Video Editing Support System Based on Video Grammar and Content Analysis

Masahito Kumano and Yasuo Ariki
Dept. of Electronics and Informatics,
Ryukoku University,
Seta, Otsu-shi, 520-2194 JAPAN
kumano@rins.ryukoku.ac.jp

Miki Amano and Kuniaki Uehara
Dept. of Computer and
Systems Engineering,
Kobe University, Nada,
Kobe, 657–8504 JAPAN

Kenji Shunto and Kiyoshi Tsukada
Mainichi Broadcasting
System, Inc.
17–1 Chayamachi, Kita-ku,
Osaka, 530–0013 JAPAN

Abstract

Video editing is the work to produce the final videos with certain duration by finding and selecting appropriate shots from the material videos and connecting them. In order to produce the excellent videos, this process is generally conducted according to the special rules called "video grammar". In this paper, we propose an intelligent support system for the video editing where metadata are extracted automatically and then the video grammars are applied to the extracted metadata.

1 Introduction

In digital age, a large quantity of broadcast contents is strongly required to be created and reused. Although a non-linear video editing is available owing to the digitization, the video editing is still a bottleneck because a lot of skills and works are required. In order to solve this problem, an intelligent support system for video editing is proposed in this paper for efficient production of highly qualified contents.

The video editing is a work to produce the final videos with certain duration by finding and selecting appropriate shots from the material videos and connecting them. This work is appreciated as an abstraction process of the time and space of the story content. In order to produce the excellent videos, the abstraction process is generally conducted according to the special rules called "video grammar".

The video grammar is composed of rules to extract appropriate shots and to connect them such as "A panning shot follows and is followed by 1 second fixed shot" or "A medium (size) shot follows a loose (size) shot". In order to make these rules applicable, the metadata such as shot size or camera work included in the shots have to be extracted and catalogued. The purpose of this study is to develop an intelligent support system for video editing where these metadata are extracted automatically and then the video grammars are applied to them.

2 Related work

Chiueh, et. al. [1] proposed an interactive video authoring system which employs an edit history abstraction. The edit history abstraction is the generalization of the edit decision list data structure. The edit history abstraction is based on a branching history model, which keeps track of the development paths associated with multiple design alternatives. The system is useful to detect a composed stream's shot and scene boundaries, but it offers no assistance to select an appropriate shot.

Girgensohn, et.al. [2] proposed a semi-automatic video editing system that can find usable video clips and select in and out points automatically. Editing rules are used to optimize the length of each clip to be included. However users have to decide the kind of occurrence before and after any given shot by themselves. That is, the system does not have any editing rules or advices for shot transition.

The video skimming system which is proposed by Sundaram [3] uses video grammar for skimming. Video grammar refers to the arrangement of shots to represents the meaning of the shot sequence, but these grammatical rules are only used to condense scenes.

1051-4651/02 $17.00 (c) 2002 IEEE

Our work may be considered as an intelligent version of commercially available editing software where metadata can be automatically extracted and the video grammar can be applied to efficiently produce the excellent edited video.

3 Video grammar

The video grammar is a group of rules to judge the shot connection. The rules are described in the same manner as conventional sentence grammars. A basic element, to which the video grammar is applied, is a group of shots. We shall explain, at first, the definition of shots. The cut is defined as a physical continuous section where the camera starts at the beginning and stops at the end. On the other hand, the shot is defined as a logical continuous section where the shot size or camera work is uniquely defined within the cut. Therefore, one or more shots are included in one cut.

The shot size is defined according to the distance from the camera to objects. The shot size is classified into loose shot (LS), medium shot (MS) and tight shot (TS). TS and LS are the shots taken by approaching to or leaving from the object respectively compared with MS. A full shot is the shot where all the objects are included and is used as a master shot at the editing process. The following video grammar is available concerned with these shot sizes:

Rule (1) Two shots cannot be connected each other, where their shot sizes are extremely different, such as TS and LS.

Rule (2) The start shot of the scene must be a master shot.

