Trajectory Convolution for Action Recognition

Yue Zhao[1], Yuanjun Xiong[2], Dahua Lin[1]

[1]CUHK – SenseTime Joint Lab, The Chinese University of Hong Kong [2]Amazon Rekognition

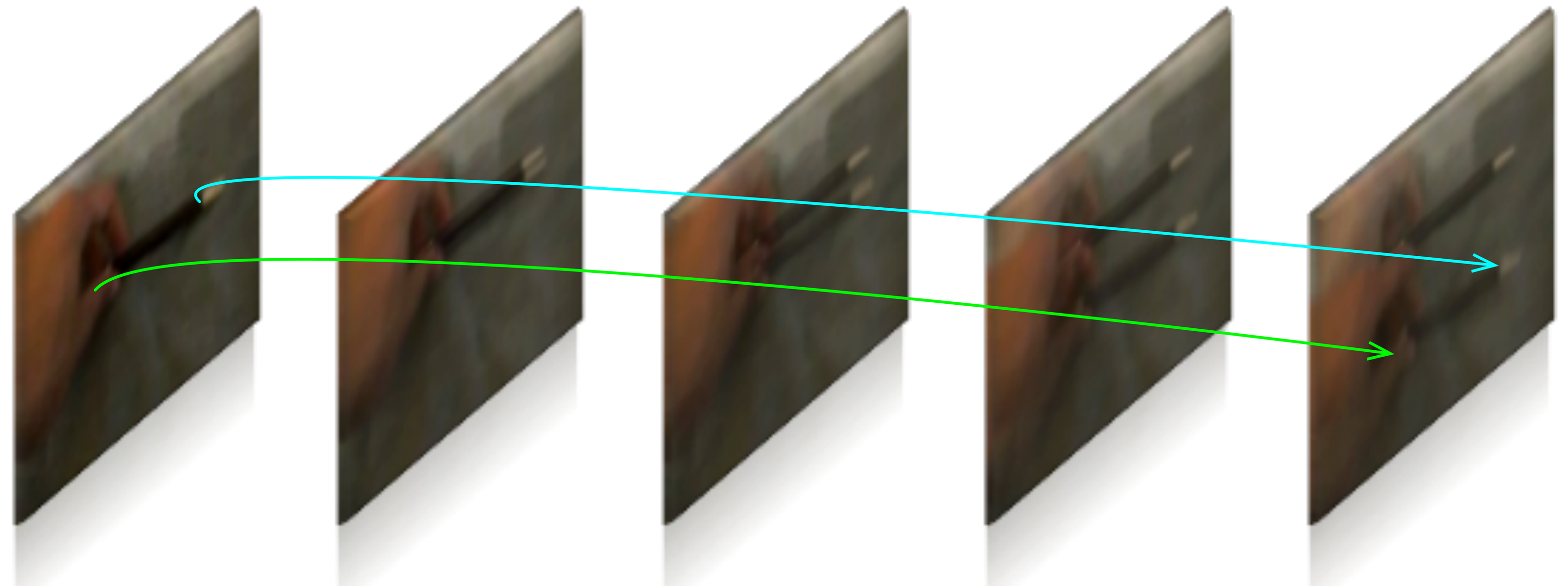
Neural Information Processing Systems Foundation

