Scene Segmentation of Wedding Party Videos by Scenario-based Matching with Example Videos

Kazuki Sawai[*]
Graduate School of I.S.,
Nagoya University
ksawai@murase.m.is.nagoya-u.ac.jp

Tomokazu Takahashi
Faculty of Econ. & Info.,
Gifu Shotoku Gakuen Univ.
ttakahashi@gifu.shotoku.ac.jp

Daisuke Deguchi
Graduate School of I.S.,
Nagoya University
ddeguchi@is.nagoya-u.ac.jp

Ichiro Ide
Graduate School of I.S.,
Nagoya University
ide@is.nagoya-u.ac.jp

Hiroshi Murase
Graduate School of I.S.,
Nagoya University
murase@is.nagoya-u.ac.jp

ABSTRACT

We propose a method for scene segmentation of a wedding party video. Recently, it has become popular to take videos of a wedding ceremony and its party. Especially, because of its length, each scene of a wedding party video needs to be indexed with each event for efficient browsing. The proposed method segments a wedding party video into scenes of events by scenario-based matching with example videos that are synthesized by combining scenes from other wedding party videos according to a scenario.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Video; H.3.1 [**Content Analysis and Indexing**]: Indexing methods; I.5.4 [**Applications**]: Computer vision

General Terms

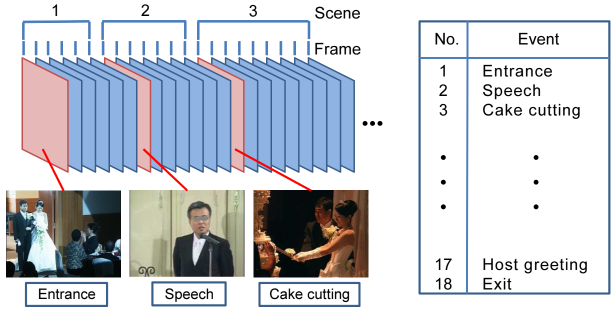
Algorithms

Keywords

Wedding party, scene segmentation, dynamic time warping

1. INTRODUCTION

Wedding is one of the most important events in a person's life. With the rapid popularization of video cameras, it has become popular to take videos of a wedding ceremony and its party. Especially, wedding parties takes a long time compared to wedding ceremonies. Therefore, to quickly browse video segments of particular events, each video segment needs to be indexed with each event in the wedding party. Currently, this process is performed as a part of services by bridal agencies. However, there is a demand to perform this automatically because it takes much cost in terms of effort and time.



(a) Video

No.	Event
1	Entrance
2	Speech
3	Cake cutting
•	•
•	•
•	•
17	Host greeting
18	Exit

(b) Scenario

Figure 1: Video of a wedding party and its scenario.

As shown in Fig. 1 (a), a wedding party video consists of various scenes. Here, we use the term "scene" to represent a video segment associated with an event. Generally, a wedding party is organized along a scenario which describes the order of events as shown in Fig. 1 (b). The proposed method segments a wedding party video into scenes and indexes them by associating the scenes with the events based on the scenario.

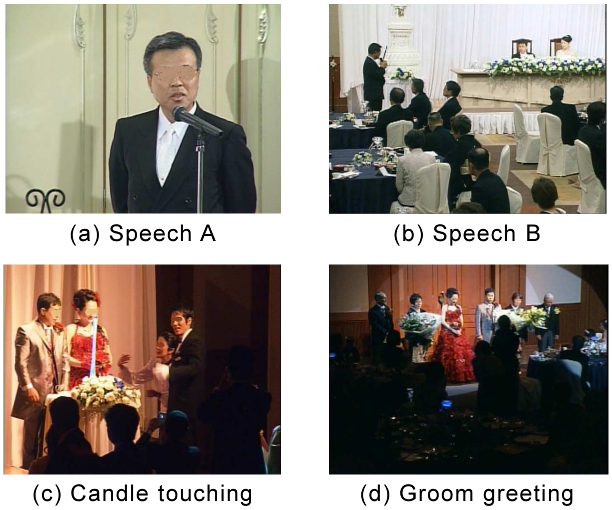
Wedding parties share the following characteristics:

- The same kinds of events are performed in various ways and length.
- Different kinds of events with similar characteristics could be performed.

