VideoMaster: A Multimodal Micro Game Video Recreator

Yipeng Yu[*] , **Xiao Chen** , **Hui Zhan**
Tencent
yypzju@163.com, {evelynxchen, huizhan}@tencent.com

Abstract

To free human from laborious video production, this paper proposes the building of VideoMaster, a multimodal system equipped with four capabilities: highlight extraction, video describing, video dubbing and video editing. It extracts interesting episodes from long game videos, generates subtitles for each episode, reads the subtitles through synthesized speech, and finally re-creates a better short video through video editing. Notably, VideoMaster takes a combination of deep learning and traditional computer vision techniques to extract highlights with fine-to-coarse labels, utilizes a novel framework named PCSG-v (probabilistic context sensitive grammar for video) for video description generation, and imitates a target speaker's voice to read the description. To the best of our knowledge, VideoMaster is the first multimedia system that can automatically produce product-level micro-videos without heavy human annotation.

1 System Architecture

VideoMaster includes four components as shown in Figure 1. The system receives a long raw video of a mobile MOBA game named "Honor Of Kings" from E-sports live websites. The final output is a re-created short video with subtitles and dubbing which is more appealing.

1.1 Highlight Extraction

As video frames of a game are usually well structured and have timely broadcasts, both deep learning [Ren *et al.*, 2015; LeCun *et al.*, 1998; Zhang *et al.*, 2019; He *et al.*, 2016] and traditional methods [Thanh *et al.*, 2009; Bradski and Kaehler, 2008; Lowe, 2004; Derpanis, 2010; Rublee *et al.*, 2011; Leutenegger *et al.*, 2011] are adopted in our work. As shown in Figure 2, each raw video is split into frame sequences with a frame rate of 2, then event labels and attribute labels can be figured out after the element information (Broadcast, KDA, ROI, etc.) of each frame are calculated. A highlight occurs 5 seconds before and after an event.

1.2 Video Describing

Video describing is to generate text description for each highlight video episode in details [Xiao *et al.*, 2022]. Each generated text description is used as the subtitle for each highlight. At first a relation graph of all the 101 game champions is built. Then we collect a small amount of natural video description, and holistically incorporate them to enrich the grammars and text of the proposed PCSG-v framework. PCSG-v is a variant of PCFG [Zhang and Krieger, 2011; Wang *et al.*, 2017a]. It is a 4-tuple $G = (N, \Sigma, R, S)$, where

- N is finite set of non-terminal symbols.
- Σ is a finite set of pseudo terminal symbols. Each pseudo terminal symbol γ comes from $\gamma \rightarrow [c_1]\lambda_1[p_1] \,|\, \lambda_2[p_2] \,|\, \cdots \,|\, [c_m]\lambda_m[p_m]$ with $m \geq 1$. c is an alternative Boolean value depends on context of events and attributes or context of the value of preceding pseudo terminal symbols, namely the text generated before. λ can be a number, word, phrase, clause, or sentence selected from video description collection and champion relation graph. p_i is the probability of λ_i to be selected, and $\sum_{i=1}^{m} p_i = 1$. If c_i exists and it is $False$ then λ_i and p_i will be deleted, and value of p_i will be divided into equal parts and then given to other λ. c makes text selection and collocation more accurate and diverse, and p makes text selection more diverse with preference.
- R is a finite set of probabilistic and context-aware production rules of the form $\alpha \rightarrow [d_1]\beta_1^1[q_1] \,|\, \beta_2^1\beta_2^2[q_2] \,|\, \cdots \,|\, [d_n]\beta_n^1 \cdots \beta_n^t[q_n]$ with $n \geq 1$ and $t \geq 1$. d is an alternative Boolean value depends on context of events and attributes. q_i is the probability of β_i sequence to be selected, and $\sum_{i=1}^{n} q_i = 1$. $\alpha \in N$, and $\beta_i \in (N \cup \Sigma)$ for $i = 1 \cdots n$. Most of the rules are translated from sentences of video description collection. if d_i exists and it is $False$ then β_i sequence and q_i will be deleted, and value of q_i will be divided into equal parts and then given to other β sequences. d makes sentence structures and sentence combination more accurate and diverse, and q makes sentence structures and sentence combination more diverse with preference.
- and $S \in N$ is a distinguished start symbol.

