Better understanding movies: towards high-level features for content and structure analysis and synthesis

Marc Christie*, Vicky Kalogeiton†

* Univ Rennes, Inria, CNRS, IRISA,
† LIX, CNRS, Ecole Polytechnique, IP Paris

1 Summary

This thesis proposal lies at the intersection of the computer vision field and the computer graphics field and relies on deep learning technologies. The project aims at developing new methods for understanding the contents and high-level filmic structures behind video scenes. It will consist in identifying the filmmaker's decisions (what to portray, and how to portray it), building a *director signature* that would include aspects of film genre (western, action, romance,...), film style (framing/editing/image depth/camera static/dynamic behavior) and narrative construction (temporal arrangement of shots and scenes). It will also consist in exploiting the filmmaker's decisions (or director signature) to compose and create novel sequences (typically from 3D contents).

In essence, this thesis proposes to extract and formalize key cinematographic features at a degree that would enable understanding the content, style and structure of a movie and to exploit it in interactive 3D applications.

2 Introduction

The world is in constant motion. Drones fly over vast areas, but they cannot understand if somebody is swimming or drowning. We have built self-driving cars, but without vision, they cannot tell the difference between a truck and a racing ambulance that needs to overtake all vehicles. The common need among these limitations is video understanding. Interestingly, humans can naturally understand the correlation among various video events, their triggers and motivations. For instance, by analysing the scene before the ambulance's departure, we can understand that it carries someone injured. This ability is particularly prominent in films and TV shows due to their structured nature. Specifically, cameras in films are deliberately placed to ensure spatial and temporal continuity and guide the audience's interpretation of events. Moreover, best practices for scene configuration or camera placement are repeated across movies of various genres, as they are effective in conveying actions or emotions. Therefore, for true video understanding, we need to go beyond the factual events and understand the reason they take place: **the decisions behind the scenes from a filmmaker's perspective**. In real applications, this can help convert movies to high-level structures (and even to audio or textbooks), by including descriptions, motivations and some emotional dimensions. These structures enable sequence and structure comparisons across movies, with applications to film/video recommendation, but also tracing the evolution of film techniques, a key challenge in many streaming platforms. Analysing this can also benefit storytelling in other domains (such as virtual reality or 3D animations) or can even help bridge the gap between amateur directors and big studios that have all equipment, personnel, and editing pipeline to edit a scene.

3 Objective

The goal of this thesis is to develop methods for understanding the **decisions behind the scenes from the filmmaker's perspective**. We aim at understanding and transferring the director choices for a particular scene. It will consist of two tasks.

*University Rennes, Inria, CNRS, IRISA
†LIX, CNRS, Ecole Polytechnique, IP Paris

First, we will focus on **movie-style and structure understanding**, for instance, by examining whether directors are using the same style patterns for their movies. We will consider several cinematographic features [6, 15] such as frame composition, speed of speech, dialogue duration, interchange of dialogues, colors, lighting, shot length, soundtrack analysis. Following our previous work [8], we will develop pipelines to extract these high-level features and learn how to combine them to reveal the particularities of each director, and build film signatures than can enable high-level comparisons (see our recent work on film tropes [7]).

Second, we will focus on **creating novel cinematographic sequences of specific styles**, as an attempt to capture the director's staging as a particular intention at a specific instant. Specifically, we are interested in understanding how to rely on the film signature to re-design a sequence with a different contents, different structures or different style. For this, we will use automatic editing techniques [15] and cast the problem as a next-shot prediction task (from current scene analysis, current shot, and high-level characteristics, we will determine the optimal next shot). This would be useful for interactive videos (the audience could control the camera on the fly) and could be a step towards better content-based recommendation systems.

For these, we will exploit recent advances in multimodal transformers [11] by leveraging cinematographic conventions. Indeed transformers have shown their capacity to generalize structures on from multimodal contents (see recent contributions in classification [26], summarizing [23] and generative tasks [10]). In parallel, we have also demonstrated how high-level features (cinematographic and narrative) are related to elements of structure, content and style [7, 6, 8, 13, 24].

4 State of the art

Video understanding is a long-standing problem, and despite incredible advances, obtaining the best video representation is still an active research area. Videos require employing effective spatio-temporal processing of RGB and time streams to capture long-range interactions [10, 20], while focusing on important video parts [22] with minimum computational resources [26].

Story Understanding targets automatic understanding of human-centred storylines in videos. It has been formulated in several ways, e.g., by learning character interactions [20, 21] or relationships [18], creating movie graphs [28]; text-to-video retrieval [3]. Many works [3, 18, 28] highlight the importance of knowing the characters present in a scene [5, 17] and their intention [10] for understanding the story.

