Video Editing Support System Based on Video Content Analysis

M.Kumano and Y.Ariki
Ryukoku University
{kumano,ariki}@rins.ryukoku.ac.jp

K.Shunto and K.Tsukada
Mainichi Broadcasting System, Inc.
{shunto, tsukada}@mbs.co.jp

Abstract

The video editing is a work to produce the final videos with certain duration by finding and selecting appropriate cuts 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". The purpose of this study is to develop an intelligent support system for video editing so that metadata are extracted automatically and then the video grammars are applied to the extracted metadata. In this paper, we describe the metadata extraction such as camera work, camera tempo, camera direction, face and shot size.

1. Introduction

In digital age, a large quantity of broadcast contents is strongly required to be created and reused. Although a nonlinear 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 cuts 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 cuts 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"[1]. In order to make these rules applicable, the metadata such as shot size or camera work included in shots have to be extracted and catalogued. The purpose of this study is to develop an intelligent support system for video editing so that these metadata have to be extracted automatically and then the video grammars to be applied to them.

The famous conventional technique for video editing is video skimming developed at CMU [2] which can summarize the document videos by extracting important words from speech data and collecting the corresponding video segment. However their video data to be processed is not material videos but broadcast videos with the speech data. When the material video is processed for editing, the speech data is not always available and further more the material video includes the redundant retakes.

Other famous conventional technique for video editing is a video parser developed at MITRE [3] which can separate the CM and each news stories included in the news program by utilizing the key phrases as well as speaker change and key words extracted from CC. This work was also done on the broadcast videos that are well edited.

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

2. Video grammar

The video grammar is a group of rules to judge the shot connection. A basic element, to which the video grammar is applied, is a group of shots. Therefore we describe at first the definition of shots and their types and then the rules are described in the same manner as conventional sentence grammars.

2.1. Shot size

Figure 1 shows the definition of cuts and shots. In the figure, 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.

The shot size is defined according to the distance to objects from the camera. The absolute shot size is defined as "full figure", "knee shot", "waist shot", "bust shot", "up" and "close up" for human size. The size difference indicates the mental distance so that we feel intimacy to the objects as the shot size becomes large. On the other hand, the relative shot size is defined as 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. 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 grammars concerned with this shot size are available;

- Two shots with the same shot size can not be connected each other when the objects are same.
- Two shots can not be connected each other when their shot sizes are extremely different such as TS and LS.
- The start shot of the scene must be a master shot.

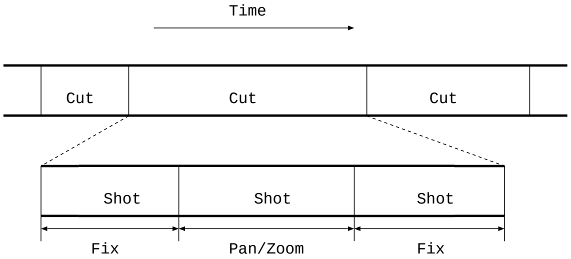


Figure 1: Definition of cuts and shots

2.2. Shot duration

Shot duration plays an important role to convey the meaning of the shot. For example, if all the shots are adjusted to have the same duration, the audience can not catch the important shot. If shots with the slow movement continue long time, the audience becomes tired. Inversely, if shots with rapid changes are connected shortly, the audience can note catch the director or editor's intension. In order to avoid these situations, the following grammars are available about the shot duration.

- The duration of a fixed shot is up to 15 seconds.
- The starting shot and ending shot are 2 seconds longer than normal shot.
- The durations of LS, MS and TS are about 6, 4, 2.5 seconds respectively.

2.3. Camera work

Camera work indicates camera movement such as pan, zoom and follow which tracks moving objects. The camera work also includes a fixed shot that has no ambiguities. On the other hand, the pan, zoom and follow include the ambiguities so that they are used in the limited situation. For example, they are used when the object is too large to present by the fixed shot, or when to present the space expansion.

The zoom has special effect to present the camera work approaching to the object so that after zooming the camera will be inside the object in the most cases when the object is a house or building. In order to avoid the camera work ambiguities, the following grammars are available.

- Before and after the pan and zoom shots, the fixed shot continues more than 1 seconds.
- The tempo of the pan and zoom is continuous.
- The shots following the pan and zoom are restricted according to the context.

2.4. Direction

Directions indicate moving direction of objects, face direction, eyes direction and camera direction. These directions have to be consistent between consecutive shots. For example, when a man is walking from left to right, the following new TS (tight shot) has the same direction of the man walking taken in the previous shot. In the same way, when two persons are facing each other in the original shot, the following new TSs, which take each person in close-up, have to have the same face directions as the original shot.

