8—17

Automatic Useful Shot Extraction for a Video Editing Support System

Masahito Kumano and Yasuo Ariki *
Faculty of Science and Technology
Ryukoku University

Abstract

Video editing is a work to produce final videos with certain duration by finding and selecting appropriate shots from raw material videos and connecting them. In order to produce excellent and intelligible videos, the editing process is generally conducted according to the special rules called "video grammar". The purpose of this study is to develop an intelligent support system for video editing based on automatically extracted metadata. This paper proposes a method to automatically segment the raw video materials into useful sections and useless sections based on the video grammar as a part of the video editing support system, using the camerawork and cut point information.

1 Introduction

In the coming digital age, a lack of video contents makes a serious problem and consequently a large quantity of broadcast contents is strongly required to be created and reused. To solve this problem, an efficient and new video editing technique or system is required because it consumes a lot of works. The goal of this study is to develop an intelligent support system for video editing based on a video grammar[1]. One of core problems in developing such kind of the system is to extract useful video sections from the raw video materials.

Girgensohn, et.al. [2] proposed a semi-automatic video editing system that can find usable video clips and determine the in- and out-points automatically. However the estimation of the inappropriate section is restricted the extraction of dark sections.

In this paper, a method is proposed to automatically estimate the useless sections such as "hand shake" and "failure camerawork" and to estimate useful sections.

2 Video Editing Support System

Figure 1 shows the video editing support system we are developing based on the video grammar. The video grammar is a group of rules shown in Table 1 to judge the clip connection. The clip is defined as the section clipped, based on the video grammar, from the raw video material. In the figure, the editing process is composed of three phases; phase1, 2, 3.

Phase 1 works as video analysis such as "cut point extraction", "camerawork analysis", "shot extraction", and "shot size indexing". We have already reported the shot size indexing process in [3]. The most fundamental connection rule is based on "Shot size" such as rules(1)-(5) in Table 1 .

The shot size is defined depending on the object size and classified into loose shot (LS), medium shot (MS) and tight shot (TS). The TS and LS are the shots taken by approaching to or leaving from the object respectively compared with the MS. Figure 2 shows the examples of these shot sizes and relative relations. Obviously the shot size is given to the camera-fixed section.

Phase 2 works as extraction of clips to be connected according to the meta information to the useful section. Figure 3 shows two types of clips presumed according to the video grammar such as rules(6)-(10) in Table 1.

*Address: 1-5 Yokotani, Seta-ohe-cho, Otsu, Shiga 520-2194 Japan. E-mail: {kumano,ariki}@rins.ryukoku.ac.jp

