An Automatic System for Editing Dance Videos Recorded by Multiple Cameras

No Institute Given

Abstract. As social media has matured, uploading video content has increased. With physical performances like dance performances, such videos are better and easier to understand than static content. Multiple videos with intense performances, such as dance performances, are difficult to integrate into high-quality videos without knowledge of video editing principles. In this study, we present a system that automatically edits the dance performance videos taken from multiple viewpoints into a more attractive and sophisticated dance video. Our system can crop the frame of each camera appropriately by using the performer's behavior and skeleton information. The system determines the camera switches and the cut lengths following a probability model of general cinematography guidelines and of knowledge extracted from expert experience. The system automatically edits the dance video of four performers taken from multiple viewpoints, and 10 experts in video production evaluated the generated videos. As a result, the system tended to be better evaluated than other automatic editing methods.

Keywords: Video editing, Dance, Computational Cinematography; Automation

1 Introduction

As social media has progressed, many videos have been uploaded to the Internet for private or public purposes. With physical performances like dance performances, such videos are better and easier to understand than static content. High-quality videos are likely to be shared by many people. This leads to increased opportunities for performer work. Therefore, video contents are an important tool for a performer to become famous. Sponsored professional dance videos are regularly uploaded on social media, attracting fans and raising the recognition of performers. However, these attractive and sophisticated videos are usually achieved by professional techniques used in movies and TV series. Thus it is difficult to create a high quality dance video for amateur performers. One reason is the lack of knowledge of video photography and editing. However, the principle of video editing that adapts to the home video in daily life cannot deal with the video which includes the intense movement such as dance performance.In this study, we focus on video editing, and propose a system that automatically generates more attractive and sophisticated dance videos, even for non-specialists. The system creates a single dance movie by automatically editing performance videos taken from multiple locations according to a probability

model created based on principles of video editing and principle extracted from preliminary interviews.

The remainder of this paper is organized as follows. Section 2 introduces related works. In Section 3, we explain our system: we describe the principles of video editing, explain the system outline, and explain the probability model based on the principles. In Section 4, we conduct evaluation experiments and discuss the results and considerations. Section 5 discusses our system and we summarize this paper in Section 6.

2 Related Works

There are many existing works on automatic video editing, such as the method of Heck et al.[1]. However, few have focused on dance video editing. In this study, we propose an automatic editing system to focus on dance performance and show performance better. As related research, we survey some studies on automatic video editing and editing support systems.

Arev et al.[2] use insight to share the focus of people's attention and determine where important actions in the scene are being done. The system selects the camera by combining its function and the cinematography guidelines and determines the timing of cutting. Ranjan et al.[3] propose an automatic editing system for meetings that applies television production principles. Zsombori et al.[4] propose a system that automatically generates video for events based on media annotations and highlights specific people during events. Lu et al.[5] present a video summarization technique to discover a story of an egocentric video. Given a long input video, their method selects a short video that shows an essential event. Jain et al.[6] re-edit widescreen video by using pan, cut, and zoom based on gaze data. Shin et al.[7] propose a system for converting the contents of a blackboard-style lecture video into a readable interactive lecture note with a graphic in corresponding text. Kumar et al.[8] present a system that focuses on each performer from stage-performance video to crop the video and presents a divided-screen video. For panoramic video, Sun et al.[9] propose a system to automatically extract the region of interest and control the virtual camera. Roininen et al.[10] describe how to model the shot cut timing of professionally-edited concert videos. Mate et al.[11] present an automatic video remixing system that intelligently processes user-generated content in combination with sensor information. Truong et al.[12] propose QuickCut, which can quickly create narration videos. The system treats narrated videos of appropriate sizes as clips, taking into account the length of narration and the length of motion. Leake et al.[13] propose a system for efficiently editing video of dialogue-driven scenes. Jeong et al.[14] propose a method to enhance a user's dance video by selecting seven emotion categories to be communicated to viewers.

We also mention research on movement and emotion as a related study. Nam et al.[15] state that the speed of motion of an object is the most influential factor to emotions. Montepare et al.[16] clarify that walking characteristics such as the

swing amount, step length, weight and walking speed of the arm differentiated the emotion expressed by pedestrians.

From above, we can see that action and emotion are closely related. Our system focuses on the behavior of performer and makes dance video more attractive and sophisticated by editing.

3 Automatic Dance Video Editing System

We look at the process of editing videos taken from multiple viewpoints into a single movie. An editor selects a video clip and then determines how long the clip will be and selects the next video clip. This is equivalent to selecting a camera at a time and transitioning. This means that if there is two successive shots that takes into account the camera before the transition and the appropriate camera work according to the movement of the performer, it can be regarded as an estimation problem. This can be solved as a problem to estimate the appropriate camera work depending on the movement of performer while selecting the camera every moment. In this paper, we define the minimum time unit as one beat (synonymous with a quarter-note) and treat as an optimization problem which selects an effective camera for each beat. We aim to generate more attractive and sophisticated dance videos by using the constraints and evaluation functions based on preliminary interviews and the principles of video editing.

3.1 Principles of Video Editing

In the field of professional video editing, there are some principles of editing, and editors take into account these principles and make edits based on their own experience. Similarly, in order to generate a more attractive and sophisticated dance video by editing, it is necessary to do edits according to the dance performance based on the principle of video editing. Therefore, we enumerate the principles of general video editing that can adapt to dance videos from the guidelines of cinematography, and the empirical principles that can adapt to dance videos extracted from interviews with video production experts. In this paper, we define a shot far from the subject as a long shot, and a shot near the subject as an up shot. We define a video of length one beat as one shot and a series of consecutive shots from the same camera as one cut.

