2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 978-1-6654-9548-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/MIPR54900.2022.00055

Bridging Script and Animation Utilizing a New Automatic Cinematography Model

Zixiao Yu[1], Enhao Guo[2], Haohong Wang[3], Jian Ren[1]
[1]Department of ECE, Michigan State University, East Lansing, MI 48824-1226.
Email: {yuzixiao, renjian}@msu.edu
[2] Henry M. Gunn High School, 780 Arastradero Rd, Palo Alto, CA 94306
benny927guo@gmail.com
[3]TCL Research American, 2025 Gateway Place, Suite 460, San Jose, CA 95110
Email: haohong.wang@tcl.com

***Abstract*—For animated film-making, automatic cinematography is an effective approach for junior filmmakers to speed up the process. The virtual camera placement in the 3D environment can be automatically conducted by auto-cinematography algorithms. In the literature, the algorithms proposed were mainly optimized from aesthetic point of view, thus the scene captured by virtual cameras may not effectively represent the intention of the script sometimes. In this paper, a new model that incorporates the fidelity and aesthetic models in a unified framework was proposed, so that the visual presentation of the input script and the compliance of the generated video with given cinematography specifications can be jointly considered during the optimization process. In this way, the problem of virtual camera placement is translated into an optimization problem that can be solved through dynamic programming that optimizes the computational efficiency. Experimental results shows that the proposed algorithm has quality improvement up to 35% compared to the earlier methodology and the content of the video is easier to understand. More video examples can be reviewed https://youtu.be/0PUdV6OeMac.**

I. Introduction

With the latest developments in AI technology, a significant amount of the animation production process can be automatically taken care of by computer program. In near future, general people may gain capability to create films on their own by simply writing their stories into a script and then rely on the software to do the rest [1]. In this process, transforming the cinematography stage into an automatic process, also known as *Auto Cinematography*, is very attractive because the automation process does not require much corresponding film-making knowledge and is able to dramatically reduce the production time for camera placement in the 3D virtual environment.

In [2], [3], cinematic rules or conventions were used as constraints to select virtual cameras during the auto cinematography process. In [4], the additional information about the director's guidance for each shot was given and used in the auto cinematography optimization process, and in [5], the camera usage and behaviors were learned from the existing films to enhance the camera selection process. A common assumption that these works have made is that the aesthetic aspects of cinematography can be utilized for the camera selection and optimization, however, how to assure that the scenes captured reflect the intention of the script has not been seriously considered. Given the fact that sometimes even a slight mismatch between script and outcome video may confuse viewers, it is worth putting the script in a more significant position during the auto cinematography process.

In this paper, we argue that bridging script and scene, and connecting visual and text, may significantly improve the capacity for auto cinematography. Therefore, in addition to the existing aesthetic model described in the literature, a new quality assessment model, called the fidelity model, is proposed for the cinematography optimization process to determine whether the same content is expressed by two different media, video and script, is fully aligned. Therefore, the proposed methodology aims to achieve the following two goals: (i) the output animation maintains reasonable fidelity of the script, and (ii) the outcome video film follows cinematic rules with cinematographic aesthetics.

To estimate the proposed method, an animation production framework, named Text2Animation (T2A), was developed to implement the fidelity model by incorporating the latest video understanding advances [6]. By combining the aesthetics model with the fidelity requirements into a unified computational framework, the original automatic cinematography problem can be mapped into an optimization problem that seeks to select the best options of camera placements and achieve the aesthetic quality expectations and the consistency of content. The optimization problem can be formulated in a way that dynamic programming can be utilized to achieve an efficient solution.

To the best of our knowledge, T2A is the first framework in the literature that introduces the fidelity distortion factor in the optimization process of auto cinematography from a new perspective.

