Rendition-Based Video Editing for Public Contents Authoring

Atsuo Yoshitaka[*1] and Yoshiki Deguchi[*2]

[*1]School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Nomi, Asahidai, Ishikawa, 923-1292 Japan

[*2]Graduate School of Engineering, Hiroshima Univ.
1-4-1 Kagamiyama, Higashi-Hiroshima, Hiroshima
739-8527 Japan
**Currently with Panasonic Corp.

Abstract

Previous video editing is performed by specifying cutting points in a sequence of video frames and combining the trimmed shots together. In order to emphasize a nonverbal, emotional information of a scene effectively, we need to edit a video sequence so that it follows 'film grammar' or cinematography, which is the technique for effective expression of the scene in a film. In this paper, we propose a new framework of rendition-based video editing. Instead of specifying cutting points of a video sequence, an editor specifies the type of emotional expression that he/she likes to emphasize. Consequently, desirable transition of shot duration and shot combination are determined, and video contents are semi-automatically edited so as to follow the film grammar.

1. Introduction

General way of video editing is performed by selecting one of the candidate video data and specifying cut-in and out which correspond to start and end of the shot included in the edited video data. After trimming each of the video data, they are combined together to produce edited contents.

Especially in producing movies or TV dramas, cursory concatenation of video shots may not depict a proper atmosphere to viewers, because not only the visual content itself but also camera operation and cut planning affect viewers' impression for a scene. The 'film grammar'[1] is a set of techniques for effective expression of atmosphere, which corresponds to emotional, nonverbal information of a scene. This kind of information is emphasized by camera work and cut planning for shot length transition.

Recently, increasing number of people use digital video cameras and personal computers for editing home video. Computer-aided video editing system enabled us to perform non-linear editing, however, it is still frame-based editing. Editors need to specify cutting points in order to trim source video and concatenate shots. During this editing process, he/she needs to take the film grammar into account for the better expression of a scene especially in the sense of emotional information.

In this paper, we propose a new framework of video editing, which we call *rendition-based editing*. In the rendition-based editing, source video data is first placed in temporal order on the screen. Two or more pieces of video data may be placed in parallel in case of editing sources shot by the multiple cameras simultaneously. Unlike frame-based video editing, rendition-based editing specifies semantic content or emotional information, such as the atmosphere to emphasize, instead of specifying cutting points for source video data. After specifying such information for each scene, selection from candidate source video data and cut planning are carried out so that the resultant video data follows film grammar to depict semantic contents.

Film grammar is utilized for content-based retrieval for emotional information or scene boundary detection. Content-based retrieval is discussed in [2], where semantic content or emotional information is detected by evaluating the transition of shot length and alternation of similar shots. Film grammar is also utilized for scene boundary detection [3-4].

There are several studies on video editing. One direction is improving the functionality of editing video contents, and another is aiding editing operation itself. As an example of the former, structuring video contents by XML is discussed in [5]. Another example is in [6]. Zodiac[6] is a video editing system which improves video editing operation based on edit history manipulation. This also provides video content analysis such as cut detection, however, it is aimed at facilitating video browsing. As for the examples of the latter, there are only a few studies. One of the examples is a system named *Hitchcock*[7]. It detects unsuitable shots for viewing such as a shot with fast camera motion. The system guides an editor to avoid choosing such improper shots in order to improve the quality of resultant video, however, even this system does not take film grammar into account for aiding to depict nonverbal, emotional information.

As far as we know, only our system allows a user to edit video data by specifying semantic content or nonverbal emotional information. Primary operation in editing is to specify not cut-in/out of a shot but semantic content or emotional information that the editor wishes to depict. Corresponding to the specification, the system determines which source video to chose and plans where to cut the source video for controlling shot length transition so that the resultant video contents emphasize semantic content or emotional information. Since this process of editing by the system follows the film grammar, the proposing

system reduces the amount of task for experienced editors, and helps general users to produce video contents with sophisticated editing technique without depthful knowledge of video editing theory.

