Journal of Artificial Intelligence Practice (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2025.080212
ISSN 2371-8412 Vol. 8 Num. 2

Integrating Application of Artificial Intelligence and Digital Imaging Techniques in Television Documentary Production

Zhixian Lu[1,a,*]

[1]*Institute of Media, Shanghai Lida University, Shanghai, China*
[a]*2430156669@qq.com*
**Corresponding author*

Keywords: Artificial Intelligence, Digital Imaging, Television Documentary, Temporal-Semantic Transformer, NeRF-Diffusion Model

Abstract: In the context of simultaneous upgrading of ultra-high-definition production and immersive storytelling, the integration of artificial intelligence and digital imaging technology needs to be urgently carried out in TV documentaries to dispel the bottleneck of efficiency and expression in the traditional editing-color grading-special effects pass. In this paper, we propose a D-DocFusion model, which uses a bidirectional temporal-semantic codec Transformer to jointly align scripts, interview texts and multi-camera RAW images to generate an editable "narrative timeline map". Subsequently, the improved Hierarchical NeRF-Diffusion module was introduced, which realized 3D duplication and super-resolution redrawing of old film sources with a controlled diffusion process while maintaining photometric consistency. Then, Cross-Attention Motion Composer fuses semantic clips with camera motion vectors to automatically generate tilt-shift, time-lapse, and virtual aerial trajectories that meet the director's intent. Comparative tests on BBC Planet Earth footage and its own 12 TB documentary library showed that D-DocFusion improved overall editing-grading efficiency by 47%.

1. Introduction

Transformations in TV Documentary Via New Digital Media Since digital technology and Internet become popular, documentary audience's asking for content is rising, documentary creation is also under pressure of innovation. Traditionally, documentary production process needs to invest a lot of manpower in manual work and spend a lot of time in post-processing, but the rapid development of modern technology, especially AI and digital image processing technology breakthroughs, make the documentary production process, not only quality improvement, and the number of efficiency leap [1]. The deep integration of these technologies brings new insights and tools to the production of documentary, not only expanding the performance space of the creator, but also making the audience feel the richness of visual experience.

The introduction of digital image processing technology provides a very precise image optimization method for TV documentary production. From basic color correction to complex VFX-

heavy productions, these techniques enable production teams to optimize and refine video footage to improve the quality and artistry of the final image. Digital color correction technology, such as used to restore the color of older films or change the visual style of the raw footage, can thus be appliedto increase the marketability and artistic merit of a film [2]. Meanwhile, the adoption of special effects technology enables production personnel to produce some scenes that are hard to achieve or, on the plot expressiveness make up the absence of traditional shooting project, further enrich the biodiversity of documentary works presented.

Moreover, AI-based image rundown technology allows the production team to create virtual sets and characters that are realistic. In traditional documentary films, factors such as location, available footage, and cooperation with participants often affect the shooting quality. This is an unlimited creativity space for documentary makers, as it allows the creator to build all kinds of scenes and characters in the virtual environment with the help of image generation technology [3]. Not only does this open up new avenues in the visual presentation of documentaries, but it is also a great incentive for content innovation. Virtual reality, for instance, enables documentary makers to recreate events in the past and present scenes that cannot be captured in the distance, or build entirely imaginary characters or settings to help viewers experience a truly immersive way.

However, this convergence of technologies presents new challenges as well. AI and digital image processing technologies can greatly enhance the efficiency of the production process and the quality of visual effects. However, the introduction of these technologies also involves many issues related to creativity, technology ethics, and authenticity of content. One of the ideals at the heart of the non-fiction realm of documentary is authenticity [4]. This can also lead to distortion of the content or doubts about the authenticity of the content by the audience. So, how to balance the cutting-edge of technological innovation with the genuine expression of creation, and how to control the use of these high-tech tools without deviating from the unique mission of documentary, has become a solemn topic in the industry. Moreover, the application of AI technology in data processing and analysis may also bring privacy and copyright problems, which puts forward higher ethical requirements for producers and developers of related technologies.

2. Related Work

Li et al. [5] consider how high-speed data transfer and cloud storage allow globally dispersed teams to work together simultaneously, rapidly share hundreds of gigabytes of media files, and speed up the movie-making process. For instance, technologies fail to record face recognition as well as speech, which records the voice and elements to bring actors to real-time animation. Usibjonovich et al. [6] discuss the positive impact of AI in terms of enhancing the audience experience through personalized recommendations and increasing the efficiency of the content production process with its intelligent optimization. According to the study, the application of AI technology to data analysis and content generation will help the TV industry develop in the direction of intelligence and personalization.