Visual Transformers. Recent advances in image classification show improved accuracy on several vision tasks. For instance, starting from the Transformer architecture [27] for text understanding, the Vision Transformer (ViT) [9] treats the image as a sequence of patches. Standard video approaches treat videos as a sequence of images and extend 2D architectures to spatio-temporal volumes [16].

Video Transformers for single or multiple modalities. By combining spatio-temporal volumes [16] with the ViT for images [9], some approaches create 3D Video Transformer models [2, 4]. Video Transformers are used in various video tasks, e.g., classification [2, 4, 14, 26], summarization [23], retrieval [11], audiovisual classification [22], predicting goals [10]. Typically, most works propose new attention blocks either to address space-time continuity [2, 4, 26] or to learn which parts of the video are important [26]. Moreover, the self-attention operation of transformers provides a natural mechanism to connect multimodal signals. Therefore, some works exploit video transformers to account for multiple modalities, such as iterative cross-attention in Perceiver [14] or bottleneck attention in [22].

Cinematographic understanding is a well-studied field, both for analysis and synthesis [1, 25]. Several works address virtual cinematography, for instance by camera motion prediction [12, 15], by automatically learning shot boundaries [24], by analyzing visual attention [6], by automatically editing dialogue-driven scenes [19], etc. However, formalizing filming techniques into computational models remains a challenge [15, 25]. Our goal is to create new techniques that combine video understanding with cinematography for understanding the Why behind the scenes from the filmmaker's perspective.

5 Practical information

Supervision. The thesis will be co-supervised by:

- **Marc Christie**, Assistant Professor at the Mimetic team at University of Rennes 1. (my HDR application is under study by the doctoral school, with a defense scheduled in March/April 2022)
- **Vicky Kalogeiton**, Assistant Professor at the GeoViC team at the Computer Science Department (LIX) of École Polytechnique.

Starting date: September-October 2022

Duration: three years

Required skills:

- In pursuit of or having completed Master's degree in a relevant field (CS, Informatics, ...)
- Proven experience in Python
- Hands-on experience with deep learning frameworks (PyTorch)
- Experience in computer vision and computer graphics
- High level of innovation and motivation
- Excellent written and communication skills in English

References

[1] D Arijon. *Grammar of the film language*. Focal press London, 1976.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

[3] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proc. ACCV*, 2020.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

[5] Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. Face, body, voice: Video person-clustering with multiple modalities. In *ICCV-W*, 2021.

[6] Alexandre Bruckert, Marc Christie, and Olivier Le Meur. Where to look at the movies: Analyzing visual attention to understand movie editing. *arXiv preprint arXiv:2102.13378*, 2021.

[7] Jean-Peïc Chou and Marc Christie. Structures in tropes networks: Toward a formal story grammar. In *International Conference on Computational Creativity*, 2021.

[8] Robin Courant, Christophe Lino, Marc Christie, and Vicky Kalogeiton. High-level features for movie style understanding. In *ICCV-W*, 2021.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021.

[10] Dave Epstein and Carl Vondrick. Learning goals from failure. In *Proc. CVPR*, 2021.

[11] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proc. ECCV*, 2020.

[12] Chong Huang, Chuan-En Lin, Zhenyu Yang, Yan Kong, Peng Chen, Xin Yang, and Kwang-Ting Cheng. Learning to film from professional human motion videos. In *Proc. CVPR*, 2019.

[13] Yuzhong Huang, Xue Bai, Oliver Wang, Fabian Caba, and Aseem Agarwala. Learning where to cut from edited videos. In *Proc. ICCV*, 2021.

[14] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

[15] Hongda Jiang, Bin Wang, Xi Wang, Marc Christie, and Baoquan Chen. Example-driven virtual cinematography by learning camera behaviors. *ACM Transactions on Graphics (TOG)*, 2020.

[16] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proc. ICCV*, 2017.

[17] Vicky Kalogeiton and Andrew Zisserman. Constrained video face clustering using 1nn relations. In *Proc. BMVC*, 2020.

[18] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. Learning interactions and relationships between movie characters. In *Proc. CVPR*, 2020.

[19] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. *ACM Transactions on Graphics (TOG)*, 2017.

[20] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proc. CVPR*, 2019.

[21] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net++: revisiting people looking at each other in videos. In *IEEE PAMI*, 2021.

[22] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *NeurIPS*, 2021.

[23] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *NeurIPS*, 2021.

[24] Alejandro Pardo, Fabian Caba, Juan León Alcázar, Ali K Thabet, and Bernard Ghanem. Learning to cut by watching movies. In *Proc. ICCV*, 2021.

[25] Rémi Ronfard. Film directing for computer games and animation. In *Computer Graphics Forum*. Wiley Online Library, 2021.

[26] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *arXiv preprint arXiv:2106.11297*, 2021.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[28] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proc. CVPR*, 2018.