of moving pictures and associated audio information: Digital Storage Media Command and Control. ISO/IEC JTC1/SC29/WG11 (1998)
4. ETSI/EBU TS 102 812 V.1.1.1 Digital Video Broadcasting (DVB): Multimedia Home Platform (MHP) Specification 1.1. ETSI. (b2001)
5. A. McPrland, J. Morris, M. Leban, S. Rarnall, A. Hickman, A. Ashley, M. Haataja, F. deJong.: MyTV: A practical implementation of TV-Anytime on DVB and the Internet. International Broadcasting Convention. (2001)
6. ATSC Standard A/65B Program and system information protocol for terrestrial broadcast and cable (Revision B). Advanced Television Systems Committee. (2003)
7. ETSI/EBU EN 300 468 V1.4.1 Digital Video Broadcasting (DVB): Specification for Service Information (SI) in DVB Systems. European Telecommunications Standards Institute (2000)
8. J.-C. Yoon, H. Kim, S. Oh, and S. Sull.: Design of Color-Code System for Time-Stamping Broadcast Video. IEEE Trans. on Consumer Electronics, Vol. 49, No. 3. (2003) 750-758
9. B.T. Truong, S. Venkatesh, and C. Dorai.: Scene Extraction in Motion Pictures. IEEE Trans. on Circuit and Systems for Video Technology, Vol. 13, No. 1. (2003) 5-15
10. S.-C. Chen, M.-L. Shyu, W. Liao, C. Zhang.: Scene change detection by audio and video clues. Proceedings of IEEE ICME 2002, Vol. 2. (2002) 365-368
11. N. Babaguchi, Y. Kawai, and T. Kitahashi.: Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration. IEEE Trans. on Multimedia, Vol. 4, No. 1. (2002) 68-75
12. H. Kim, J. Lee, J. Yang, S. Sull, W. Kim and S. M. Song.: Visual rhythm and shot verification. Multimedia Tools and Applications, vol. 15. (2001) 227-245
13. C.-W. Ngo; T.-C. Pong; H.-J. Zhang.: Motion analysis and segmentation through spatio-temporal slices processing. IEEE Trans. Image Processing, Vol. 12, Issue 3. (2003) 341-355