The rest of the paper is organized as follows: Section II overviews the related work to this study, such as computational cinematography, video editing, and video understanding. Section III shows the framework, the problem formulation, and the dynamic programming solution of T2A. Section IV discusses impact of possible error in video understanding to the whole framework and the experimental results, and we conclude in

2770-4319/22/$31.00 ©2022 IEEE
DOI 10.1109/MIPR54900.2022.00055

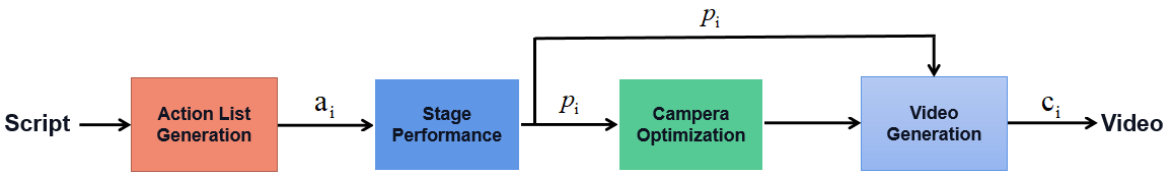

Fig. 1: Existing Framework for Auto Cinematography

Section V.

II. Related Work

The camera configuration and placement are crucial to the shooting phase of the movie production process. The views captured by the camera directly affect the quality of the production. The automatic camera placement process follows the following two approaches: first, select the best camera options using the cinematography guidelines as constraints. Here different cinematography rules can be applied to the film-editing under various scenarios, such as dialogue scenes [7] and cooking [8]. The advantage of these methods is that the generated video is more closely aligned to the aesthetic standards of the desirable cinematography guidelines. Second, learn cameras' behaviors from the existing movies [5], [9], [10]. The advantage of these methods is that the optimal camera solution can be matched by the existing successful camera settings based on the state of the characters and the scene. The challenge is that the learned results cannot be directly applied to general scenes.

With the issues and challenges mentioned in [11], almost all computational cinematography works, according to our knowledge, optimize the camera path from the cinematography guideline perspective. However, in filming practices, the guideline is not always been treated as the golden rule and most of the great films break these rules. Inspired by the recent works [12], we propose an evaluation measurement called fidelity distortion, that can objectively assess the quality of the resulting video. The fidelity stands for the consistency of the video and original script content, which has been ignored in both optimizations mentioned above. The fidelity model helps the T2A maintain consistency by using the video caption and action recognition model [6], [13] as an objective viewer, and evaluates the video content to see whether it has successfully visualized the corresponding actions in the action list. In the latest research advances, video caption [14], [15] and video scene recognition [16] are able to describe more and more details for a specific video. Although none of these technologies are perfect at the moment, we believe that as we continue to refine them, our proposed fidelity model will eventually be able to achieve all its goals.

III. Framework Design

In a typical auto-cinematography system, the whole process from the input script to the output video contains the following steps, namely, action list generation, stage performance, camera optimization, and video generation. The action list generation module analyzes the original animation script to obtain the corresponding characters' chronological action list $\{a_i|i = 1, 2, \cdots, N\}$, where a_i is the ith action object in the scene, and N is the total number of action objects. It is important to realize that multiple characters might perform simultaneously (e.g., two persons are fighting with each other, or a mom is hugging her daughter). Thus an action object may contain multiple characters in the same scene. In the stage performance step, the input $\{a_i\}$ is transformed into the corresponding stage performance data $\{p_t|t = 0, 1, \cdots, T\}$, where p_t is the character stage performance at time t and T is the total performance time determined by the action list. Specifically, for each a_i, the corresponding performance can be denoted as $\{p_{t_{a_i}}, p_{t+1_{a_i}}, \cdots, p_{t+l_{a_i}}\}$, where l_{a_i} is the action duration of a_i. There are multiple virtual cameras available in the 3D scene that can record all the views for every character from various angles. In the camera optimization step, all the available views are considered to calculate the optimized camera $\{c_t\}$ for each time t. The video generation step assembles all the video frames captured by camera $\{c_t\}$ at time t and outputs the final video.