The organization of the paper is as follows: film grammar that we apply for video editing and feature extraction from source video data is described in section 2. After that, we discuss the framework of rendition-based video editing and its implementation in section 3. Experimental results for evaluating the usability and efficiency of the proposed system are shown in section 4. Concluding remarks are lastly described in section 5.

2. Film Grammar and Feature Extraction

In this section, we describe film grammar we apply for rendition-based video editing. In addition, we also describe motion feature extraction from source video data, which is utilized for the selection of source video and the determination of cutting points in editing process.

2.1 Film Grammar

Film grammar which is described in [1] widely ranges over various techniques related to camera work and cut planning. Part of the film grammar, which relates to techniques for emphasizing or effective expression of emotional atmosphere, refers to cut planning (i.e., shot length transition) and the alternation of similar shots.

More concretely, we take the following film grammar into account in rendition-based editing.

(i) Action scene
- consisting of a sequence of short-length shots
- each shot is visually dynamic (more object motion)

(ii) Tension rising scene
- the length of shots is gradually shortened

(iii) Conversation scene
- alternation of shots shooting one person and another
- shot switching based on speaker in dialogue.

(iv) Calm scene
- consisting of a sequence of long-length shots
- each shot is visually static (less object motion)

(v) Tension releasing scene
- the length of shot becomes gradually longer

2.2 Feature Extraction

Visual feature is extracted from raw video data prior to rendition-based editing. Visual motion of video data is one of the factors of film grammar in case of action and calm scene. In this section, we explain the process of measuring visual motion of a shot based on spatiotemporal projection or video tomography[8].

The idea of spatiotemporal projection is illustrated in Fig. 1. Spatiotemporal projection of a video shot is obtained as a vertical slice where $x=a$ and a horizontal slice where $y=b$, respectively. In the proposing system, spatiotemporal projection is obtained for gray-scaled video frames which have a frame size of 160x120 pixels. For better stability of the extraction, the

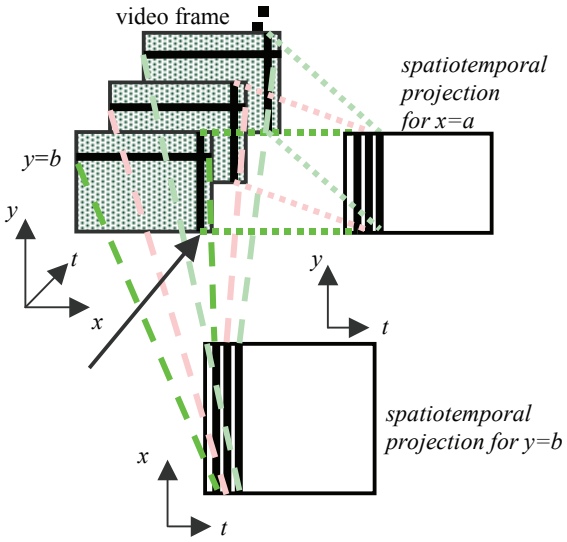


Fig. 1. Spatiotemporal Projection of Video Data

vertical and horizontal slice of the spatiotemporal projection are obtained as the averaged luminance among x=1, 79, 158, and y=1, 59, 118, respectively.

Since the movement of object is dynamic in an action scene, the luminance of the pixel at a certain point may differ frame by frame. It results in intermittent edges in an image obtained by spatiotemporal projection. Therefore, the degree of visual motion of a shot in an action scene is evaluated based on the density of edges, $D(s_k)$, which is defined as follows.

$$D(s_k)=Ev(s_k)/L(s_k)$$

$Ev(sk)$ denotes the number of edges appeared in shot s_k , and $L(s_k)$ denotes the length of the shot s_k. Here, an edge which has length less than $l_h/12$ is regarded as a *noise* that does not correspond to the motion of an object. The l_h corresponds to vertical/horizontal size of a video frame.

