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

Mirco Planamente - Andrea Bottino - Barbara Caputo

POLITECNICO
DI TORINO

Task Description

- First Person Action Recognition

Task Description

- First Person Action Recognition

- More challenging with respect to Third Person:
  - Ego-motion
  - Object Occlusions

Related Works

Two Stream Approach :

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

Related Works

Two Stream Approach :

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

2D Backbone + Recurrent Neural Network (RNN)

3D CNNs

Related Works

Two Stream Approach :

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

2D Backbone + Recurrent Neural Network (RNN)

3D CNNs

Self-Supervision → Representation Learning (order prediction, odd-one-out, jigsaw3D, ..)

Our Contribution

A single stream architecture called SparNet

A set of motion prediction self-supervised pretext tasks in the specific domain of egocentric action recognition

- Motion Segmentation (MS)
- Optical Flow Classification (OFC)

Main Idea - Motion Prediction

GENERAL IDEA

APPLICATION

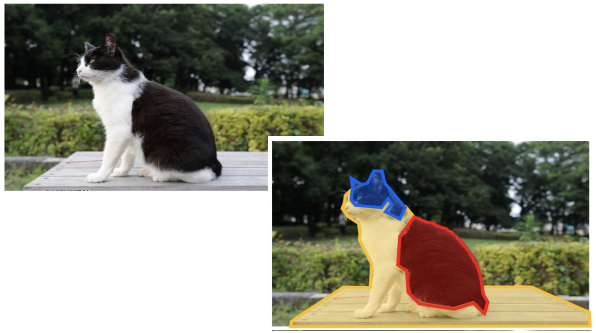
Main Idea - Motion Prediction

GENERAL IDEA

APPLICATION

Main Idea - Motion Prediction

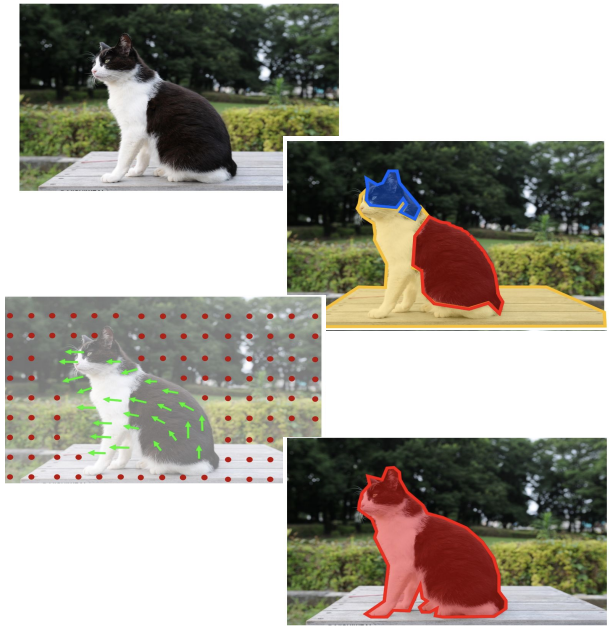
GENERAL IDEA



APPLICATION

Main Idea - Motion Prediction

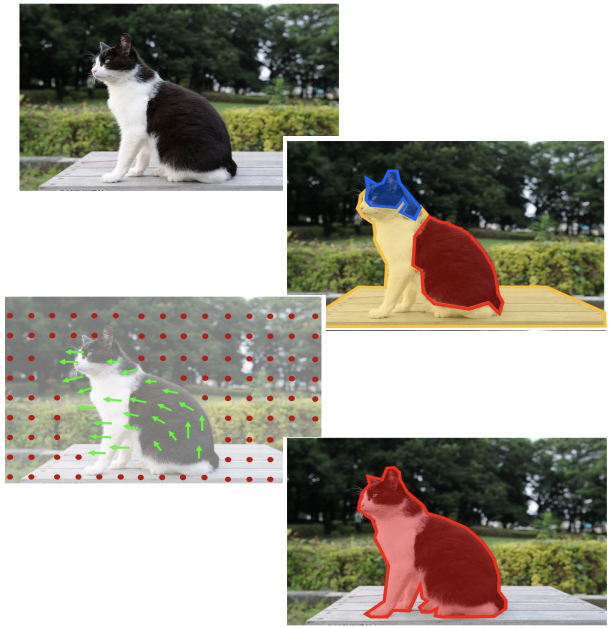
GENERAL IDEA



APPLICATION

Main Idea - Motion Prediction

GENERAL IDEA

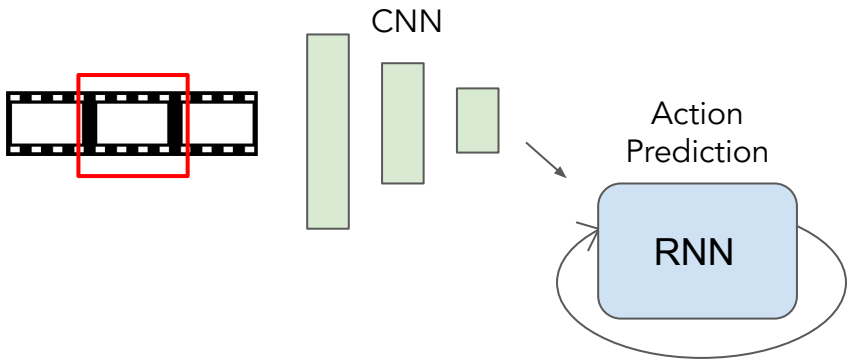


APPLICATION

Main Idea - Motion Prediction

GENERAL IDEA

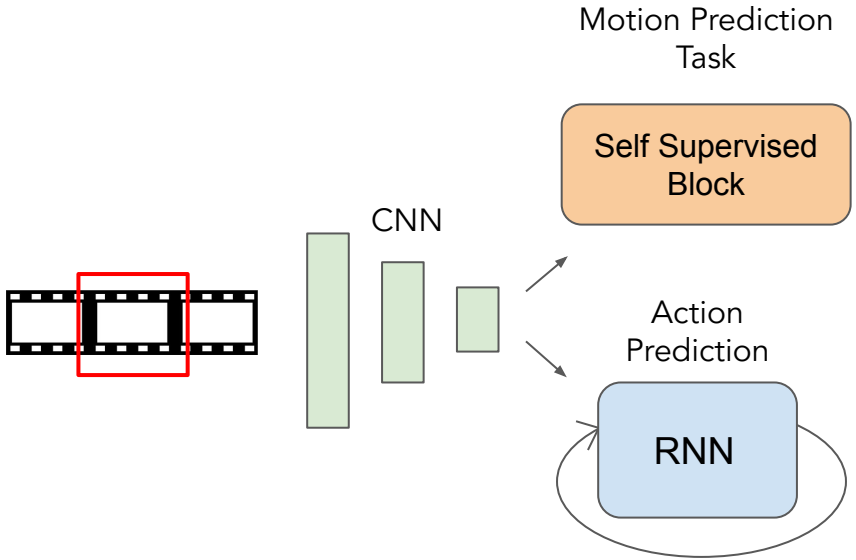
APPLICATION



Main Idea - Motion Prediction

GENERAL IDEA

APPLICATION



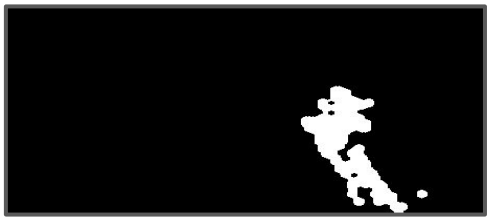
Motion-based Self-Supervised Tasks

Motion Segmentation (MS)

Optical Flow Classification (OFC)

Motion-based Self-Supervised Tasks

IDT[3]

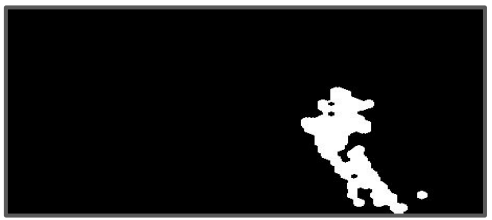


Motion: Yes or No?

