Action Recognition

Computer Vision (SJK02)

Universitat Jaume I

Examples (demos)

Pose estimation and action recognition

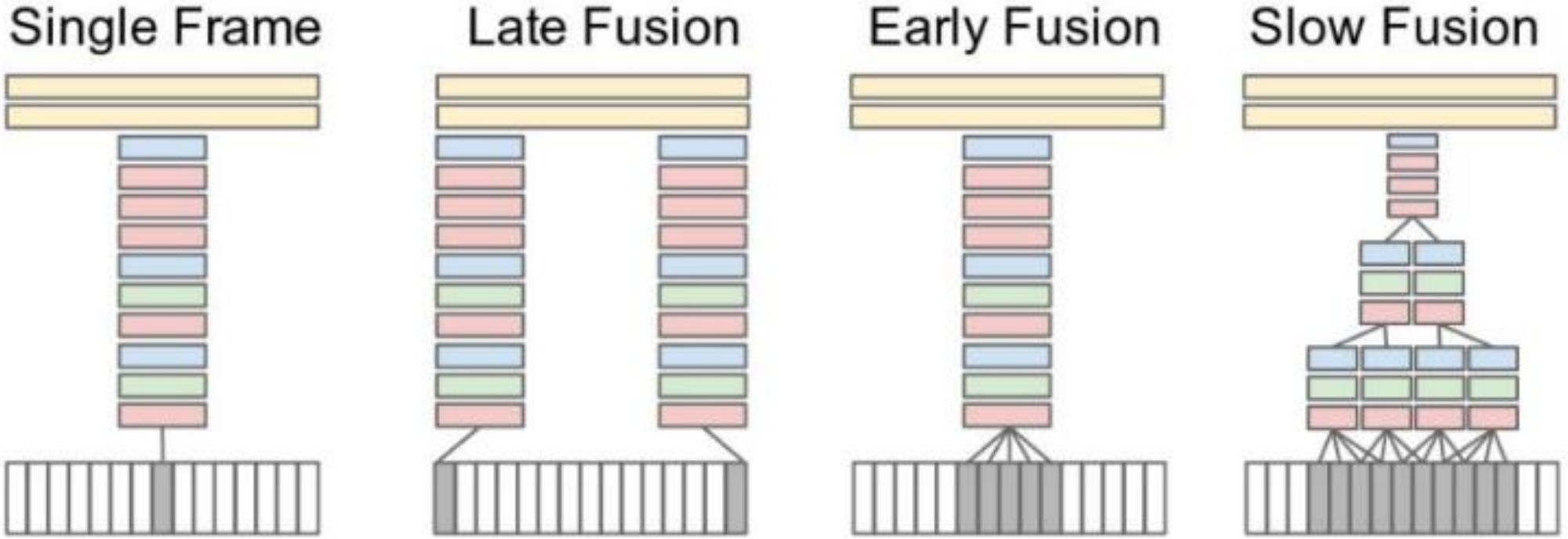
EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition

Choice #1: CNNs

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (CVPR 2017)

Large-scale video classification with convolutional neural networks (CVPR 2014)

What about the temporal information?



