What time is it? Temporal grounding in movies

Vicky Kalogeiton <vicky.kalogeiton@polytechnique.edu>, Ivan Laptev <ivan.laptev@gmail.com>
Location: Computer Science Laboratory, École Polytechnique & Willow Group, Department d'Informatique de l'École Normale Supérieure

Motivation

Humans can instinctively understand the timeline in a film and sort events, even without explicit mentions. They can relate visual cues to corresponding contextual information that could be multi-modal (audiovisual, dialogue), and draw on background knowledge when interpreting and grounding a stream of scenes or videos. For instance, imagine you are watching the movie Titanic: In the beginning (1996), an old lady (Rose) is talking with the crew that searches the wreck of Titanic. They recover a safe that contains a drawing of a woman wearing a necklace. Suddenly, the story jumps to 1912 where a group of people is playing poker and Jack wins a Titanic ticket (first two frames in Figure 1 (top)). Such a dramatic change of scenes plays an important role in the storytelling [Brown21]. Typically, a film is composed of a well-designed series of scenes with transitions, where the underlying storyline based on characters' interactions determines the order of the scenes [Kukleva20]. Therefore, recognizing scenes in films, understanding their content and sorting them in chronological order is essential for a wide range of storytelling applications (Figure 1). However, most computer vision approaches focus on answering specific questions [Pardo21,Tapaswi16]. Only a couple of recent works [Fu22,Yang22] target understanding time; these, however, focus on images and do not account for the relative order of sequences, especially for long-range reasoning [Wu21].

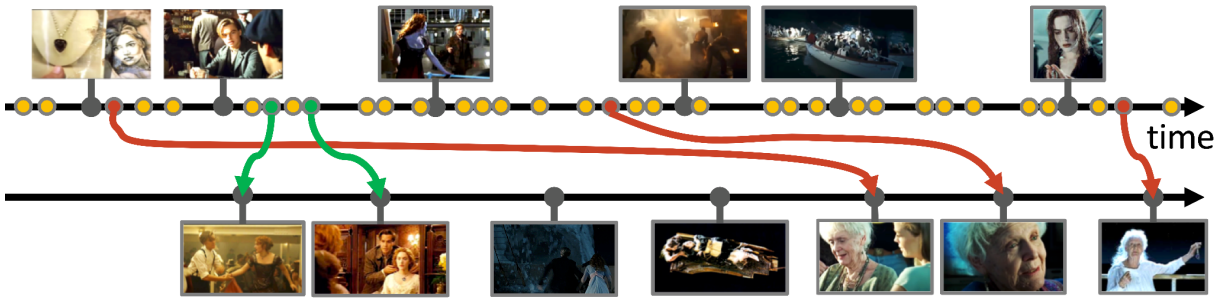

Figure 1. Storyline in edited videos. (top) Titanic scenes in filmed order; (bottom) Temporal grounding in sequences: scenes ordered chronologically, from past (green arrow) to present (red arrow)

Project description

This project falls into the cutting-edge video understanding domain and aims at understanding the storyline in movies. For this, we will formulate the problem of temporal grounding in videos. Specifically, given long-range video scenes, we will propose state-of-the-art methods to order all scenes chronologically (Figure 1 (bottom)) by exploiting multimodal cues (video, audio, text).

Requirements

We are looking for strongly motivated candidates with an interest in machine learning and computer vision. The project requires a strong background in applied mathematics and excellent programming skills (mostly in Python). If we find a mutual match, the project can lead to a joint PhD in video understanding at Ecole Polytechnique and in the Willow Group of Inria Paris.

References

[Brown21] Brown A., Kalogeiton V., Zisserman A. Face, Body, Voice: Video Person-Clustering with Multiple Modalities. In ICCV-W 2021
[Fu22] Fu X., Zhou, Chandratreya, Vondrick C., Roth D. There's a Time and Place for Reasoning Beyond the Image. In ACL 2022
[Kukleva20] Kukleva A., Tapaswi M., Laptev I. Learning interactions and relationships between movie characters. In CVPR 2020
[Pardo21] Pardo A., Caba F., Alcázar J.L., Thabet A.K., Ghanem B. Learning to Cut by Watching Movies. In ICCV 2021
[Tapaswi16] Tapaswi M., Zhu Y., Stiefelhagen R., Torralba A., Urtasun R., Fidler S. MovieQA: Understanding stories in movies through question-answering. In CVPR 2016
[Yang22] Yang C., Xie W., Zisserman A. It's About Time: Analog Clock Reading in the Wild In CVPR 2022
[Wu21] Wu C.Y., Krahenbuhl P. Towards long-form video understanding. In CVPR 2021