Optical Flow Classification (OFC)

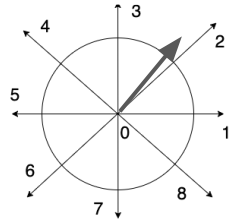
Motion-based Self-Supervised Tasks

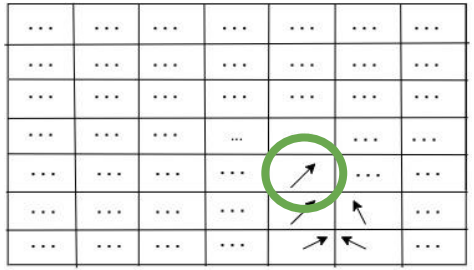
IDT[3]

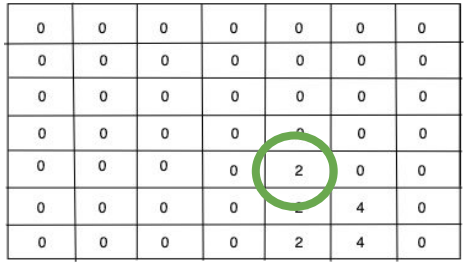


Motion: Yes or No?

Label Rules







Results - GTEA61

Method	GTEA-61
EleAttG[5]	66.77
TSN[6]	69.93
Ma et al.[7]	73.02
Ego-RNN[1]	79.00
LSTA[4]	80.01
SparNet-MS	80.51
SparNet-OFC	81.17
SparNet-MS-OFC	**81.39**

GTEA-61 (split2)	Single Stream (SS)	SS+MS	SS+OFC
Ego-RNN[1]	63.79	68.97	68.10
LSTA[4]	65.80	66.96	67.24

Results - GTEA61

Method	GTEA-61
EleAttG[5]	66.77
TSN[6]	69.93
Ma et al.[7]	73.02
Ego-RNN[1]	79.00
LSTA[4]	80.01
SparNet-MS	80.51
SparNet-OFC	81.17
SparNet-MS-OFC	**81.39**

GTEA-61 (split2)	Single Stream (SS)	SS+MS	SS+OFC
Ego-RNN[1]	63.79	68.97	68.10
LSTA[4]	65.80	66.96	67.24

Results - GTEA61

Method	GTEA-61
EleAttG[5]	66.77
TSN[6]	69.93
Ma et al.[7]	73.02
Ego-RNN[1]	79.00
LSTA[4]	80.01
SparNet-MS	80.51
SparNet-OFC	81.17
SparNet-MS-OFC	**81.39**

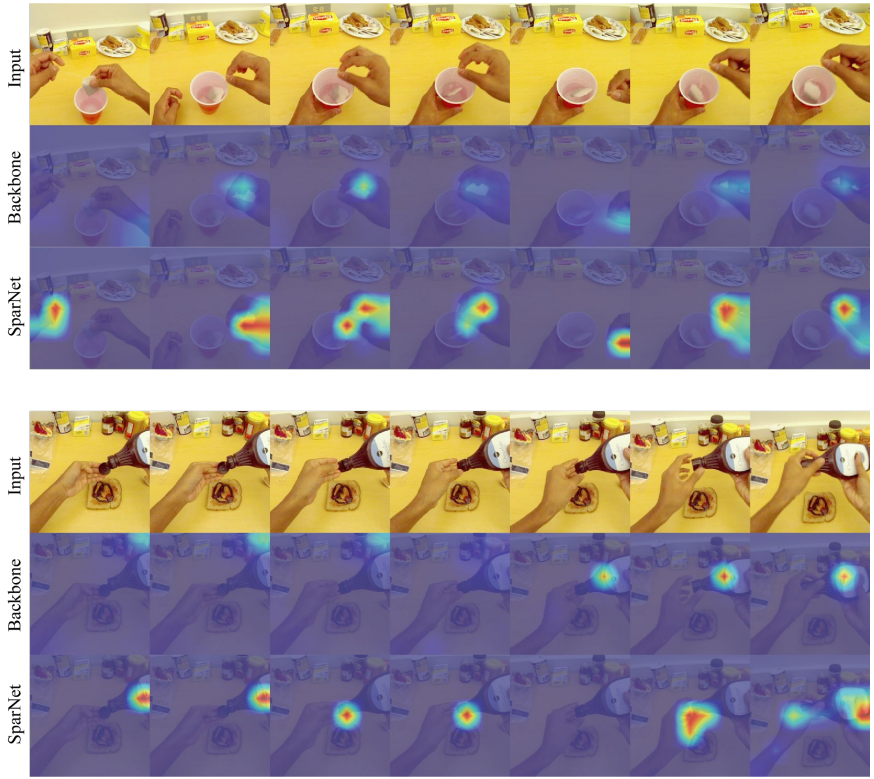
GTEA-61 (split2)	Single Stream (SS)	SS+MS	SS+OFC
Ego-RNN[1]	63.79	68.97	68.10
LSTA[4]	65.80	66.96	67.24

