Considering conversation scenes in movie summarization

Masashi Inoue[1][0000-0002-9364-3114] and Ryu Yasuhara[2]

[1] Tohoku Institute of Technology, Yagiyama Kasumicho 35-1, Sendai, Japan
m.inoue@acm.org
http://www.ice.tohtech.ac.jp/ inoue/index.html
[2] Yamagata University, Jyonan 4-3-16, Yonezawa, Japan

Abstract. Given that manual video summarization is time consuming and calls for a high level of expertise, an effective automatic video summarization method is required. Although existing video summarization methods are usable for some videos, when they are applied to story-oriented videos such as movies, it sometimes becomes difficult to understand the stories from the generated summaries because they often lack continuity. In this paper, we propose a method for summarizing videos that can convey the story beyond the sequence of extracted shots so that they can fit user perception patterns. In particular, we examine the impact of conversation scenes in movie storytelling. The evaluation of summarized videos is another challenge because existing evaluation methods for text summarization cannot be directly applied to video summarization. Therefore, we propose a method for comparing summarized movies that maintains the integrity of conversation scenes with those that do not. We demonstrate how preserving conversational aspects influences the quality of summarized videos.

Keywords: Video summarization · movie summarization · evaluation · storytelling · conversation

1 Introduction

A summary video presents the important parts of a video usually by combining short video segments extracted from the original video. However, it is difficult and time consuming to prepare a summary video manually. To address this problem, various automatic summarization methods have been studied [6]. Among the various types of video that exist, it is relatively easy to generate a summary video where the contents are stylized and have a less story-oriented nature, such as sports videos in which redundant and highlighted sections are identifiable through machine-processable, low-level features. In the case of story-oriented videos such as movies and dramas, it is difficult to determine the important sections for generating a relevant summary using a computer.

In this research, we aim to provide a method for automatically generating a summary video for story-oriented videos for the purpose of increasing understanding and enjoying. Movies and dramas tell a story, but it is unclear which sections in the video are involved in the progress of the story. In fact, important segments are difficult to determine based on low-level features alone such as audio and visual information. Therefore, there are semantic video summarization methods proposed. The difficulty in semantic summarization is that the model used for representing deep semantics are often complex and obtaining high-level feature is costly when they are created manually. Therefore, it is desirable to estimate some shallow semantic features from from low-level features. As a computationally derivable shallow semantic feature, we focus on conversation scenes. Although conversation scenes were used for video abstraction [3], the evaluation of the generated summary videos remains as a major problem. Therefore, we proposed a text description-based and crowdsourcing method for quantitative evaluation. The role of conversation in movie for effective summarization is clarified through the experiment.

2 Materials

There are few story-oriented video data sets that can be used for the evaluation of video summarization owing to copyright restrictions. In terms of data for the experiment, accessible data with Creative Commons (CC) or Public Domain (PD) licenses are desirable for their reproducibility and usability. Therefore, in this research, we use a publicly available data set that is a collection of public domain movies. In addition to the video files, title text, and video description text, we utilized automatically assigned conversation section information and genre information provided in the original data set [7][3]. The videos are movies with a CC license that are hosted on Internet Archive [4]. The genre information is given as 22 genre tags defined in the Internet Movie Database (IMDb) [5].

All $1,722$ movies in the dataset were plotted based on both the frequencies of utterances in the movie and on the average duration of utterances (Figure 1. We defined three groups as the three clusters found after plotting. The groups were (G1) with many conversations, an intermediate group (G2), and a group with few conversations (G3). We selected five works for evaluation from each of the three groups of (Table 1).

3 Method

3.1 Base process

Video segmentation The summarization is carried out in four steps: video segmentation, feature extraction and importance assignment, and summary video generation. In the first step, there are differences in the granularity of divisions such as frame, shot, scene, and sequence. Among them, we used the shot unit as

[3] http://www.ice.tohtech.ac.jp/ inoue/moviedialcorpus/index.html

[4] https://archive.org/index.php

[5] http://www.imdb.com/

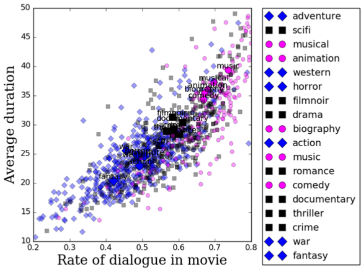

Fig. 1. Distribution of utterance frequencies and utterance duration.

Table 1. List of target movies.