Principle 1. Switch camera according to beat. *Video Production: Disciplines and Techniques*[17] contains the following description. "*edits should be made at appropriate points at the end of a sentence on the narration track in a documentary, for example, or on the beats of the music.*" Dancing to music is common, and most of the time it is adapted to the beat of the music. Therefore, the switching timing of the camera must be adjusted to the beat.

Principle 2. The maximum length and minimum length of a cut are changed according to user preferences. *Video Production: Disciplines and Techniques*[17] contains the following description. "*a shot lasting more than three seconds is viewed by some producers and directors as 'boringly long.'*" As

for the length of a cut, it cannot be stated unconditionally that the shorter the cut length, the more attractive and sophisticated it is. Therefore, setting the maximum and minimum cut lengths will be entrusted to the user. This paper specifies that the difference between maximum and minimum lengths increases so that the cut length is selected in a wide range.

Principle 3: Switch to a camera with different composition in order to prevent jump cut. *Grammer of the Edit*[18] states that "*Editing together two shots of similar camera angles will cause a jump at the cut point. Differing camera angles and framing will help prevent the 'jump cut' in the mind of the viewer.*". The jump cut is an edit that combines two shots of the same subject from slightly changed camera positions. This type of editing provides an effect to skip time. Switching to a camera with different composition can prevent jump cuts.

Principle 4: The closer the shot, the shorter the cut length. The more distant the shot, the longer the cut length. *Grammer of the Edit*[18] states that "*In deciding the length of a shot, it is essential to give the eyes enough time to read and absorb the visual information. If you are questioning the 'proper' duration for a shot, then you could describe, in your mind, what you are seeing in the shot.*". The appropriate time for cutting is equal to the time that the situation can be explained. In a closer shot, the information is clear enough to tell a viewer in a short time, so the cut length is short. The more distant shot has more information, so the cut length must be longer.

In the above, from cinematography guidelines we extracted and enumerated principles that can adapt to dance performance videos. Next, we interviewed four experts: three people who were involved in video production work and one person who graduated from the video body department. We had them watch the dance video prepared in advance. Moreover, we interviewed them about being careful when taking a dance and editing technique. We defined the following two items based on the comments obtained.

Principle 5. Make the main camera recognizable. From the previous interview, we received the following comments."*It is good to be able to recognize that this is the main cut the creators want to show.*", "*It is important for creators to think that this cut is main.*", "*Put a side dish inside a long shot.*" A 'side dish' refers to a shot that emphasizes movements, in contrast to a long shot that is easy to get overall information. From these comments, it is important that the main camera is designated and the viewer can recognize the main camera.

Principle 6. Decide what to emphasize and when to vary the pace. From the previous interview, we received the following comments. "*I will emphasize the delicious part by up shot and make sure that this cut does not continue to some extent*" , "*If you cut video that focuses on performer's feet at every time he kicks his feet, a good impression will fade. Choosing a better shot with priority is good.*" , "*I think it is important to consider how much the balance is the best to repeat between up shot and long shot.*" , "*Slow and high speed. A viewer is tired of switching the cut at a constant speed. We sometimes make an extreme change in the number of cuts according to the tune.* "

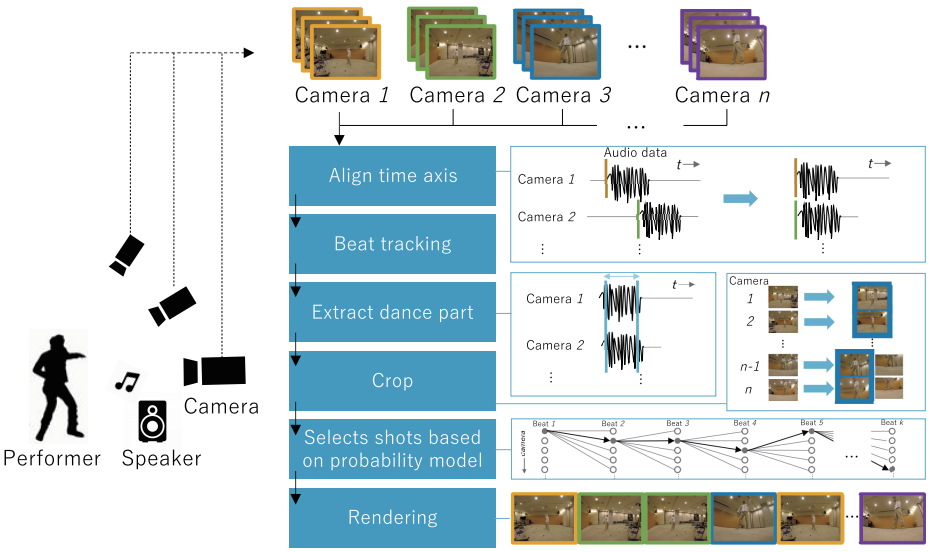

Fig. 1. System overview.

We will build a system based on these six principles. We interpreted several principles as follows in order to construct the system.

In Principle 3, the composition was assumed to be the size of the motion vector on the video. By increasing the difference in the size of the motion vector between the cameras that are transited, the video switches to different compositions. In Principle 5, we predict that the camera will be recognized as the main camera by increasing the percentage of only one video. In Principle 6, what is an attractive shot depends on the case. Therefore, we assume that the attractive part of this time has intense movement (the reverse situation is technically compatible). Our system balances the relatively intense movement and non-intense movement area throughout the whole.

