Leveraging over depth in egocentric activity recognition

I-RIM Conference 2019
October 18 - 20, Rome, Italy
ISBN: 9788894580501
DOI: 10.5281/zenodo.4782200

Leveraging over depth in egocentric activity recognition

Mirco Planamente
Italian Institute of Technology & Politecnico di Torino
Turin, Italy
mirco.pl.93@gmail.com

Paolo Russo
Dept Comput and Control Manag Engineering Sapienza Rome University
Rome, Italy
paolo.russo@diag.uniroma1.it

Barbara Caputo
Italian Institute of Technology & Politecnico di Torino
City, Country
barbara.caputo@polito.it

***Abstract*—Activity recognition from first person videos is a growing research area. The increasing diffusion of egocentric sensors in various devices makes it timely to develop approaches able to recognize fine grained first person actions like picking up, putting down, pouring and so forth. While most of previous work focused on RGB data, some authors pointed out the importance of leveraging over depth information in this domain. In this paper we follow this trend and we propose the first deep architecture that uses depth maps as an attention mechanism for first person activity recognition. Specifically, we blend together the RGB and depth data, so to obtain an enriched input for the network. This blending puts more or less emphasis on different parts of the image based on their distance from the observer, hence acting as an attention mechanism. To further strengthen the proposed activity recognition protocol, we opt for a self labeling approach. This, combined with a Conv-LSTM block for extracting temporal information from the various frames, leads to the new state of the art on two publicly available benchmark databases. An ablation study completes our experimental findings, confirming the effectiveness of our approach.**

***Index Terms*—computer vision, egocentric vision, rgb-d fusion, activity recognition.**

I. Introduction

Within the current fast-growing range of successful applications of computer vision, those related to action and activity recognition from videos are among the most popular. This in turn drives research towards algorithms able to perform these tasks in various scenarios, from video summarization to action understanding, to automatic video indexing and retrieval, human robot interaction and many others. Although most of the work done so far has focused on third-person action and activity recognition, the recent diffusion of wearable devices able to acquire seamlessly videos has elicited attention towards first-person videos. First person activity recognition, i.e. the recognition of fine-grained actions performed by the camera's wearer such as pouring, opening, lifting etc, is more challenging than third person action recognition due to the lack of information about the actor's pose. Moreover, sharp movements by the wearer make it unfeasible the use of tracking algorithms, while at the same time it causes big shakes in videos.

Within this context, several authors pointed out the importance of using 3D information, going beyond the use of RGB data only. This has been done using 3D point clouds [1] or depth images [2], [6] in various forms, combined with standard images. This paper follows this trend, and presents a deep architecture that uses depth information to prime RGB-based first-person activity recognition. Specifically, we propose to use depth information as an attention mechanism to guide the network towards learning selective features, able to recognize robustly the hand motion patterns as well as the object being manipulated. We do so by blending depth and RGB images together so to weight the RGB channels according to the depth information, and thus guide the deep learning of activity recognition according to this knowledge. After this first step, we implement a CNN-RNN network trained in a weakly supervised fashion, as proposed in [3], but without the optical flow branch, as the initial depth blending already enables a Conv-LSTM block (as proposed in [3]) to capture effectively important knowledge about the hand-object interactions. We instead further embrace this weakly supervised scenario, and we introduce a Mean Teacher model [4] that allows to leverage over multiple iterations and makes ultimately the overall architecture more robust. Figure 1 illustrates the overall architecture.

Identify applicable funding agency here. If none, delete this.

II. Experiments

A. Experimental Setup

All the experiments reported in this section are performed on two RGB-D first person datasets: the Grasp Understanding dataset (GUN-71, [5]) and the Wearable Computer Vision Systems dataset (WCVS, [6]). The two datasets are very different, presenting different challenges, hence they allow us to show the robustness of our method. Indeed, the first dataset (GUN-71) has many classes and the actions are labeled with respect to the grasp made by each subject, independently of the object used. As opposed to this, the structure of the second dataset (WVCS) represents a great challenge due to the large intra-class variations caused by multiple users and scenarios, although in the setting we use it consists of only 4 classes. We compare our method with other existing approaches for first person action recognition and more generic algorithms.