Figures 2 (a) and (b) show sample images of the same kind of events in different wedding parties. Although these images are captured in the same kind of events, they differ in composition, lighting, background, and subject. Therefore, the proposed method selectively uses appropriate audio

[*]Currently at NTT COMWARE CORPORATION.
[*]Area chair: Qi Tian

MM'11 November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.



(a) Speech A

(b) Speech B

(c) Candle touching

(d) Groom greeting

Figure 2: Sample images of wedding party events.

and visual features for each kind of events. The proposed method also employs a segmentation strategy that could accept the difference of the length in the same kind of events. On the other hand, Figs. 2 (c) and (d) show sample images of different events in the same wedding party. Their characteristics are similar although these images are captured in different kinds of events. To improve segmentation accuracy in such cases, the proposed method uses the information of the order of events written in the scenario.

2. RELATED WORK

Cheng et al. [1] proposed a method for scene segmentation of wedding ceremony videos. This method segments a wedding ceremony video into scenes associated with events using audio and visual features. To obtain the features, various techniques are used, such as speech/music classification [3] and camera flash detection. The proposed method also uses these techniques because their features would also be effective for scene segmentation of wedding party videos. The method in [1] constructs a single common model of temporal variation of the features for each kind of events to achieve scene segmentation with a model-based matching method. However, it is difficult in case of wedding parties to construct a single model for each kind of events because the events are performed in various ways and length. Therefore, the proposed method takes not the model-based approach but an example-based approach.

The proposed method calculates the dissimilarity between two video segments based on a DTW (Dynamic Time Warping) method. The DTW method is a sequence matching method that could be applied to two sequences with different lengths, and has also been used for video segment matching [2, 4, 6]. When two video segments $\mathbf{X}$ and $\mathbf{Y}$ are represented as sequences of feature vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_P)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_Q)$, the dissimilarity is calculated by using the DTW method as below. First, the dissimilarity between two vectors $\mathbf{x}_p$, $\mathbf{y}_q$ is defined as

$$d(p,q) \quad = \quad \sqrt{\textstyle\sum_{n=1}^{N} (x_{p,n} - y_{q,n})^2}, \qquad (1)$$

where $x_{p,n}$ and $y_{q,n}$ represent the n-th element of $\mathbf{x}_p$ and $\mathbf{y}_q$, respectively. The dissimilarity $D(\mathbf{X}, \mathbf{Y})$ is calculated as $D(P,Q)$ by the following recurrence formula:

$$D(1,1) \quad = \quad d(1,1), \qquad (2)$$

$$D(p,q) \quad = \quad \min \begin{cases} D(p-1,q) + \omega_1 d(p,q) \\ D(p-1,q-1) + \omega_2 d(p,q) \\ D(p,q-1) + \omega_3 d(p,q) \end{cases} . \qquad (3)$$

Here, ω_1, ω_2, and ω_3 are weight factors for the matching.

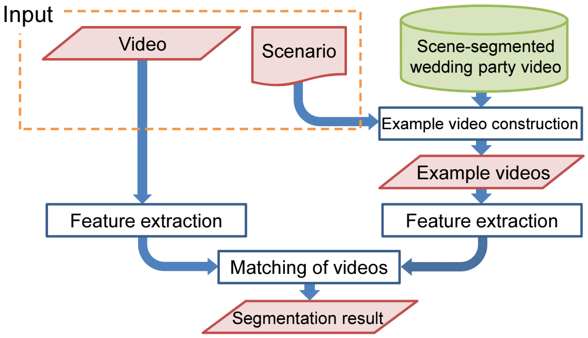


Figure 3: Flow diagram of the proposed method.