3.2 System Overview

The system overview is shown in Fig. 1. The system flow is as follows.

1. A user puts dance videos taken from multiple places into our system. All videos must be recorded with the same song playing during shooting. However, it is not necessary to shoot at the same time.
2. The system extracts the audio data of each video and takes the correlation function of the music data used in the dance and it. The system cuts the movie according to where the correlation is maximized. As a result, the time axis can be aligned since the start of all movies is at the start of the song.

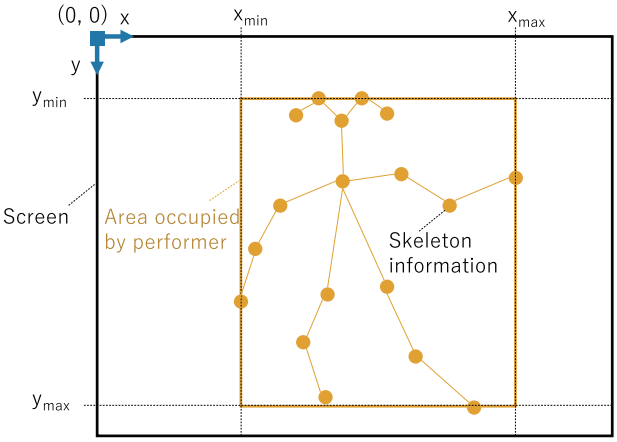

Fig. 2. Area occupied by performer on the screen.

3. The system calculates the beat positions in the song using the beat tracking method proposed by Bock et. al.[19].
4. The system extracts the dance performance part. In current status, the system does not perform automatic detection of choreographed parts of performers. Therefore the system needs to know in advance how many beats there are from the beginning of the song to when the performer starts to dance, and how many more beats until the performer stops dancing.
5. The system aquires the performer's skeleton information using the OpenPose library by Cao et al.[20]. The difference between the maximum (x_{max}) and minimum (x_{min}) values in the x-axis direction of the obtained skeletal information is defined as the area occupied by the performer by multiplying the difference between the maximum (y_{max}) and minimum (y_{min}) values in the y-axis direction (Fig. 2). The system calculates R, the percentage of the screen area S_a occupied by the performer.

$$R = \frac{y_{max} - y_{min}}{x_{max} - x_{min}} / S_a \tag{1}$$

6. The image is cropped so that all calculated ratios R fall between 0 and 1.0 and the average value is 0.55 to 0.65. The center of the crop is the center position of the area occupied by the performer. Moreover, in the upper 40 percent camera, which is up shot of all cameras, the system produces the cropped videos centered on the nose and the cropped videos centered on the average position of the heel position of both feet. The system calculates the average value of optical flow per camera and treats a video with a large average value as one of the shots (Fig. 3).

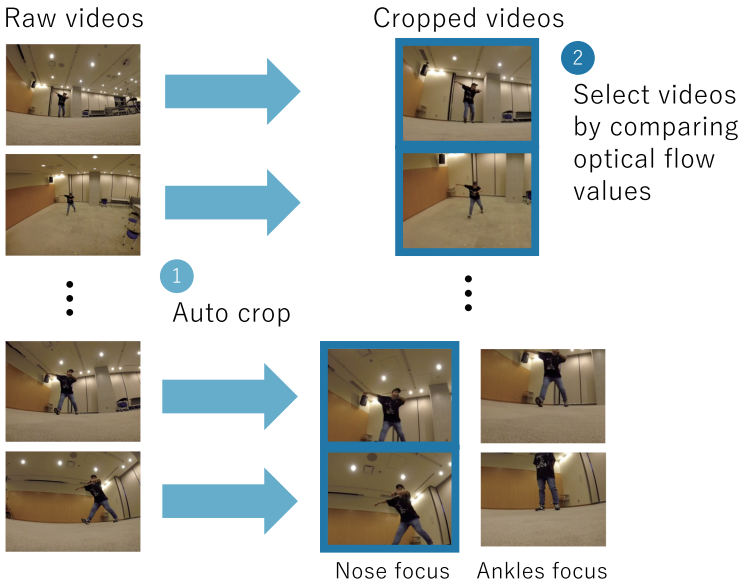

Fig. 3. Auto crop function flow.

7. A video of one beat is defined as one shot. The system selects shots based on the probability model for each shot and saves the order in which the evaluation function value for the selected shot is maximized.
8. The system generates a single movie based on the saved order.

3.3 Probability model

The following probability distributions are generated based on principles 2 to 6.

Principle 2. The maximum length and minimum length of a cut are changed according to user preferences. In this paper, the minimum and maximum cut lengths are 2 and 12 beats, respectively. If the camera before the transition is continuously smaller than the minimum length, the probability of transition to the same camera is 1, and the other is 1^{-20}. If the camera before the transition is continuously bigger than the maximum length, the probability of transition to the same camera is 1^{-20}, and the other is uniform distribution:

$$p\left(x_t^n | x_{1:t-1}^m\right) = \begin{cases} 1.0 \ (n = m) \\ \varepsilon \quad (otherwise) \end{cases} \tag{2}$$

Principle 3: Switch to a camera with different composition in order to prevent jump cut. The optical flow P_t at each time t is calculated with the Gunnar Farneback method. We compute P_t^n, the average value of P_t for all $0 < t < t_{max}$ for camera n. We normalize P_t^n so that $\sum_n P_t^n = 1$.

$$p\left(x_t^n\right) = \frac{P_t}{\sum_n P_t^n} \tag{3}$$

Principle 4: The closer the shot, the shorter the cut length. The more distant the shot, the longer the cut length. The system squares the average of the optical flow of each camera calculated by Principle 3, and the normalized one is expressed by probability distribution. The square of the average value is experimentally determined to increase the difference. From the previous camera location, the probability value of the current camera location is assumed to be moved to another camera according to the calculated probability distribution. The destination camera number is determined by a uniform probability. By doing this, the higher the degree of up shot, the easier it is to move to other cameras.