W x H x 3

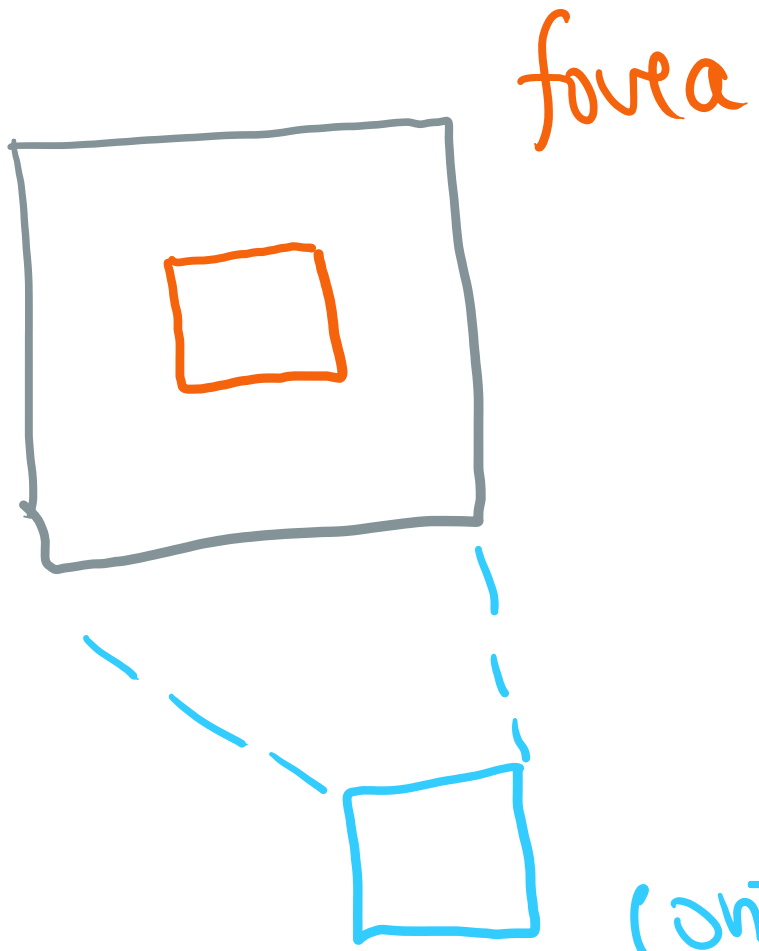
W x H x 3

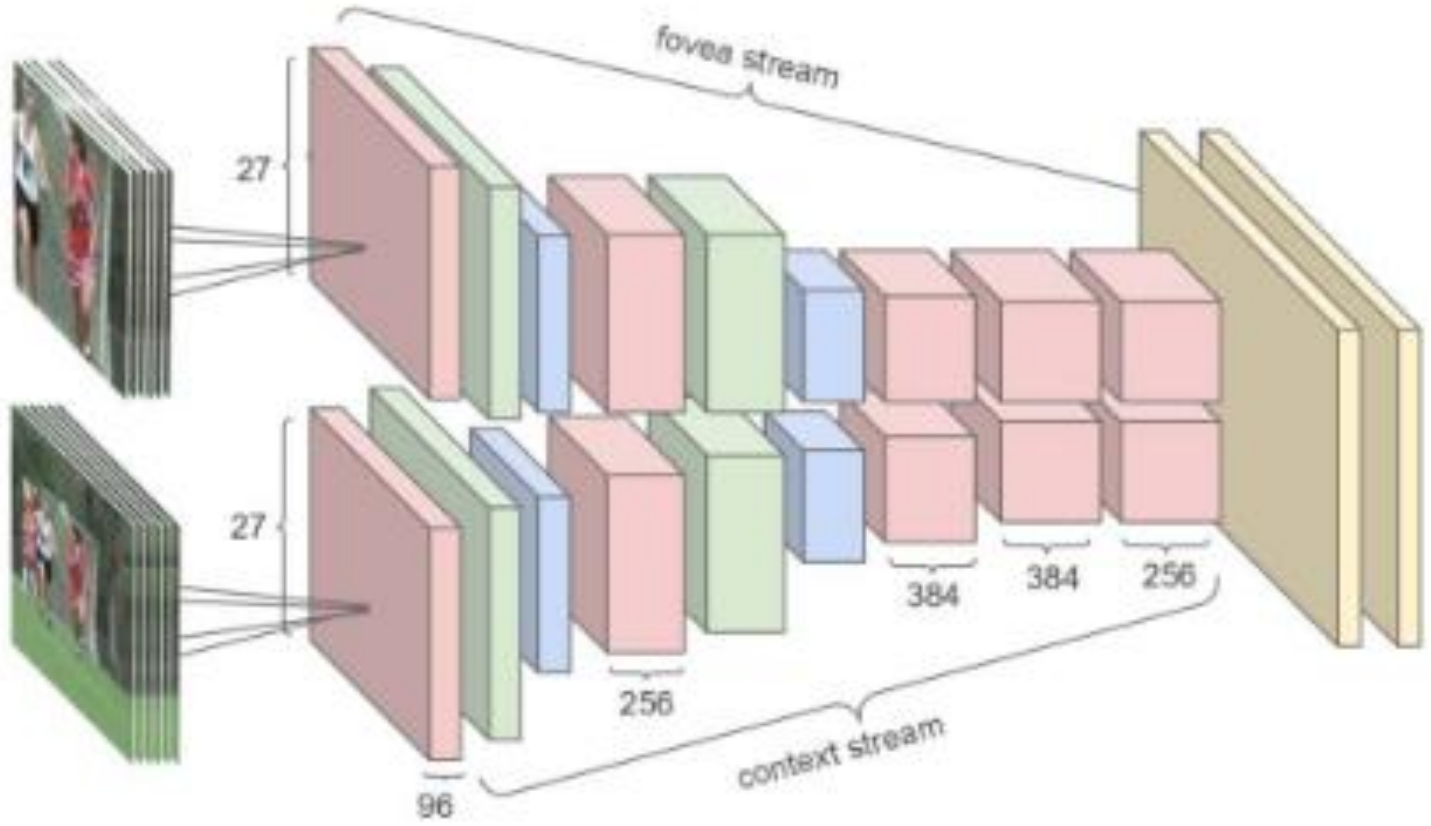
W x H x 3

Shared params

W x h x 3 x T

Multiresolution: fovea + context



