Representation Learning of Next Shot Selection for Vlog Editing

Yuqi Zhang Bin Guo Nuo Li Ying Zhang Qianru Wang Zhiwen Yu
Department of Computing, Northwestern Polytechnical University
{yuqizhang, linuo}@mail.nwpu.edu.cn, {guob, izhangying, zhiwenyu}@nwpu.edu.cn, qr369wang@gmail.com

Abstract

As vlog has become increasingly popular on video-sharing platforms, more amateurs have participated in vlog creation. One critical step of attractive vlog creation is multi-shot assembly, which is to compose shots to form a coherent and engaging narrative. The process highly requires cinematographic knowledge and remains challenging for inexperienced users. Therefore, we aim to achieve automatic cinematographic-aware shot assembly for vlog editing. Our key idea is to automatically mine the cues of shot assembly from large-scale well-edited vlog data. To this end, based on cinematographic theories and experiments, we find that semantics and scale are crucial for shot assembly. Accordingly, we propose a Two-stream Cinematographic-aware Contrastive (TCC) model to learn the representation that discriminates between the good next shot choice against other options. Quantitative results clearly demonstrate the effectiveness of the proposed methods against other alternative baselines.

1. Introduction

In recent years, vlog has become a trending form of video recording individual life on the video-sharing platform as Youtube [1] and Bilibili [2]. The vlog market in China amounted to 221 billion yuan by 2020 and is estimated to reach 516 billion yuan in 2023 [3]. Due to the popularity of this kind of media and the ubiquity of filming devices, video creation has attracted numerous inexperienced users' interests. However, due to their limited expertise and technical skills, these videos usually have poor quality and minimal attention on the platform. One of the critical steps in generating an attractive vlog is multi-shot assembly, which is to compose several shots to form a continuous narrative by considering cinematographic elements like scale, camera movement and etc. It highly demands cinematographic knowledge and editing skills, presenting a significant challenge for inexperienced users. Therefore, how to automatically achieve cinematographic-aware shot assembling for vlog editing is a key and unexplored issue.

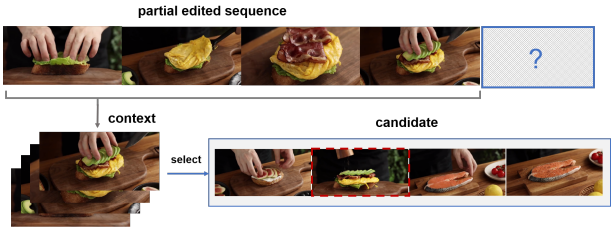

Figure 1. Next shot selection task.

Multi-shot assembly has been well studied for various types of videos from movies [2], interviews [9], live performances [14], social gatherings [1] to animation [6]. For a video type, specific and fixed cinematographic rules are usually pre-defined to assist in video editing. However, due to the unique characteristic of vlogs(i.e., short duration and fast switching techniques), it makes the above pre-defined rule-based methods hardly applicable. Also, depending on fixed predefined rules, it is difficult to imitate what professional editors do and barely produce appealing videos. Recently, some automatic methods generating videos with transcripts employ the text narrative and some fixed rules [17, 18, 19]. These works focus on text-visual matching rather than cinematographic relationships among shots, while the latter is crucial to video aesthetics. In summary, different from previous works, this paper aims to learn the cinematographic cues of multi-shot assembly from well-edited professional-generated vlogs(PUGVs) without predefined rules or transcripts for automatic vlog editing.

Similar to [2], we model the process of multi-shot assembly in a sequential manner and propose a new task of next shot selection (NSS) in vlog editing, which is to recommend the next shot from a candidate shot list given the partially edited shot sequence as a context, shown as Figure 1. This step repeats until the video meets the purpose of creation. According to editing theory [3, 15, 11, 4], we consider and verify two crucial aspects during selection: **semantic**

[1]https://www.youtube.com/
[2]https://www.bilibili.com/
[3]https://www.statista.com/statistics/874562/china-short-video-market-size/

continuity and **smoothness of scale transition**, and then propose a two-stream cinematographic-aware framework to solve the NSS task for automatic vlog generation. Finally, we collect a PUGV video dataset (proVlog) to demonstrate the superiority of the TCC method against other baselines.

