Scalable Hierarchical Summarization of News Using Fidelity in MPEG-7 Description Scheme

Jung-Rim Kim, Seong Soo Chun, Seok-jin Oh, and Sanghoon Sull

School of Electrical Engineering, Korea University,
1 Anam-dong 5ga Songbuk-gu, Seoul, Korea
{jrkim,sschun,osj,sull}@mpeg.korea.ac.kr

Abstract. The notion of fidelity is an attribute in MPEG-7 FDIS (Final Draft International Standard [1]) that can be used for scalable hierarchical summarization and search [2]. The fidelity is the information on how well a parent key frame represents its child key frames in a tree-structured key frame hierarchy [1-5]. The use of fidelity was demonstrated for scalable hierarchical summarization [2] based on the low-level features such as color, but the temporal information was not used. Content of a video such as news and golf is temporally well structured and it is desirable to utilize such information. In this paper, we demonstrate the use of fidelity for the summarization of a well-structured news by using temporal information as well as low-level features.

1 Introduction

Nowadays, the speed of network grows up and its bandwidth becomes wide. So there are much more chances of making access to multimedia data. Although improvements are being made on the Internet, the size of multimedia data is too large to deliver complete data. Because of this problem, the study on multimedia compression, transferring and indexing has become important. Also, since the amount of multimedia data increases fast, it is necessary that we should be able to search and access/navigate them easily.

Study on content-based retrieval has been extensively researched through various indexing schemes, but the research on multimedia access is still insufficient and being developed. One of the useful methods for access and navigation of multimedia content is summarization.

The summarization helps us understand the whole content of a video by showing a set of the key frames/clips representing the whole video. This functionality is especially useful since the size of a video is very large in general and thus users might not want to spend much time watching the whole video. Furthermore it might be difficult to deliver the whole video with limited bandwidth. Therefore, there is a need for the scalable summarization schemes and MPEG-7 MDS (Multimedia Description Scheme) has been developed to provide such schemes.

S.-K. Chang, Z. Chen, and S.-Y. Lee (Eds.): VISUAL 2002, LNCS 2314, pp. 239–246, 2002.
© Springer-Verlag Berlin Heidelberg 2002

Among a variety of video contents, news content is typically structured by time or by topics and thus can be hierarchically well described by the scalable summarization scheme for easy browsing and summary. The scalable description allows users to select the parts of the news video depending upon their preference or available bandwidth. In this paper, we describe the notion of fidelity in MPEG-7 Summarization Description Scheme (DS) and propose an efficient method for the scalable hierarchical news video summarization based the MPEG-7 DS.

This paper is organized as follows. Section 2 introduces related works and the notion of fidelity. Section 3 describes the appropriate notion of fidelity and algorithm for news summarization, and Section 4 demonstrates the experimental results for scalable summarization of news. Finally, Section 5 provides conclusions of the paper.

2 Related Work and Fidelity

2.1 Related Work

Recently, there have been several approaches for the video summarization [6-9]. D. DeMenthon *et al.* [6] proposed scalable summarization using curve simplification. They developed a method for summarizing video by splitting a trajectory curve in the high dimensional feature space for the key frames. Y. Gong *et al.* [7] proposed optimal video summarization algorithm using the singular value decomposition of the feature vectors. S. Uchihashi *et al.* [8] introduced a video summarizing scheme using shot importance. As the shot length is longer, the importance of each shot is assumed to be larger, and as it becomes shorter, its importance diminishes. Mark T. Maybury *et al.* [9] showed summarization of broadcast news using audio, visual information and closed-captioned text. They summarized news video by key frame selection through the audio and video correlation, and annotated the summarized frames using closed-captioned text.