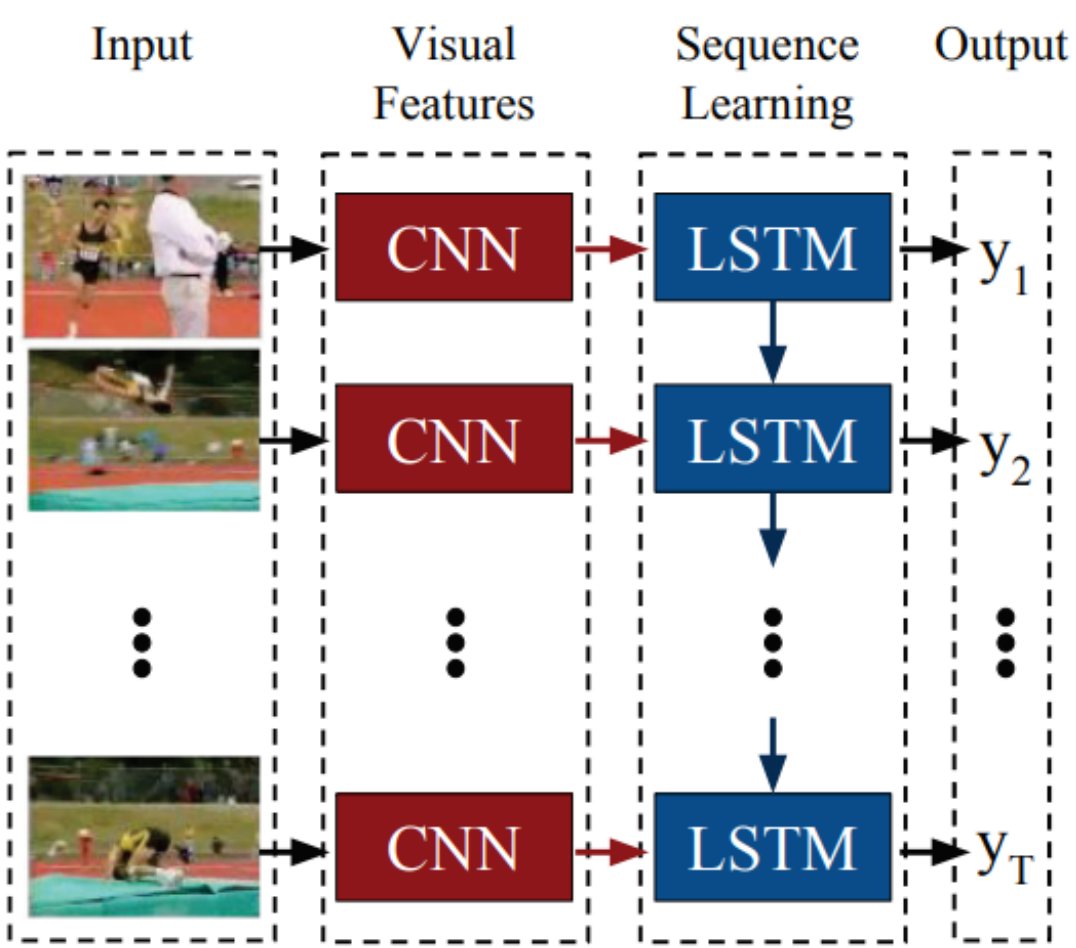


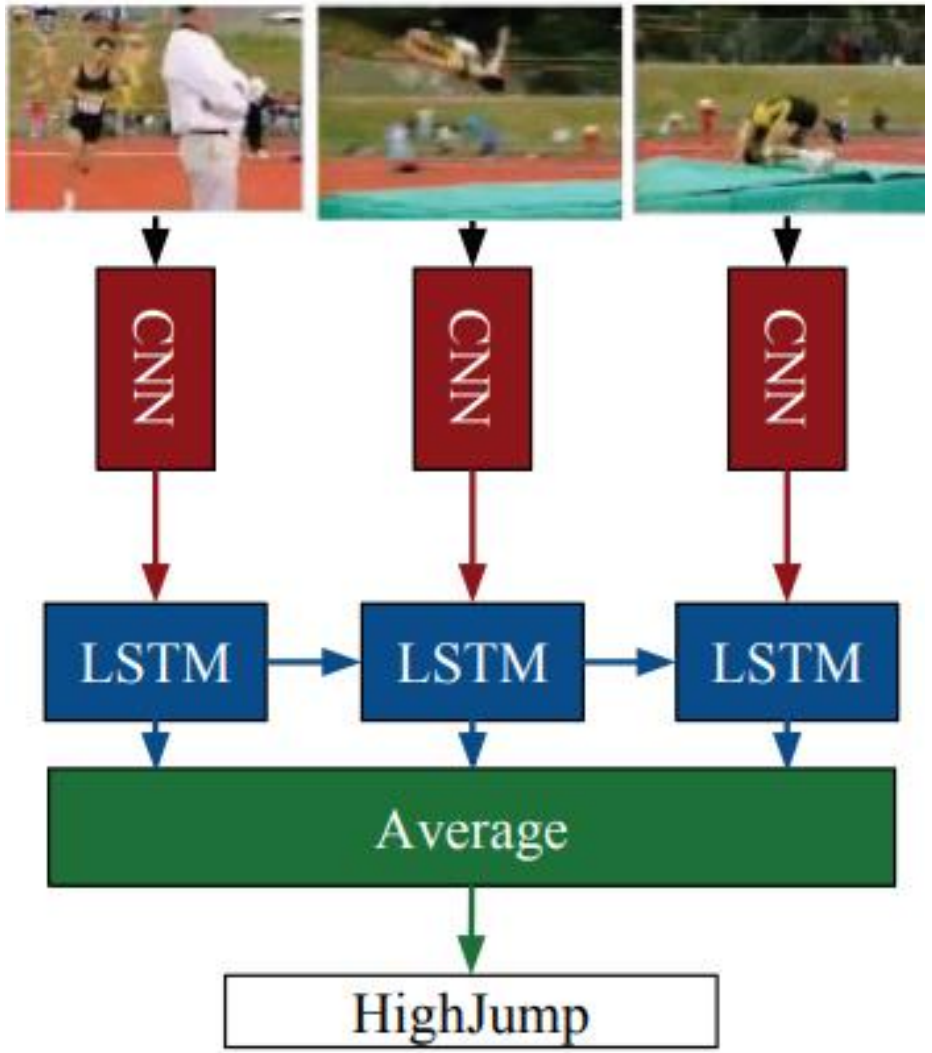
Issue

Generally, temporal structure is ignored (or not well captured)

Choice #2: CNNs + LSTMs

Long-term recurrent convolutional networks for visual recognition and description [CVPR 2015, PAMI 2017]


Input
Visual Features
Sequence Learning
Output
CNN
CNN
CNN
LSTM
LSTM
LSTM
y_1
y_2
y_T