Principle 5. Make the main camera recognizable. The system increases the percentage of a single camera in the video generated at the end. The main camera was determined as follows:

1. Use OpenPose[20] to detect skeletal information.
2. The amount of skeletal information not obtained in a frame is b_f; the amount of skeletal information available is B; the total number of frames is F; and the raet of undetected skeletal information (ND for 'no data') is expressed in the following formula:

$$ND = \frac{1}{F}\sum_{f=1}^{F}\frac{b_f}{B} \tag{4}$$

   Moreover, the variance values in the x direction and y direction of the skeleton number in all frames are denoted as $V(x_{all})$ and $V(y_{all})$.
3. The main camera is the camera that maximizes this formula:

$$m = \frac{V(x_{all}) + V(y_{all})}{ND} \tag{5}$$

The system selects the main camera in a probability of 60% and others in the uniform probability distribution. We use the normalized value of m as the probability model to determine a shot at the first beat of the video.

Principle 6. Decide what to emphasize and when to vary the pace. The system uses a shot with an intense movement when a relatively intense movement is carried out throughout the entire videos. The average value of the optical flow per beat of every camera is calculated, and the median value is used as a threshold value. If the average value is higher than the threshold, it is a shot which includes an intense movement. If the average value is lower than the threshold, it is a shot which does not include an intense movement. The average value of the optical flow per beat is normalized and expressed as a probability distribution. If it does not include an intense movement, the system normalizes the maximum value of the optical flow minus the average value and uses it as a probability distribution. If it includes an intense movement, the system

normalizes the average value of the optical flow and uses it as a probability distribution.

Probabilities corresponding to the above Principles 2 to 6 are P_2, P_3, P_4, P_5, P_6. The system generates a video in which the following evaluation function O is maximized. However, the coefficient follows the following equation, $a_2 + a_3 + a_4 + a_5 + a_6 = 1$.

$$O = \log(a_2 P_2) + \log(a_3 P_3) + \log(a_4 P_4) + \log a_5 P_5 + \log(a_6 P_6) \tag{6}$$

4 Experiment

4.1 Procedure

We conducted an expert evaluation experiment to investigate whether the principles were reflected in the system and whether it was capable of generating more attractive and sophisticated video. We shot the dance performance of four performers, two male and two female dancers, as an editing material. The performer's genres were all different: BREAK, HIPHOP, WACK and POP, in order of movement intensity. Performers created a choreography that combined basic movements for each genre. The BPMs for the songs used were 109, 129, 93, and 92, respectively, and the choreography ranged from the start of the song to 16∼64 beats. We used 10 cameras (HERO5 Sessions, by Gopro) and a speaker (Computer MusicMonitor, by Bose). The layouts of the cameras are shown in Fig. 4. The height of each cameras is indicated by its color in the diagram. We adjusted each camera so that the subject was at the center of the screen. The layout of the cameras in BREAK and HIPHOP are on the left in Fig. 4, the layout in POP and WACK on the right. In order to prevent other cameras from entering between the subject and the camera, shooting was performed in two separate cameras.

We prepared seven different videos from each dance movie. Considering the coefficients used in the system, we decided to vary them one at a time, to prevent the number of combinations from exploding. For the first video, we set the following ratios: $a_2 : a_3 : a_4 : a_5 : a_6 = 3 : 2 : 5 : 3 : 3$. For videos 2∼6, we changed the coefficients one at a time to zero.

For the seventh video, we created a simple automatic editing system that generates a single video as a baseline. It operates with the following algorithm:

1. The system calculates the average value of the optical flow of every frame and calculates the average of them of each beat.
2. The system normalizes the average value for each camera.
3. For each beat, the system selects a camera with the maximum value of normalized optical flow.

By doing this, a shot which includes relatively intense movement becomes easier to be chosen.

We prepared a total of 28 videos from four performer's dance videos in the seven ways described above.

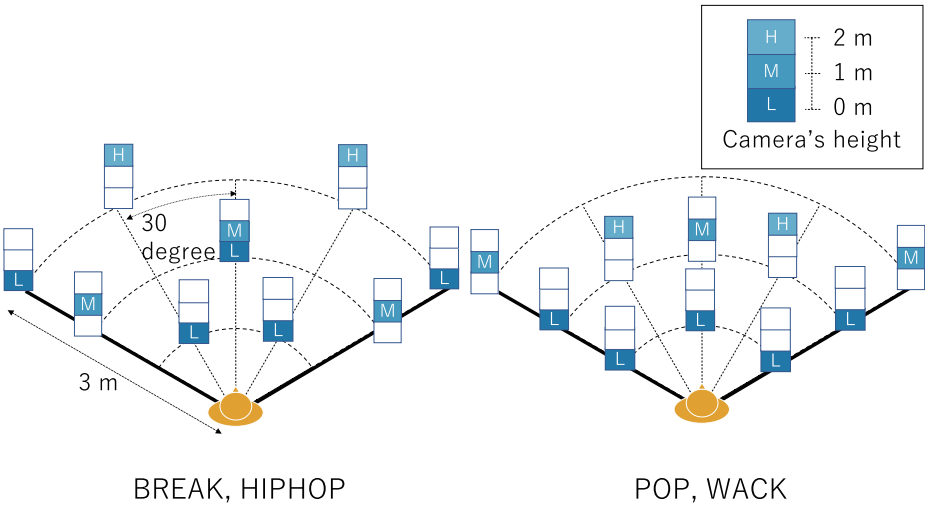

Fig. 4. Layout of the camera.

