Self-Supervised Joint Encoding of Motion and Appearance for First Person Action Recognition

Mirco Planamente - Andrea Bottino - Barbara Caputo

First Person Action Recognition

Related Works

Two Stream Approach :

- Appearance Stream (RGB)
- Motion Stream (Optical/Warp Flow)

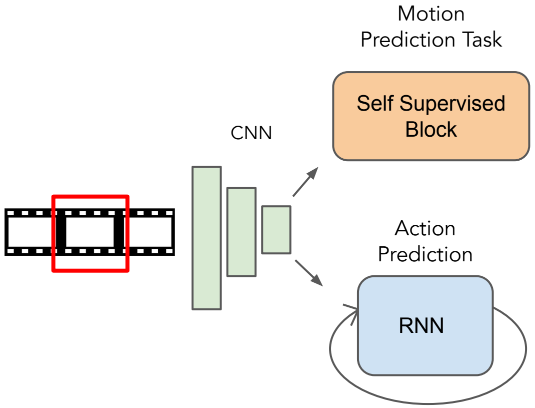
2D Backbone + Recurrent Neural Network (RNN)

3D CNNs

Our Contribution

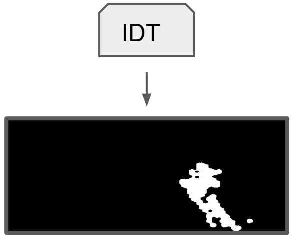
Single stream architecture called SparNet that exploits a set of motion prediction self-supervised pretext tasks in order to learn jointly Motion and Appearance information.

SparNet Overview



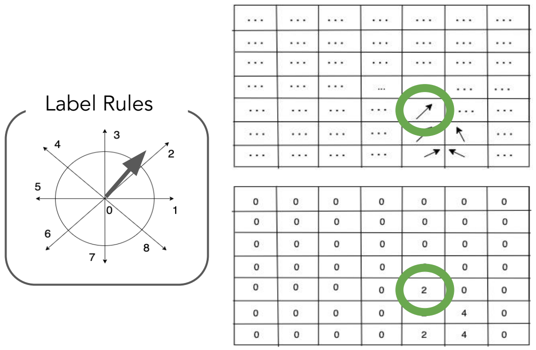
Motion-based Self-Supervised Tasks

Motion Segmentation (MS)



Motion: Yes or No?

Optical Flow Classification (OFC)



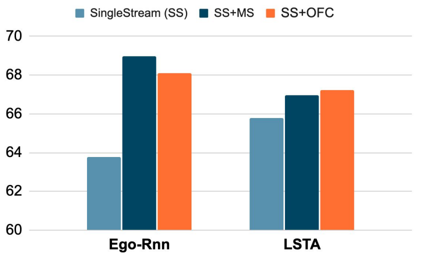
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0		0	0
0	0	0	0	2	0	0
0	0	0	0		4	0
0	0	0	0	2	4	0

Experiments

EGTEA+	
RULSTM [43]	60.20
Ego-RNN [3]	60.76
LSTA [2]	61.86
3DConv MTL [17]	65.70
Two-stream I3D + STAM [19]	65.97
Baseline	63.96
SparNet-MS	66.15
SparNet-OFR	64.22
SparNet-OFC	67.36
SparNet-OFR+OFC	**67.52**
SparNet-MS+OFC	67.44
SparNet-MS+OFC (11 frames)	**69.80**

GTEA-61	
EleAttG [12]	66.77
TSN [45]	69.93
Ma et al. [47]	73.02
Ego-RNN [3]	79.00
LSTA [2]	80.01
Baseline	80.18
SparNet-MS	80.51
SparNet-OFR	80.14
SparNet-OFC	81.17
SparNet-OFR+OFC	80.51
SparNet-MS+OFC	**81.39**

FPHA	
H+O [44]	82.43
Gram Matrix [46]	85.39
ST-TS-HGR-NET [48]	93.22
Baseline	94.32
SparNet-MS	96.41
SparNet-OFR	95.07
SparNet-OFC	96.41
SparNet-OFR+OFC	96.35
SparNet-MS+OFC	**96.70**



Qualitative Results