In the eyes direction, when the first shot takes a person looking up, then the following shot takes the gazed object in a direction from top to down inversely. The camera direction to the object has to be kept consistent to avoid the ambiguities before and after the shot size is changed.

3. Video content analysis

3.1. Metadata for video edition

All of the metadata required for application of video grammars are shown in Figure 2. In the figure, camera works are extracted from the material videos at first. When the camera work is fixed, persons are detected. If one person is detected, the shot size is classified into one of seven types. Face direction and eye direction are detected according to the shot size and also person is tracked if he is moving. If non-person is detected, the camera direction and shot size are identified. If the camera work is detected at the camera work extraction, it is classified into the categories of pan, zoom and follow. Furthermore the tempo is also extracted from the camera work parameters. Hereafter we describe some of the techniques employed in our system to produce the metadata.

3.2. Camera work

3.2.1. Camera work extraction.

The camera work is extracted from each frame by dividing the frame into a group of macro blocks and computing

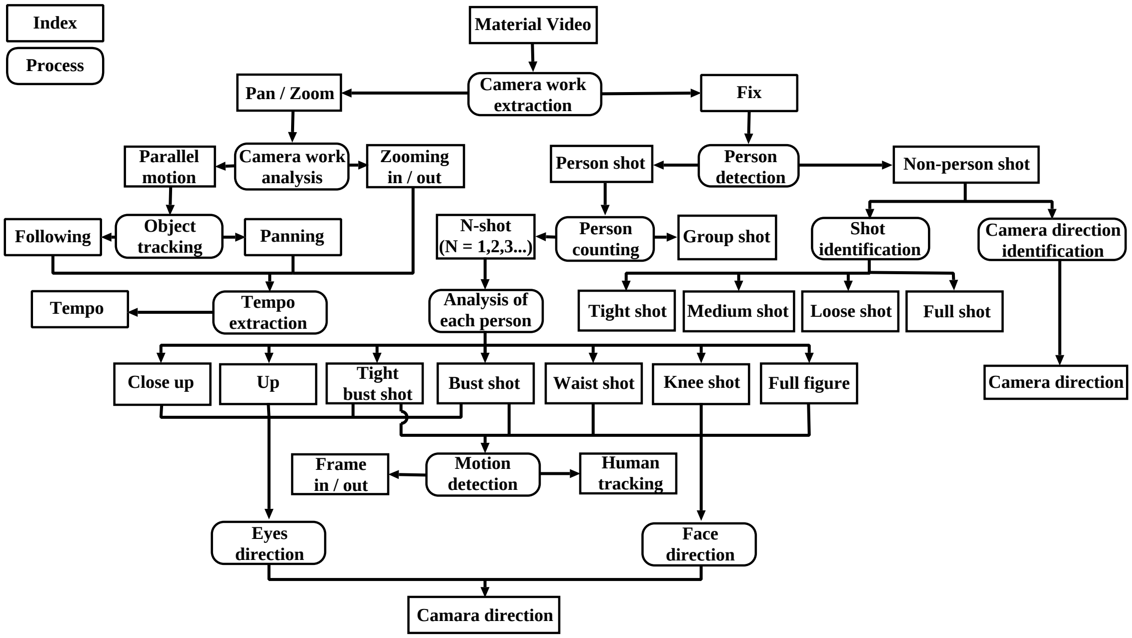


Figure 2: Metadata extraction flow

the moving vectors u, v at each macro block between the consecutive frames. Based on these moving vectors, the camera parameters are estimated by a least mean square method to Affine transformation parameters at each frame as shown in Eq.(1)[2].

$$\begin{aligned} u(x_{i,k}, y_{i,k}) &= (a_i - 1)x_{i,k} + b_i y_{i,k} + c_i \\ v(x_{i,k}, y_{i,k}) &= d_i x_{i,k} + (e_i - 1)y_{i,k} + f_i \end{aligned} \quad (1)$$

where $a_i, b_i, c_i, d_i, e_i, f_i$ are the Affine parameters and $x_{i,k}, y_{i,k}$ are the positions of macro block k at frame i.

The Affine transformation estimates one global motion at each frame so that local motions that deviate from the global motion disturb the estimation itself. In order to solve this problem, we employed robust estimation method that can remove the local motions and estimate the global motion more accurately.