2. Related Work

Existing works of multi-shot assembly can be mainly divided into rule-based and transcript-based approaches.

The rule-based methods predefine cinematographic rules for a specific scenario, like movies [2], interviews [9], live performances [14], social gatherings [1], animation [6] and so on. For instance, continuity check in animation(e.g., jump cut, opposite motion, and reverse ordering of characters) [6], and "start wide" rule in interview video, which refers to beginning with a wide-scale shot [9]. However, these methods rely on fixed rules for one specific field and could hardly adapt to varying situations. Therefore, it cannot utilize different strategies for different scenarios like a professional editor and create an engaging vlog.

The automatic approach, the transcript-based method, is to generate a corresponding video, given the textual description and some fixed rules. The organization of the shot sequence follows the textual narrative and the rules simultaneously [17, 18, 19]. However, these methods mostly focus on text-visual matching and still rely on predefined simple rules. And the cinematographic associations among shots in the visual modality are not explored. Recently, Pardo et. al. [10] leverage audio-visual cues in the movies to learn cut-trigger patterns. Although it achieves good results, the model learns the cinematographic associations in an implicit manner, which is difficult for the users to understand.

3. Preliminary

3.1. Problem statement

3.1.1 Types of Shot Assembly

In video production, shot assembly aims to arrange and combine shots together in a sequence to convey a cohesive and engaging narrative, which is driven by various cues [11, 4]. For example, one of the commonly used techniques in the movie, Reaction Cut, is to combine the shot of one event and the shot showing the subject's reaction to it altogether, to emphasize the emotional impact. According to cinematographic theories [11, 4, 3, 15], semantics play an important role in conveying continuous narrative, and the scale transition presents different perspectives and improves the overall visual aesthetic. Thus we introduce semantic-driven assembly and scale-driven assembly.

- Semantic-driven assembly: Arrange shots in a logical order that follows the narrative, actions, or events depicted in vlogs. It ensures a coherent flow of visual information and narrative continuity.
- Scale-driven assembly: Arrange shots based on their scales, such as the transition between extreme close shots, close shots, medium shots, etc. It provides various content views, highlights specific details, and improves the visual variety and overall aesthetic appeal.

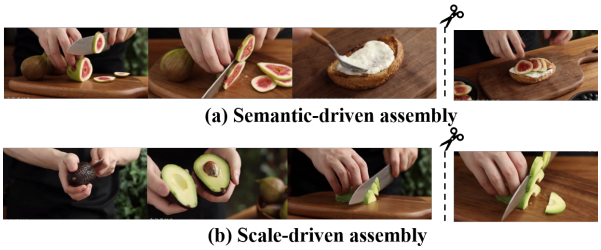

Figure 2. Examples of shot assembly.(a) semantic-driven assembly. The first three shots show actions of making a sandwich (preparing the ficus carica and spreading yogurt on toast), and the next shot presents an action of putting the fruit on the yogurt toast, which has a strong and long-range semantic dependency on the context sequence. (b) scale-driven assembly. The last shot of the context presents the action of cutting an avocado in an extremely close shot, and the next shot is the same action in a close shot, which is scale-coherent with the last shot of the context sequence.

3.1.2 Task Definition

Based on the workflow of editing, we simplify the multi-shot assembly as a sequential process. The user provides a short and ordered sequence $X = (x_1, x_2, ..., x_n)$ as the context and unordered candidate shots $C = \{c_1, c_2, ..., c_m\}$ as the candidate set. Given the context sequence X and shot candidate set C, we aim to select the optimal next shot c^* from C, which is closest to the context in the feature space.

$$c^* = \arg\min_{c_i} \{D(f_1(X), f_2(c_i)) | c_i \in C\} \quad (1)$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are two functions that could project the context sequence and the candidate shots respectively into a joint feature space, and during the inference phase, we apply $D(a, b)$ to measure the difference between a and b. In this paper, we adopt cosine distance.