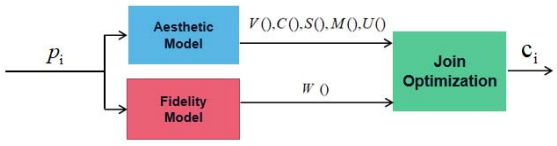

Fig. 2: The updated camera optimization model in Text2Animation (only reflects the camera optimization module shown in Fig. 1)

In this work, we make a bold assumption that a mathematical model can be found to approximate the fidelity relationship between a video and its associated script (i.e., a_i). In other words, with any action a_i and any selected camera $a_{t_{a_i}}$, the fidelity between the action and the video generated from this camera at time t can be obtained from this approximation model. In Fig. 2, the camera optimization module has been updated in our proposed framework with three new modules, namely, aesthetic model, fidelity model, and optimization. The task of the aesthetic model is to provide a quality evaluation from an aesthetic point of view for each admissible virtual camera at time t to the optimization engine, the task of the fidelity model is to provide fidelity evaluation for each admissible virtual camera at time t to the optimization engine, and the optimization engine considers all inputs and makes the optimal choice for the camera selection.

A. Aesthetic Model

In this section, we discuss the aesthetic evaluation model, which emphasizes the distortion caused by camera planning and measured by the cinematography guidelines.

Character Visibility: This cost evaluates the character visibility in the c_t and it is determined by two factors: (1) r_k, the size of the character k in the frame to total the frame size. (2) $I(c_t, k)$, the different weight to different characters and camera combinations during the calculation. Thus the cost function for character visibility $V(c_t)$ can be represented as:

$$V(c_t) = \sum_{\text{character } k=0 \to K-1} I(c_t, k) \cdot r_k. \tag{1}$$

Character Action: It describes whether the character has a_t (action at t) or not. The cost function can be represented as:

$$A(c_t) = \begin{cases} 0 & c_t \text{ bounded character } k \text{ has action at time } t \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

Camera Configuration: The camera configuration distortion depends on the action type that can be derived from the action object $\tilde{a_t}$ of time t. We use the $\phi_C()$ function to describe this distortion calculation process [2]. Thus the cost function of camera configuration can be represented as:

$$C(c_t) = \phi_C(p_{c_i}, d_{c_i}, \tilde{a_t}), \tag{3}$$

where p is camera position, and d is shooting direction.

Screen Continuity: It is the summary of each single character position change in the frame. The cost function is defined as follows:

$$S(c_t, c_{t-1}) = \sum_{char} \sum_{k=0}^{K} v(k, c_t) \cdot \phi_S(p(k, c_t) - p(k, c_{t-1})), \tag{4}$$

where $p(k, t_i)$ and $p(k, t_{i+1})$ represent the kth position in the frame captured by c_t and c_{t+1}, while $v(k, c_t)$ determines whether character k is visible in the view of c_t or not.

Moving Continuity: It represents if an ongoing action changes of the direction of a character before or after the view change. The moving continuity cost quantifies the penalty in this aspect as follows:

$$M(c_t, c_{t-1}) = \sum_{char} \sum_{k=0}^{K} v(k, c_t) \phi_M(m(k, c_t) - m(k, c_{t-1})), \tag{5}$$

where $m(k, c_t)$ is the motion direction vector of the character in frame captured by c_t. $\phi_M()$ is penalty function and it increases as these diverge from each other.

Shot Duration: We allow the average shot duration $\overline{u}$ to be set for each scene to control the shot duration distribution or use the default value. The $\phi_U()$ is an non-linear penalty function with 0 at $x = \overline{u}$. Let p be the longest allowable shot duration which is defined as the penalty of the frames in the range $[t-q, \cdots, t]$ for cameras changing, and we have:

$$U(\overline{u}, c_t, c_{t-1}, \cdots, c_{t-q}) = \phi_U(\overline{u}, c_t, c_{t-1}, \cdots, c_{t-q}). \tag{6}$$

By adding all the factors mentioned above together, the total aesthetic distortion D_a can be calculated by the following equation:

$$\begin{aligned} D_a = & \sum_{t=0}^{T} [\omega_0 \cdot V(c_t) + \omega_1 \cdot C(c_t, \tilde{a_t}) \\ & + \omega_2 \cdot A(c_t) + \omega_3 \cdot S(c_t, c_{t-1}) + \omega_4 \cdot M(c_t, c_{t-1})] \\ & + \sum_{t=q}^{T} (1 - \omega_0 - \omega_1 - \omega_2 - \omega_3 - \omega_4) \cdot \\ & U(\overline{u}, c_t, c_{t-1}, \cdots, c_{t-q}), \end{aligned} \tag{7}$$

where $\omega_0, \omega_1, \omega_2, \omega_3, \omega_4$ are the weights for each distortion component within a range of 0 to 1.