The consecutive frames with the same camera work such as fix, pan, zoom-in and zoom-out obtained by the above mentioned method can be grouped as a segment (shot) as shown in Figure 3. We call this segmentation as camera work segmentation. In this camera work segmentation, the major camera work is determined based on majority rule within the consecutive frames in order to avoid the unstable value of the camera parameters extracted at each frame.

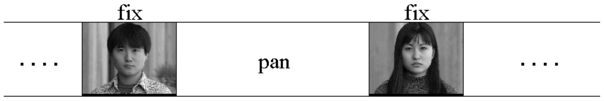


Figure 3: Camera work segmentation

3.2.2. Camera tempo.

A tempo is a parameter to present the speed of the camera work and is computed by Eq.(2). Here $a_i, b_i, c_i, d_i, e_i, f_i$ are the camera parameters obtained at frame i. $Zoom_i$ is the zoom ratio at frame i and Pan_i is the camera moving ratio.

Figure 4 shows the difference between the fast tempo and the slow tempo. In the figure, the upper part indicates the fast tempo video where the zoom ratio and pan ratio are almost same even the scene is different in the same video. On the other hand, the lower part indicates the slow tempo video where again the zoom ratio and pan ratio are almost same. However the ratio is different between the different videos, the fast tempo videos and the slow tempo videos.

$$\begin{aligned} zoom_i &= |(a_i - 1)(e_i - 1) - b_i d_i| \\ pan_i &= \sqrt{c_i^2 + f_i^2} \end{aligned} \quad (2)$$

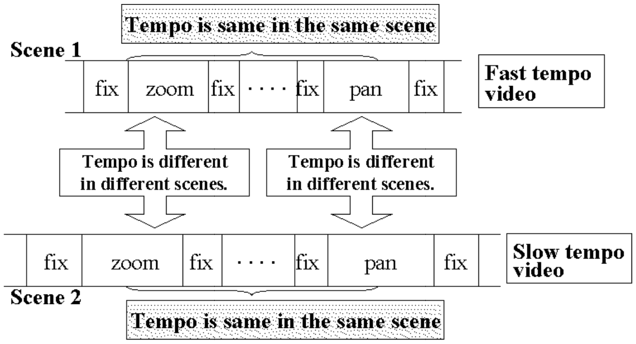

Tempo is same in the same scene
Scene 1
fix
zoom
fix
· · · ·
fix
pan
fix
Fast tempo video
Tempo is different in different scenes.
Tempo is different in different scenes.
fix
zoom
fix
· · · ·
fix
pan
fix
Slow tempo video
Scene 2
Tempo is same in the same scene

Figure 4: Camera tempo

3.2.3. Camera direction.

For the fixed camera shot, the camera direction is extracted in order to keep the consistent direction of the objects before and after the shot size changes. The camera direction, in which the camera is set to the objects, is mainly classified into the left direction and the right direction.

To obtain the camera direction, the edge lines shown in Figure 5 plays an important role. For example, two lines in Figure 5 (a) indicate that the vanishing point locates in the left hand side so that the camera is judged to be in left direction. On the other hand, two lines in Figure 5 (b) indicate that the camera direction is right.

Therefore the edge lines are extracted at first by Hough transformation on the edge image (the first frame in the fixed camera shot) that is obtained by canny operator. Then the camera direction is decided based on the histogram of the extracted line directions.

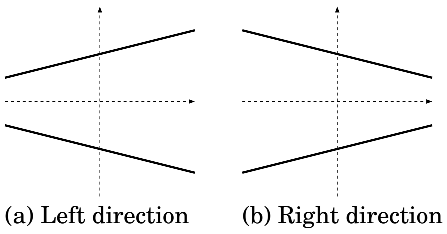

(a) Left direction (b) Right direction

Figure 5: Camera direction

3.3. Face detection

3.3.1. Extraction & direction.

For the fixed camera shot, it is checked whether the shot includes person or not by detecting human faces. If the shot includes a number of persons, their faces are all detected and counted. The shot size such as "close up", "up", "tight bust shot", "bust shot", "waist shot", "knee shot" and "full figure" are decided as well as the face direction. For the shot without person, the relative shot size is decided such as "tight shot", "medium shot" and "loose shot".

In order to detect human faces, we employed fuzzy pattern matching based on skin color similarity and hair color similarity[4]. The skin color similarity at each pixel of an input image is computed as probability of the color value in the skin color probability distribution that was constructed in advance using many skin color data. The hair color similarity is also computed in the same way.

Figure 6 (a), (b) and (c) show an input image, skin color similarity map and hair color similarity map respectively. In the figure, the dark value indicates the higher similarity.