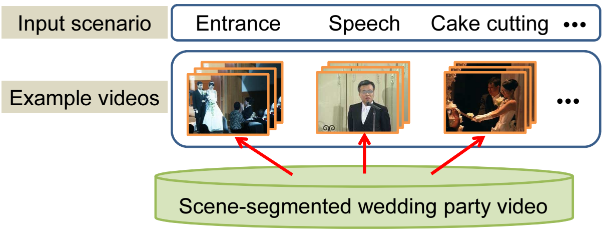


Figure 4: Construction of example videos.

3. PROPOSED METHOD

Figure 3 shows the flow diagram of the proposed method. The proposed method consists of the following three parts.

3.1 Construction of example videos

Figure 4 illustrates the construction of the example videos. As a preparetory process, multiple wedding party videos other than the input video are segmented into scenes of events manually. Thus we obtain a large number of labeled scenes $\{\mathbf{S}_{e,i_e} | e = 1, \ldots, E, i_e = 1, \ldots, I_e\}$. Here, E represents the number of events, I_e, the number of labeled scenes of an event e. The proposed method constructs example videos by arranging the labeled scenes along the event sequence $\mathbf{e} = (e_1, \ldots, e_M)$ described in the input scenario.

3.2 Feature extraction

Multiple audio and visual features are used. A feature vector is constructed from the features extracted from each unit segment consisting of T frames; a video is represented as (number of frames)$/T$ feature vectors. The proposed method extracts the features as follows:

Face detection

Faces are useful information to recognize events. The proposed method uses the average and variance of the number, size, and position of faces as visual features. A famous face detector [7] is used to detect the faces.

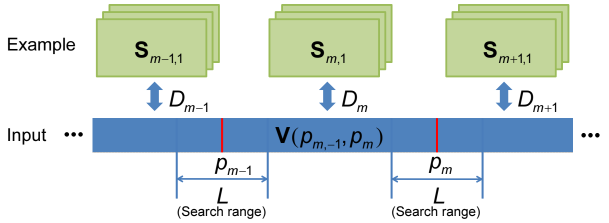


Figure 5: Matching between input and example videos.

Camera flash detection

Many camera flashes are observed in important events. Therefore, the number of flashes and the maximum difference of pixel intensities before and after the flashes are extracted by using a flash detection method [5].

Pixel intensities

Various lighting effects are used depending on events. Also, most male guests wear black suits and cover a large area of a frame in particular events. Therefore, the proposed method uses the average and the variance of pixel intensities.

Speech/music classification

Music plays loudly in some events. On the other hand, a person speaks in other events. The proposed method classifies sound sequences into speech, music, and silence by using the method proposed in [3], then uses the classification result as an audio feature. The root mean square and zero crossing frequency of speech power are used in the classification, which are also used as audio features.

3.3 Matching between videos

3.3.1 Dissimilarity between video segments

The proposed method calculates the dissimilarity between two video segments based on the DTW method described in Section 2. The effective features for matching between example and input videos depend on events. Therefore, to improve the accuracy of the matching, the proposed method uses a weighted Euclidian distance instead of Equation (1):

$$d(p,q) \quad = \quad \sqrt{\textstyle\sum_{n=1}^{N} w_{e,n}(x_{e,p,n} - y_{q,n})^2}, \qquad (4)$$

where $x_{e,p,n}$ is the n-th element of the p-th vector of a scene of an event e in example videos, $y_{q,n}$, the n-th element of the q-th vector of a video segment of an input video, $w_{e,n}$, the n-th element of the weight vector for the event e.

The weight vectors $\mathbf{w}_e(e = 1, \dots, E)$ are learned beforehand for each event from a large number of labeled scenes in example videos.

3.3.2 Scene segmentation

Scene segmentation is performed by the following steps:

Step 1. **Initial segmentation**

We represent the beginning and the ending times of the m-th event in the input scenario as p_{m-1} and p_m ($p_{m-1} < p_m, m = 1, \dots, M$), respectively. Here, p_0 and p_M are given as the beginning and the ending times of the input video. The initial segment positions $\mathbf{p}_0 = (p_0, \dots, p_M)$ are calculated by the following equation:

$$\mathbf{p}_0 = \arg\min_{\mathbf{p}} \textstyle\sum_{m=1}^{M} \frac{(p_m - p_{m-1} - \mu_m)^2}{\sigma_m^2}, \qquad (5)$$

Table 1: Event names and number of scenes (Total: 216).