MPEG-7 also provides video summarization schemes in MDS. The multimedia content description in MPEG-7 is divided into two parts. One part is about the description of the structural aspects of the content that describes the audio-visual content from the viewpoint of its structure. It represents the spatial, temporal or spatio-temporal structure of the audio-visual content and can be described on the basis of perceptual features using MPEG-7 Descriptors for color, texture, shape, motion, audio features, and semantic information using Textual Annotations. And the other is about the description of the content conceptual aspects that describes the audio-visual content from the viewpoint of real-world semantics and conceptual notions. It involves entities such as objects, events, abstract concepts and relationships. Based on such descriptions, MPEG-7 gives description schemes for navigation and access of multimedia content that facilitates browsing and retrieval of audio-visual content by defining summaries, partitions and decomposition, and variations of audio-visual material.

A brief explanation about fidelity in MPEG-7 description schemes is given in the following section to be used for scalable hierarchical summarization.

2.2 Fidelity

The fidelity is the information on how well a parent key frame represents its child key frames in a tree-structured key frame hierarchy [1-5]. We can construct the tree-structured hierarchy using the key frame shown as Fig. 1 based on the relationship between a parent key frame and its children using fidelity. The definition of fidelity e_α of a node α having the parent node p_α is proposed in [2] as

$$e_\alpha = 1 - \max_{x \in T_\alpha}(d(p_\alpha, x)), \tag{1}$$

where $d(\ )$ denotes normalized distance/dissimilarity from 0 to 1 and T_α is the hierarchy rooted at the node α.

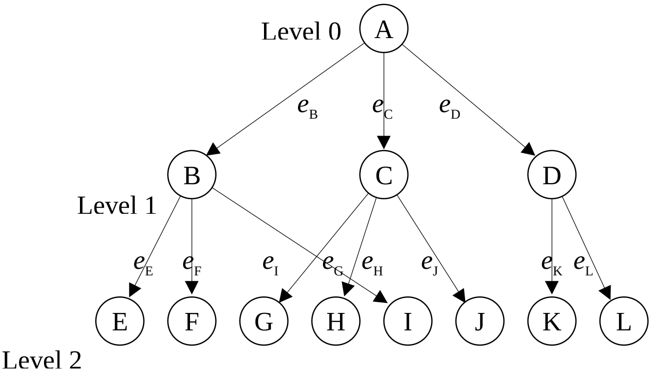

Fig. 1. An Example of the Key Frame Hierarchy with Fidelity

3 Scalable Hierarchical Summarization of News Using Fidelity

In this section, we describe the use of fidelity using both the low-level feature and the temporal information of news to construct key frame hierarchy.

3.1 Scalable Hierarchical Algorithm Using Fidelity

The scalable summarization algorithm using fidelity is a *max-cut* finding algorithm proposed in [2], [5]. This algorithm maximizes the minimum edge cost cut by cut-line in the hierarchy so the fidelity after splitting the hierarchy becomes maximal. It can be summarized as follows:

The root node is inserted into the summarization set *K* at first, and a node β not in *K* with the minimum fidelity between a node α in *K* and itself is inserted into *K*. Until the number of elements in *K* becomes equal to that specified by a user, the inner loop of the algorithm is repeatedly performed.

```
add root_node to K;
while ( card(K) < N ) {
        let <$\alpha$,$\beta$> be a least cost edge
            such that $\alpha \in$ K and $\beta \notin$ K;
        add $\beta$ to K;
}
```

Fig. 2. Scalable Hierarchical Summarization Algorithm

For example, suppose $e_B < e_C < e_D$ in Fig. 1. By the max-cut algorithm in Fig. 2, we can choose two nodes that represent the whole hierarchy with maximal fidelity. At first, the root node, A, is selected, and then we have chance to choose one of three nodes B, C, and D. By the algorithm, we should select node B because of the above condition $e_B < e_C < e_D$. Then, we obtain two hierarchies rooted at A and B, maximizing the fidelity of the hierarchies.