The shot duration plays an important role to convey the meaning of the shot. For example, if shots with the slow movement continue for a long time, the audience becomes tired. On the other hand, if shots with rapid changes are connected shortly, the audience cannot understand the director's intention. In order to avoid these situations, the following grammar is available for the shot duration.

Rule (3) Durations of LS, MS and TS must be about 6, 4, 2.5 seconds respectively.

The camera work means camera movement, such as pan, zoom and follow. The follow means that the camera tracks moving objects. It is difficult to identify an important object in the shot in pan, zoom and follow. In order to avoid the difficulty, the following grammar is available.

Rule (4) Pan and zoom shots should be surrounded with fixed shots, which continues more than 1 seconds.

We have used 27 production rules based on the video grammar, as described above, for our video editing support system.

4 Video editing support system

4.1 System organization

According to the video grammar, we determined attributes which are associated with each cut of material video. The attributes are shown in Table 1.

Table 1. Table of attributes

Attribute	Explanation	Example
Id	Identification number of the cut	19, 20
Scene	Name of Scene	Scene1
Master	The cut can be a master shot, or not	0, 1
Shot	Head shot size of the cut	TS, LS
Shot End	Last shot size of the cut	MS, -
Camera	Camera work	Zoom, Fix
Start Frame	The head frame number of the cut	1000, 1501
End Frame	The last frame number of the cut	1500, 2000
Camera Start	Frame number where camera work has started	1250, -
Camera End	Frame number where camera work has done	1350, -
Use	The cut had already been used or not	0, 0

The first column of the Table 1 shows the name of the attributes. The next column indicates the explanation of the attributes. Examples of the attributes are shown in the last column. For example, the shot with ID 20 in Figure 1 is a restaurant exterior of Scene 1. The shot could be a master shot, therefore the attribute of Master is 1. The head shot size (Shot) of the cut is a loose shot and this cut is a fixed shot, therefore the shot size at the end (Shot End) of the cut is the same as Shot. The shot whose ID is 19 could not be a master shot, therefore the attribute of Master is 0. The head shot size of the cut is tight shot. This cut is a zoom cut and the shot size changes into middle shot at the end of the cut. So, attribute Shot, End Shot and Camera become TS, MS and Zoom, respectively.

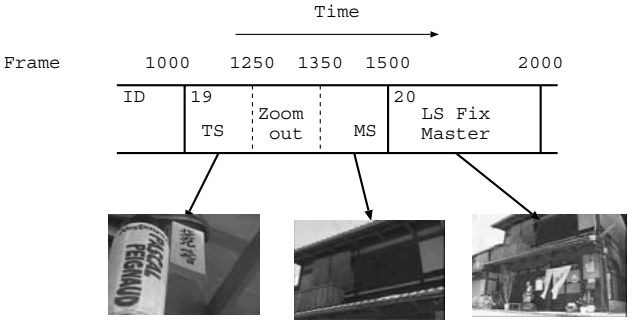

Figure 1. Example of shots

From now on, we shall explain the video editing support system based on these attributes and video grammar. The system chooses the shot, which comes after any given shot, from the material video according to the video grammar. Each cut of material video is stored in the MySQL database

labeled with attributes. The engine of the video editing system is forward-chaining production system, written in Prolog. The overview of the system is shown in Figure 2. The interface between production system and MySQL database is written in Java.

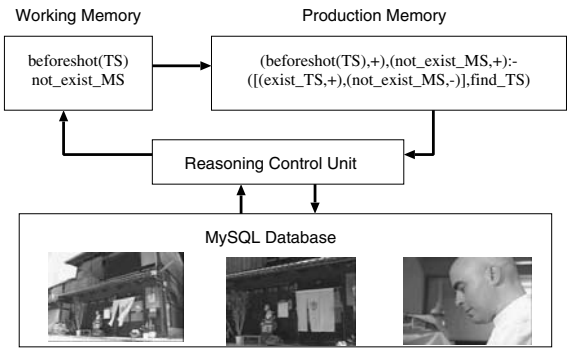

Figure 2. Video editing support system

The working memory is a collection of items. The production memory is a collection of production rules. The reasoning control unit determines the conflict set, resolves the conflict and fires the appropriate rule.