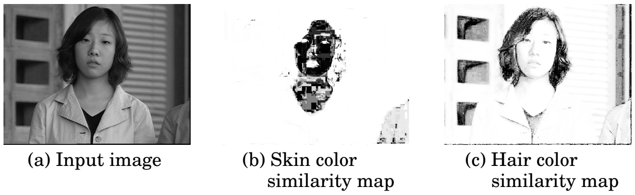

(a) Input image (b) Skin color similarity map (c) Hair color similarity map

Figure 6: Skin color and hair color similarity map

To detect human faces from these similarity maps, two-dimensional head figure model is constructed in advance as shown in Figure 7. It is composed of 11 × 13 cells and the ratio of skin color similarity to hair color similarity is computed at each cell by averaging the value at each pixel within the cell. Finally cells are averaged within the image and the head figure model is completed. In order to detect human faces with right and left direction, three head figure models are constructed; front, right and left directional models. These three models are applied to two similarity maps obtained from the input image as shown in Figure 6 and human faces and the directions are detected by a pattern matching method.

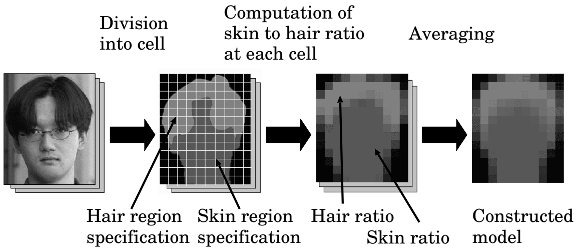

Division into cell
Computation of skin to hair ratio at each cell
Averaging
Hair region specification
Skin region specification
Hair ratio
Skin ratio
Constructed model

Figure 7: Construction of head model

3.3.2. Shot size.

Shot size with human faces is decided by extracting the skin region and by computing the occupancy ratio of skin color similarity to the whole input image as shown in Figure 8. When the skin region is large, the shot size is classified into "close up" or "up" according to the skin color occupancy ratio. On the other hand, when the skin region is small, the shot size is classified into "tight bust shot", "bust shot", "waist shot", "knee shot" and "full figure" according to the face region size after extracting the individual face region.

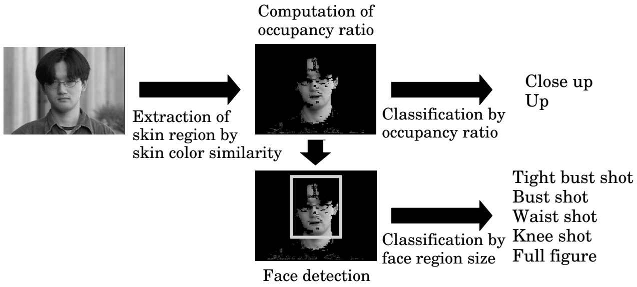

Computation of occupancy ratio
Extraction of skin region by skin color similarity
Classification by occupancy ratio
Close up
Up
Tight bust shot
Bust shot
Waist shot
Knee shot
Full figure
Classification by face region size
Face detection

Figure 8: Classification of shot size

4. Experiments

4.1. Camera work

We have carried out an experiment of camera work extraction for 20 minutes material video taken by a professional camera man and segmented the video into the individual shot. Each frame in the material video is digitized into Motion JPEG format with 640 × 480 pixel size from a beta cam tape through SGI Impact compressor.

The result is shown in Table 1. In the table, the recall is defined as the ratio of the number of correctly segmented sections to the total number of sections existing in the material video. On the other hand, the precision is defined as the ratio of the number of correctly segmented sections to the number of exactly segmented sections in the material video. The precision is a little lower but the recall is satisfactory.

Table 1: Result of camera work extraction and segmentation

Camera work	Recall (%)	Precision (%)
Panning	100.00	99.62
Zooming in	91.55	97.01
Zooming out	100.00	89.47
Follow (person)	99.60	88.93
Follow (non person)	100.00	92.45

4.2. Tempo extraction

We carried out a tempo extraction experiment for the same data described in section 4.1. The results are shown in Table 2. From the panning and follow, the averaged moving amount was extracted and from the zooming in and zooming out, the averaged zoom ratio was extracted. From the table, it can be seen that the fast tempo and the slow tempo are well extracted.

Table 2: Result of tempo extraction

Camera work	Averaged moving amount	Averaged zoom ratio
Panning	7.300	1.011354
Zooming in	7.094	1.051854
Zooming out	7.915	0.991216
Follow (person)	5.686	1.006881
Follow (non person)	7.468	1.005186

4.3. Camera direction

Camera direction was extracted from video data composed of 30 shots including news video, indoor and outdoor scenes. The result is shown in Table 3. In the table, "right" and "left" indicate the right direction and the left direction of the camera. The accuracy of the camera direction is not so high partly because of the edge noise after application of canny operator and partly because of the multiple vanishing points more than two included in the image.