On the contrary, the motion of object in static scenes results in the edges being longer and flatter in the image of spatiotemporal projection. Based on this tendency, we evaluate the degree of flatness of edges in an image of spatiotemporal projection in order to evaluate static shots. During the process of tracing an edge, we apply a constraint for the order of tracing as shown in Fig. 2(a). As shown in the figure, one lost pixel, i.e., intermittent edge, is allowed as far as another pixel is found at one of the position “2”, “4”, “6”, “8”, or “10”. Until no pixel is found at any of 10 positions or the tracing reaches to the end of a

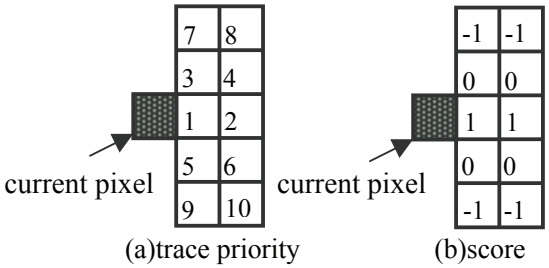


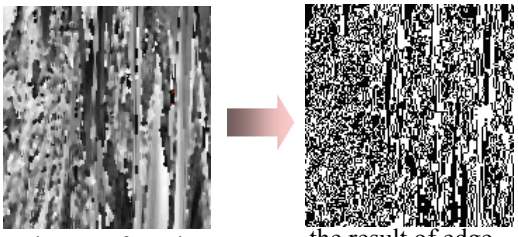
Fig. 2. Flat edge tracing

shot, scores shown in Fig. 2(b) are accumulated. Here, we denote the total score of edge tracing for a shot as $Sum(s_k)$, and the number of traced pixels as $N(s_k)$. We define the flatness, i.e., the

$$Vc(s_k) = (Sum(s_k) / N(s_k) + 1) / 2$$

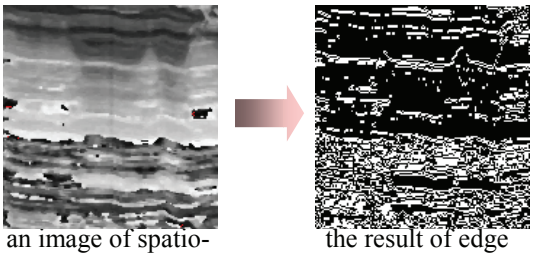
static degree of a shot s_k, $Vc(s_k)$, as follows.

When an edge is flat, score to be accumulated is always 1. Therefore, $Vc(s_k)$ takes maximum value of 1. In case where an edge is continuously slanted in the direction of "7" or "9" as illustrated in Fig. 2(a), $Vc(s_k)$ takes minimum value of 0. Fig. 3 shows an example of edge detection for a motion scene and a static scene.



an image of spatio-temporal projection — the result of edge detection

(a) motion scene



an image of spatio-temporal projection — the result of edge detection

(b) static scene

Fig. 3 examples of edge detection

3. Editing Decision Rules

In this section, we describe a set of editing decision rules which is referred to in rendition-based video editing. When a user specifies a semantic content or emotional atmosphere of the scene, the proposed system selects one of the candidate video data based on visual motion or voice detection. Then it is trimmed and included in the resultant video, where the length of shots is tailored so that it follows the film grammar. The ranges of parameters below are determined based on the statistical values in commercial movies we analyzed.

3.1 Action Scene

When a user specifies to render "action scene" in editing, the system selects the busiest shot if there are two or more pieces of candidate video data placed in parallel on the user interface (see Fig. 4). Here, the shot is selected based on the value of $D(s_k)$. The length of the shot included in the resultant video is shortened if it exceeds maximum length of a shot in action scenes. Default limit of shot length is set to 60 frames (2 sec.), however, this may be adjusted to the arbitrary length between 30 to 90 frames (i.e., 1 to 3 sec.).

3.2 Tension rising scene

When a user specifies to render a sequence of shots as 'tension rising scene', the length of shot is gradually shortened based on the film grammar.