3.2 Hierarchy for News Summarization

Sull *et al.* proposed a key frame hierarchy with fidelity using only low-level features of the key frames based on equation (1) [2]. However, since such low-level features cannot represent the conceptual aspect of contents well, the result sometimes does not represent semantically meaningful summarization.

Structured news content often contains the degree of importance related to time: The news content is structured as two major different parts composed of an anchor shot where one or two anchors reports events, and the event shots between two successive anchor shots. Furthermore the headline news shown at the beginning is important relative to the upcoming news stories shown thereafter. In general, the importance of news decreases as it is approaching to the end of the news. Taking both anchor shots and their time into account, we can show the contents or information of the news more effectively. Instead of equation (1), we propose a fidelity e_α' at a node α as follows:

$$e_\alpha\text{'} = we_\alpha + (1-w)e_\alpha(t), \tag{2}$$

where w is the weight for the fidelity based on low-level feature in the range from 0 to 1 and $e_\alpha(t)$ is the *temporal fidelity* at a node α given by time position/code or by temporal distribution between a parent frame and its descendents as

$$e_\alpha(t) = \frac{\alpha(t)}{\tau}, \tag{3}$$

where $\alpha(t)$ is time code of α and τ is the total time length of the video content. The $e_\alpha(t)$ increases as the temporal position of α increases, allowing us to obtain temporally earlier nodes first when applying the summarization algorithm shown in Fig. 2.

Also in [2] they did not consider the temporal order for clustering, and thus it ignored the temporal relationship between a parent key frame and its child key frames.

For example, the key frames in the bottom level consist of 3 shots such as {E, F}, {G, H, I, J}, and {K, L}, and their parent key frames are B, C, and D, respectively shown in Fig. 1. The key frame B represents I though there is little temporal relationship between them. If this scheme is applied to news content, the anchor shots whose low-level features are almost same will be classified into the same cluster. For an effective summarization, it is desirable to initially extract the anchor shots and the event shots between two successive anchor shots, and then construct the key frame hierarchy.

The key frame hierarchy is typically constructed by a 4-level bottom-up method. The bottom level, level 3, consists of key frames, level 2 consists of the anchor frames and the key event frames representing the event frames between two successive anchor frames. The key event frames are positioned to the level 1 using only temporal fidelity, and the key frame that represents whole video becomes root. The overall algorithm is shown as Fig 3, and Fig. 4 is an example hierarchy based on this algorithm.

```
1. Detect shots.
2. Extract key frames in each shot using the low-
level feature vector (level 3).
3. Separate the anchor shots and the event shots.
4. Cluster the successive anchor and event shots re-
spectively using the algorithm proposed in [5]
(level 2).
5. Extract all event frames from level 2 (level 1).
6. Set the key frame, for example title frame, to
the root key frame of the hierarchy (level 0).
```

Fig. 3. An algorithm for News Summarization

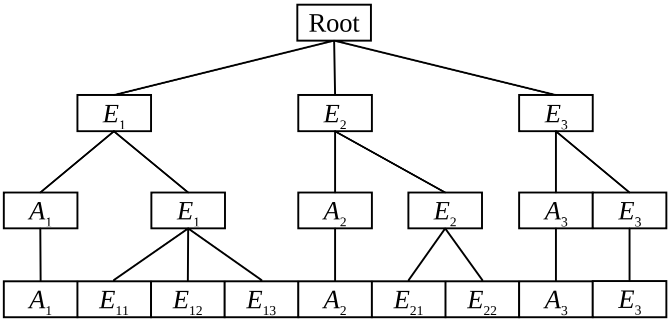

Fig. 4. An Example of the Key Frame Hierarchy of News (A_n : Anchor shot, E_n : Event shot)

4 Experimental Results

In this section, we describe the experimental result of the key frame hierarchy and the summarization result of news using our proposed algorithm.