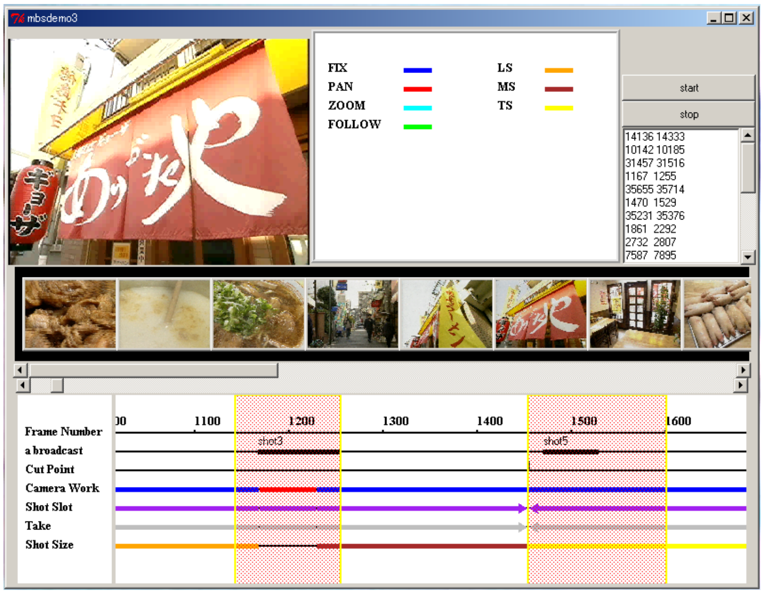
Table 3: Result of camera direction

Camera direction	Correct	Total	Correct ratio (%)
Right	7	10	70.0
Left	16	20	80.0
Total	23	30	76.7

4.4. Face detection

An experiment was carried out to extract the human faces from 31 frames that are the starting frames of 31 shots taken from a professional cameraman. Thirty-four human faces are included in these 31 frames. The size of the frames is 640×480 pixels. The result is shown in Table 4. In the table, the "human faces" indicates the number of shots where human faces are correctly counted in 31 frames. On the other hand, the "shot size" and "face direction" indicate the number of faces whose size and direction are correctly decided among 34 faces.

In the evaluation, extracted faces are judged as correct in a case where the extracted facial region overlaps more than half of the true facial region.



```
14136 14333
10142 10185
31457 31516
1167  1255
35655 35714
1470  1529
35231 35376
1861  2292
2732  2807
7587  7895
```

Figure 9: Example of metadata visualization

Table 4: Result of face detection

Type	Correct	Total data	Correct ratio (%)
Human faces	17	31	54.8
Shot size	31	34	91.2
Face direction	26	34	76.4

4.5. Indexing system toward a video editing

The metadata mentioned above are automatically extracted from the material video and stored into metadata database that can be accessed by a video editing support system. Although, the video editing support system is at present under developing, we can visualize automatically the extracted metadata as shown in Figure 9.

In the lower part of the figure, metadata are listed such as cut point, camera work (fix, pan, zoom and follow) and shot size. In the middle, the shots corresponding to the uniform camera work are listed. The upper left is a movie to be edited hereafter. The video editing support system accesses the metadata shown in the lower part and searches the shots that can be connected to the underlying shot shown in the upper left corner by utilizing the video grammar.

5. Conclusion

In this paper we described a video grammar and the video editing support system that we are now developing. To enable the system work well, the metadata such as camera work, camera direction, face extraction and face direction were extracted from material video data that is different from the broadcast video data.

Although we focused on the metadata, we are at present developing the video editing support system that utilizes the metadata and video grammar. The future work will be the accuracy improvement of the metadata and accomplishment of the support system.

6. References

[1] Daniel Arijon, “Grammar of the Film Language,” Focal Press Limited, London, 1976.

[2] Smith, M.A. and Kanade, T., “Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques,” CVPR1997, pp. 775–781, 1997.

[3] “Merlino, A., Morey, D. and Maybury, M. , Broadcast news navigation using story segmentation,” (http://www.acm.org/sigmm/MM97/papers/morey/index.html).

[4] Wu, H., Chen, Q., and Yachida, M., “Face Detection From Color Images Using a Fuzzy Pattern Matching Method,” PAMI-21(6), pp. 557-563, 1999.