Here, Lo_{si} and Lt_{si} are the length of the original shot s_i and that of the resultant shot trimmed from the s_i ($i=1,..,n_s$), where the scene of tensed atmosphere consists of a sequence of source shot $s_1,..,s_{ns}$. The ratio of shot length decrement factor, r, is specified by an editor between 30 to 50, but it is forced to satisfy $Lt_{si}<Lo_{si}$ for $i=1,..,n_s$. The length of shot is gradually decreased by the ratio r, where the initial length of the shot is determined as bt_{init} by the editor. Lt_{si} is determined so that it follows the constraint below.

$Lt_{si}=bt_{init}-r/(i-1)$, $i>1$

3.3 Conversation scene

In rending a conversation scene, most of the shots are chosen so that each of them is accompanied by a salient voice. Therefore, it is mandatory to detect whether a shot contains speakers or not. The system detects a section where a person is speaking by ZCR (zero-cross ratio)-based voice detection. In rending a scene specified as 'conversation', a shot containing salient voice is automatically chosen from two or more source video data placed in parallel, and switching one data to another is performed in accordance with the alternation of speakers.

3.4 Calm scene

A calm scene is rendered so that long shots are concatenated, each of which is visually static. In case where two or more candidates of video data are specified in parallel, shots which take higher value of $Vc(s_k)$ among the candidates are chosen in turn so as to be included in the resultant video. In editing a calm scene, no shot is trimmed since trimming of a shot may diminish the atmosphere of calmness. Even if short shots are chosen as parts of a resultant video data based on the evaluation of $Vc(s_k)$, though they may not be appropriate for the calm scene from the viewpoint of the length, they are also included in the resultant video data and they may be eliminated manually, if necessary.

3.5 Tension releasing scene

In order to emphasize the atmosphere of releasing from tension, a sequence of shots is rendered so that the length of a shot becomes gradually longer. The length of a shot in the resultant video is determined as follows.

We denote the length of the original and the trimmed shot from s_i ($i=1,..,n_r$) as Lr_{si} and Lo_{si}, respectively, where a scene of released from tension consists of a sequence of source shot $s_1,..,s_{ns}$. The initial length of the first shot is denoted as br_{init}, and r_r is the increment ratio, where $r_r=40/n_r$. The length of shot in a scene of releasing from tension is determined as follows:

$Lr_i=br_{init}+(i-1)r_r$

When Lo_{si} is larger than Lr_{si}, the part of the shot, which has a time stamp from Lr_{si} to Lo_{si}, is cut out from s_i. On the contrary, if any one of the Lr_{si} exceeds Lo_{si} for $i=1,..,n_r$, r_r is decreased so that it satisfies $Lr_{si} \leq Lo_{si}$.

4. System Evaluation

4.1 User Interface

The user interface of the prototype system is shown in Fig. 4. In Figure 4, three pieces of video data at the top of the interface are displayed as source video data, which are placed by the editor. Temporal position of video data is adjusted by shifting it to right or left with a pointing device on screen. Semantic content or emotional information which the editor likes to emphasize is displayed in the form of icons above the source video data on top. The selection from the candidate source video data and trimming is automatically performed by the system, following the editing decision rule described in section 3. The resultant video data that is automatically edited by the system is shown at the bottom of the interface. It is also possible to adjust cutting points manually and/or to specify a certain shot so that it is forced to be included in the resultant video.