CNN
CNN
CNN
LSTM
LSTM
LSTM
Average
HighJump

Benefit & Issue

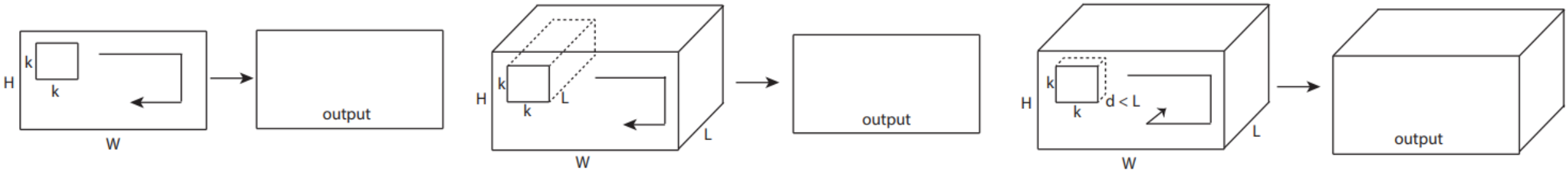
Can model order and long-range dependencies

May miss (potentially critical) low-level motions

Choice #3: 3D CNNs

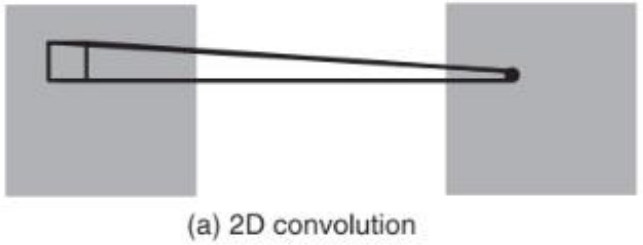
Learning spatiotemporal features with 3d convolutional networks [ICCV 2015]
3D Convolutional Neural Networks for Human Action Recognition [PAMI 2013]

