Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Demonstrations Track

An AI-Empowered Visual Storyline Generator

Chang Liu[1] , **Zhao Yong Lim**[1] , **Han Yu**[1] , **Zhiqi Shen**[1,2*] , **Ian Dixon**[3] , **Zhanning Gao**[4] , **Pan Wang**[4] , **Peiran Ren**[4†] , **Xuansong Xie**[4] , **Lizhen Cui**[5,6] and **Chunyan Miao**[1]
[1]School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore
[2]Alibaba-NTU Singapore Joint Research Institute, Singapore
[3]Wee Kim Wee School of Communication and Information, NTU, Singapore
[4]Alibaba Group, Hangzhou, China
[5]School of Software, Shandong University (SDU), China
[6]Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, China
zqshen@ntu.edu.sg, peiran.rpr@alibaba-inc.com

Abstract

Video editing is currently a highly skill- and time-intensive process. One of the most important tasks in video editing is to compose the visual storyline. This paper outlines Visual Storyline Generator (VSG), an artificial intelligence (AI)-empowered system that automatically generates visual storylines based on a set of images and video footages provided by the user. It is designed to produce engaging and persuasive promotional videos with an easy-to-use interface. In addition, users can be involved in refining the AI-generated visual storylines. The editing results can be used as training data to further improve the AI algorithms in VSG.

1 Introduction

Promotional videos have become increasingly popular in E-commerce platforms as a means to recommend products to customers. To make an engaging and persuasive video, it is necessary to perform video editing, which includes visual storyline composition (i.e. selecting and sequencing the footage) and post-processing. In traditional film making, the composition of visual storylines is performed by experienced directors. Since this manual process is often costly and time-consuming, large-scale video-based product promotion campaigns in e-commerce are infeasible.

With the development of AI techniques in recent years, algorithms for automatically generating visual storylines start to emerge [Sigurdsson *et al.*, 2016; Choi *et al.*, 2016; Zhong *et al.*, 2018; Xu *et al.*, 2019; Liu *et al.*, 2019; Liu *et al.*, 2020]. However, these algorithms have different criteria for sequencing and selecting visual materials (VMs) (i.e. images and video footage). For instance, [Xu *et al.*, 2019] considers temporal relation and [Choi *et al.*, 2016] takes dynamics into account. To help the user apply these algorithms to generate visual storylines from which they can further refine, we outline Visual Storyline Generator (VSG) in this paper. It is an easy-to-use system that automatically generates engaging visual storyline by artificial intelligence (AI) techniques using images and video footage as input. It also supports the refinement edits to the generated storylines which can be used to enhance the AI algorithms in the future.

2 The AI Algorithms

VSG supports a number of visual storyline generation algorithms. The WundtBackpack (WBP) [Liu *et al.*, 2019] sequencs images to maximize the perceived persuasiveness. Its Learnable Wundt Curve (LWC) estimates the perceived persuasiveness measured by informativeness, emotion, and aesthetics. They are measured by the dissimilarity between VMs, arousal and aesthetics. LWC is formulated as the difference between two sigmoid functions - reward and punishment:

$$\begin{aligned} R(x) &= \frac{R_{\max}}{1+e^{-\rho_R(x-d_R)}}, \\ E(x) &= R(x) - P(x), \end{aligned} \tag{1}$$

where R_{max}, ρ_R and d_R are the maximum value of the reward, the slope of the reward function, and the minimum stimulus to be rewarded, respectively. The punishment $P(x)$ is also similarly defined with $R(x)$ as a sigmoid function. x is the linear combination of dissimilarity, arousal and aesthetic score, which is extracted by corresponding algorithms [Wang *et al.*, 2004; Kim *et al.*, 2018; Talebi and Milanfar, 2018]. The LWC can learn with a small amount of training data. A dynamic programming-based searching is then employed to construct the sequence with the maximum LWC score.