Introduction

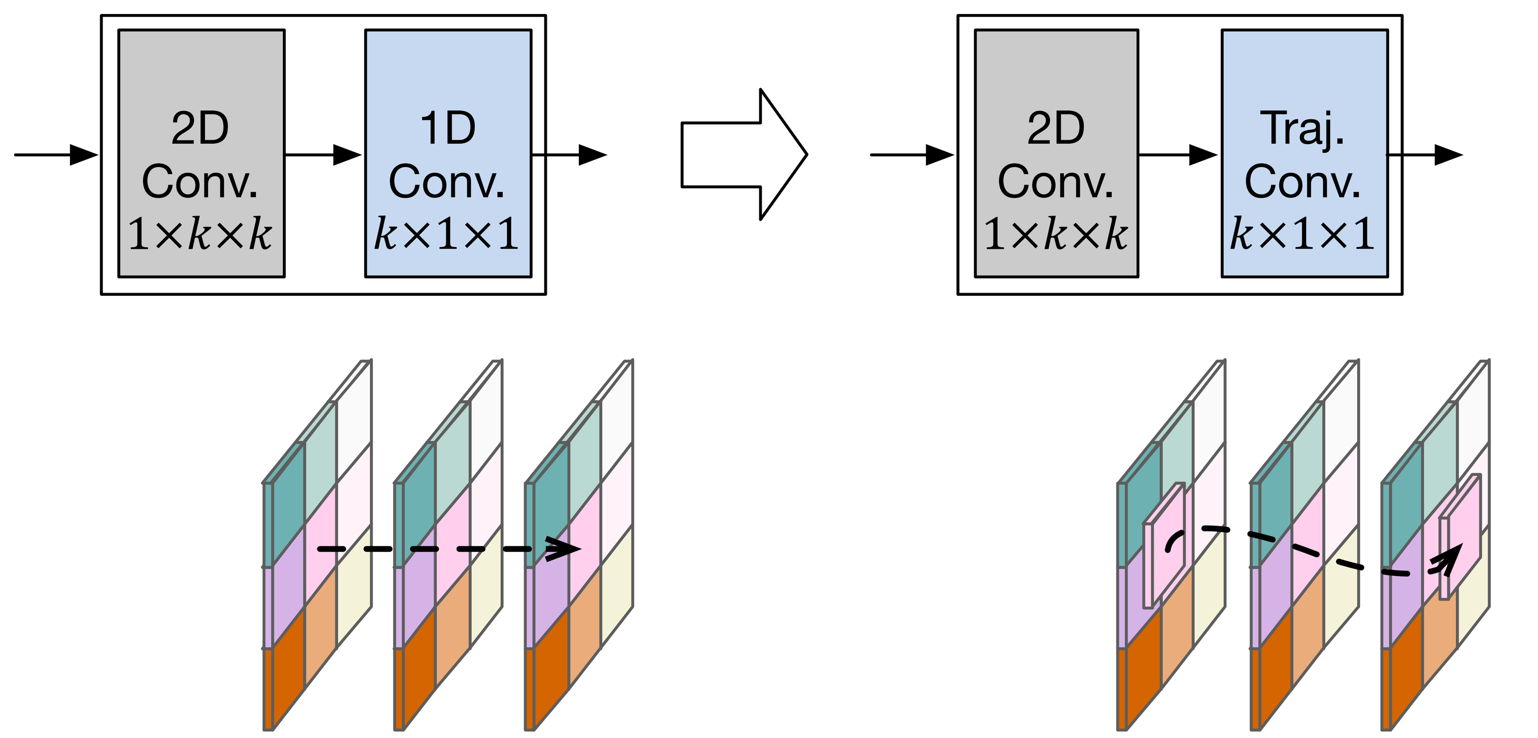
- How to leverage the temporal dimension is one major question in video analysis.
- The temporal convolution in Separable-3D networks, however, comes with an implicit assumption - the feature maps across time steps are well aligned.
- This assumption can be overly strong in action recognition because of *motion*.

Main idea

- The trajectory convolution operates along the *trajectories* that trace the pixels corresponding to the same physical points (e.g. nib), rather than at fixed pixel locations.