Event name (Number of scenes)
Bouquet casting (17), Bride and groom entrance (5),
Bride and groom exit (15), Cake cutting (5),
Candle touching (15), Congratulatory telegram (8),
Exit for changing costumes (7), Groom greeting (14),
Guest speech (22), Host greeting (14), Intermission (19),
Message from bride to parents (15), Other show (19),
Photo session (3), Singing show (22), Slide show (16)

where μ_m and σ_m^2 represent the mean length and the variance of an event e_m, respectively, which are calculated beforehand from labeled scenes of example video for each event.

Step 2. **Updating by video segment matching**

Figure 5 shows an overview of the video segment matching. $\mathbf{S}_{m,i}$ denotes the i-th labeled scene of an event e_m in the example videos. On the other hand, $\mathbf{V}(p_{m-1}, p_m)$ denotes a video segment from p_{m-1} to p_m in the input video. This step finds the position of each event that minimizes the dissimilarity between input and example videos by searching for p_m from a range in a width of L. The positions $\mathbf{p}_{t+1} = (p_0, \dots, p_M)$ are obtained by the following equation:

$$\begin{aligned}\mathbf{p}_{t+1} = \arg\min_{\mathbf{p}_t} \min_{\mathbf{i}} \textstyle\sum_{m=1}^{M} \Big\{ & \frac{(p_m - p_{m-1} - \mu_m)^2}{\sigma_m^2} \\ & + \alpha_m D(\mathbf{S}_{m,i}, \mathbf{V}(p_{m-1}, p_m)) \Big\} \,. \end{aligned} \qquad (6)$$

Here, α_m represents the weight of the event e_m, which is learned beforehand from example videos for each event so that Equation 6 can be minimized.

Step 3. **Iteration**

The segmentation process increments t by one and returns to Step 2 if the difference between $\mathbf{p}_{t+1}$ and $\mathbf{p}_t$ is greater than a threshold, otherwise outputs $\mathbf{p}_{t+1}$ as the results.

4. EXPERIMENT

Two experiments were conducted on actual wedding party videos. The first experiment evaluated the accuracy of scene segmentation. On the other hand, the second experiment investigated the effectiveness of each of the visual and the audio features used in the proposed method.

4.1 Setup

15 videos from different wedding parties were used for the experiments. Each video was 1.2 to 1.8 hours long and included 11 to 19 events. Table 1 shows event names and the number of scenes included in the videos. For example, 17 scenes of different bouquet casting events were included in the 15 videos. The unit video length for the segmentation was set to $T = 10$ sec. and the range width for the video segment matching was set to $L = 150$ sec. For each video, we used the video for input and the other 14 videos for the construction of example videos.

To evaluate the segmentation accuracy, we calculated the success rate by the following equation:

$$\text{success rate [\%]} = \frac{\text{number of correct units}}{\text{number of all units}} \times 100. \qquad (7)$$

As a comparative method, we used a method that performs only the initial segmentation step described in 3.3.2.

Table 2: Segmentation accuracy.

Method	Proposed	Comparative
Success rate [%]	82.6	69.8

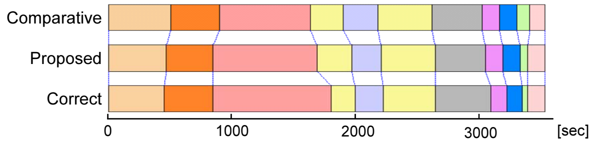


Figure 6: Example of successful segmentation results.

