Scene Duplicate Detection from News Videos Using Image-Audio Matching Focusing on Human Faces

Haruka Kumagai*, Keisuke Doman†, Ichiro Ide*, Daisuke Deguchi‡ and Hiroshi Murase*
*Graduate School of Information Science, Nagoya University, Nagoya, Aichi, Japan
Email: {ide,murase}@is.nagoya-u.ac.jp
†School of Engineering, Chukyo University, Toyota, Aichi, Japan
Email: kdoman@sist.chukyo-u.ac.jp
‡Information and Communications Headquarters, Nagoya University, Nagoya, Aichi, Japan
Email: ddeguchi@nagoya-u.jp

***Abstract*—As one tool for structuring a massive volume of archived news videos based on their semantic contents, this paper proposes a method to detect scene duplicates from news videos. A scene duplicate is a pair of video segments taken at the same event from different viewpoints. Referring to the audio channel is effective to detect scene duplicates regardless of viewpoints, but it cannot be relied on when external audio sources (e.g. narrations, sound effects) overlap the original one. In contrast, the image channel can be useful in most cases, although significant difference in viewpoints affect the detection. The proposed method integrates the information from these two channels in order to improve the accuracy of scene duplicate detection from news videos. The performance of the proposed method was evaluated through an experiment with actual broadcast news videos. As a result, we obtained the higher detection accuracies in both recall and precision. Therefore, we confirmed the effectiveness of the proposed method.**

***Keywords*-Scene duplicate detection; difference of viewpoints; collection of speeches; news video**

I. INTRODUCTION

In recent years, it has become easy to archive a massive volume of broadcast videos. Among them, news videos are especially worthy to be archived since various real-world events are reported in them. In order to allow easy access to news videos in an archive as reference data, it is necessary to handle them based on their contents. As one tool for such a purpose, in this paper, we focus on a technique for detecting a "scene duplicate" which is a pair of video segments taken at the same event from different viewpoints.

As shown in Fig. 1, a scene duplicate is a type of near duplicate. Each of the types are briefly described below:

Strict Near-duplicate: This is a pair of video segments taken at the same event from the same viewpoint, as shown in Fig. 1(a). The differences caused by editing and/or captioning are ignored.

Object Duplicate: This is a pair of video segments taken at different events, as shown in Fig. 1(b). Viewpoints do not matter for this type.

Scene Duplicate: This is a pair of video segments taken at the same event from different viewpoints, as shown in Fig. 1(c). This type is often observed when several broadcast stations shoot videos at the same event with different cameras.

In this paper, we focus on detecting only scene duplicates. If scene duplicates could be detected automatically, we can watch news videos taken from different viewpoints for the same event [1], [2]. We can also automatically make a video collection of speeches by a particular person combined with a face detection technology, which is one of the possible applications of this research. Therefore, detecting a scene duplicate is important to efficiently understand various real-world events.

There have been many methods for detecting near-duplicates [3], and some of them aim at detecting scene duplicates. For example, Takimoto et al. proposed a flash-based method [4]. This method, however, cannot be applied to shots without flashes. On the other hand, Wu et al. proposed a method based on the discontinuity of feature point trajectories [5]. This method evaluates the speed and the change of motion of a subject, which makes it robust to the difference of viewpoints. This method, however, has mainly two problems:

Problem 1: It is difficult to detect scene duplicates containing a non-uniform background texture, as shown in Fig. 2. In such a case, the speed and the change of the motion of a subject cannot be evaluated correctly, since most feature points exist in the background region, and consequently, the number of feature points extracted from the subject is relatively small.

Problem 2: It is difficult to detect scene duplicates if their viewpoints are significantly different from each other, which is the fundamental problem of an appearance-based method. Again, in such a case, the speed and the change of the motion of a subject cannot be evaluated correctly.