B. Fidelity Model

The fidelity model is the essential element of T2A, which assures that the generated video matches the video input script. In an ideal case, there is a human-like agent that has comprehension intelligence similar to humans. However, even the current state-of-the-art model is still far from being widely utilized due to the low performance and accuracy. To address this challenge, an approximation method as shown in Fig. 3, is considered to compare each action generated from the action list module. Therefore, the original text-video matching problem is approximated and converted into a text-text matching problem with the assumption that the video action recognition engine can achieve a reasonable quality.

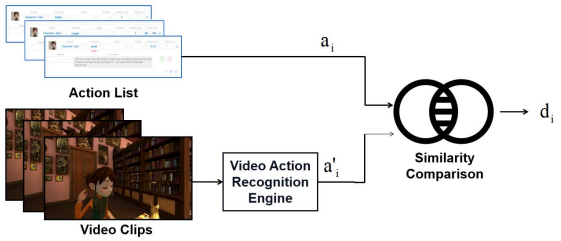

Fig. 3: Using a video action recognition engine to convert a video clip into a text, and then compare the similarity between two actions in text format.

Let us denote by m_j the jth camera of the admissible camera set, a'_i the word or phrase that describes the action obtained from the video action recognition engine (which is able to recognize the video generated from input action a_i), d_i the similarity between the textual description of a_i and a'_i, then d_i can be measured by:

$$d_i = \begin{cases} 0 & \frac{G(a_i) \cdot G(a'_i)}{\|G(a_i)\| \times \|G(a'_i)\|} \leq \mathsf{Th}_G \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

where Th_G is the threshold to admit that a_i and a'_i refer to the same action, and function G is the Glove word embedding model of [17]. In this way, the fidelity level of a generated video and its corresponding input script can be approximated by the average of all the action similarities obtained as:

$$F_j = \frac{1}{N} \sum_{i=1}^{N} d_{i,j}. \tag{9}$$

It is important to emphasize that the equation above is just an approximation of the real fidelity level, as noise may have been introduced during the whole process by the video action recognition engine and the textual similarity comparison mechanism. The human-involved simulation observed accuracy (accept rate) is 87.3% in our experiment by comparing the outcome calculated by equation (9) and human subjective judgment with a pre-defined threshold for acceptance.

The full (or partial) occlusion of characters dominant the performance of the action recognition. Hence it is intuitive to investigate whether the degree of occlusion of characters according to all admissible camera parameters is correlated to the $d_{i,j}$ obtained in equation (8). Let us denote by $o_{i,j}$ the occlusion percentage of all characters involved in a_i shot by camera m_j, where J is the total number of admissible cameras. Then the average occlusion O_j can be calculated by:

$$O_j = \frac{1}{N}\sum_{i=1}^{N} o_{i,j}. \tag{10}$$

As shown in Fig. 4, there is a high correlation between F_0 and O_0 (i.e., the fidelity and occlusion level measured by the 0th camera of the admissible cameras), and their relationship can be approximated by a linear function.

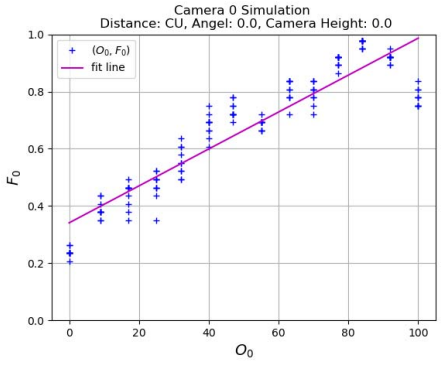

Fig. 4: The plot of pairs of (O_0, F_0) in dots for the 0th default camera and the linear function $F_0 = \alpha O_0 + \beta$ to fit the dots

Let us denote by D_f the fidelity distortion; then D_f can be modelled and represented by a function of the selected camera at each time t, as the object occlusion level is determined once the camera is specified at a certain timestamp. Therefore, D_f can be calculated by:

$$D_f = \frac{1}{T}\sum_{t=0}^{T}[\alpha O(c_t) + \beta], \tag{11}$$

where $O(c_t)$ is the occlusion measurement function for all subjects and objects for the selected camera c_t at time t, α and β are the parameters that can be derived by fitting the (O_j, F_j) pairs by a linear function.