However, the WBP does not guarantee a good viewing experience, which is important for user retention. Thus, the system also includes the Shot Composition, Selection and Plotting (ShotCSP) algorithm [Liu *et al.*, 2020], which provides a principled approach to leverage important film-making principles to produce improved viewing experience while achieving high perceived persuasiveness. It is an algorithmic equivalent of the following three film-making principles: 1) the progression of shot "proximity" from wide shot to close up, 2) logical story sequence and 3) graphic discontinuity. It first composes shots from the input (i.e. the set of VMs), which

*Contact Author
†Contact Author

can be considered as candidate sub-sequences for composing the final sequence. Shots are then selected based on perceived persuasiveness measured by the LWC and semantic distance with the products. Finally the best sequence is generated by considering semantic distance SED, salient region ratio SRR and similarity SIM of selected shots, based on the three film-making principles.

In addition, [Choi *et al.*, 2016] which calculates dissimilarities and plots penalties, encouraging the sequence of video clips to follow the proposed general plot of the stories; and [Zhong *et al.*, 2018] which uses a two-stream RNN with submodular optimization to compose storylines are also incorporated into the VSG system.

An A/B test was performed with 114 users. We collected all the images and videos from the introduction pages of 40 products from Taobao and Tmall, and ask respondents to compare the flow of logic, viewing experience and perceived persuasiveness between the baselines and the ShotCSP. A total of 1,102 responses were collected. Table 1 reports the results of the A/B test. Overall, ShotCSP is perceived to produce significantly better visual storylines than the baselines.

Questions	Baslines	ShotCSP Better	ShotCSP Slightly Better	Almost the Same	Baseline Slightly Better	Baseline Better
Has Logic	[Liu *et al.*, 2019]	17.93%	26.90%	25.00%	18.75%	11.41%
	[Zhong *et al.*, 2018]	24.22%	25.78%	28.39%	13.80%	7.81%
	[Choi *et al.*, 2016]	20.54%	21.35%	32.70%	17.84%	7.57%
Viewing Experience	[Liu *et al.*, 2019]	15.49%	29.08%	25.82%	18.75%	10.87%
	[Zhong *et al.*, 2018]	21.09%	28.91%	28.12%	15.36%	6.51%
	[Choi *et al.*, 2016]	17.52%	25.61%	32.35%	17.25%	7.28%
Perceived Persuasiveness	[Liu *et al.*, 2019]	14.40%	28.25%	32.96%	16.34%	8.03%
	[Zhong *et al.*, 2018]	17.46%	29.37%	31.48%	16.14%	5.56%
	[Choi *et al.*, 2016]	15.98%	25.62%	34.99%	16.25%	7.16%

Table 1: Real-world performance evaluation.

3 The VSG System

The system architecture is shown in Figure 1. The dashed arrows denote processes requiring user confirmation. As the system allows users to make the final decision to accept or make revise the generated storylines, the mode of operation of the AI technology is *human-in-the-loop* [Yu *et al.*, 2018].

The interface design of VSG system (Figure 2) adopts a simple user interaction design to reduce the technical barrier for users. Clicking the "Create Video" button starts the workflow of generating visual storylines:

1. The images and videos are uploaded to the system database through drag-and-drop operations. The order for selecting or uploading the VMs is not important since the algorithms will perform sequencing automatically.

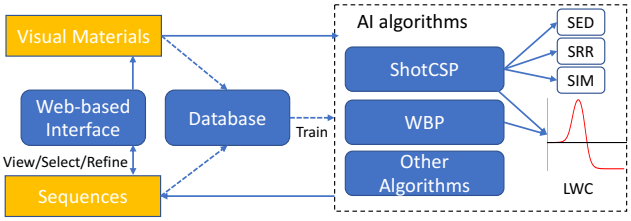

Figure 1: The system architecture of VSG.