In this paper, we propose a solution to the above-mentioned problems by limiting the target to face shots in news videos, and developing a method to detect scene dupli-

978-0-7695-5437-2/14 $31.00 © 2014 IEEE
DOI 10.1109/ISM.2014.43

(a) Strict Near-duplicate: Pair of video segments taken at the same event from the same viewpoints

(b) Object Duplicate: Pair of video segments taken at different events (Viewpoints do not matter)

(c) Scene Duplicate: Pair of video segments taken at the same event from different viewpoints

Figure 1. Types of near duplicate video segments.

(a) Frontal viewpoint

(b) Diagonal viewpoint

Figure 2. Example of a scene duplicate with a small number of feature points detected in the person area.

cates from news videos by integrating the information from the audio and the image channels. Referring to the audio channel is effective to detect scene duplicates regardless of viewpoints, although it cannot be relied on when external audio sources (e.g. narrations, sound effects) overlap the original one. In contrast, the image channel can be useful in most cases, but significant difference in viewpoints affect the detection. By integrating the information from the two channels, we expect to solve Problem 2 in Wu et al.'s method, and improve the accuracy of the scene duplicate detection. In addition, we expect to solve Problem 1 by comparing the image features in the face region and the background region separately.

The paper is organized as follows. First, Section II describes the proposed method for detecting scene duplicates from news videos. Next, Section III reports the results of experiments to evaluate the effectiveness of the proposed method, and Section IV provides discussions. The paper concludes with a summary and future work in Section V.

II. PROPOSED METHOD

The proposed method expects as input, a pair of sets of images and audio from a shot and a closed caption (CC) corresponding to the story that contains the shot, as described in Fig. 3. Note that a story is defined here as the minimum semantic unit of a news video involving one event.

The process flow of the proposed method is shown in Fig. 4. First, the input is filtered according to the similarity of CCs. Next, a scene duplicate is detected from the filtered input pair according to the integrated similarity between audio and image features. If the similarity of either the audio or the image features is higher than a threshold, the input

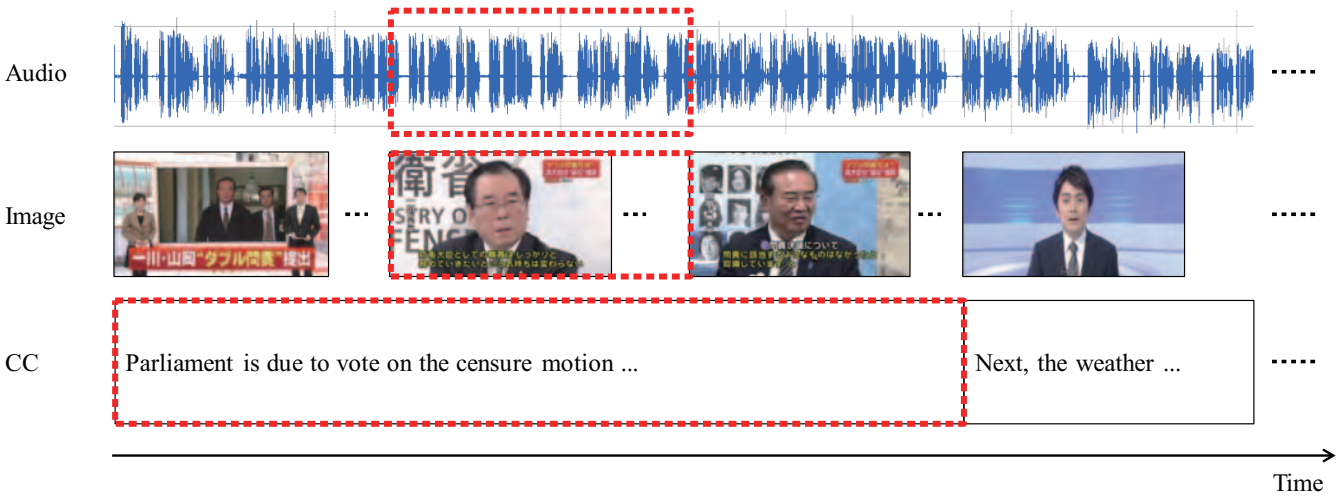


Figure 3. Input data of the proposed method for one scene. The dashed line indicates the input data for each media.