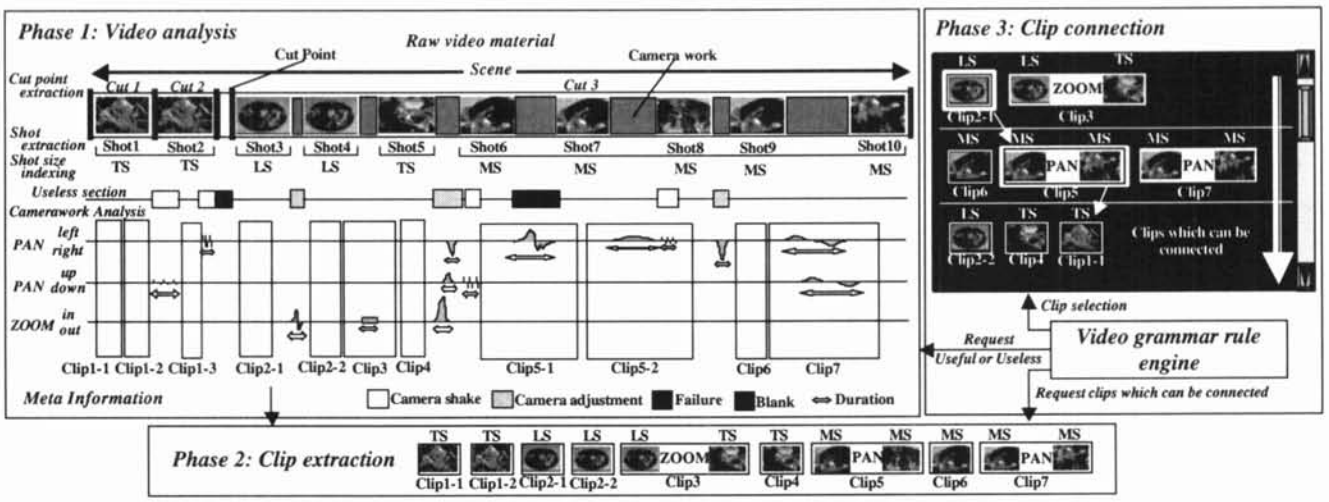

Figure 1: A simple video editing support system

Table 1: The extract of video grammars

rule(1):	The start shot of the scene must be a master shot(Usually it is LS) which reveals the whole figure of a scene.
rule(2):	MS can not be connected to MS because it is redundant.
rule(3):	TS can be connected to TS.
rule(4):	Two shots with the same shot size can not be connected each other when the objects are same.
rule(5):	Two shots can not be connected each other when their shot sizes are extremely different such as TS and LS.
rule(6):	Before and after the pan and zoom shots, the fixed shot continues more than 1 seconds.
rule(7):	The duration of a fixed shot is up to 15 seconds.
rule(8):	The durations of LS,MS and TS are about 6,4 and 2.5 seconds respectively.
rule(9):	The movement of the pan and zoom should be stable.
rule(10):	The starting shot and ending shot are 2 seconds longer than normal shot.

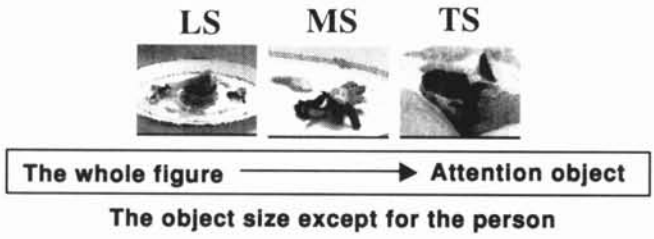

Figure 2: Shot size and relative relations of them

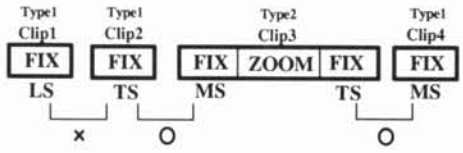

Figure 3: Clip connection

The type1 clip includes only fix section such as clips1,2 and 4. The type2 clip includes a camerawork section and camera-fixed sections such as the clip3 in Figure 3 according to the rule(6) in Table 1. The duration of clips are mainly decided by rule(8) when clips are extracted from the raw video material.

Phase 3 is an interactive board which supports the video editing by presenting an available list of clips to be connected to the previous one according to the video grammar. In clip connection based on the shot size, the fix section of each clip is compared with those in other clips shown in Figure 3. By this system, a human editor can concentrate on the connection of listed clips and is freed from the inefficient work.

This paper focuses on phase 1 system because the video analysis is inefficient and occupies the most part of the video editing.

3 Discrimination Flow of Useful and Useless Section

The useful section includes camera-fixed section, stable camerawork section such as zooming & panning, and finally "camera-follow" section where the cameraman follows moving objects by his camera. On the other hand, the useless section is defined as "hand shake", "camera adjustment", "failure camerawork" and "blank section". Three types of useful sections and four types of useless sections are discriminated by the speed of the camerawork parameters. The most difficult problem is the discrimination among "camera-follow", "hand shake", "failure camerawork" and "camera adjustment", because their camerawork sequences are almost similar.