C. Joint Optimization

We can optimize both aesthetic and fidelity aspects by using a weighting factor λ between [0, 1] to bridge both models, the total distortion can be represented as follows:

$$D = (1-\lambda)D_a + \lambda \cdot D_f. \tag{12}$$

When λ is set to a value close to 1.0, the fidelity distortion is more important, otherwise for λ close to 0, the aesthetic distortion becomes dominant.

Our goal is to find the optimal solution $\{c_t^*\}$ such that $\{c_t^*\} = \text{argmin}_{c_t} D^*$. To implement the algorithm for solving the optimization problem, we define $z_k = c_k$ and a cost function $D_k(z_{k-q}, \cdots, z_k)$, which represents the minimum total distortion up to and including the kth frame, given that $z_{k-q}, \cdots, z_k$ are decision vectors for the $(k-q)$th to kth frames. Therefore $D_T(z_{T-q}, \cdots, z_T)$ represents the minimum total distortion for all frames, and thus

$$\min_z D(z) = \min_{z_{T-q},\cdots,z_T} D_T(z_{T-q}, \cdots, z_T) \tag{13}$$

The key observation for deriving an efficient algorithm is the fact that given $q+1$ decision vectors $z_{k-q-1}, \cdots, z_{k-1}$ for the $(k-q-1)$st to $(k-1)$st frames, and the cost function $D_{k-1}(z_{k-q-1}, \cdots, z_{k-1})$, the selection of the next decision vector z_k is independent of the selection of the previous decision vectors $z_1, z_2, \cdots, z_{k-q-2}$. This means that the cost function can be expressed recursively as

$$\begin{aligned} & D_k(z_{k-q}, \cdots, z_k) \\ = & \min_{z_{k-q-1},\cdots,z_{k-1}} \{D_{k-1}(z_{k-q-1}, \cdots, z_{k-1}) \\ & + \frac{\lambda}{T}[\alpha O(c_k) + \beta] + (1-\lambda)\{[\omega_0 \cdot V(c_k) \\ & + \omega_1 \cdot C(c_k) + \omega_2 \cdot A(c_k) \\ & + \omega_3 \cdot S(c_k, c_{k-1}) + \omega_4 \cdot M(c_k, c_{k-1})] \\ & + (1-\lambda)(1-\omega_0-\omega_1-\omega_2-\omega_3-\omega_4)\cdot \\ & U(\overline{u}, c_k, c_{k-1}, \cdots, c_{k-q})\}\}. \end{aligned} \tag{14}$$

The recursive representation of the cost function above makes the future step of the optimization process independent from its past step, which is the foundation of dynamic programming. The problem can be converted into a graph theory problem of finding the shortest path in a directed cyclic graph (DAG). The computational complexity of the algorithm is $O(T \times |Z|^{q+1})$ (where $|Z|$ is the cardinality of Z), which depends directly on the value of q. For most cases, q is a small number, so the algorithm is much more efficient than an exhaustive search algorithm with exponential computational complexity.

IV. Experimental Results

In this section, the details of the simulation fidelity model are demonstrated, and the proposed framework is evaluated by comparing the proposed camera optimization framework with the state-of-the-art solution without fidelity model.

The video action recognition engine is built following the work of sequence to sequence neural network for video caption [18] with additional attention layer; it is first trained and tested with the MSR-VTT public data set [19] and then trained using our own data set for 2000 epochs. The Adam optimizer [20] is utilized with a batch size of 128, and the learning rate starts with 0.0004 and gets decreased every 200 epochs by multiplying it by the decay factor of 0.8. It takes around 8 hours to finish the whole training process based on the NVIDIA GTX 2080 TI GPU.

Fig. 5: Sample frames for the whole action maintaining chronological order from left to right. The optimization using the fidelity model obtains the camera settings (bottom), representing the entire action in the frame.