Results - EGTEA+ & FPHA

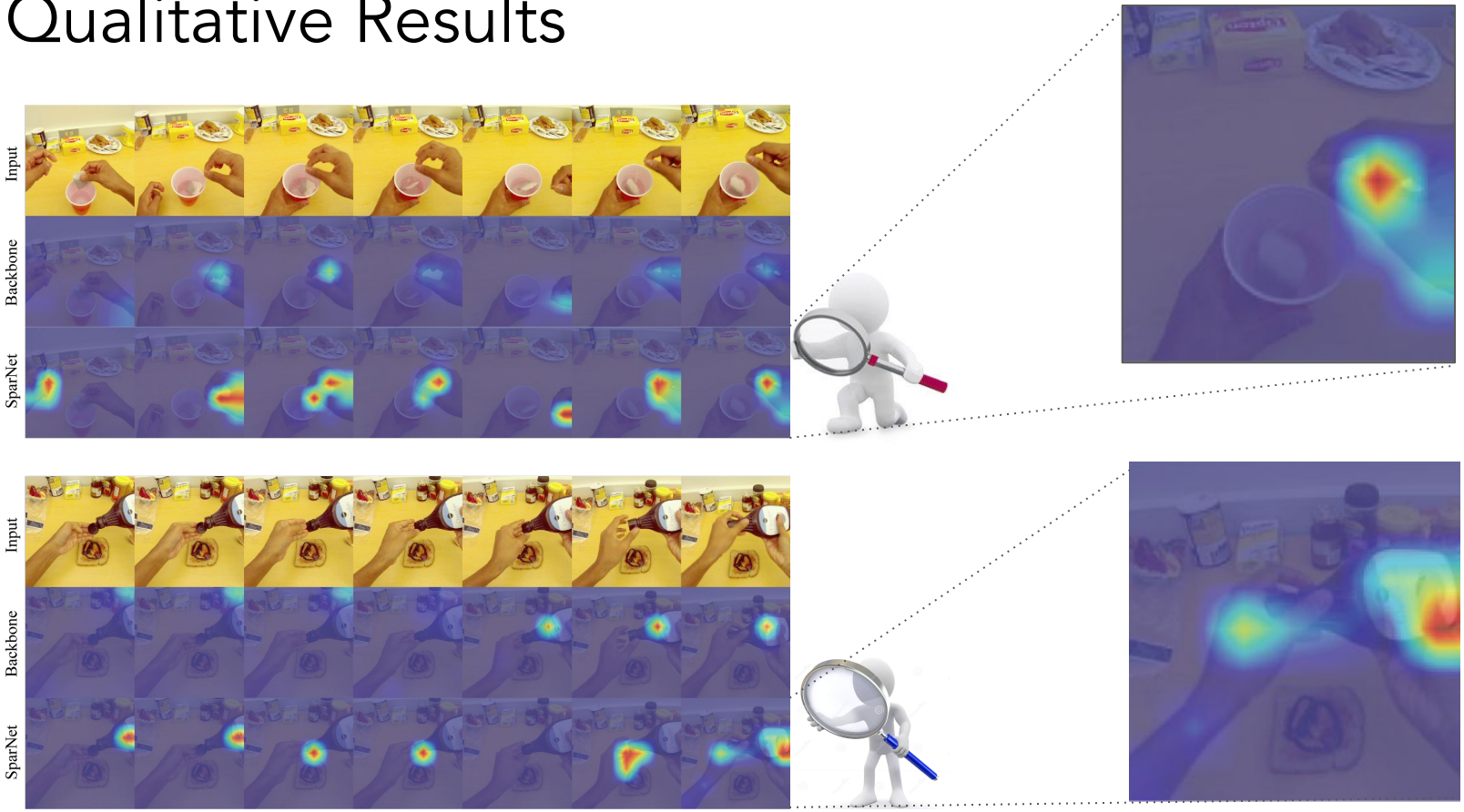
Method	EGTEA+
RULSTM[8]	60.20
Ego-RNN[1]	60.76
LSTA[4]	61.86
3DConv MTL[9]	65.70
Two-stream I3D + STAM[10]	65.97
SparNet-MS	66.15
SparNet-OFC	67.36
SparNet-MS-OFC	67.44
SparNet-MS-OFC (11Frames)	69.80