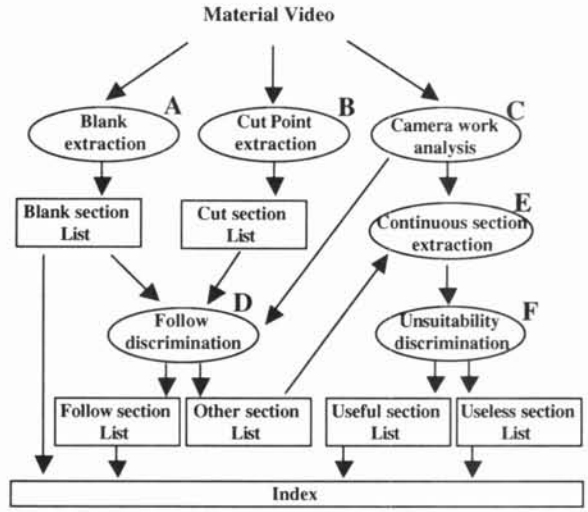

Figure 4: Discrimination process of useful and useless section

Their discrimination process is shown in Figure 4. First, the cut points extraction(B) and camerawork analysis(C) are carried out for the raw video material using histogram intersection and intensity-projection respectively. These processes are done in real time. Then a raw video material is divided by these cut points and the divided sections are grouped into the "camera-follow section" and the other "miscellaneous section"(D). The "blank section"(A) is easily discriminated by the amount of intensity-projection. The "camera-follow section" is discriminated by using camerawork density and camerawork instability because the camera is continuously tracking the moving object in the "camera-follow section". The camerawork density is computed by the camerawork value, and the camerawork instability is computed by the normalized variance of the camerawork value within every shifted short window. In continuous section extraction process(E), continuous camerawork becomes one continuous section and discontinuous camerawork inside a short window is merged to neighboring continuous section.

Finally, the "miscellaneous section" is segmented into "hand shake", "rapid change camerawork" ("failure camerawork" and "camera adjustment"), "stable camerawork" section and the "fixed" sections using the camerawork density and camerawork instability. "Hand shake" section has the feature of camerawork sparseness, and the "failure camerawork" and "camera adjustment" have the feature of high density camerawork and its rapid change.

4 Discrimination Method

4.1 Cut Point Extraction

As a technique of cut detection, we have developed a new technique based on Eq.(1) which computes the histogram intersection[4] as shown in Eq.(2). Let $h'_{f,i}$ denote a histogram bin of class i at frame f. Then the normalized histogram bin $h_{f,i}$ is defined as $h_{f,i} = h'_{f,i} / \sum_j h'_{f,j}$ $(i, j = 1 \cdots I : I = Q^3)$, provided that Q is the number of classes in R,G,B color space.

$$Cut(a) = \min_b HI(h_a, h_b) - \max_c HI(h_a, h_c) \quad (1)$$

$$HI(h_a, h_b) = \sum_{i=1}^{I} min(h_{a,i}, h_{b,i}) \quad (2)$$

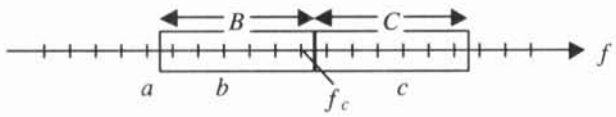

Figure 5: Extraction of cut point f_c

Figure 5 shows a configuration of cut point extraction. Two windows with B and C frames are set on future time at the frame a. When a condition $Cut(a) > \theta_c > 0$ is satisfied, the cut point f_c is determined as $f_c = a + B$, provided that $b = a + m(m = 1 \cdots B)$, $c = a + B + n(n = 1 \cdots C)$. Table 2 shows the result of cut point extraction under the conditions of $Q = 8$, $\theta_c = 0.2$, $B, C = 6$.

In the table, the Correct ratio is defined as the ratio of the number of correctly extracted cut points to the total number of cut points existing in the material video. On the other hand, the accuracy is defined as the ratio of the number of correctly extracted cut points to the number of exactly extracted cut points in the material video. Both of the correct ratio and the accuracy are satisfactory.