A. *Value of Fidelity Model*

To evaluate the value of the fidelity model, we compared the proposed framework with an existing auto-cinematography solution [4] which uses an aesthetic model in the optimization. The experiments indicate that the proposed solution achieved better performance for the various scenarios as shown in the following:

Entire action: In animation, the character's action can take seconds to complete thus may cross dozens of frames. As shown in Fig. 5, the video frames generated without the fidelity model (as shown in the top row) were compared with the ones with the fidelity model (as shown in the bottom row), and the sample frames for the whole action were demonstrated in chronological order from left to right. It is very clear that the bottom row can represent a good abstraction of the entire movement of characters' actions while the top row cannot because there is a referee built in the system to keep the video frames easy to understand.

Full-body: In animation, some of the character movements need the virtual camera to be able to capture the whole body of the character to provide a better viewing experience. Fig. 6 shows one of the examples, we can see that the action "push" of the character is better captured by the full body camera.

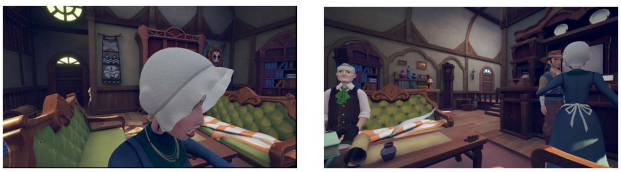

Fig. 6: Captured "push" action scene from the script: Lead the witch finder inside, the optimization was done only for D_a (left), the optimization was done based on $D_a + D_f$ (right)

Single-character: The use of a single character shot is also very important in animation. For example, to express the emotion of anger, a single close-up of the character's face is better to convey this emotion to the audience (as shown in Fig. 7)

Fig. 7: Captured "angry" emotion scene from the script: Lead the witch finder inside, the optimization was done only for D_a (left), the optimization was done based on $D_a + D_f$ (right)

Fig. 8: Captured dialogue scene from the script: interrupt the conversation, the optimization was done only for D_a (left), the optimization was done based on $D_a + D_f$ (right)

Multi-character: Multi-characters interaction in animation is very common. Take dialogue scenes as an example, for the audience to understand the relationship between the characters, it is sometimes necessary (as shown in Fig. 8) for both sides of the dialogue to appear in the frame at the same time.

To better demonstrate the performance of various models, three example videos of comparison results can be found in video starting from 02:33 https://youtu.be/MMTJbmWL3gs (Fidelity Model Comparison) to experience the differences between the animation videos with and without the fidelity model in the joint optimization.

B. *Ablation Studies*

For ablation studies,the influence of distortion weight coefficients (λ) and main parameters (V, C, S, U) on the opti-

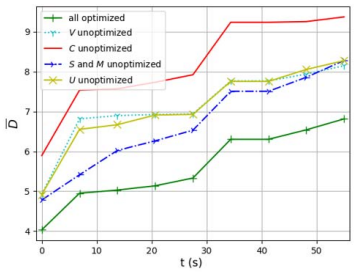

Fig. 9: Aesthetic component impact on the optimization results ($\lambda = 0.1$)

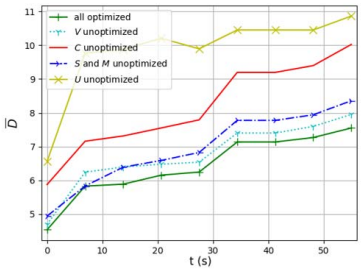

Fig. 10: Aesthetic component impact on the optimization results ($\lambda = 0.5$)

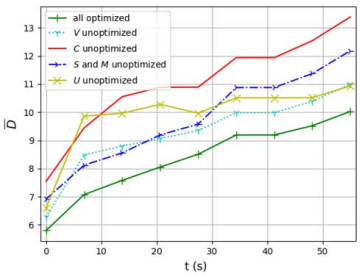

Fig. 11: Aesthetic component impact on the optimization results ($\lambda = 0.9$)

mization results are compared separately for the optimization process.