4.2 Segmentation accuracy

Table 2 shows the segmentation accuracy of the proposed method. This result indicates that the scenario-based matching with example videos improved the segmentation accuracy. From this result, we confirmed the effectiveness of the proposed method. Figure 6 shows an example of successful segmentation results. Each colored segment represents each segmented scene. In this case, the success rates of the comparative method and the proposed method were 84.5% and 95.7%, respectively. Similarly, Fig. 7 shows an example of segmentation results with low success rates. In the case where the same kinds of events occurred successively, it is difficult to achieve accurate segmentation by video segment matching based on the dissimilarity of visual and audio features because similar features are extracted from them. Other approaches would be required to solve this problem, such as scene boundary detection using speech detection of the MC and handclap detection. In the case where the initial segmentation result differed largely from the correct one, the result may fall into a local solution before reaching the correct answer. If estimated times of events are described in the wedding party scenario, use of this information could be effective to improve the accuracy.

4.3 Effectiveness of visual and audio features

Figure 8 shows the segmentation accuracy when only one kind of feature was used. For each kind of features, the accuracy was higher than that of the comparative method. The proposed method which is a combination of features, was even better than any of the features used indivisually. From this result, we confirmed the effectiveness of the features used in the proposed method.

5. SUMMARY

We proposed a method for scene segmentation and indexing of a wedding party video. The proposed method segments the video into scenes associated with events by scenario-based video segment matching between input and example videos. Experiments were conducted on actual wedding video party videos. As a result, the proposed method achieved an accuracy of 82.6% for scene segmentation.

Future work will include further improvement of the segmentation accuracy by use of a scene boundary detection technique and the information of estimated times of events planed in the wedding party scenario.

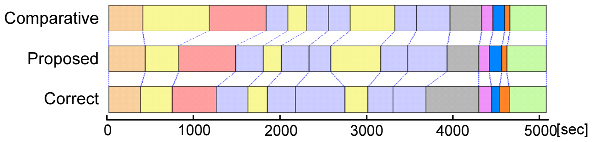


Figure 7: Example of failed segmentation results.

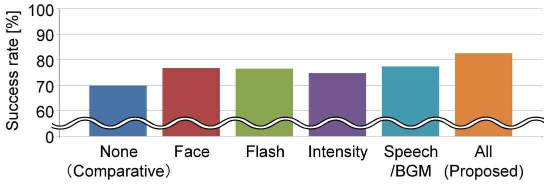


Figure 8: Success rates for each kind of features.

6. ACKNOWLEDGMENTS

Parts of this work were supported by the Grants-in-Aid for Scientific Research. The videos used in the experiments were provided by ClimBcube Co., Ltd.

7. REFERENCES

[1] W.-H. Cheng, Y.-Y. Chuang, and M.-C. Tien. Semantic-event based analysis and segmentation of wedding ceremony videos. In *Proc. Int. Workshop on Multimedia Information Retrieval 2007*, pages 95–104, March 2007.

[2] C.-Y. Chiu, C.-H. Li, H.-A. Wang, C.-S. Chen, and L.-F. Chien. A time warping based approach for video copy detection. In *Proc. IAPR 18th Int. Conf. on Pattern Recognition*, pages 228–231, September 2006.

[3] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on RMS and zero-crossings. *IEEE Trans. Multimedia*, 7(1):155–166, February 2004.

[4] J. Sato, T. Takahashi, I. Ide, and H. Murase. Change detection in streetscapes from GPS coordinated omni-directional image sequences. In *Proc. IAPR 18th Int. Conf. on Pattern Recognition*, pages 935–938, August 2006.

[5] M. Takimoto, S. Satoh, and M. Sakauchi. Identification and detection of the same scene based on flashlight patterns. In *Proc. 2006 IEEE Int. Conf. on Multimedia and Expo*, pages 9–12, January 2006.

[6] H. Uchiyama, D. Deguchi, T. Takahashi, I. Ide, and H. Murase. Ego-localization using streetscape image sequences from in-vehicle cameras. In *Proc. 2009 IEEE Intelligent Vehicles Symposium*, pages 185–190, June 2009.

[7] P. Viola and M. Jones. Robust real-time face detection. *Int. J. Computer Vision*, 57(2):137–154, May 2004.