Deep Temporal Pyramid Design for Action Recognition

Ahmed Mazari, Hichem Sahbi

LIP6 CNRS, UMR 7606, UPMC, Sorbonne Université, Paris, France

Context

Goal:

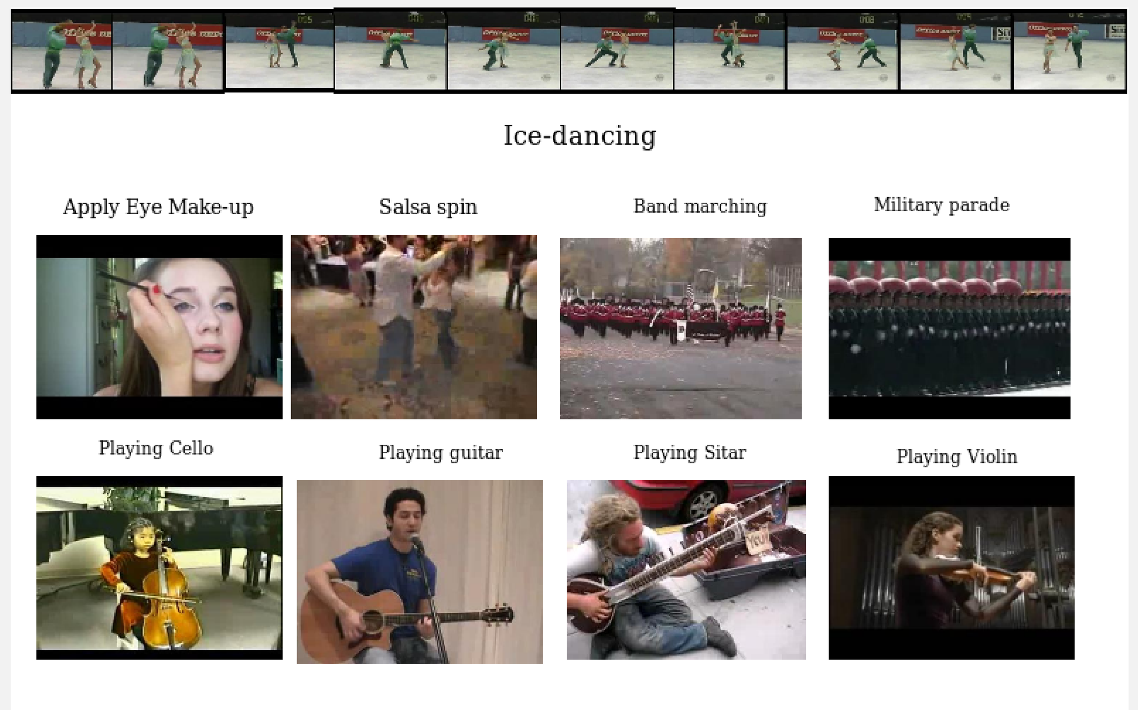
- Describe variable length videos while preserving their temporal structures
- Capture the granularity of action categories in videos

Methodology:

- Design an aggregation method at different levels of granularity
- Select representations
- Generalize multiple kernel framework on temporal pyramid

Dataset : UCF-101 (split-2)

- 9586 training and 3774 test videos
- 101 actions



UCF-101 dataset (trimmed videos)