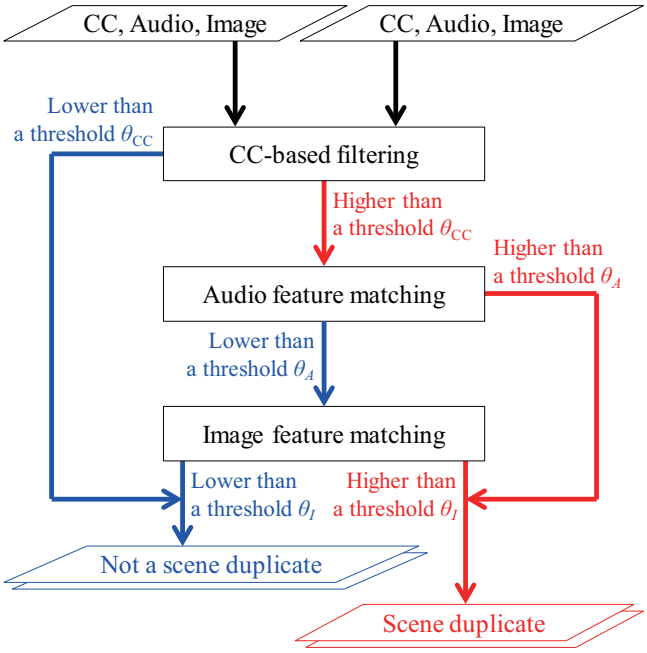


Figure 4. Process flow of the proposed method.

pair is regarded as a scene duplicate.

We design the proposed framework considering the characteristics of each step. As mentioned in Section I, the audio and the image channels have different characteristics, so we expect that the audio feature matching and the image feature matching steps should compensate for each other's weak points. CC-based filtering with a very low threshold is introduced as a pre-processing step in order to discard non-scene duplicates with high recall and small computation power. Audio feature matching and image feature matching follow. The former can detect a scene duplicate with high precision if external audio sources (e.g. narrations, sound effects) do not overlap the original ones. The latter can accurately detect a scene duplicate without significant difference in viewpoints. In the proposed method, the former is first applied and then the latter is secondly applied because the computational cost of the former is lower than that of the latter. Details of each step are described below.

A. *CC-based filtering*

First, morphological analysis is performed. Next, nouns and undefined words are extracted. Then, a feature vector is composed based on the term frequency for each input CC. Finally, the cosine distance between the feature vector pair is calculated as the similarity of the input CCs. If the similarity is lower than a threshold θ_{CC}, the input pair is discarded as not a scene duplicate.

B. *Audio feature matching*

First, in order to reduce noise, the input audio signals are filtered by a band-pass filter with a pass-band of $[0.3, 4]$ kHz considering the frequency band of human voice. Next, the signals are resampled at the rate of 8 kHz in order to reduce the processing time. Furthermore, the volume is normalized. Finally, the cross-correlation between the signal pair is calculated by shifting one of them. We used the maximum correlation as the similarity of the input audio pair. If the similarity is higher than a threshold θ_A, the pair is regarded as a scene duplicate.

C. *Image feature matching*

Details of the image feature matching process are as follows:

(a) Detect a face from an input frame.

(b) Split regions according to the face region.

(c) Crop the upper and the lower parts of each region (The center region is the face region, and the others are background regions).

Figure 5. Process flow of the region segmentation.

1) Region segmentation: First, as shown in Fig. 5(a), the face of a subject is detected from each frame of the input shots. Next, as shown in Fig. 5(b), each input shot is horizontally divided into three regions based on the detected face region. Then, as shown in Fig. 5(c), the background region and the face region are obtained by cropping the upper and the lower parts of each region, in order to avoid the overlap with superimposed captions. Note that only the first frame of the input shot is used for the similarity calculation in the background region, assuming that the appearance change in background regions should be small within a shot.

