3D Pose Motion Representation for Action Recognition

3D Vision Project Proposal
Supervised by: Bugra Tekin, Federica Bogo, Taein Kwon
March 9, 2018

GROUP MEMBERS

Jingtong Li

Ye Hong

Shengyu Huang

I. DESCRIPTION OF THE PROJECT

Action recognition is one of the most fundamental problems of computer vision. Human pose features provide valuable cues for recognizing human actions. To this end, [1] recently proposed an efficient motion descriptor based on 2D pose features. However, this lacks depth information which is crucial for recognizing fine-grained actions. The objective of this project is extending aforementioned work to 3D pose feature space [2] and evaluating our method on Penn Action Dataset [3].

II. WORK PACKAGES AND TIMELINE

Our project is pretty straightforward. Firstly, we will train a model on several public pose action datasets to extract 3D heatmaps for the human joints [2] for each frame, and then finetune the model to adapt to Penn Action Dataset. Secondly, we will obtain our 3D PoTion representation by temporally aggregating the dense 3D heatmaps of the human joints. This is achieved by 'colorizing' each of them depending on the relative time of the frames in the video clip and summing them. Thirdly, the fixed-sized 3D PoTion representations for each video clip are going to be trained to recognize human actions using different neural network architectures and compared against the state-of-the-art.

The training codes from [2] are open-source and the second step is trivial. [1] achieved state-of-the-art classification results based on 2D PoTion using a shallow 2D CNN, we are also expected to achive acceptable classification results based on 3D PoTion using a shallow 3D CNN.

We plan to finish implementing and evaluating the whole pipeline by **middle April** and add some modifications to the model according to the results afterward. We three would work together at the beginning and may work separately when we are all clear about the whole task. Python will be selected as the programming language and PyTorch as the Deep Learning framework. The final model shall run only on PC/Mac/Desktop.

III. OUTCOMES AND DEMONSTRATION

The expected classification results should be at least as good as that of [1] since our pose descriptor is encoded with more information. We plan to give an offline demo of our model at the end of the semester, presenting the obtained 3D PoTion descriptor, the chosen neural network architecture as well as the classification results on several datasets compared against the state-of-the-art. Several characteristic misclassified samples shall also be presented to identify the shortcomings of our method.

REFERENCES

[1] Choutas et al. Potion: Pose motion representation for action recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Weiyu Zhang, Menglong Zhu, and Konstantinos Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *International Conference on Computer Vision(ICCV)*, 2013.