The subjects were 10 experts (8 males and 2 females) who work in video production or who were educated on video production. They had experience of between 1 and 20 years (on average, 8.6 years) since they were involved in video production. The subjects watched 14 videos for 2 performers and evaluated them in seven stages according to a Likert scale. For each item, the subject selected 1 if they could not agree with it at all, and 7 if they could agree very much with it. They watched the videos in the random order.

In addition to four questions on editing principles (items #1–4), the evaluation items have two items that are considered important in editing (#5–6), and one item that measures the goodness throughout editing (#7). We show the evaluation items as follows.

Q1: The video switches were between shots with different compositions.
Q2: The closer the shot, the shorter the cut length. The more distant the shot, the longer the cut length.
Q3: I can recognize the main camera.
Q4: The video has a varied pace to emphasizecertain cuts.
Q5: I feel dynamic.
Q6: I can grasp the overall situation.
Q7: This editing is attractive.

We conducted a questionnaire using Google Forms. The questionnaire was expected to take about 30 minutes. The subjects could take a break even while answering the questionnaire. In addition, subjects could review the video multiple times and reevaluate it to the previous video. While the subject watched

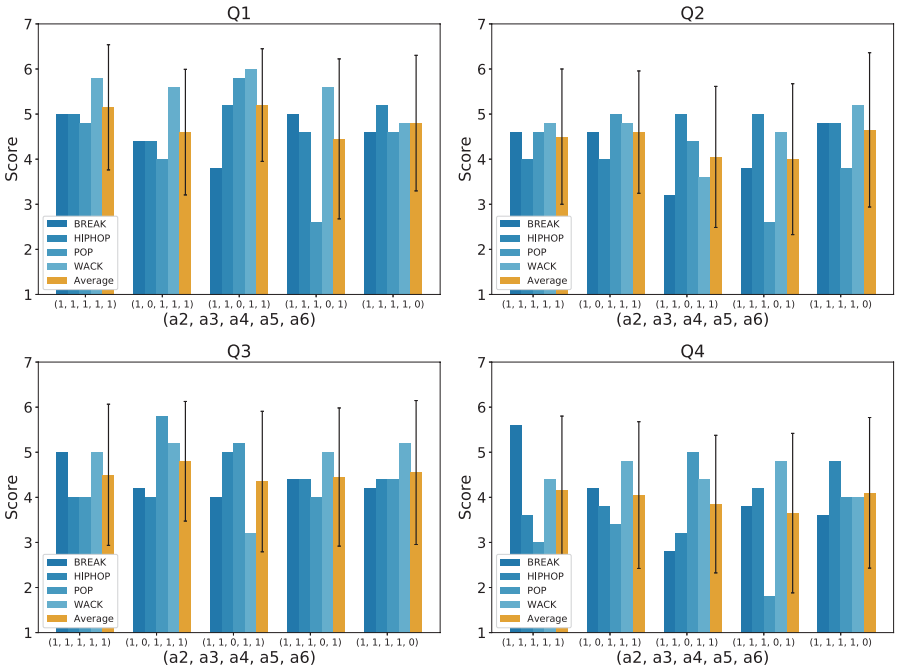

Fig. 5. Result of Q1∼Q4.

the video, the browser screen was maximized. When watching movies, the subjects wore earphones or headphones and we asked them to listen to the audio. After the subject finished all answers, we received their oral feedback about the system.

4.2 Results and Consideration

The results of Q1∼Q4 based on editing principles are shown in Fig. 5. The vertical axis indicates the average value of the answers to Q1∼Q4 given by all subjects. The vertical bars indicate standard deviation. The horizontal axis represents valid and invalid coefficients. It is expressed as 1 if the coefficient of constraint is more than 0, and 0 if it is 0.

In Q1, we assessed the difference between the average score for each constraint by using ANOVA. There was no significant difference; however, we look at all the coefficients (1,1,1,1,1) in Q1 are enabled, and the coefficients (1,0,1,1,1) corresponding to Q1 are disabled, and the score (1,0,1,1,1) is lower than the score (1,1,1,1,1) in all performers. This indicates that constraints are functioning based on principle due to an invalid constraint.

In Q2, we assessed the scores by using ANOVA. There was no significant difference; however, we look at all the coefficients (1,1,1,1,1) in Q2 are enabled,

and the coefficients (1,1,0,1,1) corresponding to Q2 are disabled, and the score (1,1,0,1,1) is lower than the score (1,1,1,1,1) in three of four performers. Constraints are not functional depending on performance.

In Q3, we assessed the scores by using ANOVA. There was no significant difference. We look at all the coefficients (1,1,1,1,1) in Q3 are enabled, and the coefficients (1,1,1,0,1) corresponding to Q3 are disabled, and there was no trend of change. In this time, the main camera was chosen with a probability of 60 percent as a constraint to recognize the main camera. However, it was not sufficient to recognize it as a main camera. It is necessary to increase the probability in order to make the difference in this item significant.