2) Region matching: The similarity S_f between the face region pair, and the similarity S_b between the background region pair are calculated.

Face region matching: First, each face region is vertically divided into three sub-regions; top, middle and bottom, as shown in Fig. 6. This is to prevent image feature matching

Figure 6. Dividing the face region into sub-regions.

between different facial parts. Next, considering the possible occlusion by microphones and so on, the similarity of each sub-region pair except for the bottom sub-region is calculated using Wu et al.'s method [5].

Wu et al.'s method calculates the similarity based on the discontinuity of feature point trajectories in videos. The trajectory represents the speed and the change of a subject's motion, which makes it robust to the non-significant difference of viewpoints.

Then, the similarities are combined as follows:

$$S_f = \alpha S_{f_2} + (1-\alpha) S_{f_1}, \tag{1}$$

where S_f is the similarity of the face region pair in the input shot pair, S_{f_2} and S_{f_1} are the similarities calculated by applying Wu et al.'s method [5] to the middle sub-region pair and the top sub-region pair, respectively. The parameter α is a weight coefficient with a value of [0, 1]. The proposed method ignores the similarity of bottom sub-regions, because they are often occluded with a microphone and noisy.

Background region matching: The similarity of the background region pair is calculated based on their color histograms. Here, as shown in Fig. 7, the brightness of a scene duplicate in news videos may differ depending on adjustments or editions by broadcast stations. In order to cope with this problem, the similarity S_b between the background region pair is calculated by subtracting the Bhattacharyya distance between the pair of Hue-Saturation histograms from the maximum distance of 1, as follows:

$$S_b = 1 - \sqrt{1 - \sum_{u=1}^{m} \sqrt{p^{(u)} q^{(u)}}} \tag{2}$$

where m is the total number of bins in each histogram, and $p^{(u)}$ and $q^{(u)}$ are the value of the u-th bin in each normalized histogram, respectively.

3) Integration of the similarities: The similarities S_f and S_b are integrated as follows:

$$S = \beta S_f + (1-\beta) S_b, \tag{3}$$

(a) Darker frame (b) Brighter frame

Figure 7. Example of a scene duplicate with different brightness.

Table I
EXPERIMENTAL DATASET. EACH NEWS PROGRAM WAS RECORDED FROM JUNE 7 TO 13, 2012.

Station	Program	Broadcast Time
NHK	NEWS	From 12:00 to 12:20
FNN	SPEAK	From 11:30 to 11:55
NNN	STRAIGHT NEWS	From 11:30 to 11:55
JNN	NEWS	From 11:30 to 11:55
ANN	NEWS	From 11:45 to 12:00

where the parameter β is a weight coefficient with a value of [0, 1]. If the similarity S is higher than a threshold θ_I, the input pair is detected as a scene duplicate.

III. EXPERIMENT

We evaluated the detection performance of the proposed method in the following experiment.

A. Dataset

We prepared experimental data as follows, considering that one of the important applications of this research is to automatically create a video collection of speeches. We recorded broadcast news videos from June 7 to 13, 2012 from five major broadcast stations in Japan, as shown in Table I. 678 face shots in total were manually extracted from the video, so the candidates of scene duplicates were ${}_{678}\mathrm{C}_2$ = 229,503 pairs of shots. We manually detected 840 scene duplicates as ground-truth. 262 stories containing the 678 face shots were also manually segmented. The parameters α and β and the thresholds θ_{CC} and θ_A were determined according to preliminary experiments.

B. Method

We compared the detection accuracy between the proposed method and the original method by Wu et al. [5]. We calculated the recall and the precision as evaluation criteria.