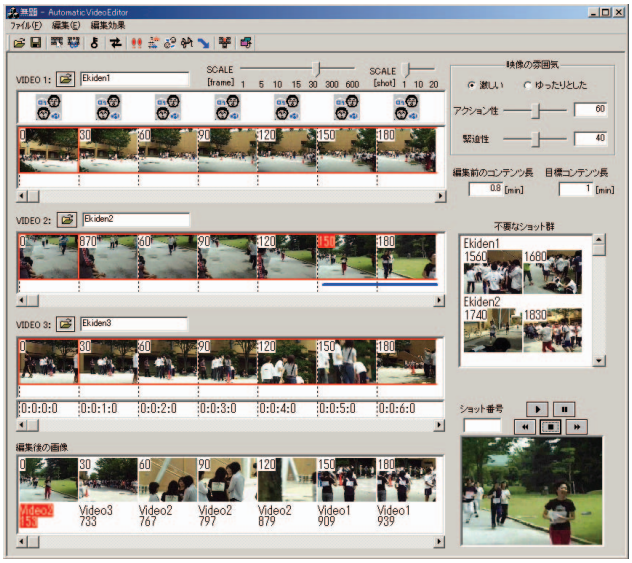


Figure 4. User Interface of the prototype system

4.2 Experimental Result

We experimented to evaluate the quality of generated contents and compared the proposed method with frame-based editing by the time required for editing operation.

Five university students are participated as test subjects for the experiments. Table 1 shows the result for evaluating the quality of generated contents. The value is the average score of the five subjects, which ranges from 1 through 5. The score of 1 corresponds to the case where a test subject did not catch the semantic contents or emotional atmosphere intended by the editor, whereas score 5 means the edited video is very easy to understand semantic contents or emotional information intended by the editor. Here, we regard score 3 or above as acceptable. In most of the cases except 'conversation', video data is edited enough for a viewer to catch emotional information emphasized by editing. The reason for insufficient result in editing a conversation scene was due to the error of speaker tracking.

Table 2 shows the time efficiency in editing work. The values in the table denote time consumed (in minutes) for producing a scene from the same source video. The time cost with the proposing system includes the time for manual adjustment of cut-in/out and shot selection by an editor for reforming the scene as close as his/her intention. This result shows that the proposed method reduces editing time for depicting semantic contents or emotional information.

Table 1. Evaluation of the quality of editing

	action	tensed	conversa-tion	calm	released from tension
Score	4.0	3.0	2.2	3.6	3.6

Table 2. Comparison of operation time (min.)

	action	tensed	conversa-tion	calm	released from tension
frame based	9.5	14	17	10	11
content based	2.3	3.3	6.0	4.9	4.7

5. Conclusion

We proposed a framework of rendition-based video editing, where video data is edited by specifying semantic content or emotional information instead of frame-based cut-in/out. Selection from candidate video data and determination of cut-in/out are automatically carried out based on film grammar. This method of editing helps editors to produce video contents with less time-consuming operations.

References

[1] D. Arijon, "Grammar of the film language", Silman-James Press, 1976.

[2] A. Yoshitaka, T. Ishii, M. Hirakawa, and T. Ichikawa, "Content-Based Retrieval of Video Data by the Grammar of the Film", Proc. of International Symposium on Visual Languages, pp. 310-317, 1997.

[3] A. Yoshitaka and M. Miyake, "Scene Detection by Audio-Visual Features", Proc., IEEE International Conference on Multimedia and Expo., CD-ROM, pp. 49-52, 2001.

[4] B. T. Truong, S. Venkatesh, C. Dorai, "Film Grammar Based Refinements to Extracting Scenes in Motion Pictures", Proc., IEEE International Conference on Multimedia and Expo., CD-ROM, 2002.

[5] C. Roisin, T. T. Thuong, L. Villard, "Integration of structured video in a multimedia authoring system", Proc. of the Eurographics Multimedia'99, Springer Computer Science, pp. 133-142, 1999.

[6] T. Chiueh, T. Mitra, A. Neogi, C-K Yang, "Zodiac: A history-based interactive video authoring system", ACM/Springer-Verlag Multimedia Systems journal, special issue on Multimedia Authoring and Presentation Techniques, Vol. 8, pp. 201-211, 2000.

[7] A. Girgensohn, et al., "A semi-automatic approach to home video editing", Proc. of UIST '00, ACM Press, pp. 81-89, 2000.

[8] A. Akutsu and Y. Tonomura, "Video tomography: an efficient method for camerawork extraction and motion analysis", Proc. of the ACM International Conference on Multimedia, pp. 349-356, 1994.