In this system, the production rule "if **condition** then **action**" will be represented by **condition –: action**. The **condition** consists of a list of item patterns, which refers to working memory items. Each item has a state showing + or -, that represents the presence or absence of the corresponding item in the working memory. For example, if there is a fact that "the preceding shot is TS" , the working memory contains "beforeshot(TS)" and this is represented by (beforeshot(TS),+).

The **action** consists of a list of individual actions and Prolog predicate to access MySQL database. Each Prolog predicate returns true or false, and the rule is only fired when the Prolog predicate returns true. For example, assuming that the working memory contains forward-chaining items such as "beforeshot(TS)" and "not_exist_MS," then the following production rule is applicable as,

(beforeshot(TS),+),(not_exist_MS,+)–:

([(exist_TS,+),(not_exist_MS,-)],find_TS)

This rule means that if the preceding shot is TS and there is no MS in MySQL database, then the next shot should be TS. The prolog predicate "find_TS" accesses MySQL database to check whether there are any tight shots that have not yet been edited. If the tight shot exists, "find_TS" returns true, then the production rule will be fired. By firing, the new forward-chaining item "exist_TS" is added to working memory and the "not_exist_MS" is removed.

4.2 Editing process

The outline of the editing process is shown in Figure 3. The shadow parts in the material video depict the shots that have already been used to edit. First, the system chooses the cut whose ID is 20, according to the rule (2). Because of the cut is a fixed loose shot, the system extracts 6 seconds from the material cut. See rule (3). Since MS can only be connected to LS because of rule (1), then the next shot must be ID 17 or 18. If there are many cuts satisfying the rule (1), the shot that has the nearest shot to the preceding shot should be selected to resolve the conflict.

The second shot will be the cut whose ID is 18. Second shot is extracted 4 seconds from the material cut. See rule (3). The valid shot size to connect MS is either TS or LS according to the rule (1). Since the priority of the transition from MS to TS is higher than that of the transition from MS to LS, so TS is chosen.

The ID 19 is chosen as the third shot. While editing a pan or zoom shot, such as the cut whose ID is 19, the frame is extended 1 second before and after the camera movement. It should be extracted from the material cut, according to the rule (4). In this case, the Shot End is MS, therefore the next shot should be TS.

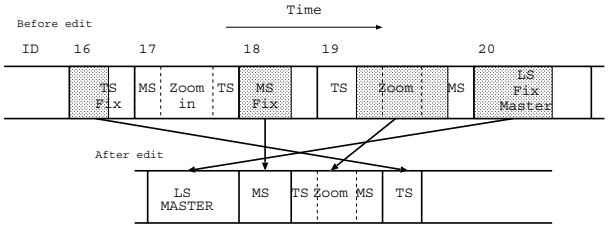

Figure 3. Editing process

5 Video content analysis

5.1 Metadata for video edition

All of the metadata required for the application of video grammars are human face direction, walking direction, eye direction, camera work, camera direction, camera tempo and shot size. Among them, the camera work and shot size are important because these two attributes are employed in the present video editing support system as shown in Table 1. We describe here the methods to extract the camera work and to determine the shot size.

5.2 Camera work extraction

The shot was defined in section 3 as a logical homogeneous section where the shot size or camera work is

uniquely defined within the cut. According to this definition, the camera work parameters of panning and zooming are automatically extracted from the material video. They are defined by translating value and expanding/reducing scale between consecutive frames respectively. If no values of the camera parameters are detected, then the corresponding section is regarded as the fix section.

As a technique to extract the camera parameters, we employed a method described in [4] which computes the translation and expansion/reduction on the gray value projection computed by projecting the gray values of the image into a horizontal direction and vertical direction. This method can quickly compute the camera parameters.