3.2. Preliminary Experiment

Since there is no publicly available PUGV video dataset, we collect 205 well-edited PUGV videos from the Bilibili platform. Then we analyze the dataset to find out whether there are certain patterns of shot assembly driven by semantic and scale transition.

Since the collected videos are narrated cooking vlog videos, they describe steps of dishes, and scattering shots are combined based on their semantic associations.

For scale, we verify the pattern of scale transitions between adjacent shots in the dataset, and results are shown in Figure 3. It clearly shows that Close shot and Medium shot are commonly used and Long shot is barely applied. The Medium Shot tend to succeed a Close shot. For an Extreme Close shot, it is frequently followed by a Close shot.

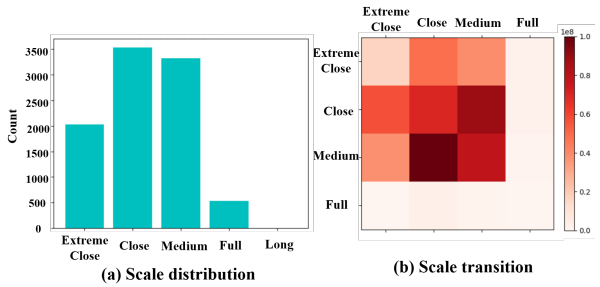

Figure 3. Shot scale distribution and transitions.

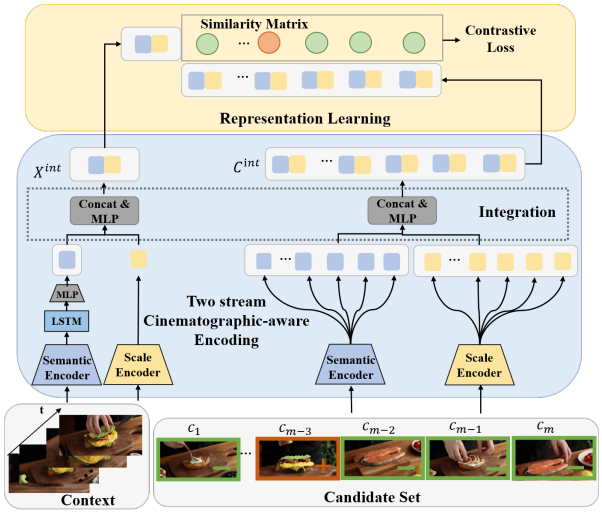

Figure 4. Overview of method.

4. Two-stream Cinematographic-aware Contrastive Framework

According to cinematographic theories and preliminary experiment results, the long-range semantics and short-range scale information are crucial for choosing a good next shot. Inspired by this, we propose a two-stream cinematographic-aware contrastive model to learn a feature space discriminating between a good next shot choice against bad ones for one specific context. As shown in Figure 4, it contains Two-stream Cinematographic-aware Encoding and Representation Learning.

4.1. Two-stream Cinematographic-aware Encoding

We propose a two-stream cinematographic aware encoding scheme to jointly represent the context and candidate shots, including Semantic Encoding, Scale encoding.

Semantic Encoding. We adopt ResNet50 [8] pretrained in ImageNet [13] as semantic encoder $h(\cdot)$ to encode semantics of each shot. To represent the semantics of the context sequence, we encode each shot and apply the LSTM method to obtain long-range semantics shown in Eqn 2.

$$\boldsymbol{X}^t = LSTM(h(\{x_1, x_2, ..., x_n\})) \tag{2}$$

The semantic representation of each candidate is denoted as $\boldsymbol{c}^t = h(c_i), \boldsymbol{c}_i^t \in \boldsymbol{C}^t, i = 1, 2, .., m$.

Scale Encoding. we adopt a ResNet50 model pretrained on the MovieShots [12] and fine-tuned on our vlog dataset as the scale encoder $l(\cdot)$, to better characterize the scale type (e.g., Extreme Close shot, Close shot, Medium shot, Full shot). For the scale embedding of the context, we only infer from the last shot of the context sequence to capture the short-range scale association shown in Eqn 3.

$$\boldsymbol{X}^a = l(x_n) \tag{3}$$

The scale representation of candidates is denoted as $\boldsymbol{c}_i^a = l(c_i), \boldsymbol{c}_i^a \in \boldsymbol{C}^a, i = 1, 2, .., m$.