Group	Title	Genre	Duration (min:sec)	Task ID
G1	Dreaming Out Loud	Comedy	65:10	1
	Hey! Hey! USA	Comedy	87:56	2
	The Ghost Walks	Comedy	63:26	3
	Windbag the Sailor	Comedy	81:30	4
	Texas, Brooklyn and Heaven	Comedy	76:18	5
G2	Cosmos: War of the Planets	Sci-Fi	89:03	1
	The Great Commandment	Drama	80:12	2
	First Spaceship on Venus	Sci-Fi	78:31	3
	A Star Is Born	Drama	110:53	4
	Night Alarm	Drama	61:08	5
G3	Svengali	Horror	81:08	1
	Under California Stars	Western	72:20	2
	Hollywood Man	Action	84:34	3
	Pecos Kid	Western	54:09	4
	Night of the Living Dead	Horror	95.52	5

a video section in which a single camera shoots consecutively. For segmentation, we used PySceneDetect [6] as a tool for detecting a shot change by observing the change in the amount of the difference of HSV histograms between frames. The precision and recall values were 0.885 and 0.830 respectively.

Importance Scoring In order to select the shots included in the summary from the segmented video, the importance is calculated for each shot. Ma et al., combined low-level features to determine importance scores to estimate people perception at a higher cognitive level [4]. We adopt their method and leverage a model that estimates the part that attracts people's interest. The method of Ma et al. does not target videos with a narrative nature, but their model is considered useful for wide variety of videos.

[6] https://github.com/Breakthrough/PySceneDetect

Video Generation The selection of the video segments to be included in the summary can be considered as a 0–1 knapsack problem selecting shots that maximize the obtained importance score so that it fits within the limited duration of the post-summarized video. In this research, a summarized video is generated with a dynamic programming algorithm [5].

3.2 Conversation integration

When a conversation scene is divided into shots as the basis for the summary video, and if the shot in the middle of the scene is not rated as important, there is the possibility that the information during the conversation drops out, making it difficult to understand the contents. Therefore, the proposed method explicitly uses the conversation section information and the divided shots are grouped into conversation units. For example, for a scene S consisting of shots $S = \{s_1, s_2, ...s_5 s_6..., s_n\}$, if it is assumed that three cut points in the second to fifth shots are within the conversation section , $s_2, ...s_5$ are merged and the set of shots after the integration are $S' = \{s_1, \hat{s_1}, s_6..., s_n\}$ where $\hat{s_1} = \{s_2, s_3, s_4, s_5\}$.

4 Evaluation

4.1 Evaluation Procedure

An evaluation of the video summary is performed via a subjective evaluation method that shows the automatically generated videos to the human assessors and collects their evaluations. The evaluation indices consist of: informativeness (information quality) and enjoyability (entertainment quality). Informativeness is an index which shows how much the summary video preserved the information necessary for understanding the content compared to the original video, and enjoyability is an index which indicates the degree of satisfaction with the summary video [4].

In order to measure the informativeness, it is necessary for the subject to know the contents of the original video. If the assessors have already seen the movies, they can use their knowledge of the respective films. However, the movies collected in this research consist of many old or less-known releases, and it is unlikely that the participants have watched them before the task. In addition, considering the movie running time, the burden of viewing the summary video after viewing the original video becomes too large. Therefore, in this research, we provided text information explaining the outline of the movie to enable users to grasp the content of the original movie after watching summaries. By comparing the information given by watching summarized video and the text explanation that is considered as the ground truth, it becomes possible to estimate the degree of informativeness of the summaries. For the text describing the outline of the movie to be used, the story-line (outline) information of the movie has been taken from the IMDb movie database where the information is created by user postings. Conversely, enjoyability can be judged without any additional information. Assessors were asked if they enjoyed the summary video they just watched as a video, and they provided a score in 10 steps.Because the target movies are in English, we collected assessors who understand English using a crowdsourcing service (CrowdFlower[7]). There are five video set (tasks) and 10 assessors were assigned for each task. That is, 50 assessors participated in total. In addition to the scores of informativeness and enjoyability, the subjects were asked to answer questions on the summary videos in complete sentences so that they could be used for analysis. An answer was collected by setting free description columns asking what type of movie it was, what type of information was lacking to understand the content, and why they could or could not enjoy the video.

Each participant watched only one of the videos generated by the proposed method and that by the conventional method. In other words, each participant responded to three works in total, one for each crowdsourcing task ID from each group (G1, G2, or G3). Ten crowdworkers participated in each task (evaluation of three summaries) and each group assigned 50 scores for both evaluation measures.

4.2 Results

Table 2 shows the average scores of informativeness and enjoyability for each summary. When considering conversation, the average score of informativeness improved by 0.56 points, 0.30 points, and 0.48 points, respectively, for each group compared with the case that did not consider conversation. Likewise, the average enjoyability score improved by 0.74 points, 0.58 points, and 0.68 points, respectively, for each group. When individual films were considered, the results were mixed. From these results, it can be predicted that it became easier to understand the contents of the movie by generating the summary video grouping the conversational shots, which also confirmed that the naturalness of the video can also be preserved.