Ramagundam et al. [7] examine how are transforming content creation and audience engagement for linear television. The proposed system uses machine learning based forecasting to be able to beataudience number prediction simply by analyzing viewers as well as recommendations/ showing prediction. Lees et al. [8] examine the impact of deepfakes on documentary filmmaking, discussing its implications on authenticity and viewer perception. “Mainstream documentary filmmakers are acutely aware of their responsibility for the use of deepfakes,” research cites them as writing, and the review of research concludes that they are not only eager to label well such use so as to keep intact viewers’ right to know and trust. Zhongwei et al. [9] report on the use of artificial intelligence — specifically ChatGPT in writing movie and television scripts. The article discusses how AI impacts

the art making process generally, while examining its strengths and weaknesses with respect to semantic generation and idea emergence.

Han et al. [10] propose a method for dynamic improvement of artificial intelligence algorithms, which is used to enhance the precision and efficiency of selecting innovative strategies for film and television works. By employing dynamic factors a primary series of innovation strategies were set, and MATLAB simulation analysis was conducted. He [11] considers about the development direction of artificial intelligence in the film and television industry and how to take advantage of the digital optimization of business and economy. According to the study, AI technology has great potential in terms of content creation, content distribution and interaction with the audience.

3. Methodologies

3.1. Bidirectional Temporal-Semantic Codec

We introduced a multi-head attention mechanism in the original multimodal alignment to deal with heterogeneity between different modalities such as text and images. The traditional attention mechanism can only compute a query-key relationship, while multi-head attention allows the model to learn the correspondence between different modalities from multiple perspectives, expressed as Equation 1:

$$\alpha_{i,k,m}^{(\mathrm{mathcalStomathcalF})} = \mathrm{softmax}\left(\frac{\left(s_i W_q^{(1)}\right)\left(s_{k,m} W_k^{(1)}\right)^{\intercal} + \left(s_i W_q^{(2)}\right)\left(v_{k,m} W_k^{(2)}\right)^{\intercal}}{\sqrt{d}}\right), \quad (1)$$

This Equation is extended by introducing two different query-key matrices, $W_q^{(1)}$ and $W_q^{(2)}$, to allow the model to account for multiple different modal alignments at the same time. In the multi-head attention mechanism, each head can learn a different set of features, which helps to capture the complex relationships between different modalities and enhance the flexibility and accuracy of the model in multimodal data alignment.

Here, we provide temporal information for each input semantic vocabulary by positional encoding $P_{time}(t_i)$. The purpose of this is to allow the model to better understand the causal relationships and temporal dependencies between events occurring at different points in time, expressed as Equation 2:

$$s_i = E_{scr}(\omega_i) + P_{scr}(i) + P_{time}(t_i), \quad (2)$$

Where each semantic unit is not only represented as an embedding vector $E_{scr}(\omega_i)$, but also enhanced by temporal position encoding $P_{time}(t_i)$ to ensure that the model can capture the contextual meaning of each word in the overall story when processing time series data. This is an effective way to deal with the problem of long-time dependencies.

3.2. Hierarchical NeRF-Diffusion

To make the generation of 3D scenes more detailed, we optimized the NeRF model with hierarchical volume rendering. By breaking up the scene into volumetric pyramids of multiple resolutions, you can reduce the amount of computation while preserving detail. Each layer of the voxel pyramid model is trained to optimize its rendering at a specific resolution, and then a certain fusion is fused to obtain the final rendering result, as shown in Equation 3:

$$I_r = \int_{t_n}^{t_f} \exp\left(-\int_{t_n}^{s} \sigma\big(r(u)\big)du\right) c(r(s))\sigma\big(r(s)\big)ds + L_{sambient}(r), \quad (3)$$

This is a volumetric rendering equation that calculates lighting and color information for each point in the scene by integral. In this formula, the first part calculates the light falloff from the start point of t_n to the end point of t_f, while the second part calculates the contribution of the environment light, $L_{sambient}(r)$ represents the influence of the environment light on the scene.

The core idea of the diffusion model is to gradually reduce noise through the process of restoring an image from a completely noisy state to a clear image, and we apply this idea to the 3D reconstruction of old film sources, using a conditional diffusion model to ensure consistency of luminosity and texture during the restoration process, as shown in Equation 4:

$$z_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(z_t - \beta_t \epsilon_\theta(z_t, t, y)\right) + \sigma_t \xi, \tag{4}$$

The Equation describes the inverse process in a diffusion model, where z_t is the current noise image and ϵ_θ is the neural network that predicts the noise. Through the back-diffusion process, the model is able to gradually restore a noise-free image, thereby improving the quality and detail of the image.

To ensure the luminosity consistency of the reconstructed image, we optimize the lighting effect of the model by minimizing the error between the rendered image and the real image.

