Generating Engaging Promotional Videos for E-commerce Platforms (Student Abstract)

Chang Liu,[1] Han Yu,[1] Yi Dong,[1] Zhiqi Shen,[1,2*] Yingxue Yu,[1] Ian Dixon,[3] Zhanning Gao,[4] Pan Wang,[4] Peiran Ren,[4] Xuansong Xie,[4] Lizhen Cui,[5,6] Chunyan Miao[1,2]

[1]School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore
[2]Alibaba-NTU Singapore Joint Research Institute
[3]Wee Kim Wee School of Communication and Information, Nanyang Technological University, 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, Jinan, China
*Corresponding author: zqshen@ntu.edu.sg

Abstract

There is an emerging trend for sellers to use videos to promote their products on e-commerce platforms such as Taobao.com. Current video production workflow includes the production of visual storyline by human directors. We propose a system to automatically generate visual storyline based on the input set of visual materials (e.g. video clips or still images) and then produce a promotional video. In particular, we propose an algorithm called Shot Composition, Selection and Plotting (ShotCSP), which generates visual storylines leveraging film-making principles to improve viewing experience and perceived persuasiveness.

Introduction

To make promotional videos, a film-making team firstly takes photos or video clips of the products. These visual materials (VMs) are then selected and sequenced to form a visual storyline. Finally, the visual storyline is composed into a video with visual effects. This manual process is costly and time-consuming, which makes large-scale video-based product promotion campaigns infeasible in e-commerce.

To address this issue, we propose a system to automatically generate promotional videos. The system takes a set of VMs as input. First the VMs are preprocessed, and then a visual storyline is generated based on film-making principles. Visual effects are selected and composed into the final video. In addition, the background music can be considered to make shot switching fits the rhythm. The system is designed to achieve an engaging viewing experience and perceived persuasiveness. Specifically, an algorithm called *Shot Composition, Selection and Plotting (ShotCSP)* is proposed as a component in the system. A evaluation involving 96 users shows that ShotCSP outperforms the best state-of-the-art method WBP (Liu et al. 2019) with 61.16% higher viewing experience and 89.12% higher perceived persuasiveness.

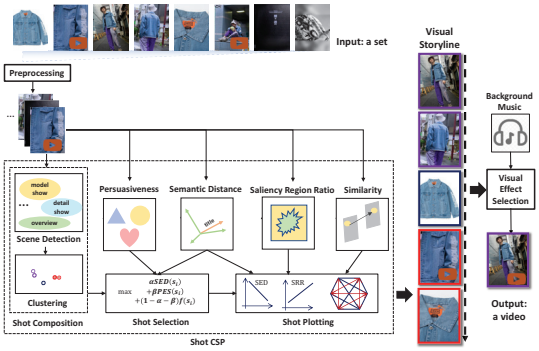

Figure 1: An overview of the proposed AI promotional video generation system.

The Proposed Systems

The proposed AI promotional video generation system is shown in Figure 1. In the preprocessing step, images and videos are cropped into one unified proper size based on its saliency region. Duplicate VMs are also removed. The ShotCSP step generates a visual storyline based on film-making principles to provide an engaging viewing experience and motivate the purchase of products. In ShotCSP, three film-making principles are adopted:

- **The development of shot "proximity" from wide shot to close up**. The visual storyline should start with a wide shot and gradually narrow to a close-up shot to orientate viewers to the physical location of the scene. It is useful when we intend to focus viewers' attention on a certain object (e.g., a character or a product).
- **Logical story sequence**. There should be a logical flow in the storyline, and it should be easy for viewers to follow.
- **Graphic discontinuity**. Graphically discontinuous montage editing is more compelling to viewers compared to graphically continuous editing, thus leading to a more engaging viewing experience.

Figure 2: Example storylines generated by different algorithms (shown as flat sequences).