Additionally, by analyzing the textual responses provided by the assessors, we were able to determine the qualitative differences among the two methods. One of them is the usage of proper nouns. A proper noun appearing in the video such as the name of a character or a place name is considered to be an important element for understanding and explaining the contents. By considering the method that did not leverage the conversations, the explanation using a proper noun in the video increased 6.0 %. As an example, if a participant watched a summary video that did not consider conversations, the explanation used nouns like "reporter goes to NY," whereas in the text explaining the summary video that considered conversation descriptions could be found similar to: "Eddie Taylor leaves Dallas, Texas and his newspaper job with an inheritance for New York." Therefore, explanations using proper nouns seem to be more concrete. Another example shows that without considering conversation consistency in the summary video, the description is plain: "Zombies start attacking

[7] https://www.crowdflower.com

Table 2. Result of human assessments on a ten-point scale.

		Group 1		Group 2		Group 3	
		Info	Enjoy	Info	Enjoy	Info	Enjoy
1	Considering conversation	5.0	4.8	4.1	4.2	4.4	4.3
	Not considering conversation	5.2	4.1	4.8	2.9	5.3	4.0
2	Considering conversation	6.2	5.9	6.2	4.4	5.8	5.8
	Not considering conversation	4.9	3.5	6.4	5.5	6.0	5.0
3	Considering conversation	7.6	6.7	7.4	6.4	7.0	5.1
	Not considering conversation	5.6	5.5	5.4	4.2	4.3	4.1
4	Considering conversation	7.4	6.3	7.4	7.0	6.8	6.1
	Not considering conversation	7.7	6.4	6.8	5.7	6.2	5.0
5	Considering conversation	6.7	5.4	6.7	5.7	8.1	7.9
	Not considering conversation	6.7	5.9	6.9	6.5	7.9	6.7
Average	Considering Conversation	6.58	5.82	6.36	5.54	6.42	5.84
	Not considering conversation	6.02	5.08	6.06	4.96	5.94	4.96

a girl and kill the guy she's with." Conversely, if conversation units are taken into account, participants are able to understand the relationships between characters: "Barbara and her brother Johnny decided to visit their parents' grave."

5 Conclusion

In this research, we proposed a method for the automatic generation of a summarized video based on a story-oriented video such as a movie. In the proposed method, the continuity of the information is maintained by preserving the conversation segments in a summary video to achieve semantic cohesion. Owing to the fact that the automatic evaluation of summarized videos is difficult and human evaluation by comparing original and summarized videos is time consuming, we proposed a text description-based and crowdsourcing methods for summary video evaluation. As a result, a subjective comparison by crowd assessors of summarized videos showed that the proposed method was rated as superior in terms of informativeness and enjoyability. The compression rate of the summarized videos were about 30 % in our setting. The influence of compression rate should be investigated.

Future topics include the expansion of evaluation measures and improvement of the conversation scene extraction. The evaluation measures used in this research are informativeness and enjoyability. In the task of video search, based on a user study, there are 28 evaluation criteria suggested [1]. Their applicability to the summarization task can be considered. Regarding the conversation scene detection, improvements of VAD algorithm that was applied for creating the dataset used in this study by incorporating additional noise classes [2]. Investigation of the improved conversation scene information in summarization quality is an interesting future work.

References

1. Albassam, S.A.A., Ruthven, I.: Users' relevance criteria for video in leisure contexts. Journal of Documentation **74**(1), 62–79 (2018)
2. Kosaka, T., Suga, I., Inoue, M.: Improving voice activity detection for multimodal movie dialogue corpus. In: IEEE 7th Global Conference on Consumer Electronics (GCCE 2018). Nara, Japan (2018)
3. Lienhart, R., Pfeiffer, S., Effelsberg, W.: Video abstracting. Commun. ACM **40**(12), 54–62 (Dec 1997)
4. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the Tenth ACM International Conference on Multimedia. pp. 533–542. MULTIMEDIA '02, Juan-les-Pins, France (2002)
5. McDonald, R.: A study of global inference algorithms in multi-document summarization. In: Proceedings of the 29th European Conference on IR Research. pp. 557–564. ECIR'07, Rome, Italy (2007)
6. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. **3**(1) (Feb 2007)
7. Yasuhara, R., Inoue, M., Suga, I., Kosaka, T.: Large-scale multimodal movie dialogue corpus. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 414–415. ICMI 2016, Tokyo, Japan (2016)