“Creativity Injection into Ai-powered Multimedia Storyboards”

by Bartolomeo Vacchetti

Synthesis

In recent years, deep learning has revolutionized the field of computer vision, enabling machines to interpret and understand the visual world with unprecedented accuracy. This impressive technology empowers computers to recognize objects, understand scenes, and even generate realistic images and videos. The recent boost impacted many fields, from autonomous vehicles and medical diagnostics to facial recognition and photo and video editing. One area of video editing that so far has not been investigated in depth is the operation of cutting and concatenating multiple video shots to create a scene. This PhD thesis focuses on integrating machine learning and deep learning approaches to automate the task of concatenating various shots into a meaningful scene. In other words we want to study and teach algorithms what are video editing structures and elements. By video editing structures and elements, we mean the concatenation of shots used to represent a scene or a specific moment in a movie. To complete this complex task we have focused our efforts to address the following three objectives: (i) learn to extract editing sequences from videos; (ii) analyze the correlation between the editing structures and their corresponding videos; (iii) generate new editing patterns and storyboards. A storyboard is a sequence of sketched frames used to pre-visualize a movie, essentially serving as blueprints for the final video. In other words, it is the visual representation of an editing sequence.

To address the different research objectives we have made the following contributions: (i) created a shot size dataset and trained an algorithm to perform shot size classification; (ii) developed a methodology to analyze the correlation between the editing structures and their corresponding videos; (iii) developed a methodology to generate images with the shot size constraint; (iv) developed a methodology to convert textual data into editing sequences.

Shot size classification consists of labeling images into shot sizes. It is an important step because shots are the building blocks of editing sequences. Shots can be categorized by their size based on how close the camera is to the subject. To achieve this, we have created a shot-size dataset and tested different deep-learning algorithms and machine-learning techniques to develop an effective methodology.

The results of our research efforts are a shot size dataset with 10 545 images and an algorithm that has an overall accuracy of 80%, which rises up 99% with an assumption that doesn’t distort the model performance. These results show that deep learning models are able to correctly classify images into shot classes. Since we are able to work with basic video elements we can use these simple elements to study more complex movie features and structures.

To analyze what type of correlation there is between editing sequences and videos we have developed a novel methodology to analyze short sequences of shots. Our approach was developed and tested on the Cinescale dataset, in which the shots are

characterized only by the shot size, and the Anatomy of Video Editing (AVE) dataset, in which the shots are characterized by also other features. This methodology groups similar sequences of shots based on shot size and other features, exploiting their structural similarities to understand how they reflect in the respective videos. On the Cinescale dataset we tested our approach in different scenarios and analyzed in depth a 16 classes case, a classifier achieved 96% overall accuracy in classifying the sequences labeled by our methodology. On the AVE dataset the accuracy drops to 92.8%, but the classes considered are 50. These results show not only that there is a correlation between what is shown in the video and the sequence of shots used to represent it, but also that deep learning models can correctly identify these structures.

In order to generate new editing sequences and storyboards, we have divided this complex operation into two separate tasks. The first one focuses on image generation with the additional constraint of shot size, essentially recreating single shots from prompts. We have evaluated our approach quantitatively using CLIP-T and DINO scores, and qualitatively, with a survey on 55 subjects. The second task involves creating sequences of shots that can be used to represent movie scenes. Specifically we input textual prompt and as output our methodology gives a sequence of shots that can be used to represent it. Our methodology, which was trained with sequences extracted from the Condensed Movie Dataset, achieved a 0.9280 average cosine similarity score on the training set and 0.8140 on the test set. To sum it up, the first task focuses on recreating individual shots that compose videos, while the second aims to recreate the editing structures that form the overall video or storyboard.