In simple terms, PCSG-v is used to generate text descrip-

[*] Yipeng Yu is the corresponding author

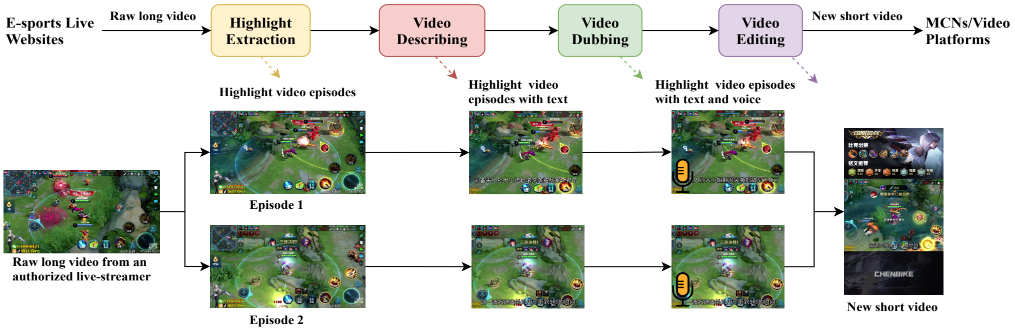

Figure 1: An overview of VideoMaster system.

tion for each short video with its labels of event and attribute provided by highlight extraction module as input. During the generation, PCSG-v will select text and infer relations between game champions from the relation graph. The more words and more production rules, the more diverse text our framework can generate.

1.3 Video Dubbing

Video dubbing reads the generated subtitles using someone's voice. To synthesize speech directly from a game commentator, we implement a text-to-speech model based on Style Tokens Tacotron [Wang *et al.*, 2018; Wang *et al.*, 2017b] which is shown in Figure 3. First, the reference encoder extracts the identify-feature from the input log-mel spectrogram to generate the reference embedding. Then, in the style token layer, the attention module learns the similarity between the input reference embedding and the randomly initialized tokens, to output a set of weights, representing the contribution of each token to the style embedding. Finally, the resulting style embedding is passed to the encoder-decoder with the input text sequence. Note that the style layer can be jointly trained with the encoder-decoder by optimizing the reconstruct loss of the encoder-decoder, where the log-mel spectrogram of the training target is used as the ground-truth. The speech synthesis model is trained on the public mandarin speech dataset THCHS-30 and audios from the game commentator (only 301 corpus about 10 minutes).

1.4 Video Editing

Video editing assembles the multimodal materials into sequences and creates new videos which are more fascinating. The following editing operations are taken in our system based on FFmpeg:

- **Cutting.** Highlight video episodes are cutting from raw live videos. The cuts are supposed to be narratively complete and fast-paced to make the flow of video ideal.
- **Transition.** Transitions are introduced to make the switch between video episodes more natural and keep the pace of the video controlled.
- **Subtitle.** The text description is presented in the form of subtitle, which is appearing along with the video scenes.
- **Dubbing.** The audio commentary is added in sync with the subtitle to achieve visual-audio consistency.
- **Music.** Music which have the same emotion as the videos are played as background music. Video soundtrack is of great importance in setting the mood and evoking emotions from audience [Li *et al.*, 2021].
- **Sticker.** Funny stickers are inserted into videos according to the frame labels and hero coordinates.
- **Backdrop.** The video episodes are cropped and put on a background portrait image, which makes the produced videos adapt to mobile devices.

2 Results and Demonstrations

2.1 Highlight Extraction

We randomly sampled 1000 highlight videos and recorded their true labels, then a mean precision of 92.95% and recall

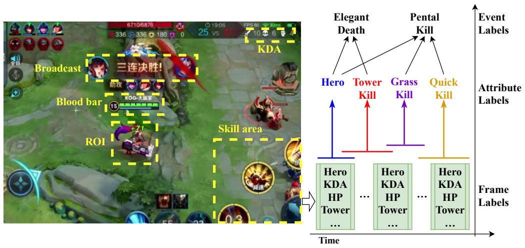

Figure 2: An illustration of highlight extraction and fine-to-coarse tagging. KDA is kill-death-assist and HP is health points.

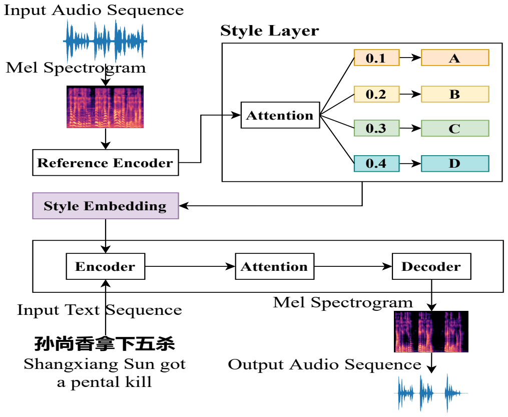

Figure 3: Model diagram of speech synthesis. A, B, C and D are style embedding vectors.

