INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Video Summarization Leveraging Multimodal Information for Presentations

Hanchao Liu, Dapeng Chen, Rongjun Li, Wenyuan Xue, Wei Peng

IT Platform Chief Expert Office, Huawei Technologies Co., Ltd., China

{liuhanchao2, chendapeng8, lirongjun3, xuewenyuan1, peng.wei1}@huawei.com

Abstract

This demonstration introduces a video summarization system, leveraging multimodal information to efficiently extract essential contents from presentations. In contrast to existing methods focusing primarily on daily life videos and solely utilizing visual information, our system extracts multimodal information, including speech, text, and visual information from videos of presentations. Specifically, the proposed system extracts crucial slide texts from key-frames as queries to filter speech transcripts. By piecing together the video clips corresponding to the filtered speech transcripts, our system outputs the final video summarizations. The evaluation on ICCV 2017 videos demonstrates the effectiveness of the proposed system compared with the lead-3 baseline.

Index Terms: multimodal, video summarization

1. Introduction

Presentation videos, such as academic reports and instructional videos, serve as visual and auditory communication media for people to share opinions and knowledge. Locating desired content from this type of source of information has become increasingly challenging and time-consuming [1, 2]. It is critical to develop video summarization techniques to extract key content from videos and compress their duration to assist people in this domain.

Existing video summarization methods [3, 4] developed for daily life videos, such as travel vlogs, are not applicable for presentation videos. In presentation videos, most semantic information is contained in the speakers' speech, with secondary information presented in the slides. This pinpoints the necessity of treating the speech and texts from the slides as major sources of information for video summarization. Nonetheless, current video summarization methods solely utilize visual information, leaving valuable multimodal information untouched.

To overcome this limitation, we propose a presentation video summarization system that utilizes multimodal (speech, text, and vision) information from the input videos to generate high-quality video and text summaries. The system first converts the video's speech and slide visual information into text transcripts, using an automatic speech recognition (ASR) model and an optical character recognition (OCR) model, respectively. Next, a query-focused text summarizer selects the important sentences from ASR texts with the OCR texts as queries. Finally, the video summary is created by assembling clips corresponding to the selected sentences. Since there is currently no publicly available dataset for presentation video summarization, our system is evaluated on a portion of the ICCV 2017 academic report videos. Figure1 illustrates an example summarized by our system. The evaluation results indicate that our system can reduce about 80% video length on average and preserve more critical semantic contents than the lead-3 [5] baseline.

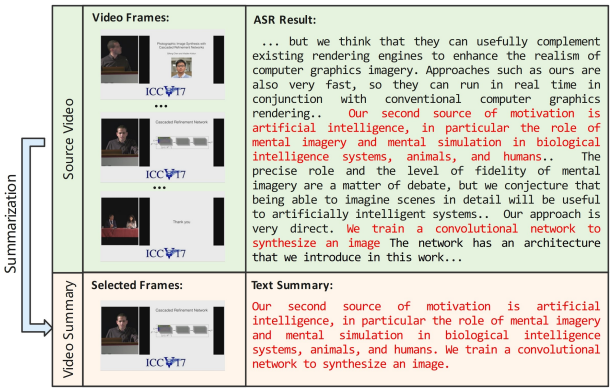

Figure 1: *Illustration of presentation video summarization. Texts in red font are extracted sentences.*

2. System Architecture

The proposed method leverages the mutual consistency between multi-modal information to enhance the most critical information in the video presentation. It consists of 5 modules: (1) A video segmentation module to extract key-frames and video segments. (2) An ASR module to convert the speech signal to text, obtaining ASR texts. (3) An OCR module to extract text from key-frames, obtaining OCR texts. (4) A query-focused text summarizer employs OCR texts as queries to summarize the ASR texts, resulting in extracted sentences. (5) A video summarizer to integrate video clips according to the extracted sentences. Figure 2 illustrates the overall flowchart of our system.

2.1. Main modules

Video Segmentation module. To efficiently extract semantic information from the video and minimize summarization performance degradation due to long inputs, we apply a frame clustering algorithm that facilitates key-frame extraction and enables video segmentation. The module extracts frame-level feature embeddings by the image encoder of the CLIP [6], which is robust w.r.t. the semantics. For each frame, the module identifies the neighbors in the temporal and feature space, which are in the same time window and have a high feature similarity. Then the module defines the sum of the neighbor similarities as the density for each frame. We assign a neighboring frame with a higher density as the parent node of the current