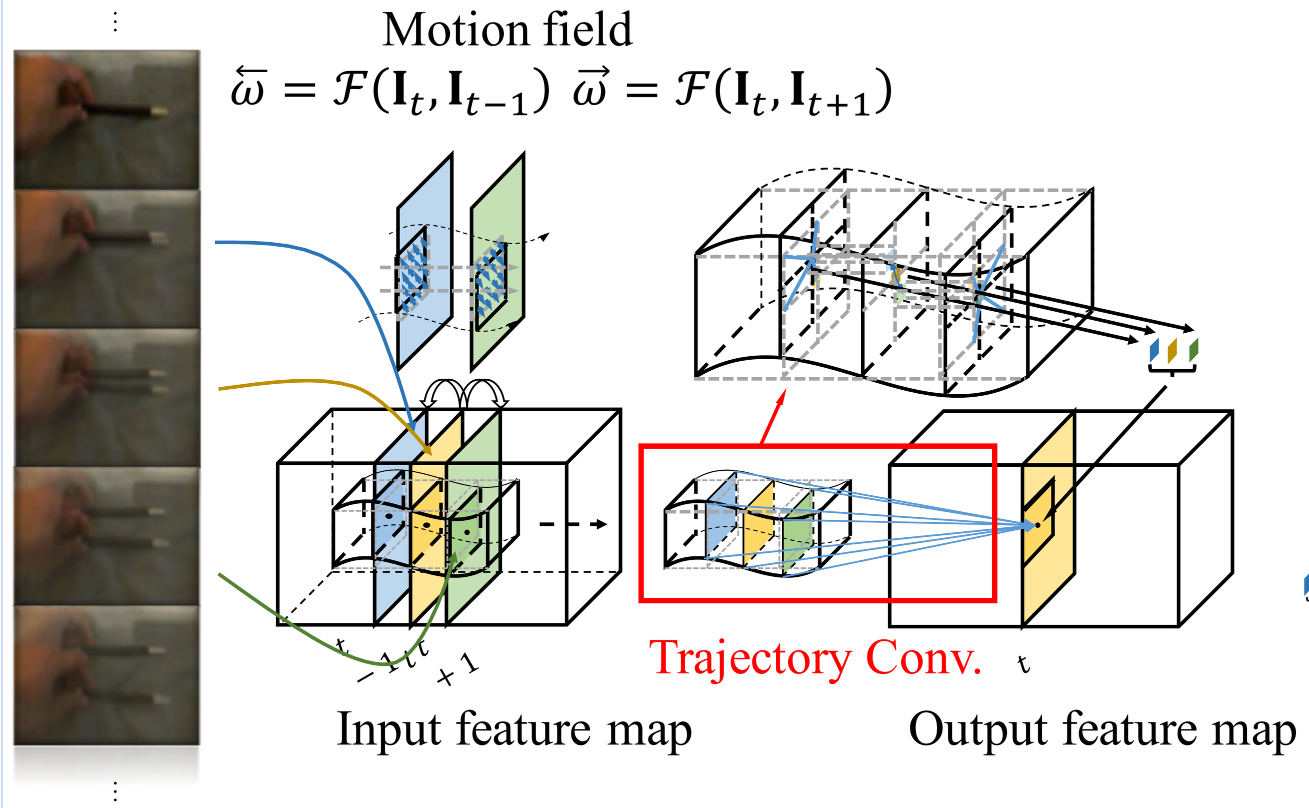
- The standard temporal (1D) convolution in Separable-3D can be seen as a special case of the trajectory convolution where all pixels are considered to be stationary over time.



Comparison with previous methods

Methods	Use deep feature?	Feature tracking?	End-to-end?
STIP + HOG/HOF/MBH/...	✗	✗	✗
DT, iDT	✗	✓	✗
Two-stream, TSN, I3D	✓	✗	✓
TDD	✓	✓	✗
TrajectoryNet (Ours)	✓	✓	✓

Trajectory Convolution



Determine $\tilde{\mathbf{p}}_{t+1}$ from motion

$$\tilde{\mathbf{p}}_{t+1} = \tilde{\mathbf{p}}_t + \overrightarrow{\omega}(\tilde{\mathbf{p}}_t)$$

Interpolate feature $\mathrm{x}(\tilde{\mathbf{p}}_{t+1})$

$$\mathbf{x}(\tilde{\mathbf{p}}_{t+1}) = \sum_{\mathbf{p}'} G(\mathbf{p}', \tilde{\mathbf{p}}_{t+1}) \cdot \mathrm{x}(\mathbf{p}')$$