Integration. To better infer the next shot, we should jointly consider both the semantic and the scale information. Thus, here we employ a simple integration strategy to fuse two streams. We concatenate the two features and fed them into a two-layer MLP to do projection. The fused representations of the context and candidate shot c_i are calculated separately as:

$$\begin{aligned} \boldsymbol{X}^{int} &= g_1(cat(\boldsymbol{X}^t, \boldsymbol{X}^a)) \\ \boldsymbol{c}_i^{int} &= g_2(cat(\boldsymbol{c}_i^t, \boldsymbol{c}_i^a)), \boldsymbol{c}_i^{int} \in \boldsymbol{C}^{int} \end{aligned} \tag{4}$$

g_1 and g_2 denote two-layer MLPs without sharing weights.

4.2. Representation Learning for Next Shot Selection

We employ contrastive learning-based representation loss to guide model to learn feature space that maximizes the similarity between the good next shot and the context based on semantic continuity or scale-change smoothness.

Positive-Negative Pair Construction. we exploit the context as the anchor, the adjacent shots following the context in the actual sequences as positive samples c^+ and other subsequent $m-1$ shots as negative sample set $N = \{c_{k_1}^-, c_{k_2}^-, ..., c_{k_{m-1}}^-\}$. The candidate shot list C is composed of positive and negative samples in shuffled order.

Contrastive Loss. We aim to project the context and candidates into a feature space where the good next shot choice c^+ is close to the context while the other alternative candidate shots are far from the context. We use infoNCE loss [7] to learn the representation as follows:

$$L_{\text{contrastive}} = -\log \frac{e^{D(\boldsymbol{X}^{int}, \boldsymbol{c}^{+,int})}}{\sum_{i=1}^{m} e^{D(\boldsymbol{X}^{int}, \boldsymbol{c}_i^{int})/\tau}} \tag{5}$$

where $\boldsymbol{c}^{+,int}$ represents the fused representation of the positive sample c^{+}, and $\boldsymbol{c}_i^{int} \in \boldsymbol{C}^{int}$.

5. Experiments

5.1. Experiment Setup

Dataset. For the collected dataset, We segment videos into shots [16] and manually filter noise data. We segment them into sequences of $n+m$ shots with the stride of one shot. The preceding n shots serve as the context, and the candidate list consists of the subsequent m shots in a shuffled order. Here, we take $n=4$, and $m=5$ in the experiments. The statistics of the dataset are shown in Table 1.

Table 1. Dataset statistics

Type	Statistic
Num. of videos	205
Avg. duration of videos (min.)	2 : 51
Total duration of videos (min.)	586 : 54
Num. of shots	8115
Avg. duration of shots (sec.)	3.54
Num. of sequences (when $\mathrm{n}=4,\ \mathrm{m}=5$)	8179

Metric. Considering the artistic nature of the task, there might be more than one appropriate candidate shot to form a smoothly combined shot assembly. During the inference phase, our model should recommend a few shots in a ranked order for users to choose, achieving diverse and satisfying editing effects. Therefore, we employ **Recall@k** ($k \in \{1,3\}$) as the evaluation metric.

Baselines. Our task is to score each candidate the possibility of being a good next shot, according to the affinity between the context and candidates in the learned feature space. As there are few works with the same goal as this paper, below we introduce the potential baselines and ablations of the proposed method.

- Random baseline: For each context sequence, this method assigns a uniform random score for each shot in the candidate list.
- AVE-NSS: This method is proposed by [2] as one downstream task for a movie dataset and we here call it AVE-NSS. They adopt a different way of positive-negative sample construction and NT-Xent contrastive loss [5]. For details, please see [2].
- TCC-semantic and TCC-scale: Respectively employ a single corresponding stream.