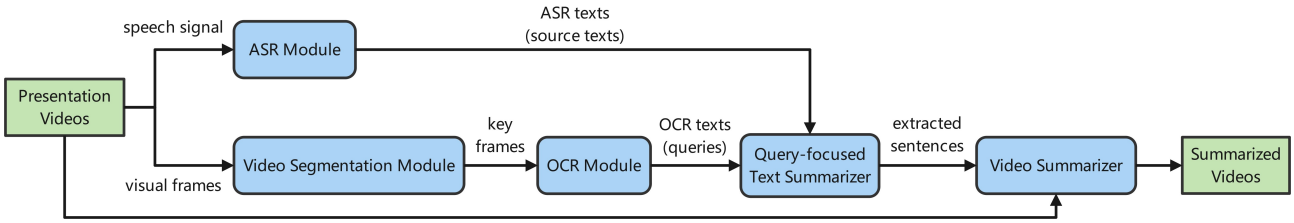

Figure 2: *Overall flowchart of our presentation video summarization system.*

frame and recursively determine parent nodes for all frames. The frames without higher density nodes in their neighborhood are key-frames. At the same time, the frames finally connecting to the same key-frame forms a video segment.

ASR and OCR modules. Converting acoustics and visual information into text is vital to the subsequent summarization module. Our system employs the existing state-of-the-art models: Whisper [7] for ASR and MASTER [8] for OCR.

Query-focused Text Summarizer. Summarizing presentation videos needs extracting video clips with crucial semantic information. Our system uses a query-focused text summarizer to select meaningful sentences from ASR transcripts. The module uses the text from presentation slides as queries, as these slides often contain valuable content that helps the audience understand the presentation better. The query-focused summarizer first segments the ASR texts based on the video segments. It then uses OCR results from corresponding key-frames as prefixes and inputs both the OCR request and segmented texts into the BERTSum [9] model, generating a query-focused summary for each text segment. All segment summaries combine into the abstractive summary of the whole video. We also mapped the abstractive summary to the original ASR sentence by a greed algorithm, where the selected sentences are with the highest ROUGE value to the abstractive summary.

Video Summarizer. After obtaining the extracted ASR sentences, corresponding video clips can be located using their timestamps. To make the final video summary look natural, our system extracts not only the video clips corresponding to the selected sentences, but also their adjacent counterparts. The seamless integration of these clips results in the eventual video summarization.

2.2. Implementation and Evaluation

Our system is developed using Python and the PyTorch framework[1] for the backend, with the Gradio app[2] employed for the web interface. Besides the video summary, our system also provides texts corresponding to both the original and summary videos as an additional reference for users. We evaluate the system on several ICCV 2017 academic report videos. The results show the system surpasses the lead-3 baseline for the average BLANC score [10] (0.41 : 0.27) and the video length is reduced approximately 80% on average. These experimental results demonstrate the effectiveness of our system.

3. Conclusion

In this paper, we propose a video summarization system to help users understand the critical content of presentation videos. Unlike existing video summarization methods that solely rely on visual information, our system harnesses multimodal (speech, textual, and vision) information from both the speaker's speech and the slides included in the video. Although our system yields satisfactory results, there is still scope for further developments. In future work, we plan to refine the utilization of OCR-based queries and incorporate a discourse parser module to accurately identify the crucial content of presentations.

[1]https://pytorch.org
[2]https://gradio.app

4. References

[1] J. A. Ghauri, S. Hakimov, and R. Ewerth, "Supervised video summarization via multiple feature sets with parallel attention," in *2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, July 5-9, 2021*. IEEE, 2021, pp. 1–6s.

[2] H. Jiang and Y. Mu, "Joint video summarization and moment localization by cross-task sample transfer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 16 367–16 377.

[3] Y. Jung, D. Cho, S. Woo, and I. S. Kweon, "Global-and-local relative position embedding for unsupervised video summarization," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV*, vol. 12370. Springer, 2020, pp. 167–183.

[4] M. Narasimhan, A. Rohrbach, and T. Darrell, "Clip-it! language-guided video summarization," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 13 988–14 000.

[5] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, R. Barzilay and M. Kan, Eds. Association for Computational Linguistics, 2017, pp. 1073–1083.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML, July 18-24*, vol. 139, 2021, pp. 8748–8763.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *CoRR*, vol. abs/2212.04356, 2022.

[8] N. Lu, W. Yu, X. Qi, Y. Chen, P. Gong, R. Xiao, and X. Bai, "MASTER: multi-aspect non-local network for scene text recognition," *Pattern Recognit.*, vol. 117, p. 107980, 2021.

[9] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, November 3-7*, 2019, pp. 3728–3738.

[10] O. V. Vasilyev, V. Dharnidharka, and J. Bohannon, "Fill in the BLANC: human-free quality estimation of document summaries," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 2020, pp. 11–20.