4.1 Key Frame Hierarchy

In our current implementation, we use the DC luminance projection introduced in [10] as low-level feature vector for key frame extraction, anchor shot detection, and clustering. The luminance projection (l_n^r, l_m^c) of *n*th row and *m*th column in *M*x*N* DC image *f* is respectively

$$l_n^r(f) = \sum_{m=1}^{M} Lum\{f(m,n)\},$$
$$l_m^c(f) = \sum_{n=1}^{N} Lum\{f(m,n)\}. \tag{4}$$

The distance/dissimilarity function, normalized to [0, 1], is also defined as

$$d(f_i, f_j) = \frac{1}{K}\left(\sum_{n=1}^{N} | l_n^r(f_i) - l_n^r(f_j) | + \sum_{m=1}^{M} | l_m^c(f_i) - l_m^c(f_j) |\right), \tag{5}$$

where *K* is a normalizing constant.

Using above feature vector and distance/dissimilarity function, we detect shot boundaries and extract key frame set *R* satisfying the following condition:

$$R = \{ f_i \in S \mid d(f_i, f_{i-1}) \le \varepsilon_k, i = 1,2,3...\}, \tag{6}$$

where S is the whole video frame set, and e_k is distortion to extract shot boundary. We also apply the equation (7) to detect a set A of anchor frames f_k in *R*:

$$A = \{ f_k \in R \mid d(f_a, f_k) \le \varepsilon_a \}, \tag{7}$$

where f_a is the reference anchor frame that a user selects, and ε_α is distortion to detect anchor frames.

Since the fidelity based on low-level feature for each anchor shot is almost 1, it is meaningless. So we applied only the temporal fidelity, i.e. *w*=0, to the fidelity of the anchor shot, and we experimentally set *w* to 0.2 or smaller value in equation (3) to construct a hierarchy for the event frames, such that the low-level feature does not affect the temporal order too much. The experimental results applied to two videos are shown as Table 1.

4.2 Summarization

Using the algorithms shown in Fig. 2 [2], we summarize two news videos by using 9 and 18 frames. Figure 5 (a) and (b) are the experimental results from the key frame hierarchy using only low-level feature, and Fig. 5 (c) and (d) show the results using low-level feature as well as temporal information. As shown in Fig. 5 (a) and (b), the 9-frame summarization and 18-frame summarization does not show a good semantic relationship between them, but in Fig. 5 (c) and (d), the 9-frame summarization gives the storyboard consisting of events purely, and in the 18-frame summarization result the anchor frames well appear between almost every event frame.

Table 1. Video Contents and Their Frames in each Level

Video	Length	Level 0	Level 1	Level 2	Level 3
News1	27m 38s	1	10	19	210
News2	27m 36s	1	9	18	138
News3	27m 39s	1	10	19	154