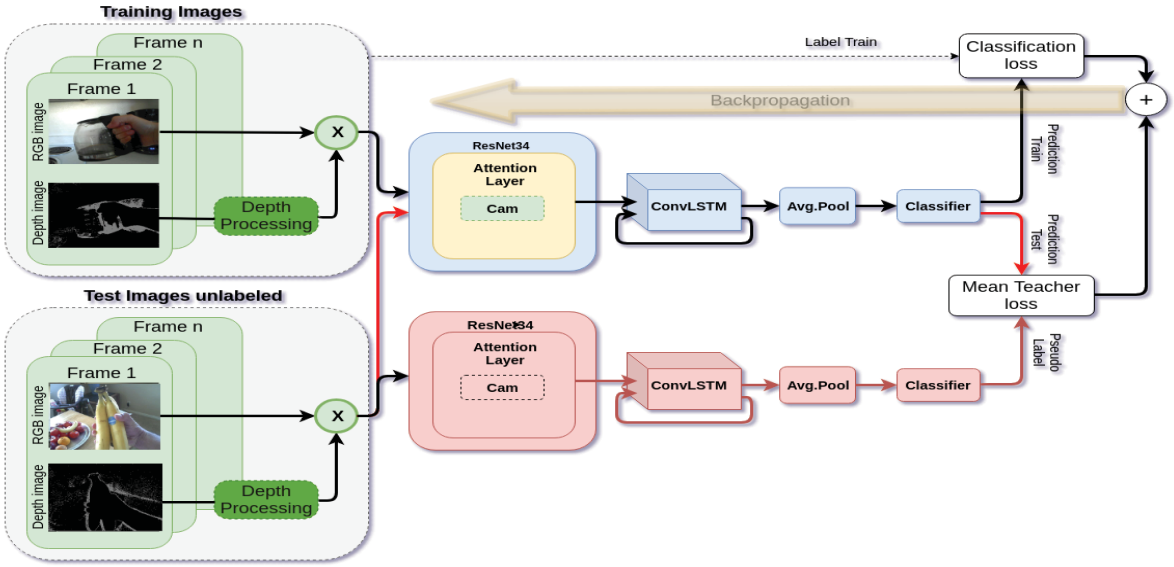

Fig. 1. Overview of the proposed architecture. The blue blocks represent the full deep model, made by backbone, Conv-LSTM layer and classifier. The red blocks show the Mean Teacher model, built from the main model through temporal mean average, which provides pseduo-labels on the test set. Green blocks show depth-RGB integration for each sequence frame (light green). Finally, the white blocks represent the two losses used during training.

B. Results

Looking at the results in Table I, we see how the accuracy of the Depth-only methods is significantly lower then the RGB-only ones, where the differences between the two modalities can reach up to 10%. Moreover, RGB + Depth integration must be carefully planned, as it can produce worse performance w.r.t RGB only. This is for instance the case for Appearance Stream, that in the RGB-only case achieves a better performance than when applied on RGB+Depth (Feature Fusion). Nevertheless, in other cases it can produce consistent improvements.

Our method provides the new state of the art on both datasets, ranging from +3% on WCVS to a +5% on GUN-71 w.r.t the current state of the art algorithms. The Mean Teacher loss is able to produce a boost of 2.5 ˜3.5% of accuracy depending on the dataset. A detailed analysis of the Depth integration is presented in the following paragraph.

We would like to finally note that in the MDNN+TSN case [7], the authors combined two different methods, obtaining a result equal to 71.83%. It is worth stressing that the use of an ensemble method reliably increases the performance compared to using a single network. In spite of this, the result achieved by MDNN+TSN is still lower than the accuracy of our method, by about a 2%.

Method	Accuracy GUN-71	Accuracy WCVS
Deep-RGB [5]	11.31	-
Best From [5]	17.97	-
CNN-RGB [6]	-	52.0
CNN(MultiWindow)-RGB [6]	-	57.00
Appearance Stream [8], [9]	26.00	60.36
Depth Stream [8], [9]	15.60	58.47
Motion Stream [9]	-	37.34
Score Fusion [2]	26.24	59.32
Feature Fusion	29.36	62.16
DCCA [10]	32.56	60.45
TSN-RGB [11]	-	66.02
TSN-Flow [11]	-	59.48
TSN-Depth [11]	-	59.32
TSN-FLow+RGB(Score Fusion) [11]	-	67.05
TSN-FLow+RGB+Depth(Score Fusion) [11]	-	70.09
MDNN [7]	33.89	65.67
MDNN + hand [7]	34.04	67.04
Proposed architecture w/o Mean Teacher	**36.46**	**70.62**
Proposed architecture with Mean Teacher	**39.07**	**73.22**

TABLE I
RESULTS ON GUN-71 AND WCVS DATASETS

REFERENCES

[1] Garcia-Hernando et al. First-person hand action benchmark with RGB-D videos and 3d hand pose annotations, CVPR, 2018
[2] Y. Tang et al, Action recognition in rgb-d egocentric videos, ICIP, 2017.
[3] S. Sudhakaran et al, Attention is All We Need: Nailing Down Object-centric Attention for Egocentric Activity Recognition, BMVC, 2018.
[4] A. Tarvainen et al, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS, 2017.
[5] G. Rogez et al, Understanding everyday hands in action from rgb-d images. ICCV, 2015.
[6] M. Moghimi et al, Experiments on an rgb-d wearable vision system for egocentric activity recognition. CVPR W, 2014
[7] Y. Tang et al, Multi-stream Deep Neural Networks for RGB-D Egocentric Action Recognition. IEEE T. on Circuits and Systems for Video Technology, 2018.
[8] K. He et al, Deep residual learning for image recognition. CVPR, 2016.
[9] K. Simonyan et al, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.
[10] G. Andrew et al, Deep canonical correlation analysis. ICML, 2013.
[11] L. Wang et al, Temporal segment networks: Towards good practices for deep action recognition. ECCV, 2016.