After the extraction of the camera parameters, the material video is segmented into sections with homogeneous camera parameters; fix, pan and zoom section. For the pan and zoom section, fix section with one second duration is attached before and after the section.

In the material video, useless sections such as camera shaking, blank and camera adjustment are included. They are excluded before the shot extraction by computing the changing speed of the camera parameters because the camera parameters are almost constant (smooth change) in useful shots.

5.3 Shot size

The camera work gives us effective information to decide the shot size. Here we separate the shot size indexing into two steps. One is shot size indexing within a cut and the other is among the cuts. The cut is defined as a physical continuous section where the camera starts at the beginning and stops at the end as described in section 3.

5.3.1 Shot size indexing within a cut

Within a cut, the camera is always working so that the camera parameters of panning and zooming give us inclusive relation in terms of the shot size. Figure 4 shows the inclusive relation between shots within one cut. For fix sections before and after panning, the shot size does not change so that FIX1 and FIX2 should have the same shot size as well as FIX3 and FIX4. On the other hand, for fix sections before and after zooming, the shot size will change according to expansion or reduction of the zooming. In the figure the shot size of FIX2 and FIX3 should be in the inclusive relation.

5.3.2 Shot size indexing among cuts

Within one cut, the shot size inclusive relation is obtained using the camera work parameters, but the exact shot size such as LS, MS and TS can not be uniquely given to each shot due to the ambiguity. In order to give the exact shot size, several consecutive cuts have to be collected within one scene and label LS (master shot) is given to the most inclusive (exterior) shot based on rule (3).

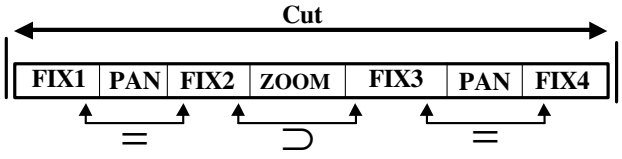

Figure 4. Inclusive relation of shots within a cut

However, the camera work between consecutive cuts has no relation so that a new method is required to estimate the inclusive relation between shots in the different cuts. We have developed this method by performing active search[5] between all the pairs of shots in different cuts. The active search is a method to search an object appearance (reference pattern) on an input image by computing the similarity. After the active search, shots with the highest similarity to other shot in different cuts are regarded to have the inclusive relation.

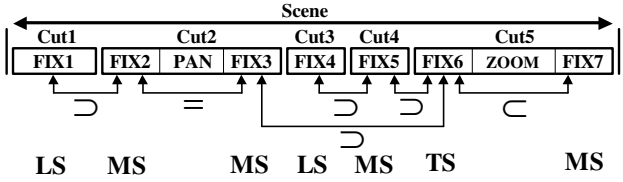

Figure 5. Shot size indexing among cuts

Figure 5 shows an example of the inclusive relation between shots among several consecutive cuts after the processing described above. In the figure, the symbols $\supset$, $=$ and $\subset$ indicate that the left side shot includes, equals and is included by the right side shot respectively. In the example, the shot size labels LS, MS and TS, which are inferred based on the inclusive relation computed by the above method, are shown at the bottom of the figure. In the case where no inclusive relation is found between the shots in the different cuts, the cut is isolated from other cuts and unknown shot label US is given to all the shots in the isolated cut.

6 Experiments

The material videos used in this study are provided by Mainichi Broadcasting System, Inc. These videos are taken for cooking shows which introduce a restaurant. Each video is containing scenes which show exterior, interior, and dishes of the restaurant. The videos are encoded into H.263 of 30 frames per second. We used four material videos and called them mat1, mat2, mat3 and mat4, respectively.