Rater	Evaluation#1				Evaluation#2	Evaluation#3	
	Human		Machine		Human vs Machine	Machine	
	Clarity↑	Naturalness↑	Clarity↑	Naturalness↑	Similarity↑	Clarity↑	Naturalness↑
1	5.00±0.00	5.00±0.00	4.46±0.50	3.21±1.00	2.58±0.95	4.21±0.59	3.08±0.82
2	4.97±0.03	4.99±0.01	4.13±0.80	3.18±1.18	3.17±0.75	4.16±0.72	3.44±0.85
3	4.77±0.22	4.77±0.18	4.07±0.65	3.21±0.96	3.83±0.58	3.70±0.54	3.34±0.53
4	4.82±0.18	4.82±0.15	4.19±0.65	3.38±1.09	3.71±1.43	4.17±0.97	3.69±1.31
5	4.84±0.16	4.85±0.13	4.26±0.64	3.50±1.08	3.58±0.25	4.29±0.25	3.78±0.46
6	4.86±0.14	4.87±0.11	4.29±0.65	3.48±1.06	2.33±0.84	4.51±0.54	3.87±0.86
Mean	4.88	4.88	4.23	3.33	3.20	4.17	3.53

Table 1: Synthetic speech evaluation. ↑ indicates the higher the better. Score is presented by mean±variance.

of 81.46% are obtained. In our case, precision is more important than recall, thus such a recall can be acceptable. Bad cases happened when video frames are not clear or blocked.

2.2 Text Description

Human rating. We randomly sampled 1000 highlight videos with generated description and labels of event and attribute for human evaluation. We recruited another 6 helpers (3 males and 3 females who enjoy playing the mobile game) to score *Fluency*, *Attractivity*, *Relevance* and *Diversity*. The rating criteria is as follows: 5-*very good*, 4-*good*, 3-*neutral*, 2-*bad*, 1-*very bad*. The mean rating scores are 4.11, 4.02, 4.47 and 4.00, respectively, which tells that our methods are able to generate good description on all of the four metrics.

Methods comparison. We also made a qualitative comparison between current methods in Table 2. Each method was measured on five metrics: Interpretability (Inte.), Controllability (Cont.), Flexibility (Flex.), Portability (Port.) and Annotation (Anno.), and each metric was scored from 1 to 5. The higher the score, the more consistent with the metric. As we can see, PCSG-v is best on all of the five metrics.

2.3 Speech Synthesis

As the purpose is to transfer the speech of human game commentator to machine, we conducted three human evaluation for video dubbing in terms of clarity, naturalness and similarity. The rating criteria for "clarity" and "naturalness" is as follows: 5-*very good*, 4-*good*, 3-*neutral*, 2-*bad*, 1-*very bad*, and that for "similarity" is as follows: 5-*almost the same*, 4-*very like*, 3-*like*, 2-*a bit like*, 1-*not like*. Six human raters were recruited to give their scores. Evaluation results are shown in Table 1. Evaluation#1 and Evaluation#2 were conducted on a small dataset of 24 testing samples. Raters don't know which speech is from human and which speech is from machine in Evaluation#1, while raters know which speech is from human and which speech is from machine in Evaluation#2. We can see that machine gets a 4.23 score between "very good" and "good" in clarity which is close to 4.88 of human, a 3.33 score between "good" and "neutral" in naturalness, and a 3.20 score between "very like" and "like" in similarity. Evaluation#3 was conducted on a dataset of 100 test samples, each test sample has a text generated by PCSG-v and a speech synthesized by machine. The results show that machine gets a 4.17 score between "very good" and "good" in clarity and 3.53 score between "good" and "neutral" in naturalness.

Method	Inte.↑	Cont.↑	Flex.↑	Port.↑	Anno.↓
End2end	1	1	1	2	5
Summarization	4	2	1	3	3
Template	5	5	2	3	2
PCFG	5	5	3	3	2
PCSG	5	5	4	4	2
PCSG-v	**5**	**5**	**5**	**4**	**2**

Table 2: Methods comparison.