Method	FPHA
H+O[11]	82.43
Gram Matrix[12]	85.39
ST-TS-HGR-NET[13]	93.22
SparNet-MS	96.41
SparNet-OFC	96.41
SparNet-MS-OFC	96.70

Qualitative Results



Qualitative Results



Conclusion

We propose a motion prediction self-supervised pretext tasks in the context of first person action recognition.

We validate our approach on several datasets.

SparNet obtained comparable results respect to the standard two stream approach, without using the Optical Flow information at test time.

Thank you!
Any questions?

References

1. S. Sudhakaran and O. Lanz. "*Attention is all we need: Nailing down object-centric attention for egocentric activity recognition.*" BMVC18 EGO-RNN
2. Deepak Pathak, Ross B. Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. *"Learning features by watching objects move."* CVPR 2017
3. Heng Wang and Cordelia Schmid. *"Action recognition with improved trajectories."* ICCV13 IDT
4. Swathikiran Sudhakaran, Sergio Escalera, Oswald Lanz "LSTA: Long Short-Term Attention for Egocentric Action Recognition". CVPR19
5. P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng,"Adding attentiveness to the neurons in recurrent neural networks,"CoRR, vol. abs/1807.04445, 2018. EleAttG
6. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool,"Temporal segment networks: Towards good practices for deep action recognition," ECCV 2016, TSN
7. M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," inProceedings of the IEEE Conference on Computer Vision and Pattern Recognition
8. A. Furnari and G. Farinella, "Rolling-unrolling lstms for action anticipation from first-person video,"IEEE Transactions on Pattern Analysis and Machine Intelligence. RULSTM
9. G. Kapidis, R. Poppe, E. van Dam, L. Noldus, and R. Veltkamp, "Multi-task learning to improve egocentric action recognition," in Proceedings of the IEEE International Conference on Computer Vision Workshops,2019, 3DConv MTL
10. M. Lu, D. Liao, and Z.-N. Li, "Learning spatio temporal attention for egocentric action recognition," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019 Two-stream I3D + STAM
11. B. Tekin, F. Bogo, and M. Pollefeys, "H+o: Unified egocentric recognition of 3d hand-object poses and interactions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019 H+O
12. X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps, "Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, Gram Matrix
13. X. S. Nguyen, L. Brun, O. L´ezoray, and S. Bougleux, "A neural network based on spid manifold learning for skeleton-based hand gesture recognition" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. ST-TS-HGR-NET
14. J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," inProceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition (CVPR20) MM-SADA