Table 2: Result of cut point extraction

Correct	M	50
Un-detection	D	3
Excessive detection	E	5
Correct ratio(%)	$M/(M+D)$	94.3%
Accuracy(%)	$M/(M+E)$	90.9%

4.2 Camera Work Analysis and Blank Extraction

As a technique to extract camera parameters, we employed a method described in [5] which computes the translation and expansion/reduction on the gray value projection. Projections into horizontal direction $P_Y(f,i)$ and vertical direction $P_X(f,j)$ are shown in Eq.(3) and Eq.(4) respectively where $Gray(f,i,j)$ is the gray value at the position i, j on the frame f with height h and width w. Blank sections can be extracted by Eq.(5).

$$P_Y(f,i) = \frac{1}{h} \sum_{j=1}^{h} Gray(f,i,j) \quad (3)$$

$$P_X(f,j) = \frac{1}{w} \sum_{i=1}^{w} Gray(f,i,j) \quad (4)$$

$$Blank(f) = \frac{1}{w} \sum_{i=1}^{w} P_Y(f,i) < \theta_b. \quad (5)$$

The amount of camera panning at frame f in left-right or up-down direction is shown in Eq.(8) and Eq.(9). Here, the projection distances in horizontal and vertical direction between consecutive frames are shown as follows, provided that $\delta_p (\delta_p = -20, -19, \cdots, 19, 20)$.

$$D_{P_Y}(f,i,\delta_p) = \{P_Y(f,i) - P_Y(f+1, i-\delta_p)\}^2 \quad (6)$$

$$D_{P_X}(f,j,\delta_p) = \{P_X(f,j) - P_X(f+1, j-\delta_p)\}^2 \quad (7)$$

$$Pan_{lr}(f) = \arg\min_{\delta_p} \sum_{\substack{i=1+\delta_p \ (\delta_p \geq 0) \\ i=1 \ (\delta_p < 0)}}^{\substack{w \ (\delta_p \geq 0) \\ w-\delta_p \ (\delta_p < 0)}} D_{P_Y}(f,i,\delta_p) \quad (8)$$

$$Pan_{ud}(f) = \arg\min_{\delta_p} \sum_{\substack{j=1+\delta_p \ (\delta_p \geq 0) \\ j=1 \ (\delta_p < 0)}}^{\substack{h \ (\delta_p \geq 0) \\ h-\delta_p \ (\delta_p < 0)}} D_{P_X}(f,j,\delta_p) \quad (9)$$

The amount of zooming at frame f can be also computed. The description is omitted here due to the complexity within the limited space.

4.3 Discrimination of Follow Section

The camera-follow section has the feature of sparse density and instability camera work. Therefore discrimination of camera-follow section is carried out by the product of density D' and instability I' at every c section called "cut section" as shown in Eq.(10).

$$F_c = D'_c \cdot I'_c \quad (10)$$

The density of a certain section is computed by Eq.(11) and Eq.(12). The function $f(x)$ shows the existence of the camera work x and $D(x)$ shows the normalized density within the section of C frame length.

$$f(x) = \begin{cases} 0 & (x = 0) \\ 1 & (x \neq 0) \end{cases} \quad (11)$$

$$D(x) = \frac{1}{C} \sum_{k=1}^{C} f(x(k)) \quad (12)$$

Finally the density score of the section is computed as $D'(Pan_{lr}, Pan_{ud}, Zoom) = \{D(Pan_{lr}) + D(Pan_{ud}) + D(Zoom)\}/3$ because the pan and zoom appear sporadically in the camera-follow section.

Next the instability score of the section is computed by $I'(Pan_{lr}, Pan_{ud}, Zoom) = \{I(Pan_{lr}) + I(Pan_{ud}) + I(Zoom)\}/3$. $I(x)$ is the normalized sum of variances in every window as shown in Eq.(13). Here C is the length of the cut section and W is the length of the window in the cut section. Also μ_i is the mean of $x(k)$ in the window beginning from frame i.

$$I(x) = \frac{1}{(C-W)W} \sum_{i=1}^{C-W} \sum_{k=i}^{W+i-1} (x(k) - \mu_i)^2 \quad (13)$$