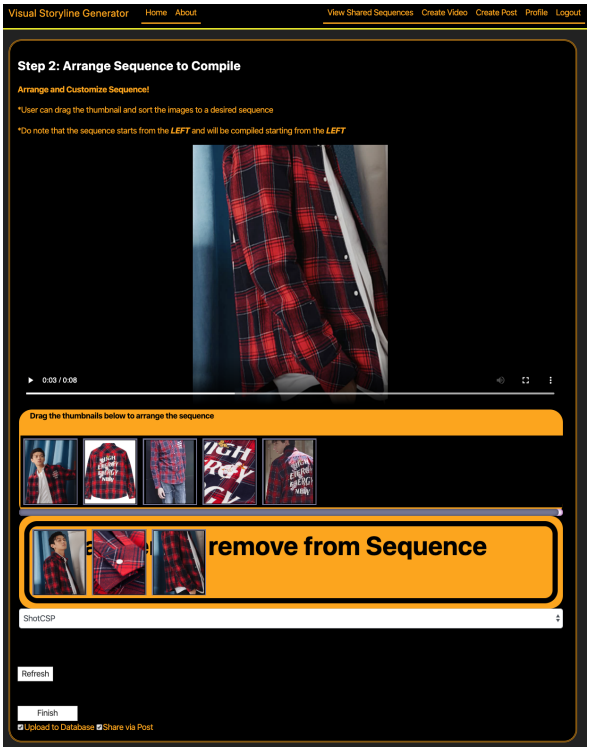

Figure 2: A screenshot of the VSG system.

2. The users then review the uploaded VMs and provide a name for the generated sequence.

3. The visual storyline is generated and edited. The user views alternative AI-generated sequences and chooses the favorite one. Based on the selected sequence, the user can perform further adjustment by re-ordering or replacing some of the VMs in the sequence. They can easily remove a VM by dragging it to the "unused VM field" right below the algorithm selection box. They can also re-selected an unused VM by putting it back to the sequence. By clicking the "Finish" button, the visual storyline is finalized and composed into a video.

This system can serve as a crowdsourcing tool for collecting data for visual storyline composition tasks. The revisions are stored as training data for improving AI models, if it is allowed by users. A video of the system is available online[1].

4 Discussions

The VSG system enables e-commerce shop owners to compose visual storylines without professional video editing skills. To the best of our knowledge, it is the first AI-empowered online system that can automatically generate visual storylines. In subsequent work, we will incorporate explainable AI features [Yu *et al.*, 2018] into the system to enhance the users' understanding of the rationale behind the generated visual storylines.

[1]https://www.youtube.com/watch?v=bEOYFpw7etg

Acknowledgements

This research is supported, in part, by the Nanyang Assistant Professorship (NAP), AISG-GC-2019-003, NRF-NRFI05-2019-0002, NTU-SDU-CFAIR (NSC-2019-011), and Alibaba-NTU-AIR2019B1.

References

[Choi *et al.*, 2016] Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Video-story composition via plot analysis. In *CVPR*, pages 3122–3130, 2016.

[Kim *et al.*, 2018] Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee. Building Emotional Machines: Recognizing Image Emotions Through Deep Neural Networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992, November 2018.

[Liu *et al.*, 2019] Chang Liu, Yi Dong, Han Yu, Zhiqi Shen, Zhanning Gao, Pan Wang, Changgong Zhang, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao. Generating Persuasive Visual Storylines for Promotional Videos. In *CIKM*, pages 901–910, 2019.

[Liu *et al.*, 2020] Chang Liu, Han Yu, Yi Dong, Zhiqi Shen, Yingxue Yu, Ian Dixon, Zhanning Gao, Pan Wang, Changgong Zhang, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao. Generating Engaging Promotional Videos for E-commerce Platforms. In *AAAI*, 2020.

[Sigurdsson *et al.*, 2016] Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. Learning visual storylines with skipping recurrent neural networks. In *ECCV*, pages 71–88, 2016.

[Talebi and Milanfar, 2018] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, and others. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Xu *et al.*, 2019] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. In *CVPR*, pages 10334–10343, 2019.

[Yu *et al.*, 2018] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. Building Ethics into Artificial Intelligence. In *IJCAI*, pages 5527–5533, July 2018.

[Zhong *et al.*, 2018] Guangyu Zhong, Yi-Hsuan Tsai, Sifei Liu, Zhixun Su, and Ming-Hsuan Yang. Learning Video-Story Composition via Recurrent Neural Network. In *WACV*, pages 1727–1735, 2018.