A semantic segmentation approach to action recognition in video

Dana Axinte[1,2] and Marius Leordeanu[1,3]

[1]University Politehnica of Bucharest, Romania [2]Bitdefender, Romania

[3]Institute of Mathematics of the Romanian Academy, Romania

daxinte@bitdefender.com, marius.leordeanu@imar.ro

1. Introduction

Problem We aim to learn and recognize the actions an human takes in a realistic (amateur) video as the tasks proves its application in multiple domains as healthcare, video surveillance and entertainment.

Our approach:

- introducing an instance segmentation method of recognizing activities in video
- the masks of the objects within the image are passed though different network architectures to classify 100 types of activities

Dataset

- 100 action classes from Kinetics Human Action Video Dataset [1] with 220 videos per class: 200 train, 20 test, 10 frames from each 10s video
- challenging videos due their the quality: illumination, sharpness and the different view angles of an action
- the action classes are: Singular Person (eg. "drinking", "running"), Person-Person (eg "shaking hands"), Person-Object (eg "cutting apples")

2. Architectures

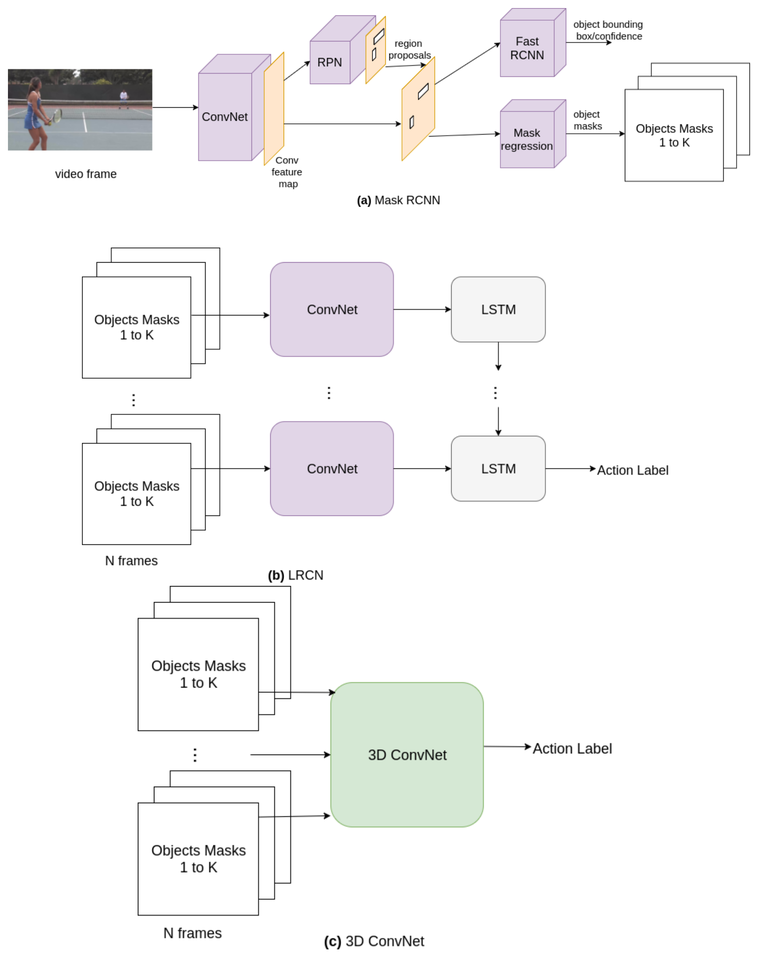

Figure 1: In (a) is presented the architecture of Mask RCNN [2] applied to a frame from a *playing tennis* action video and (b),(c) show the action recognition models used

3. Method

Instance segmentation - Mask RCNN [2]

- trained on Coco dataset [3], 81 objects classes
- output: class scores, bounding box, binary masks for each object

Objects Masks

- detect mask of the objects in each video frame
- concatenate masks of the same object classes
- vary masks sizes to reduce the number of parameters
- 10 fixed **main masks**: *person, car, bicycle, motorcycle, cat, dog, bench, surfboard, tennis racket, sports ball*

Action classification

- 3D-ConvNet [Conv 32, (3, 3, 3)]
- LRCN [Conv 32, (3, 3);LSTM 256]
- mean of masks between frames - ConvNet [Conv 32, (3, 3)]

4. Results

Frames	Masks	Size	Method	Accuracy
10	10	32x32	3D-ConvNet	11,2
10	10	128x128	3D-ConvNet	13,5
10	10	32x32	LRCN	18,6
10	10	128x128	LRCN	24,3
1 (mean) 10	32x32	CNN	11,3	
1 (mean)	10	128x128	CNN	15,0

Our results demonstrate the need for more information about the masks, since the 128X128 masks size increses the results.
CNN features passed trough RNN performed better due to the extra information and the temporal aspect the network is able to catch.
Since our representation tries to describe the activities with for only 10 objects, quick overfitting occurs as the characterization is not sufficient.

5. Qualitative Results

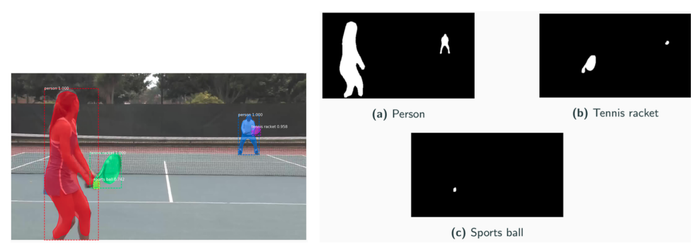

Figure 2: Mask RCNN output and the masks of the object classes of a *playing tennis* action frame

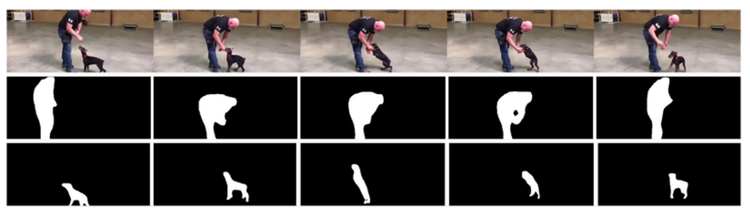

Figure 3: RGB frames and objects masks of *person* and *dog* for a *training dog* activity - 96% accuracy

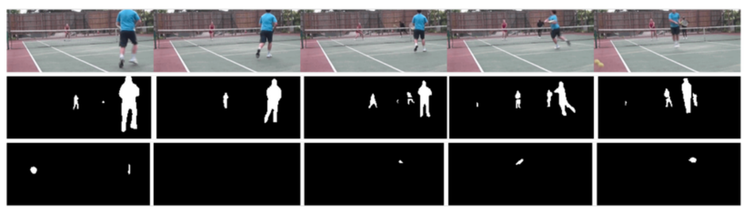

Figure 4: RGB frames and objects masks of *person* and *tennis racket* for a *playing tennis* activity - 89% accuracy

Figure 5: RGB frames and objects masks of *person* for a *jogging* activity - 2.4% accuracy

6. Conclusions and Future work

Conclusions

- Spatial information is hold by the mask representations
- The recurrent methods give the most encouraging results, but further study has to be made into the quick overfitting problem
- Difficult actions, the representation of 10 object channels is not enough to describe them

Future work

- Further study the correlation between the action classes and the fixed number of objects channels that we use to describe those actions
- Increase our top masks descriptor using more objects channels
- Improve the model by adding a CNN that would learn spatiality straight from the RGB images

7. References

[1] Kay et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[2] He et al. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV),*, 2017.

[3] Lin et al. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014.