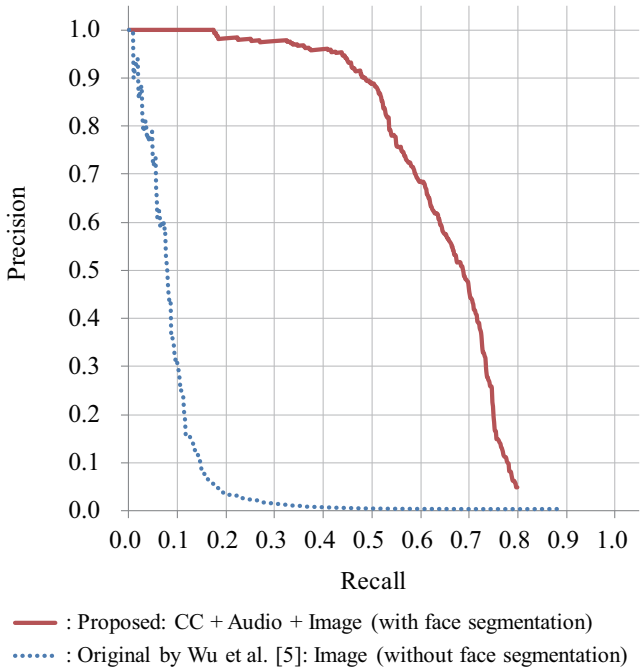


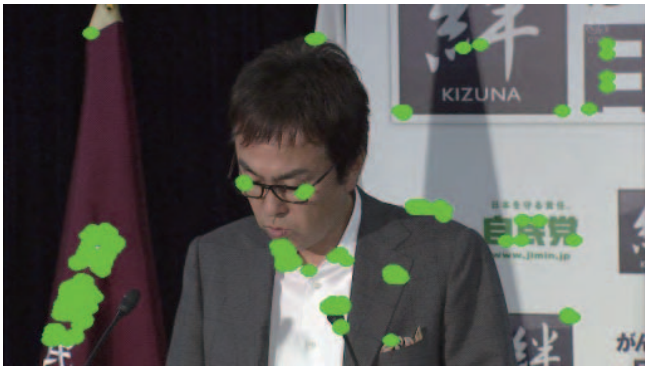
Figure 8. Comparison with the original scene duplicate detection method.

C. Result

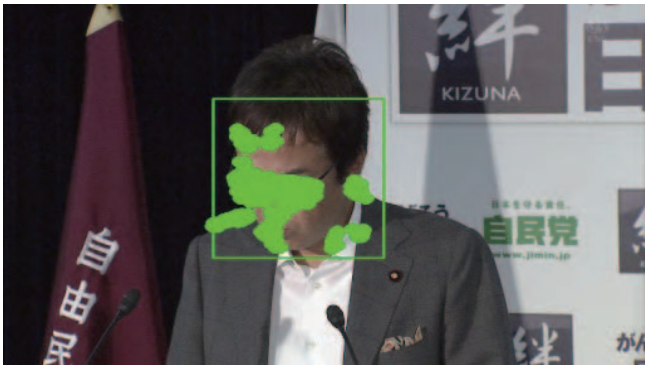
The recall-precision curve shown in Fig. 8 was obtained by changing θ_I. For reference, the maximum value of the F-measure was 0.648. Here, the F-measure is the harmonic mean of precision and recall. The proposed method outperformed the original method, which showed the effectiveness of the proposed method.

IV. DISCUSSION

As shown in Fig. 9(a), feature points were detected mostly in the background region by the original method by Wu et al. [5]. This makes the original method difficult to correctly evaluate the speed and the change of motion of a person.



(a) Original method by Wu et al. [5]



(b) Proposed method

Figure 9. Comparison of the feature points detected by the original method [5] and the proposed method.

This was the reason the recall of the original method was significantly low. On the other hand, as shown in Fig. 9(b), feature points were detected mostly in the face region by the proposed method. We consider that this allowed the proposed method to outperform the original method, as intended.

As for the effectiveness of integrating the information from the audio and the image channels, we additionally investigated the performance of two methods; comparative method 1 using only the audio channel, and comparative method 2 using only the image channel. The result is shown in Fig. 10. We can observe the improvement by audio-image integration. Figure 11 is an example of a scene duplicate which could be detected by referring to the audio channel. In this case, the audio similarity was very high, whereas the image similarity was not high due to the significant difference in viewpoints. That is, comparative method 1 can detect such a scene, whereas comparative method 2 cannot do so. Note that high precision and low recall were obtained by comparative method 1 because of setting a high threshold θ_A. This was just what we expected, and also was why both precision and recall were successfully improved by the audio-image integration.