2.4 Demonstrations

Compared with raw videos from E-sports live websites which are tedious and short of storytelling, **the resulting videos**[1] produced by our system are time-saving (from about 30 minutes to about 1 minutes), memory-saving (from about 100 MiB to about 20 MiB), more expressive and rhythmic. Cutting and reorganizing highlight episodes make video brief, fast-paced and attractive. Subtitle and audio commentary not only help viewers have a better understanding of events, but also provide funny text and charming voice. Background music helps set the mood and evoke emotions from viewers. Stickers enrich videos and show current temperature. Moreover, video editing module in our system can be easily updated according to the usage of the re-created videos.

3 Conclusions

In this paper, we present VideoMaster, an intelligent multimodal system for automatic micro game video recreation. The system is but a first attempt to build a fully functional video generator capable of cutting, subtitling, dubbing and editing without heavy manual annotation. VideoMaster is not yet perfect and has limitation and rooms for improvement. Moreover, Compared to the latest AIGC (artificial intelligence generated content) work [Ouyang *et al.*, 2022; Scao *et al.*, 2022; Yang *et al.*, 2022; Rombach *et al.*, 2022; Ramesh *et al.*, 2022; Borsos *et al.*, 2022; Singer *et al.*, 2022; Saharia *et al.*, 2022; Yu *et al.*, 2021], VideoMaster is able to generate multimodal content without losing controllability, quality, novelty and artistry, and it is more cost-effective. In the future work, VideoMaster will be applied to non-game videos.

[1] **Weiyun** (https://share.weiyun.com/e53rrZSK) or **Google Drive**

References

[Borsos *et al.*, 2022] Zalán Borsos, Raphaël Marinier, et al. Audiolm: a language modeling approach to audio generation. *arXiv:2209.03143*, 2022.

[Bradski and Kaehler, 2008] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.

[Derpanis, 2010] Konstantinos G Derpanis. Overview of the ransac algorithm. *Image Rochester NY*, 4(1):2–3, 2010.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Leutenegger *et al.*, 2011] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *ICCV*, pages 2548–2555, 2011.

[Li *et al.*, 2021] Tingtian Li, Zixun Sun, Haoruo Zhang, Jin Li, Ziming Wu, Hui Zhan, Yipeng Yu, and Hengcan Shi. Deep music retrieval for fine-grained videos by exploiting cross-modal-encoded voice-overs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1880–1884, 2021.

[Lowe, 2004] G Lowe. Sift-the scale invariant feature transform. *Int. J*, 2:91–110, 2004.

[Ouyang *et al.*, 2022] Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *arXiv:2203.02155*, 2022.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[Rublee *et al.*, 2011] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*, 2022.

[Scao *et al.*, 2022] Teven Le Scao, Angela Fan, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv:2211.05100*, 2022.

[Singer *et al.*, 2022] Uriel Singer, Adam Polyak, Thomas Hayes, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv:2209.14792*, 2022.

[Thanh *et al.*, 2009] Nguyen Duc Thanh, Wanqing Li, and Philip Ogunbona. An improved template matching method for object detection. In *ACCV*, 2009.

[Wang *et al.*, 2017a] Junjie Wang, Bihuan Chen, Lei Wei, and Yang Liu. Skyfire: Data-driven seed generation for fuzzing. In *S&P*, 2017.

[Wang *et al.*, 2017b] Yuxuan Wang, R.J. Skerry-Ryan, et al. Tacotron: Towards end-to-end speech synthesis. In *INTERSPEECH*, 2017.

[Wang *et al.*, 2018] Yuxuan Wang, Daisy Stanton, et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *ICML*, 2018.

[Xiao *et al.*, 2022] Xinyu Xiao, Zixun Sun, Tingtian Li, and Yipeng Yu. Relational graph reasoning transformer for image captioning. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022.

[Yang *et al.*, 2022] Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In *EMNLP*, 2022.

[Yu *et al.*, 2021] Yipeng Yu, Zirui Tu, Longyu Lu, Xiao Chen, Hui Zhan, and Zixun Sun. Text2video: Automatic video generation based on text scripts. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 2753–2755, 2021.

[Zhang and Krieger, 2011] Yi Zhang and Hans-Ulrich Krieger. Large-scale corpus-driven pcfg approximation of an hpsg. In *IWPT*, pages 198–208, 2011.

[Zhang *et al.*, 2019] Yaping Zhang, Shuai Nie, Wenju Liu, Xing Xu, Dongxiang Zhang, and Heng Tao Shen. Sequence-to-sequence domain adaptation network for robust text image recognition. In *CVPR*, 2019.