$\mathcal{L}_{phot}$ combines the optimization of reconstructed images and lighting to ensure that the luminosity of the image not only matches the original, but also maintains a natural transition of color by contrasting the loss of texture λ_1. This optimization helps restore the realism and visual consistency of the image, as shown in Equation 5:

$$\mathcal{L}_{phot} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} \| I_r - \hat{I}_r \|_2^2 + \lambda_1 \| c_r - \hat{c}_r \|_2^2. \tag{5}$$

3.3. Cross-Attention Motion Composer

In order to avoid the vibrations of the movement and unnatural jumps, we introduced the concept of trajectory smoothing. By smoothing the velocity and acceleration of the trajectory, it is possible to eliminate the drastic changes that can occur during motion, denoted as Equation 6.

$$\varepsilon(\tau) = \int_0^T \| \tau'(t) \|^2 \, dt + \mu \int_0^T \| \tau''(t) \|^2 \, dt + \eta \sum_{t_k} \| \tau(t_k) - p_k \|^2 . \tag{6}$$

Equation 6 smooths out the velocity and acceleration of the trajectory to avoid abrupt changes in the lens during movement. The third item ensures that the trajectory passes through the visual point of interest p_k, so that the camera movement conforms to the director's intention and enhances the narrative consistency.

4. Evaluation

In this experiment, we use the BBC Planet Earth dataset to test the proposed D-DocFusion model. This dataset contains high-resolution images of various ecological settings (i.e., forests, deserts, oceans, etc.), scripts, and transcripts of the interviews, and it is used for training and evaluation. Counter to the mentioned four methods: 1) CNN-based image captioning struggles with complex scenes and long texts; 2) Transformer-based caption generation outperforms RNNs but still lacks the deep alignment between vision and language; 3) GAN-based methods improve the quality of the text but suffer from training instability; 4) Vision-Language pretraining models significantly improve multimodal understanding, yet they all require great computational resources.

From the results shown in Figure 1, we can see that the more weight matrices, the higher the BLEU scored of all the models, indicating that the bigger expressive ability of the model gets with the increase of the number of parameters. Note that D-DocFusion consistently outperforms other methods in each of the different parameter number configurations but is especially favorable at larger weight matrices, leading to a noticeable increase in BLEU scores, confirming that the model is better equipped to find the complicated correlations and long-range dependencies across multimodal data.

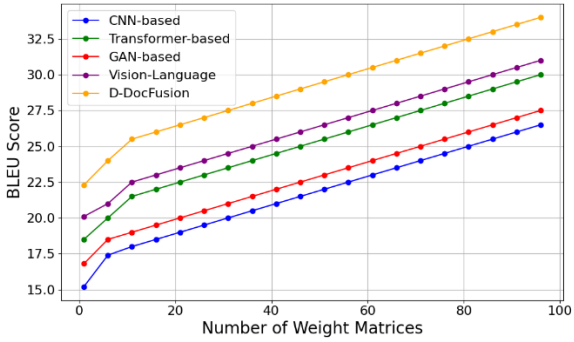

Figure 1: BLEU Score Comparison for Different Models.

5. Conclusions

In conclusion, the proposed D-DocFusion model in this paper is a novel and intelligent paradigm for TV documentary production implemented by combining multimodal learning with 3D scenes reconciliation and color styletransfer technologies. On the BBC Planet Earth dataset it achieves state-of-the-art results on the image description generation task, the best being obtained for large model parameter configurations. As such, D-DocFusion considerably improves the BLEU score with respect to the others methods. Considering the computational cost and training time of generating models in the future, how to reduce the overhead of models through technologies such as model compression and knowledge distillation is still an important research direction in the future.

References

[1] Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: a review. Artificial intelligence review, 55(1), 589-656.
[2] Li, Y. (2022). Research on the application of artificial intelligence in the film industry. In SHS Web of Conferences (Vol. 144, p. 03002). EDP Sciences.
[3] Dayo, F., Memon, A. A., & Dharejo, N. (2023). Scriptwriting in the age of AI: Revolutionizing storytelling with artificial intelligence. Journal of Media & Communication, 4(1), 24-38.
[4] Wan, Y., & Ren, M. (2021). New visual expression of anime film based on artificial intelligence and machine learning technology. Journal of Sensors, 2021(1), 9945187.
[5] Li, K. (2024). Application of Communication Technology and Neural Network Technology in Film and Television Creativity and Post-Production. International Journal of Communication Networks and Information Security, 16(1), 228-240.
[6] Usibjonovich, N. M. (2024). Analysis of the Future of Television Technology With the Help of Artificial Intelligence. Miasto Przyszłości, 54, 272-277.
[7] Ramagundam, S. (2021). Next Gen Linear Tv: Content Generation And Enhancement With Artificial Intelligence. International Neurourology Journal, 25(4), 22-28.
[8] Lees, D. (2024). Deepfakes in documentary film production: images of deception in the representation of the real. Studies in Documentary Film, 18(2), 108-129.
[9] Luchen, F., & Zhongwei, L. (2023). ChatGPT begins: A reflection on the involvement of AI in the creation of film and television scripts. Frontiers in Art Research, 5(17), 1-6.
[10] Han, J., & Shao, L. (2022). Study film and television postproduction and innovation strategy based on an artificial intelligence algorithm. Mobile Information Systems, 2022(1), 3084493.
[11] Sun, P. (2024, July). Digital Optimization of Film and Television in the Era of Artificial Intelligence. In 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC) (pp. 558-562). IEEE.