3D Pose Motion Representation for Action Recognition

Student: Shengyu Huang, Ye Hong, Jingtong Li
Supervisors: Bugra Tekin[1], Federica Bogo[1], Taein Kwon[2] [1]Microsoft Research [2]ETH Zurich
3D Vision Project, Spring 2019

1 Introduction

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 estimations. However, depth information which is crucial for recognizing fine-grained actions is lacking. The objective of this project is to extend the aforementioned work to account for depth and exploit 3D pose features for the task of action recognition.

2 Method Overview

The pipeline of this project can be roughly divided into three steps. In the first step, we estimate the 3D human pose of each frame. Here we refer to the recent work [2]. They proposed a weakly-supervised transfer learning method that uses mixed 2D and 3D labels in a unified deep neural network. Their training is end-to-end and fully exploits the correlation between the 2D pose and depth estimation sub-tasks. In doing so, the 3D pose labels in controlled lab environments are transferred to in the wild images.

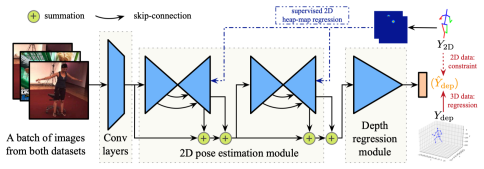

Fig. 1. Framework from [2]

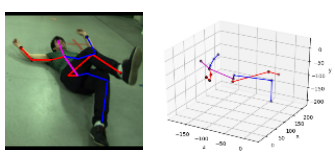

Fig. 2. 3D pose estimation

The second step is finding a motion descriptor that can well represent the 3D human pose. Analogous to [1], we temporally aggregate estimated joint probability heat-volumes and use colorization scheme to encode the relative time information. For each video, this results in a 4D descriptor, each channel represents the motion information of one joint.

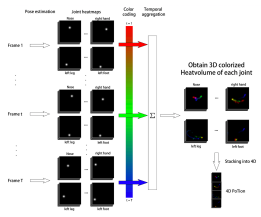

Fig. 3. 3D PoTion descriptor

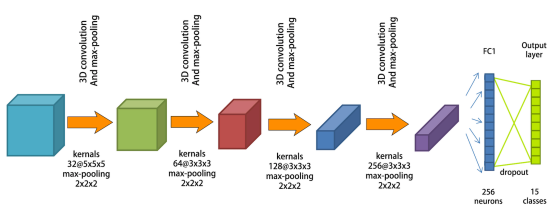

Fig. 4. Architecture of the classification network that takes as input the 3D PoTion representation of a video clip

The last step is the action recognition based on the proposed 3D PoTion descriptor. Here we exploit a shallow 3D convolutional network and show that such an efficient classifier can already accomplish descent results.

3 Data

In this project, we make use of the Penn Action dataset [3] to evaluate our proposed method. The Penn Action dataset is composed of 2326 videos in the wild with 15 different actions, among those "bowling", "gold swing", "squats", etc. The challenge on this dataset is that several body parts are missing in many actions and the image scales vary from one sample to another. The 2D joint coordinates, action labels as well as bounding boxes are provided.

Fig. 5. An example of a video sequence from the Penn Action dataset

4 Results and discussion

For pose estimation, [2] requires the human being to be centred in the input image. We firstly crop each frame based on the tight bounding boxes and then shift back the estimated coordinates so that the stacked PoTion could well represent the trajectories of each joint.

	BaseballPitch	CleanAndJerk	PullUp	StrummingGuitar	BaseballSwing	GolfSwing	PushUp	TennisForehand
Ours	15	25	42	135	15	14	73	12
	BenchPress	JumpingJacks	SitUp	TennisServer	Bowl	JumpRope	Squat	
Ours	92	15	78	13	15	37	76	

Table. 1. Results of Penn Action Dataset. The numbers are mean Euclidean distance(pixel) between the ground-truth 2D joints and the estimations of [2].

Fig. 6. Estimated 2D poses

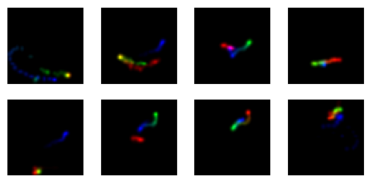

Fig. 7. 2D PoTion representation of different joints

For action recognition, we firstly experiment on 2D PoTion. We compare the results from ground truth 2D PoTion with estimated 2D PoTion and observe a 6% improvement in terms of overall accuracy. The reason is that several parts are missing in many actions in this dataset and could lead to wrong PoTion representation.

We also experiment on the different sample rates. We observe a 40% drop in terms of overall accuracy when uniformly sampling only 25 frames from each video. The reason is that the discontinuity in the trajectory could not well represent the motion information.

Finally, we experiment on our proposed 3D PoTion. We observe a 4% improvement in terms of overall accuracy compared with that from 2D PoTion.

	BaseballPitch	CleanAndJerk	PullUp	StrummingGuitar	BaseballSwing	GolfSwing	PushUp	TennisForehand
2D, gt	0.74	**1.00**	0.94	0.86	0.76	0.93	**0.99**	**0.98**
2D, est	**0.92**	**1.00**	0.97	0.73	0.35	0.94	**0.99**	0.61
3D, est	0.83	0.87	**0.98**	**0.95**	**0.98**	**0.96**	0.97	0.90
	BenchPress	JumpingJacks	SitUp	TennisServer	Bowl	JumpRope	Squat	Overall accuracy
2D, gt	0.97	**1.00**	**0.96**	**0.89**	**0.97**	**0.98**	**0.97**	**0.93**
2D, est	**0.98**	**1.00**	**0.96**	**0.89**	0.75	0.97	**0.97**	0.87
3D, est	0.82	0.91	0.79	0.85	0.96	0.97	0.96	0.91

Table. 2. Results of Penn Action Dataset. The numbers are action recognition precision of each class

5 Conclusion

This project extends the 2D PoTion[1] to 3D PoTion to also encode the depth information. We show that the depth information could aid in recognizing fine-grained actions.

Future work includes exploring the proper size of the 3D volume to fit the PoTion representation, this will determine whether coarse or fine the PoTion is, and we have already observed a drop when increasing the size from 64x64 to 96x96 in 2D cases. We only use shallow 3D CNN in the final step and the other pre-trained networks are worthy of trying.

6 References

[1] "PoTion: Pose Motion Representation for Action Recognition", Choutas et al. CVPR 2018
[2] "Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach", Zhou et al. ICCV 2017
[3] "From Actemes to Action: A Strongly-supervised Representation for Detailed Action Understanding", Zhang et al. ICCV 2013