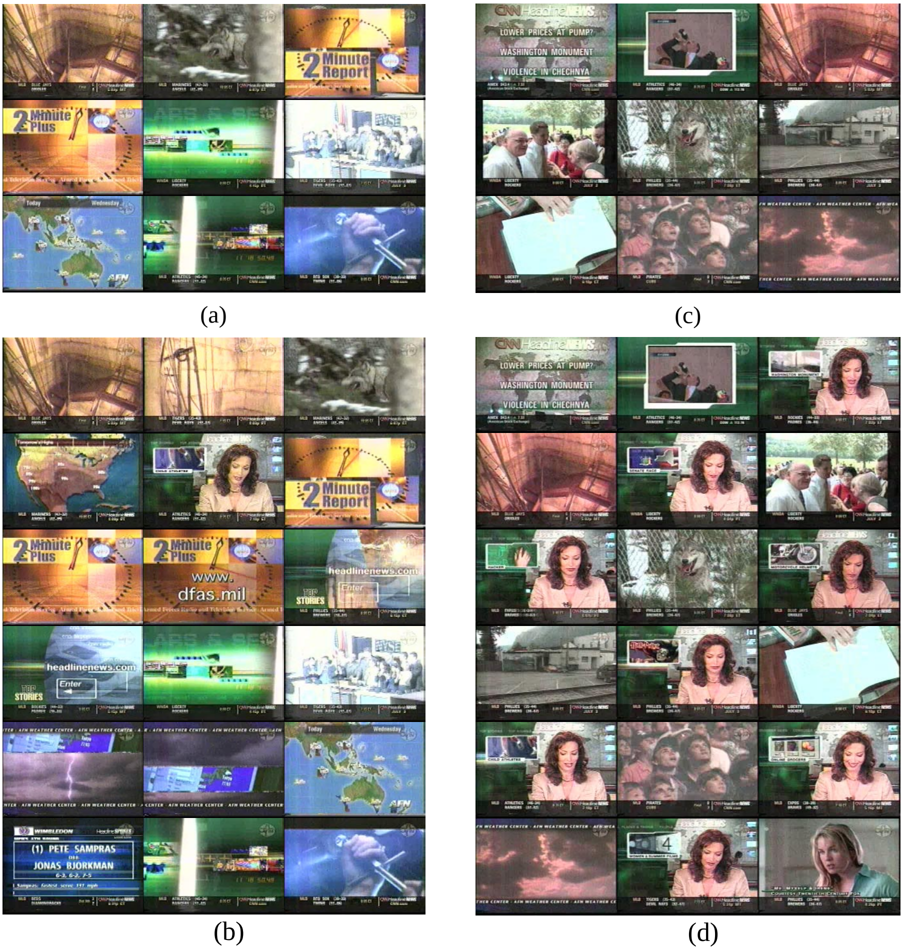

Fig. 5 Summarization results from the key frame hierarchy using fidelity. (a) 9-frame summarization of News1 based on low-level feature (b) 18-frame summarization of News1 based on low-level feature (c) 9-frame summarization of News1 based on low-level feature and temporal information (d) 18-frame summarization of News1 based on low-level feature and temporal information

5 Conclusion

In this paper, we described the use of fidelity in the MPEG-7 MDS for scalable hierarchical summarization of news. Based on the fidelity based on both low-level feature and temporal information, we constructed the semantically meaningful key frame hierarchy consisting of anchor and event frames, demonstrating a feasibility of our approach.

References

1. ISO/IEC 15938-5 FDIS Information Technology -- Multimedia Content Description Interface - Part 5 Multimedia Description Schemes. ISO/IEC JTC1/SC29/WG11 N4206 (2001)
2. Sull, S., Kim, J.-R., Kim, Y., Chang, H.S., Lee, S.U.: Scalable hierarchical video summary and search. Proceeding of SPIE2001, Vol. 4315. Storage and Retrieval for Media Database 2001, San Jose (2001) 553-561
3. Overview of the MPEG-7 standard. ISO/IEC JTC1/SC29/WG11 N4031, Singapore (2001).
4. Efficient and effective search and browsing using fidelity. ISO/IECC/JTCI SC29/WG11 M5101, La Baule (1999)
5. Improved notion of the fidelity for efficient browsing. ISO/IEC JTC1/SC29/SG11 M5442, Maui (1999)
6. DeMenthon, D., Kobla, V., Doermann, D.: Video summarization by curve simplification. Proceedings of ACM International Conference on Multimedia (1998) 211-218
7. Gong, Y. and Liu, X: Generating optimal video summaries. Proceedings of IEEE International Conference on Multimedia and Expo 2000, Vol. 3. (2000) 1559-1562
8. Uchihash, S., Foote, J.: Summarizing video using a shot importance measure and a frame-packing algorithm. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 6. (1999) 3041-3044
9. Maybury, M. T., Merlino, A. E.: Multimedia summaries of broadcast news: Proceedings of Intelligent Information Systems (1997) 442-449
10. Chang, H. S., Sull, S., Lee, S. U.: Efficient video indexing scheme for content-based retrieval. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8. (1999) 1269-1279