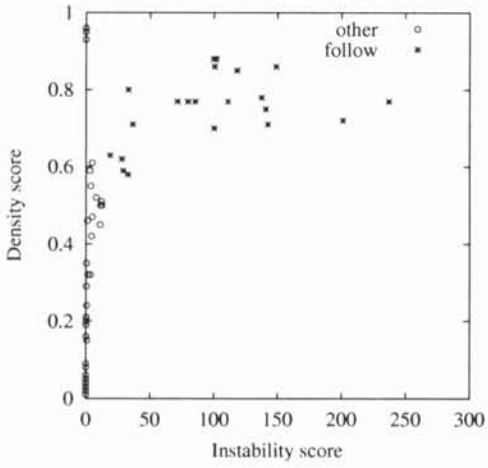

Figure 6: Feature of follow section

Table 3: The indicator of camera work

Camera work	Density	Instability
Follow	Sparseness	Big
Hand shake	Sparseness	Small
Rapid change	Denseness	Big(in short section)
Normal	Denseness	Small(in short section)

Figure 6 shows the result of D' and I' applied to cut sections included in two video materials. All the follows section show the high D' score and wide spread I' score. Therefore the follow section is discriminated by the variance of F_c. The variance σ is computed by using the section except for follow section and the section is discriminated as follow section when it is over $3 \cdot \sigma$.

4.4 Discrimination of Unsuitability

In the miscellaneous section, continuous sections are "hand shake", "rapid change camerawork" or "stable camerawork" sections. The rest of these sections are fix sections. First, the continuous section is, according to the indicator in Table 3, segmented into "hand shake" and the other sections using the camerawork density shown in Eq.(12) . Second, the other sections are segmented into "rapid change camerawork" and "stable camerawork" using camerawork instability shown in Eq.(14), where L is the length of the continuous section and substituted for C in Eq.(13).

$$I(x) = \frac{1}{(L-W)W} \sum_{i=1}^{L-W} \sum_{k=i}^{W+i-1} (x(k) - \mu_i)^2 \quad (14)$$

These thresholds for sparse, dense, big and small in Table 3 are determined by the averaged D' or I' from the raw video materials, provided that $W = 10$.

5 Experimental Result

We have carried out the segmentation experiment into the useful and useless sections for the 25 minutes raw video material taken by a professional camera man. The result is shown in Table 4. In the table, the Correct ratio is defined as the ratio of the number of correctly extracted sections to the total number of sections existing in the material video. On the other hand, the accuracy is defined as the ratio of the number of correctly extracted sections to the number of exactly extracted sections in the material video. The accuracy of hand shake is a little lower because one hand shake section is miss-discriminated to the "camera-follow" section. (*) is the number of excessive detection in "hand shake" extraction. If this number is neglected, the accuracy of "hand shake" extraction becomes 84.5%. The remaining results are satisfactory.

6 Conclusion

In this paper, a method was proposed to extract appropriate shots automatically as well as to give shot size labels to the extracted shots. At present 84% correct rate was obtained for stable camera work section and 91% accuracy was obtained for fix section in the useful section extraction. Future works will be the improvement of the correct ratio and the accuracy as well as the accomplishment of the proposed video editing support system.

References

[1] M.Kumano, Y.Ariki, K.Shunto, K.Tsukada: "Video Editing Support System Based on Video Content Analysis", ACCV2002,VolII,p.628-633,2002-01.

[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] Masahito Kumano, Yasuo Ariki, Miki Amano, Kuniaki Uehara, Kenji Shunto, Kiyoshi Tsukada: "Video Editing Support System Based on Video Grammar and Content Analysis",ICPR2002

[4] M.J.Swain and K.H.Ballard, "Color indexing",IJCV, vol.7, pp,11-32, 1991.

[5] Akio Nagasaka, Takafumi Miyatake: "Real-Time Video Mosaics Using Luminance-Projection Correlation", IEICE,Vol.J82-DII,No10,pp.1572-1580,1999.

Table 4: Results of automatic discrimination of shot sections

	Equation	Blank	Camera work				Fix
			Follow	Hand shake	Rapid change	Stable	
Correct	C	6	4	22	7	21	78
Un-detection	U	0	1	5	1	4(2)	8
Excessive detection	E	0	0	18(14)	0	2	5
Correct ratio(%)	$C/(C+U)$	100	80	81	88	84(91)	91
Accuracy(%)	$C/(C+E)$	100	100	55(85)	100	91	94