Mathematical model

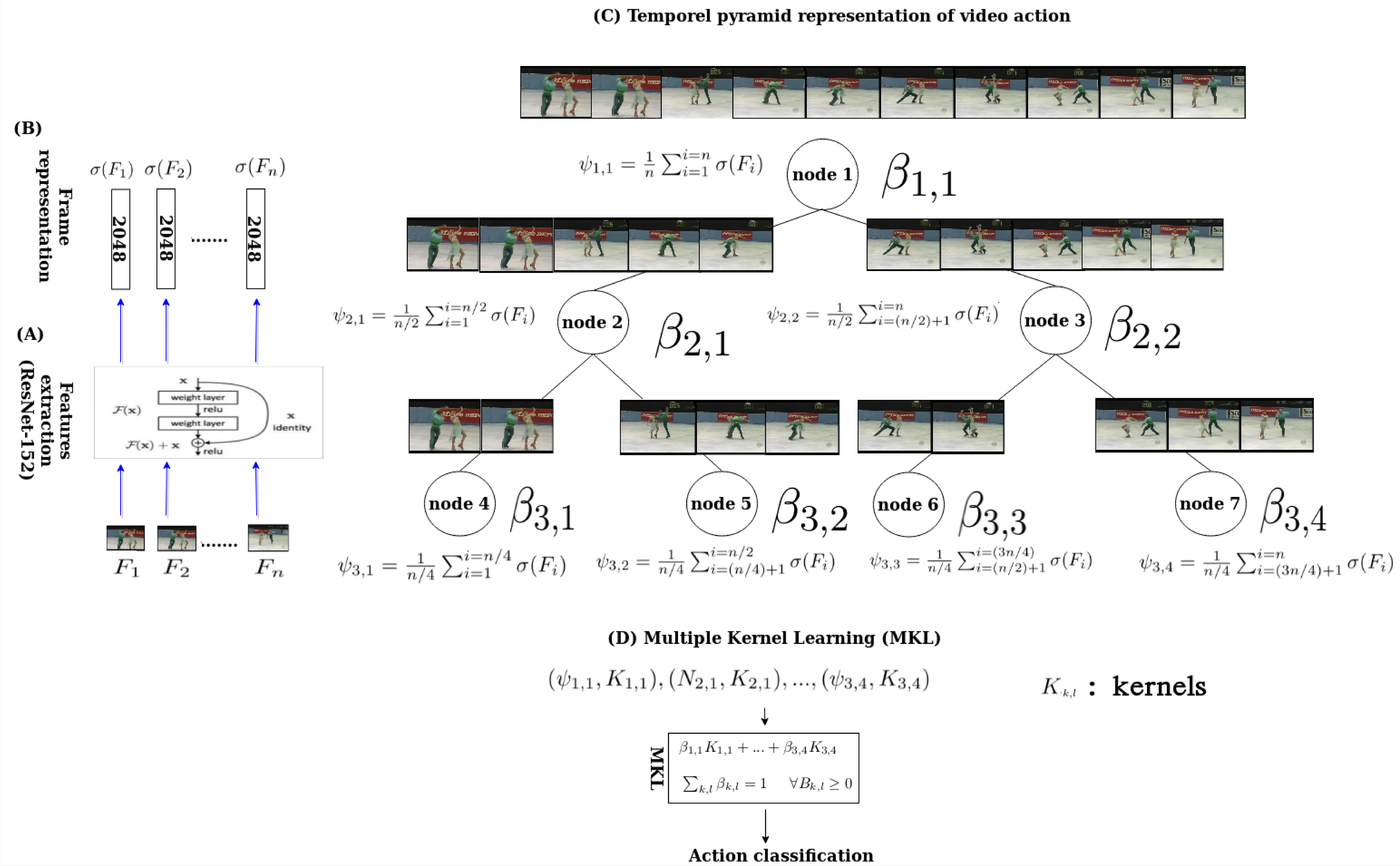
We solve the following constrained minimization problem :

$$\min_{\beta,w,b,\xi} \frac{1}{2}\sum_{k,l}\sum_{c}\beta_{k,l}\langle w_c^{k,l}, w_c^{k,l}\rangle + \sum_{j=1}^{n}\xi_j$$