Table 2. Results of automatic shot size indexing

Scene number	1	2	3	4	5	6	7	8	9	10	11	12	Total
Number of FIX sections	2	14	7	11	4	3	16	5	2	22	21	12	119
Number of correct labels	0	11	2	3	4	3	16	3	2	16	16	8	84
correct(%)	0.0	78.6	28.6	27.2	100	100	100	60.0	100	72.7	76.2	66.7	70.6
Number of USs	0	0	0	0	0	0	0	0	0	0	0	1	1
processing time(s)	5	156	30	51	13	4	34	16	3	300	144	53	809

6.1 Result of shot size indexing

We have carried out an experiment of shot size indexing for material videos. The frame size is 640×480 pixels.

The result is shown in Table2. Twelve scenes are processed. The number of fix sections and the number of correctly inferred shot size labels are listed in the table. The *correct* is defined as follows;

$$correct = \frac{Number\ of\ correct\ labels}{Number\ of\ FIX\ sections} \times 100 \qquad (1)$$

In the table, the number of USs (unknown shot) and processing time are also listed. From the table, it can be said that the correct percentage of shot size labels was 70.6% in total and the number of USs was only one. The number of correct labels for scene 1 was zero. It is attributed to errors caused by the active search for shots in different cuts.

6.2 Video editing result

6.2.1 Redundancy

In the material video, cameraman took a lot of same shots which are called retake shots and each shot continues quite long duration. In order to show that the system eliminated these redundancy, we defined "Shot" and "Frame". "Shot" is the ratio of the edited shots to the material shots and "Frame" is the ratio of the edited frames to the material frames.

$$Shot = \frac{\text{Edited shots}}{\text{Shots of material video}} \times 100 \qquad (2)$$

$$Frame = \frac{\text{Edited frames}}{\text{Frames of material video}} \times 100 \qquad (3)$$

These results are shown in Table 3. This result suggests that

Table 3. The rate of use in cut and frame

material number	Shot Rate (%)	Frame Rate (%)
material 1	44.1	12.8
material 2	72.5	11.9
material 3	53.6	16.9
material 4	41.0	15.3

total length was shortened to about 10 to 20 % and 40 % shots were edited. However, the usefulness of this system cannot be shown only by saying that the material video was edited shortly. In order to achieve an easy way to understand the content of the video, (1) the system must generate coherent output and (2) the time duration of the shot should be corresponds to information amount of the shot. So we defined "Coherence" and "Information amount" as follows.

6.2.2 Coherence

There are high-priority rules and low-priority rules in this system. For example, the priority of the transition from MS to TS is higher than that of the transition from MS to LS. Using low-priority rule is not desirable since it may make inconsequent shot transition. Therefore the coherence of the edited video is defined as "Coherence". Coherence is represented by following equations.

$$Coherence = \frac{\text{Number of using high-priority rule}}{\text{Edited cuts}} \times 100 \qquad (4)$$

These results are shown in Table 4.

Table 4. The evaluation value of quality

material number	Quality Rate
material 1	68.2
material 2	66.7
material 3	78.6
material 4	72.7

These results suggest that above 2/3 of shot transitions are coherent and it can be said that the material video has been edited into the easily comprehensible video.

6.2.3 Information amount

We verifies that certain time duration is given to each shot of the edited video. There is a relation between the length of a shot and its information amount. The information amount can be expressed with visual complexity of the shot. In the research of video skimming [3] , it is shown that the incompressibility of a shot is proportional to the visual complexity of a shot and incompressibility of a shot can be bounded by

using the Lempel-Zip compression algorithm. Therefore, if the rate of Lempel-Zip compression of a shot is calculated, the visual complexity of a shot can be defined. In other words, the shots which are hard to compress can be judged as the shots with more complexity.

Then, the key frame of all fixed shots is taken out from edited video and the rate of Lempel-Zip compression of the shot is calculated. The average of Lempel-Zip compression rate for every LS, MS, and TS is shown in Table 5.