In Q4, we assessed the scores by using ANOVA. There was no significant difference. We look at all the coefficients (1,1,1,1,1) in Q4 are enabled and the coefficients (1,1,1,1,0) corresponding to Q4 are disabled, and there was no trend of change. In this time, we assumed that the attractive part of this time has intense movement. However, the parts that should be emphasized differ by subjects. Therefore, it seems that the trend could not be taken.

The results of Q5∼Q7 are shown in Fig. 6. Recall that the baseline is the method of using only optical flow.

Next, we look at the results of two items that are considered important in the editing extracted from preliminary interviews.

In Q5, we assessed the difference between the average score by using ANOVA. There was no significant difference. The coefficients (1,1,1,1,1) in Q5 and Baseline have high score. When all coefficients (1,1,1,1,1) are valid, the score tends to be high because the overall balance is good by selecting up shot when the performer move intense. On the other hand, in Baseline, the subjects feel dynamic because the cut was switched at short intervals.

In Q6, we assessed the difference between the average score by using ANOVA. There was a significant difference ($F_{(6,139)} = 5.54, p < .05$). Moreover, we assessed the difference by using an LSD test. Here, as well, there was significant differences between the score of Baseline and all except the constraint (1,1,0,1,1) ($p < .05$). In Baseline, the cut is switched at short intervals because the length of the cut is not considered. As a result, the overall situation was difficult to grasp and it received a low score. Additionally, there were significant differences between and between the score of constraints (1,1,1,0,1), (1,1,1,1,1) and the constraint (1,1,0,1,1) ($p < .05$). This shows that using Principle 4 was effective to recognize the main camera because the difference between the short cut length and the long one became clearer.

Finally, we look at the results of an item that measures goodness throughout editing. In Q7, we assessed the difference between the average score by using ANOVA. There was no significant difference. However, Compared with our system, Baseline score tends to be low, and our system is likely to be effective. In more detail, our system's scores are higher in BREAK and HIPHOP. It was not very effective in WACK, and Baseline tended to be higher in POP. Here, we examine the average value of optical flow for each performer (Fig. 7). The vertical axis indicates the average value of the optical flow for each beat. The

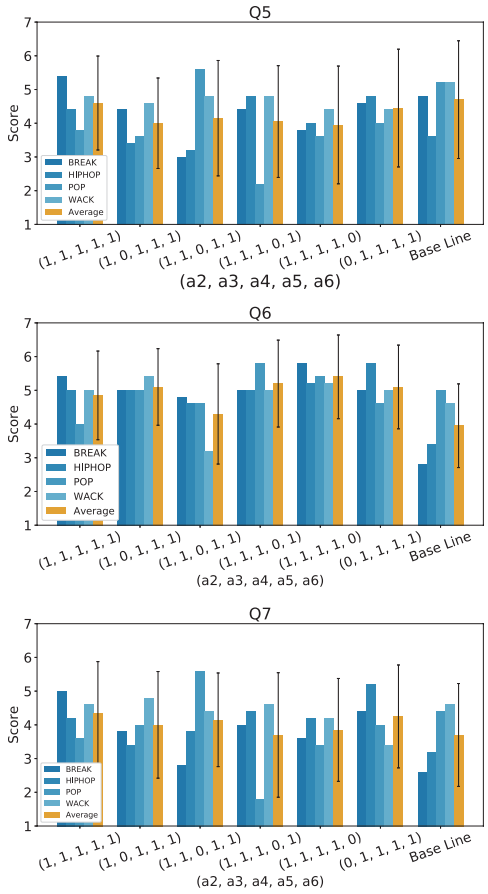

Fig. 6. Result of Q5∼Q7.

horizontal axis indicates beats. POP is a dance with a lot of movements like robotic dance. We can see that the average optical flow value is rapidly low intermittently in Fig. 7. In this system, the intense movement was taken as a part should be emphasized, and the system has the cases divided by the median of optical flow values. However, this is likely to cause the system to emphasize parts intermittently and affect other constraints. Moreover, POP has a lot of fine movement. Therefore, if there are few up shots, the video is hard to understand. Actually, in Q6, only POP has a tendency that Baseline is relatively easy to grasp. In this regard, we must consider that it is necessary to confirm the total amount of movement throughout the dance and to adjust the parameters.

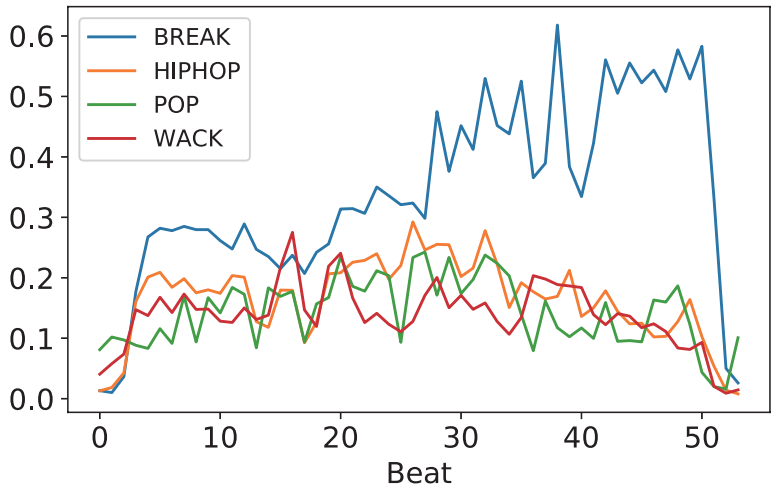

Fig. 7. Average optical flow value per beat.