5.2. Quantitive Results

As Table 2 shows, TCC and its two ablations outperform other baselines, which demonstrates the effectiveness of semantic and scale information. Compared to AVE-NSS, it shows the effectiveness of our design of the positive-negative pair construction and the contrastive loss. Besides, comparing TCC-semantic, TCC-scale against TCC, TCC gains better performance on Recall@1 but performs worse on Recall@3. Since the role of semantics and scale could vary across different shot assembly types, it could be difficult for a two-layer MLP integration to capture the dynamic importance variation.

Table 2. Quantitative results comparing with baselines

Model	Recall@1	Recall@3
Random	20.00	60.00
AVE-NSS	24.70	62.23
TCC-semantic	35.28	**74.73**
TCC-scale	31.05	**74.24**
TCC	**35.47**	74.11

Figure 5. The example of Top-1 next shot results.

We show one example presenting the top-1 next shot results of three methods in Figure 5. The original sequence presents the process of making a Cheese Roll Hot Pot. The context sequence illustrates the steps of mixing cheese and putting rolls into the hot pot. The next shot should present beef rolls in the boiling pot. While the results generated by Random and AVE-NSS present the step of adding cheese and parsley, which emerge semantic gaps with the context. Our generated next shot forms a coherent narrative with the context and the transition from an extremely close shot and a close shot brings the audience an aesthetic appeal.

6. Conclusion

In this paper, we introduce the first attempt at a learning-based shot assembly for vlog editing without predefined rules or transcripts. Specifically, we first collect a PUGV video dataset, then find that semantics and scale are crucial for automatic shot assembly. Finally, we propose a two-stream cinematographic-aware contrastive model to learn the cues of the next shot selection from well-edited videos.

Despite our contributions, there remain open challenges. It lacks further explorations on other cinematographic factors that are crucial for shot assembly (e.g., camera movement, tone, angle). Besides, it is necessary to propose an effective mechanism to dynamically assign importance to these factors based on different assembly goals, which could benefit the creative use of various artistic techniques.

References

[1] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 1, 2

[2] Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 201–218. Springer, 2022. 1, 2, 4

[3] Christopher J Bowen. *Grammar of the Edit*. Taylor & Francis, 2017. 1, 2

[4] Boris Chen, Amir Ziai, Rebecca S Tucker, and Yuchen Xie. Match cutting: Finding cuts with smooth visual transitions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2115–2125, 2023. 1, 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[6] Quentin Galvane, Rémi Ronfard, Christophe Lino, and Marc Christie. Continuity editing for 3d animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 1, 2

[7] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[9] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. *ACM Trans. Graph.*, 36(4):130–1, 2017. 1, 2

[10] Alejandro Pardo, Fabian Caba, Juan León Alcázar, Ali K Thabet, and Bernard Ghanem. Learning to cut by watching movies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6858–6868, 2021. 2

[11] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. Moviecuts: A new dataset and benchmark for cut type recognition. In *European Conference on Computer Vision*, pages 668–685. Springer, 2022. 1, 2

[12] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A unified framework for shot type classification based on subject centric lens. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 17–34. Springer, 2020. 3

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3

[14] Mukesh Kumar Saini, Raghudeep Gadde, Shuicheng Yan, and Wei Tsang Ooi. Movimash: online mobile video mashup. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 139–148, 2012. 1, 2

[15] Tim J Smith. An attentional theory of continuity editing. 2006. 1, 2

[16] Tomáš Souček and Jakub Lokoč. Transnet v2: an effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 4

[17] Miao Wang, Guo-Wei Yang, Shi-Min Hu, Shing-Tung Yau, Ariel Shamir, et al. Write-a-video: computational video montage from themed text. *ACM Trans. Graph.*, 38(6):177–1, 2019. 1, 2

[18] Zheng Wang, Jianguo Li, and Yu-Gang Jiang. Story-driven video editing. *IEEE Transactions on Multimedia*, 23:4027–4036, 2020. 1, 2

[19] Yu Xiong, Fabian Caba Heilbron, and Dahua Lin. Transcript to video: Efficient clip sequencing from texts. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5407–5416, 2022. 1, 2