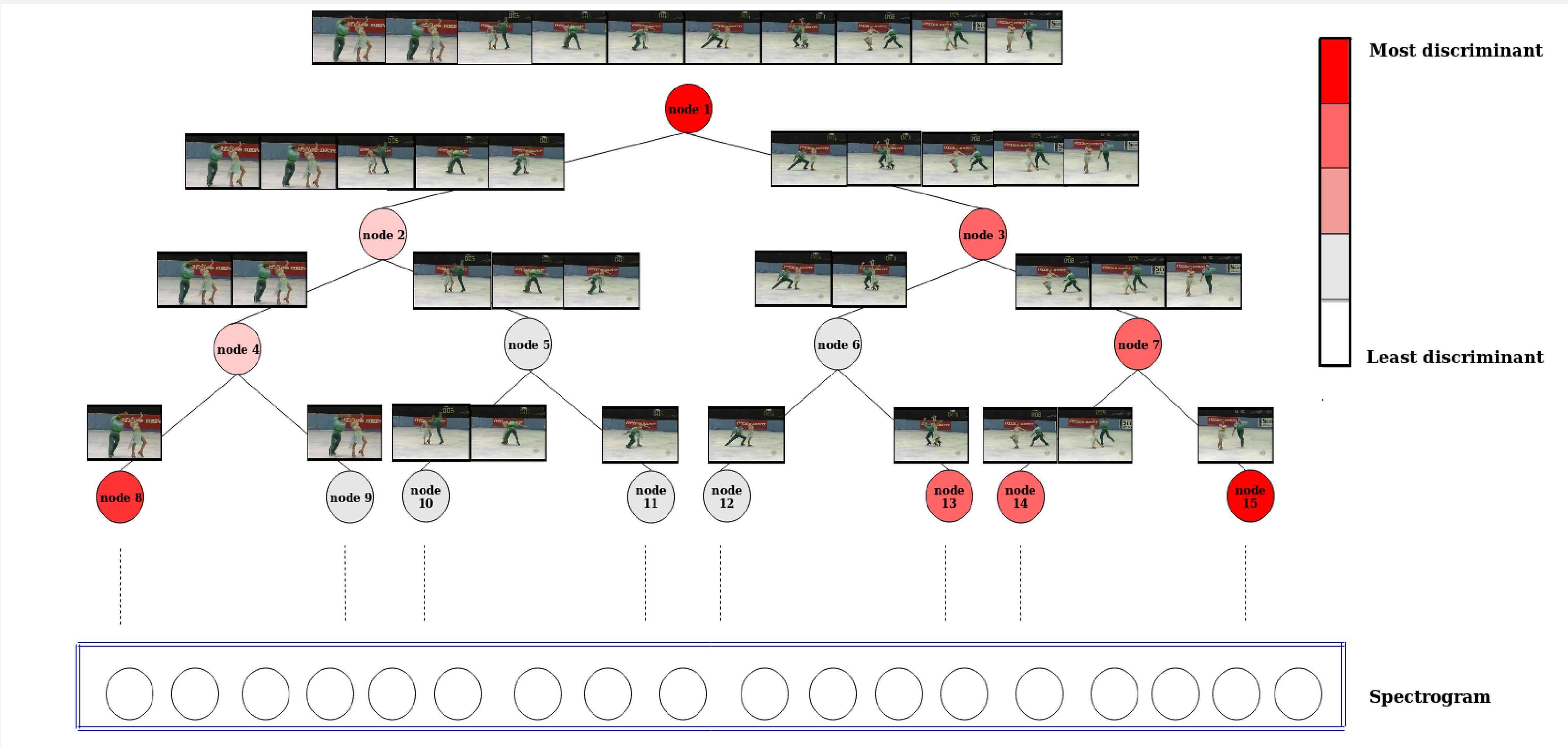
s.t $\xi_j = \max_{c' \in \mathscr{C}\backslash c} \mathscr{L}(g_c(\mathscr{V}_j) - g_{c'}(\mathscr{V}_j))$

- $\beta_{k,l}$: weights of the temporal pyramid
- $\psi_{k,l}(\mathscr{V})$: video representation associated with k-th node and l-th level
- $w_c^{k,l}$, $\mathscr{L}(.)$: SVM hyperplanes , convex loss function
- $g_c(.)$: SVM associated to action category c

Temporal pyramid aggregation scheme



Weights distribution



Results

Setting	Action recognition performance on UCF101
Global average pooling (temporal pyramid root)	66.15%
Temporal pyramid (level 2)	66.74%
Temporal pyramid (level 3)	67.14%
Temporal pyramid (level 4)	67.41%
Temporal pyramid (level 5)	67.45%
Temporal pyramid (level 6)	67.47%
Temporal pyramid + MKL	**68.58%**
Spectrogram (with resnet-18)	64.41%

Comparison with state-of-the-art

Method	Action recognition performances on UCF101
col. heatM [3]	64.38%
col. heatM [3] **+TP**	**77.34%**
Spect	64.41%
Spect **+TP**	**68.40%**
Spect + col. heatM [3]	**66.87%**
Spect + col. heatM [3] **+TP**	**74.65%**
3D 2-stream (motion) [2]	96.41%
3D 2-stream (appearance) [2]	95.60%
3D 2-stream (motion) [2] **+TP**	**97.50%**
3D 2-stream (appearance) [2] **+TP**	**95.77%**
3D 2-stream (combined) [2] **+TP**	97.94%
3D 2-stream (motion) [2] + col. heatM [3]	94.89%
3D 2-stream (appearance) [2] + col. heatM [3]	94.32%
3D 2-stream (combined) [2] + col. heatM [3]	97.02%
3D 2-stream (motion) [2] + col. heatM [3] **+TP**	**95.70%**
3D 2-stream (appearance) [2] + col. heatM [3] **+TP**	**94.60%**
3D 2-stream (combined) [2] + col. heatM [3] **+TP**	**97.56%**

Future works

- End-to-end temporal pyramid design
- Generalization of our hierarchical aggregation method to activity recognition

References

[1] C.cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning Kernels based on Centered Alignement. JMLR, 2012

[2] J. Carreira, A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR, 2017

[3] V. Choutas, P. Weinzaepfel, J. Revaud, C. Schmid. PoTion: Pose MoTion Representation for Action Recognition. CVPR, 2018

[4] K He, X Zhang, S Ren, J Sun. Deep Residual Learning for Image Recognition. CVPR, 2016