In Figs. 9, 10, 11, the optimized curves are compared with others with one selected component that is intentionally not optimized, and the setting of λ= 0.1, 0.5, and 0.9 are compared to demonstrate the impact of fidelity. It can be observed that the overall impact of the unoptimized component in λ=0.9 is smaller than that of λ=0.1, which is understandable because the former case emphasizes the fidelity factor much more than the aesthetic factor. The figures indicate that camera configuration has a very strong impact on the performance of the system with a gain of 25-32%, shot duration ranks the second with the gain of 16-35%, and visibility has the least impact.

V. Conclusions

3D Animation typically requires a professional team, sufficient funding and resources, knowledge of cinematography and film editing, and much more. Thus this field is, in general, not accessible to non-professional. In this paper, we presented T2A to reduce the animation production barrier for non-professional users. T2A used an automatic cinematography optimization method that can choose cameras and their associated settings based on a joint fidelity and aesthetic model, in which the comprehensiveness of visual presentation of the input script and the compliance of generated video with given cinematography specifications are mapped into a mathematical representation. Although the experimental results indicate both the time consumption and quality advantage of the proposed framework, we believe further investigations on the fidelity and aesthetic modeling are needed to make the solution generally applicable to a broader scope of film-making tasks.

References

[1] H.Wang, "Write-a-movie: Unifying writing and shooting," in *US Patent filed Oct.*, 2020.

[2] Q. Galvane, *Automatic Cinematography and Editing in Virtual Environments.* PhD thesis, Université Grenoble Alpes (ComUE), 2015.

[3] A. Louarn, M. Christie, and F. Lamarche, "Automated staging for virtual cinematography," in *Proceedings of the 11th Annual International Conference on Motion, Interaction, and Games*, pp. 1–10, 2018.

[4] L.Sun and H.Wang, "Director-hint based auto-cinematography," in *US Patent 11,120,638*, 2021.

[5] H. Jiang, B. Wang, X. Wang, M. Christie, and B. Chen, "Example-driven virtual cinematography by learning camera behaviors," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 45–1, 2020.

[6] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[7] M. Leake, A. Davis, A. Truong, and M. Agrawala, "Computational video editing for dialogue-driven scenes.," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 130–1, 2017.

[8] A. Truong, F. Berthouzoz, W. Li, and M. Agrawala, "Quickcut: An interactive tool for editing narrated video," in *Proc. 29th Annual Symposium on User Interface Software and Technology*, pp. 497–507, 2016.

[9] G. Abdollahian, C. M. Taskiran, Z. Pizlo, and E. J. Delp, "Camera motion-based analysis of user generated video," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 28–41, 2009.

[10] M. Gschwindt, E. Camci, R. Bonatti, W. Wang, E. Kayacan, and S. Scherer, "Can a robot become a movie director? learning artistic principles for aerial cinematography," *arXiv preprint arXiv:1904.02579*, 2019.

[11] M. Radut, M. Evans, K. To, T. Nooney, and G. Phillipson, "How good is good enough? the challenge of evaluating subjective quality of ai-edited video coverage of live events.," in *WICED@ EG/EuroVis*, pp. 17–24, 2020.

[12] M. Wang, G.-W. Yang, S.-M. Hu, S.-T. Yau, and A. Shamir, "Write-a-video: computational video montage from themed text.," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 177–1, 2019.

[13] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[14] J. Chen, L. Zhang, C. Bai, and K. Kpalma, "Review of recent deep learning based methods for image-text retrieval," in *2020 IEEE Conf. MIPR*, pp. 167–172, IEEE, 2020.

[15] A. Brown, E. Coto, and A. Zisserman, "Automated video labelling: Identifying faces by corroborative evidence," in *2021 IEEE 4th Int. Conf. MIPR*, pp. 77–83, IEEE, 2021.

[16] Z. Zheng, W. Zhong, L. Ye, L. Fang, and Q. Zhang, "Violent scene detection of film videos based on multi-task learning of temporal-spatial features," in *2021 IEEE 4th Int. Conf. on MIPR*, pp. 360–365, IEEE, 2021.

[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 conf. EMNLP*, pp. 1532–1543, 2014.

[18] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE conf. compu. vision and pattern recog.*, pp. 2625–2634, 2015.

[19] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proc. IEEE conf. on computer vision and pattern recognition*, pp. 5288–5296, 2016.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.