Camera Work Estimation from a Monocular Video Based on Optical Flow

Jinta Sakai , Chun Xie, Hidehiko Shishido and Itaru Kitahara

***Abstract*— Since camera work is considered crucial in conveying the atmosphere and impressions of movies, it is considered as a vital role in video analysis. We propose an approach to estimate camera work from monocular videos through the analysis of optical flow in a video. Our method enables the estimation of camera work from videos featuring dynamic subjects by incorporating semantic segmentation. Moreover, it is also capable of distinguishing zoom and dolly, which has not been realized by ordinal works. The method uses the relationship between image depth, optical flow, and image coordinates to perform such classification.**

I. Introduction

As each camera work significantly influences the overall impression of a movie, utilizing effective scene expression derived from camera work becomes essential. Consequently, estimating camera work from videos holds great importance in comprehending the video's structure, and the demand for such techniques is on the rise. Among the conventional methods for estimating camera work, Rao [1] introduced the Subject Guidance Network (SGNet), a learning framework for recognizing shot types. However, this approach only categorizes four camera work types: static, motion, push, and pull. Chen [2] proposed a novel deep learning-based camera motion classification framework called MUL-MOVE-Net. This method can classify nine camera work types: Static, Up, Down, Left, Right, Zoom-In, Zoom-Out, Right Rotation, and Left Rotation. However, it does not classify "Dolly," which is a fundamental and indispensable technique for camera work. Therefore, we propose a new approach to estimate camera work from monocular videos containing dynamic subjects. This approach involves segmenting dynamic regions such as humans and extracting camera work features using image depth and optical flow. The categorized camera work mainly comprises seven types: Fix, Pan, Tilt, Zoom-In, Zoom-Out, Dolly-In, and Dolly-Out. This paper introduces a method for estimating these seven basic camera work types.

II. Proposed Method

As shown in Fig. 1, while certain types of pattern are observed in the optical flow of each video with camera work, the patterns among zoom-in/dolly-in and zoom-out/dolly-out are similar. By considering the two optical flows, we found out the following differences. Since zoom enlarges/reduces the appearance of captured subjects by changing the focal length of the camera-lens, as the result, the optical flow length is proportional to the distance from the center of the zoom. On the other hand, since dolly enlarges/reduces the appearance by moving the camera position back and forward, the optical flow length varies depending on the distance between the camera and the subjects. The closer the subject is to the camera, the greater appearance change occurs with the camera movement. By using the differences, zoom and dolly can be discriminated. An overview of the proposed method is shown in Fig. 2. First, the image region of the main dynamic subject in the video, the human, is detected, and the static background region is obtained to remove the optical flow in the dynamic subject region. Then, optical flow detection and monocular depth estimation are performed on the video to obtain optical flow and depth maps. Features are extracted from these two pieces of information and learned in a random forest to estimate the camera work.

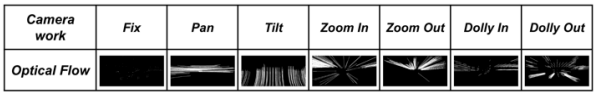

Figure 1. Optical flow features for each camera work

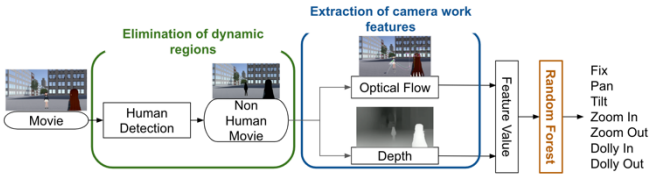

Figure 2. Overview of camera work estimation

To validate the efficacy of our approach, we conducted evaluation experiments using both our proprietary dataset and datasets sourced from prior research studies. The experiments were conducted in three different types of scenes: static scenes (3DCG, without moving subject), dynamic scenes (3DCG, with moving subject), live-action movies (footage from a real-world movie). The results of the experiments shows that our method achieves 100% accuracy in static scenes, 95% accuracy in dynamic scenes, and 67% accuracy in live-action movies. The experiment results also show that the incorporation of depth data enables the accurate discrimination between zoom and dolly movements, which has been a challenging distinction in the past.

References

[1] Rao, A., et al., “A Unified Framework For Shot Type Classification Based On Subject Centric Lens”. European Conference on Computer Vision (ECCV), 2020, pp. 17-34

[2] Z. Chen, et al., “RO-TextCNN Based MUL-MOVE-Net for Camera Motion Classification”, 2021 IEEE/ACIS ICIS Fall, 2021, pp. 182-186

J. Sakai , C. Xie , H. Shishido and I. Kitahara are with University of Tsukuba (email: kitahara[at]ccs.tsukuba.ac.jp).