Table 5. The average of the compression rate for every shot size

material number	LS (%)	MS (%)	TS (%)
material 1	79.6	73.8	75.4
material 2	93.4	80.4	74.6
material 3	85.5	78.8	76.9
material 4	88.7	88.0	91.6
material 4'	90.6	87.3	84.9
total	87.2	83.4	82.6

Table 5 shows that LS is the most complicated shot as a whole. Although the good result has come out from mat2 and mat3, the highest value has appeared in TS at mat4. This is due to the error contained in the judgment of the shot sizes by the automatic indexes. Shot sizes were corrected by hand and the rates of compression were calculated again, then the LS gets the highest value. This is shown in mat4' (Table 5).

6.2.4 General Result

The automatically edited videos prepared by our system were examined by 14 evaluators. They were expected to evaluate the videos on the following five indexes. The indexes were on a scale of 1-5. The highest ranking is 5. The primary index shows whether the shot was suitable for editing or not. The 2nd index shows the validity of the each shot duration, the 3rd one is the coherence of the contents, the 4th one is easiness of comprehension, and the 5th is the validity of the length of the edited video.

The ranking for these five indexes is shown in the Table 6. These numerical values are the average mark of 14 evaluators. The Table 6 shows that the values of the 4th and 5th indexes of mat1 are low. This fact suggests that the shots with the similar shot sizes were connected continuously because we used the rule having the low priority as shown in the Table 4. Also shots of mat1 were edited with the same time duration as the other material video, although each shot of mat1 was less information amount than the other material videos shown in the Table 5. As a result, it indicates a long and redundant impression on the whole. 'Mat2' shows the highest value as to the shot duration, easiness of comprehension, and the whole length. In each shot size, the difference of the amount of information appears as shown in Table 5. Therefore each shot had been edited with a proper duration. Mat3 obtained high evaluation in Table 4 and 5. It was, however, indicated low evaluation as for the selection of the shot, the coherence of contents, and easiness of comprehension which was shown in Table 6. It suggested that the very similar shots had been edited consecutively because there were a lot of retake shots in the material videos, nevertheless, the attribute of retake shots was not taken into account.

As mentioned above, mat4 shows a low evaluation as for the duration of shots because information amount did not correspondent to that of shot size (Table 5). In the Table 6, the mat4', improved version, gets a little high evaluation as for the duration of shots. These results suggest that these videos edited by our editing support system could be regarded as the easily comprehensible video.

Table 6. Test scores from 14 users

	mat1	mat2	mat3	mat4	mat4'
shot select	3.2	3.5	3.3	3.4	3.4
duration	3.8	3.9	3.6	3.6	3.8
story coherent	3.5	3.6	3.3	3.6	3.6
easiness of comprehension	3.3	3.9	3.4	3.6	3.5
whole length	2.6	3.9	3.6	3.3	3.8

7 Conclusion

In this paper, we described a video grammar and the video editing support system that we have been developing. To enable the system work effectively, the metadata such as camera works and shot sizes were extracted from material video data that is different from the broadcast video data.

The future work will be concentrating on the accuracy improvement of the metadata and extension of the vide editing support system to utilize the video grammars which employs human face direction, walking direction, eye direction, camera direction and camera tempo.

References

[1] Tzi-cker Chiueh and Tulika Mitra, "Zodiac: A History-Based Interactive Video Authoring System," Proc. of ACM Multimedia '98, ACM Press, pp.435–443, 1998.

[2] Andreas Girgensohn and John Borecxk, "A Semi-automatic Approach to Home Video Editing," Proc. of UIST '00, ACM Press, pp.81–89, 2000.

[3] Hari Sundaram and Shih-Fu Chang, "Condensing Computable Scenes Using Visual Complexity and Film Syntax Analysis," Proc. of ICME 2001, pp.389–392, 2001.

[4] A.Nagasaka and T. Miyatake: "Real-Time Video Mosaics Using Luminance-Projection Correlation", IEICE, Vol.J82-D-II, No.10, pp.1572-1580, 1999-10.

[5] T.Kawanishi, H.Murase, S.Takagi and Werner: "Dynamic Active search for Quick Object Detection with Pan-Tilt-Zoom Camera", ICIP01, VolIII, pp.716-719, 2001.