5 Discussion

Limitations: This section explains some limitations of our system.

This system is based on human movement. Therefore, a human must be within the frame of the camera. Moreover, it is expected that this system does not work well for multiple people because it is for one person.

This system is vulnerable to beat tracking errors. If the beat tracking system is wrong, then the timing of camera switches will not be on beat.

Video materials must be taken with fixed cameras. In a moving camera, the motion of camera is captured as an optical flow. It is necessary to build a system that considers moving cameras.

The same song needs to be used when shooting movies. If there is a video material with a different song, the system cannot align the time axis of each video material. Moreover, just like the previous reason, the video should not contain large noise too. Our system has tried only 10 cameras; if there are too few or too many cameras, it may not work well. This can be handled by increasing the material by cropping, and limiting the number of videos selected.

At the present stage, it is necessary to determine the choreography part of the dance in advance. It is possible to think of a method to automatically recognize the choreography of dance, but there are various noises during shooting. For example, some performers felt out the mood beforehand, and half of them were in a state of dancing before the start of the choreography. It is considered that the cooperation of the performer is necessary, such as limiting the behavior before the performance and preparing the gesture at the beginning of the choreography.

As we described in the previous section, the system may not work well depending on the genre of dance.

Although our system has some limitations as described above, it can be sufficiently improved in the future.

Interview: I interviewed after the evaluation. Through interviews with 10 experts, we pick up topics that overlap among multiple people.

The most common opinion is that the original material is bad and there is little change in composition. One of the causes is the size of the shot. In this time, the size of the person on the screen was cropped to the ratio of 0.1 to 1.0. We got an opinion that the change in the composition was small because there was no more up shot, such as a face, foot, and hand, and that this made it difficult to evaluate videos. This can be handled by changing the crop size. However, the fixed cropped frame is likely to not follow the performer because the performer moves intensely. In the future, we need to implement a crop function that follows the movement automatically.

Moreover, we found that the influence of the arrangement of the camera was also large. In this study, 10 cameras were placed within 120 degrees in front of the dancer, 1∼3 m depth and 0∼2 m height. However, some experts pointed out that it is difficult to create a video that can produce a more clear difference from others, even when editing videos from cameras placed in this range. In the case of an expert, it is said that they set a shot that is taken out largely from Theory, such as a shot from the top of the performer and shot from the viewpoint of the performer, for example, as an accent. We received an opinion from experts that if we can improve the above two points on the material, we could make a video with more easily and distinctly differences.

In the future, it is considered worth working as a new research question on systems for collecting these good materials.

Experts pointed often out about the first cut and the last cut. Actually, our system does not consider the first cut or the last cut. For example, although the pose taken at the end of the performer's choreography was stopped, a lot of cuts with short interval was seen. It turned out that this was not good for editing. In the future, we will incorporate the concept of pose into the system, and the performer is considered to be a pose if the performer has stopped for more than a certain period of time, and we consider a method to select the camera of the best viewing position. As a comment related to the pose, there was an opinion that when the performer's movement was slow, there was a bad impression when the cut was switched. We need to investigate the relationship between the intensity of the movement and the length of the cut.

We have also received many opinions that the link with music is a very important point. We received an opinion that it was attractive editing when the viewer felt that the movement and music were matched. In the future, we need to examine the balance between the movement of a shot and the information obtained from music.

Principles and Constraints: We compared the score of when all constraints were satisfied and the score of when one constraint was omitted and

could not confirm a significant difference. However, as there were some constraints that showed trends, it is likely that those constraints are functioning. In addition, even if the constraint is omitted, the combined effects of other constraints might compensate the principle. In the future, we need to do a wider range of investigation, such as comparing a constraint which enables only one coefficient and others. Furthermore, we could not confirm the relationship between the constraints and the overall evaluation. In this regard, we need to examine the video materials so that there is a clear difference.

Interactivity: The numerical values on the system were determined experimentally in advance because it was difficult to verify all parameters. The user can also control these parameters interactively. For the minimum length and maximum length of the cut of Principle 2, the minimum and maximum length should be set to shorten the length of the cut, if a user preferring to shorten the cut length. If the user prefers a long cut, he/she can set the minimum and maximum length to increase the length of the cut overall. In principle 5, we have built a system to determine the main camera automatically. However, there is also a method that the user can select this main camera by oneself and he/she can adjust the percentage of the main camera in the video. In principle 6, the parts to emphasize are currently determined while looking at the optical flow value. However, there is a method that the user can also select a range of the parts. Our system has chosen a video that maximizes the evaluation function. However, the user can choose a video that matches the user's preferences from several candidates that the system emits. In addition, it is possible to add a function that can easily modify the generated video, such as by switching the cut and adjusting the length.

Expandability: In the future, we can cite various expandability in developing systems. At present, our system assumes that the camera is fixed. Therefore, the camera is not capable of moving. If the system is capable of moving the camera, a system can be used to shoot the dance performance at an optimum angle while moving the camera with a drone. However, the system is difficult to implement in real-time because of slow computation speed. If the processing speed is improved, it can be adapted to live performance streaming. For example, the system could acquire videos from a number of smartphones that are held by an audience who watches a performance on the street. The system can conduct a streaming distribution that automatically adjusts the cut of them and selects the attractive video.