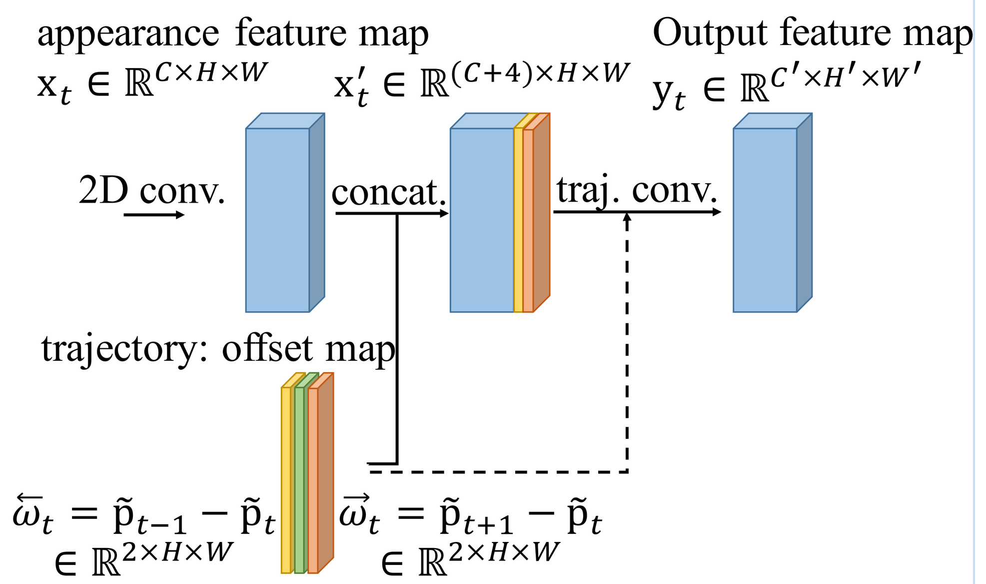
Convolute over $\mathrm{x}(\tilde{\mathrm{p}}_\tau)$

$$\tau \in \{t-1, t, t+1\}$$

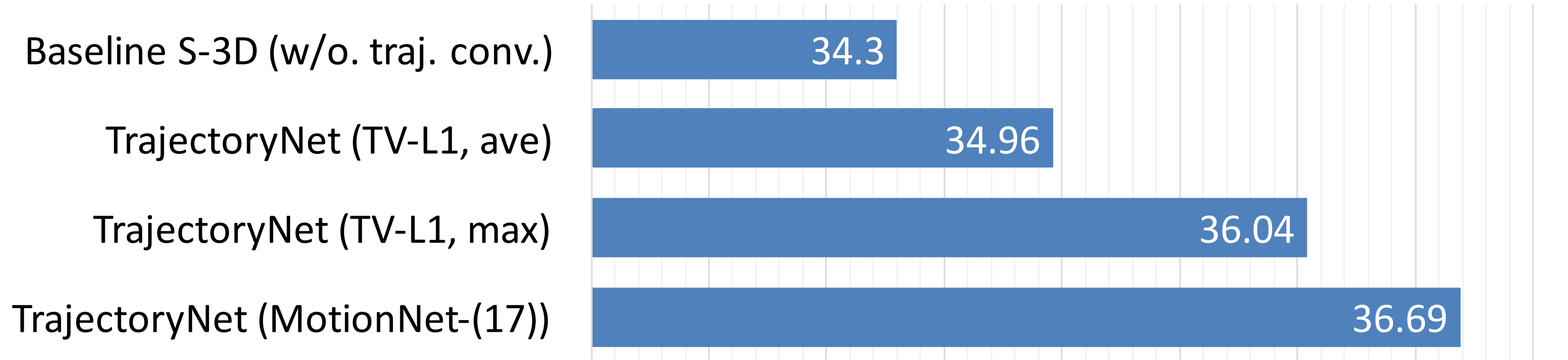
$$\mathbf{y}_t(\mathbf{p}_t) = \sum_\tau \mathbf{w}_\tau \cdot \mathbf{x}(\tilde{\mathbf{p}}_\tau)$$

TrajectoryNet

- We describe local motion patterns at each position $\mathbf{p}$ using the sequence of trajectory information in the form of coordinates of sampling offsets $\{\Delta p_\tau : \tau \in [-\Delta t, \Delta t]\}$.
- The trajectories can either be derived from pre-computed TV-L1 or be learnt from a MotionNet in an unsupervised manner.



Influence of the quality of trajectory



Main results on Something-Something-V1

Method	Backbone Network	Pre-train	Val top1
3D CNN	C3D	Sports-1M	11.5
MultiScale TRN	BN-Inception	ImageNet	34.4
ECO lite	BN-Inception + 3D-ResNet 18	Kinetics	46.4
Non-local I3D + GCN	ResNet-50	Kinetics	46.1
TrajectoryNet-MotionNet-(17) w/o. motion	ResNet-18	ImageNet	43.3
TrajectoryNet-MotionNet-(17) w/. motion	ResNet-18	ImageNet	44.0
TrajectoryNet-MotionNet-(17) w/o. motion	ResNet-18	Kinetics	47.8
TrajectoryNet-MotionNet-(17) w/. motion	ResNet-18	Kinetics	47.9

Visualization of intermediate features