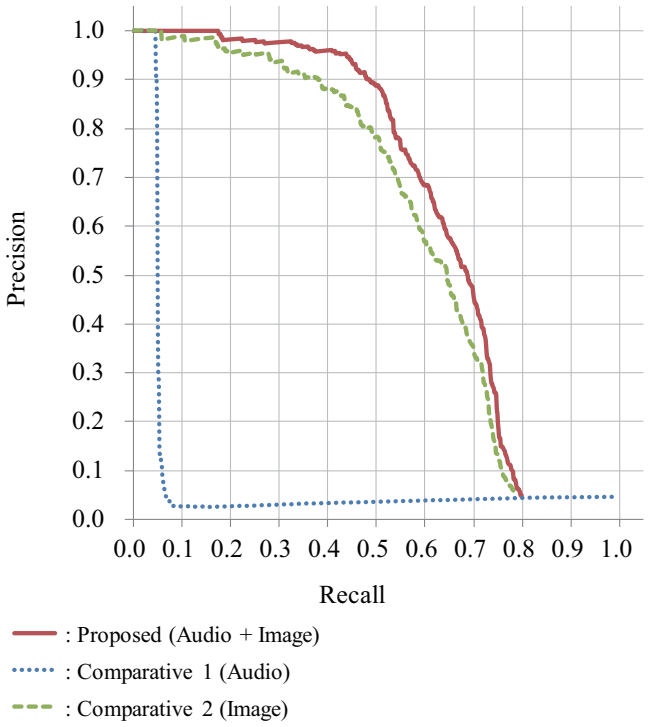


Figure 10. Effect of integrating the information from the audio and the image channels.

V. CONCLUSION

This paper proposed a method for scene duplicate detection by integrating the information from the audio and the image channels and complementing for each other's weak points. In addition, we solved the problems in Wu et al.'s method [5] by comparing the image features in the face region and the background regions independently. The performance of the proposed method was evaluated through an experiment with actual broadcast news videos. As a result, we could obtain a higher detection accuracy in both recall and precision. Therefore, we confirmed the effectiveness of the proposed method. Future work includes semantic association of news videos using the obtained scene duplicates.

ACKNOWLEDGMENT

Parts of this work were supported by JSPS Grant-in-aid for Scientific Research, and a joint research project with NII.

REFERENCES

[1] N. Katayama, H. Mo, and S. Satoh, "News shot cloud: Ranking TV news shots by cross TV-channel filtering for efficient browsing of large-scale news video archives," in *Proceedings of the 17th International MultiMedia Modeling Conference*, Jan. 2011, pp. 284–295.

(a) Image (b) Audio

Figure 11. Example of a scene duplicate which was detected using audio features.

[2] A. Ogawa, T. Takahashi, I. Ide, and H. Murase, "Cross-lingual retrieval of identical news events by near-duplicate video segment detection," in *Proceedings of the 14th International MultiMedia Modeling Conference*, Jan. 2008, pp. 287–296.

[3] J. Liu, Z. Huang, H. Cai, H. T. Shen, C. W. Ngo, and W. Wang, "Near-duplicate video retrieval: Current research and future trends," *ACM Computing Surveys*, vol. 45, no. 4, pp. 44:1–44:23, Aug. 2013.

[4] M. Takimoto, S. Satoh, and M. Sakauchi, "Identification and detection of the same scene based on flashlight patterns," in *Proceedings of 2006 IEEE International Conference on Multimedia and Expo*, Jan. 2006, pp. 9–12.

[5] X. Wu, M. Takimoto, and S. Satoh, "Scene duplicate detection based on the pattern of discontinuities in feature point trajectories," in *Proceedings of the 16th ACM International Conference on Multimedia*, Oct. 2008, pp. 51–60.