Various methods can be considered, such as generating effects according to performers' movement, or applying effects that are more emphasized on the important parts. In the interview, there were experts who said that there is knowledge about the arrangement of cameras and editing of videos, but knowledge about effects is poor. If we can propose an accurate effect to enhance the attractiveness of the video, there is a need for such an expert. Moreover, the camera work of the proposed system can also be applied to VR space. A system is considered to automatically transition the camera on VR space to show the dancing virtual character better. In addition, it is also specialized for a dance

performance. However, it can be applied to video including other sports. For example, in weight training, we can see online a number of videos to learn how to train. If there is a system that makes weight training more attractive and sophisticated, it can be adapted to such videos.

6 Conclusion

We focus on video editing, and propose a system that automatically generates more attractive and sophisticated dance videos even for non-specialists. The system creates a single dance movie by automatically editing performance videos taken from multiple locations according to the probability model created based on principles of video editing and principles extracted from preliminary interviews. We conducted an evaluation experiment. The system automatically edits the dance video of four performers taken from multiple viewpoints and 10 experts in video production evaluated the generated videos. As a result, the system tended to be better evaluated than other automatic editing method. In addition, we developed discussions based on interviews and showed the system expandability. We will further extend this system and aim to launch a web service based on the system.

References

1. Heck, R., Wallick, M., and Gleicher, M. Virtual Videography. In: ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Vol. 3(1), No. 4. ACM, New York (2007).
2. Arev, I., Park, H. S., Sheikh, Y., Hodgins, J., and Shamir, A.: Automatic Editing of Footage from Multiple Social Cameras. In: ACM Transactions on Graphics (TOG), Vol. 33(4), No. 81. ACM, New York (2014).
3. Ranjan, A., Birnholtz, J., and Balakrishnan, R. Improving Meeting Capture by Applying Television Production Principles with Audio and Motion Detection. In Proc. CHI 2008, pp. 227–236. ACM, New York (2008).
4. Zsombori, V., Frantzis, M., Guimaraes, R. L., Ursu, M. F., Cesar, P., Kegel, I., Craigie, R., and Bulterman, D. C.: Automatic Generation of Video Narratives from Shared UGC. In: Proc. Hypertext 2011, pp. 325–334. ACM, New York (2011).
5. Lu, Z., and Grauman, K. Story-driven Summarization for Egocentric Video. In: Proc. CVPR 2013, pp. 2714–2721. IEEE, New York (2013).
6. Jain, E., Sheikh, Y., Shamir, A., and Hodgins, J.: Gaze-driven Video Re-editing. In: ACM Transactions on Graphics (TOG), Vol. 34(2), No. 21. ACM, New York (2015).
7. Shin, H. V., Berthouzoz, F., Li, W., and Durand, F. Visual Transcripts: Lecture Notes from Blackboard-style Lecture Videos. In: ACM Transactions on Graphics (TOG), Vol. 34(6), No. 240. ACM, New York (2015).
8. Kumar, M., Gandhi, V., Ronfard, R., and Gleicher, M.: Zooming On All Actors: Automatic Focus+ Context Split Screen Video Generation. In: Computer Graphics Forum, Vol. 36, No. 2, pp. 455–465. The Eurographics Association, Geneva (2017).
9. Sun, X., Foote, J., Kimber, D., and Manjunath, B. S.: Region of Interest Extraction and Virtual Camera Control Based on Panoramic Video Capturing. In: IEEE Transactions on Multimedia, Vol. 7(5), pp. 981–990. IEEE, New York (2005).

10. Roininen, M. J., Leppnen, J., Eronen, A. J., Curcio, I. D., and Gabbouj, M.: Modeling the Timing of Cuts in Automatic Editing of Concert Videos. In: Multimedia Tools and Applications, Vol. 76(5), pp. 6683–6707. Springer, Heidelberg (2017).
11. Mate, S. and Curcio, I. D: Automatic Video Remixing Systems. In: IEEE Communications Magazine, Vol. 55(1), pp. 180–187. IEEE, New York (2017).
12. Truong, A., Berthouzoz, F., Li, W., and Agrawala, M.: QuickCut: An Interactive Tool for Editing Narrated Video. In: Proc. UIST 2016, pp. 497–507. ACM, New York (2016).
13. Leake, M., Davis, A., Truong, A., and Agrawala, M.: Computational Video Editing for Dialogue-Driven Scenes. In: ACM Transactions on Graphics (TOG), Vol. 36(4), No. 130. ACM, New York (2017).
14. Jeong, K. A., and Suk, H. J.: Jockey Time: Making Video Playback to Enhance Emotional Effect. In: Proc. of the 2016 ACM on Multimedia Conference, pp. 62–66. ACM, New York (2016).
15. Nam, T. J., Lee, J. H., Park, S., and Suk, H. J.: Understanding the Relation Between Emotion and Physical Movements, In: International Journal of Affective Engineering, Vol. 13, No. 3, pp. 217–226 (2014).
16. Montepare, J. M., Goldstein, S. B., and Clausen, A.: The Identification of Emotions from Gait Information. In: Journal of Nonverbal Behavior, Vol. 11(1), pp. 33–42. Springer, Heidelberg (1987).
17. Foust, J. C., Fink, E. J., Gross, L. S.: Video Production: Disciplines and Techniques. Taylor and Francis, Abingdon (2012).
18. Bowen, C. J.: Grammar of the Edit. Taylor and Francis, Abingdon (2013).
19. Böck, S., Krebs, F., and Widmer, G.: Joint Beat and Downbeat Tracking with Recurrent Neural Networks. In: ISMIR, pp. 255–261 (2016).
20. Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y.: Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In: CVPR (2017)