ShotCSP converts these principles into their algorithmic equivalents. It is composed of three steps: 1) **Shot Composition**: which groups the VMs to compose shots through scene detection, hierarchical clustering and nearest neighbor retrieval. This step is essential to guarantee a sound logic flow; 2) **Shot Selection**: which selects a set of shots. It is performed by sub-modular optimization based on the composition of scenes, the semantic distance and the perceived persuasiveness of the shots; and 3) **Shot Plotting**: which is a dynamic programming based search for finding the best sequence. The objective function of this step considers the shot "proximity", while taking logic and graphic discontinuity into account. The objective is given as follows:

Minimize:

$$\sum_{i=0}^{n-1}(\alpha\sigma(SRR(s_i) - SRR(s_{i+1})) + \beta\sigma(SED(s_{i+1}) - SED(s_i)) + (1-\alpha-\beta)SIM(s_i, s_{i+1})), \quad (1)$$

Subject to:

$$\text{if } \exists s_i, s_j, i<j, C(s_i) = C(s_j), \text{ then } \nexists s_k, C(s_k) \neq C(s_i), i<k<j, \quad (2)$$

where SRR refers to the salient region ratio and SED is the semantic distance, which measures how related a shot is to the product being promoted in the video. $SIM(s_i, s_{i+1})$ are similarities between two adjacent shots. $\sigma(x) = \max(0, x)$. $C(s_i)$ is the scene that the i-th shot s_i belongs to.

The last step in the system is to create visually-impacted videos with special effects. We define three different types of transitions: the shot switching between scenes, within a scene, and across shots. They can be identified by the output of the first step of ShotCSP. We define transition effects for each type and then apply them to a given category of products. The background music can also be considered to fit visual transitions into the rhythm by using detected beats (McFee et al. 2015) to determine the duraiton, style and intensity of visual transitions.

Evaluation and Discussion

To evaluate the performance of ShotCSP, we conducted a user study involving 368 questionnaires from 96 users. We randomly selected 40 products from Taobao.com, and collected all the images and videos from the product introduction pages. Each product is considered as a test item.

We evaluate ShotCSP against the state-of-the-art algorithm WBP (Liu et al. 2019) by pairwise comparison. Such a method of evaluation is commonly adopted by related approaches (Choi, Oh, and So Kweon 2016; Zhong et al. 2018; Liu et al. 2019). The preference score of an algorithm a compared with algorithm b is calculated by $P(a,b) = \frac{\sum_{i=0}^{|T|}\bar{p}_i(a,b)}{|T|}$, where $\bar{p}_i(a,b)$ is the mean of scores when a compared with b for test item i. T is the set of test items.

	Flow of Logic	Viewing Experience	Perceived Persuasiveness
ShotCSP	0.6360	0.6090	0.5819
WBP	0.4009	0.3779	0.3077
ShotCSP Outperformance	58.64%	61.16%	89.12%

Table 1: Preference score on the flow of logic in visual storylines, viewing experience and perceived persuasiveness.

Example visual storylines generated by ShotCSP and WBP are shown in Figure 2. ShotCSP follows an model display-detail show logic, the proximity across VMs varies for different shots. While for the WBP, the proximity acorss VMs changes rapidly, resulting in jumps and the logic flow causing poor viewing experience. As shown in Table 1, ShotCSP outperforms the WBP by at least 58.64%.

In the future we will experimentally evaluate the whole system against a more diverse set of existing approaches. We will also investigate how the choice of visual effects (e.g., using spatial video summarization) can be incorporated into visual storyline generation to improve the final videos.

Acknowledgements

This research is supported 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, J.; Oh, T.-H.; and So Kweon, I. 2016. Video-story composition via plot analysis. In *CVPR*, 3122–3130.

Liu, C.; Dong, Y.; Yu, H.; Shen, Z.; Gao, Z.; Wang, P.; Zhang, C.; Ren, P.; XIe, X.; Cui, L.; and Miao, C. 2019. Generating Persuasive Visual Storylines for Promotional Videos. In *CIKM*, 901–910.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and Music Signal Analysis in Python. In *SciPy*, 18–24.

Zhong, G.; Tsai, Y.-H.; Liu, S.; Su, Z.; and Yang, M.-H. 2018. Learning Video-Story Composition via Recurrent Neural Network. In *WACV*, 1727–1735.