2D and 3D convolutions

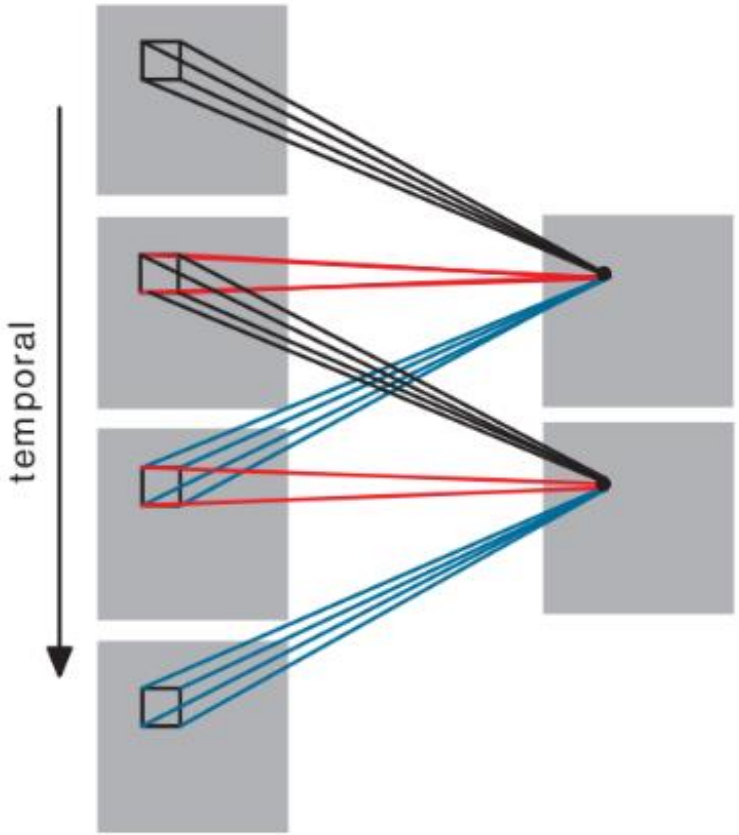


(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

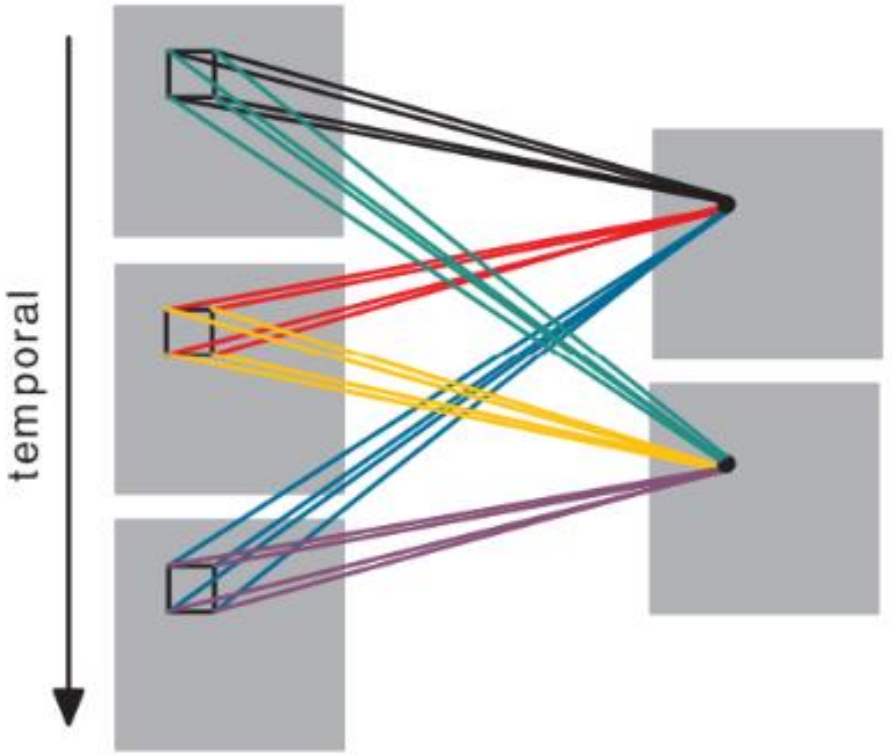
2D and 3D convolutions



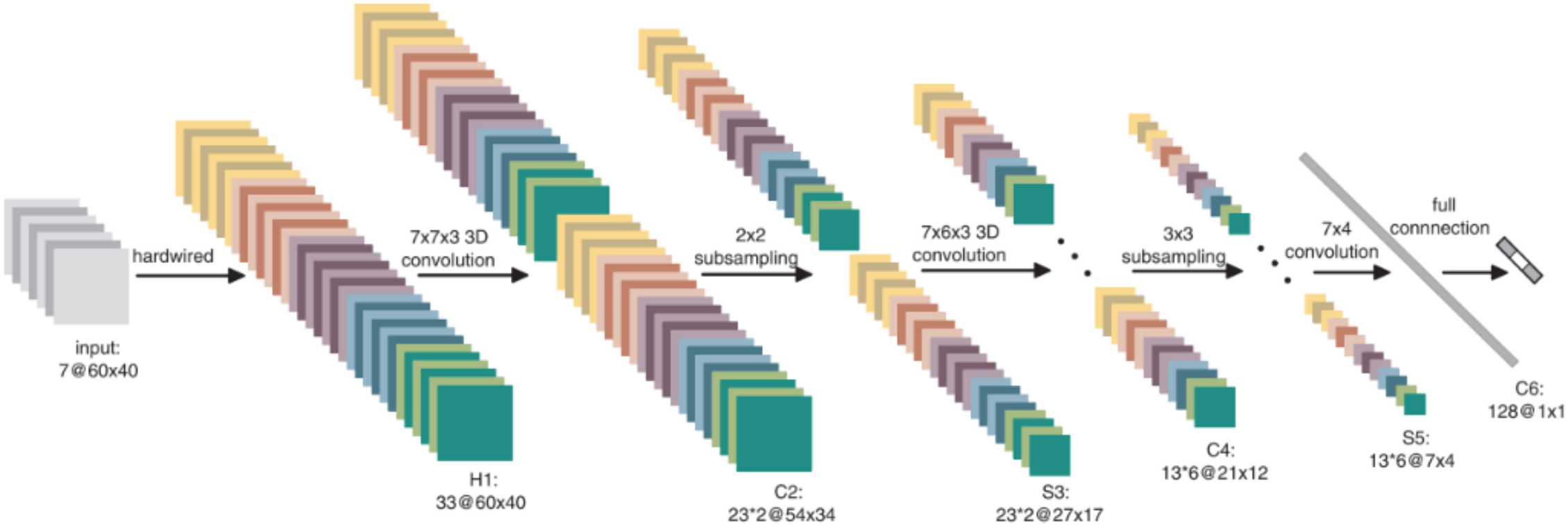
(a) 2D convolution



(b) 3D convolution



Multiple 3D convolutions for multiple features


input:
7@60x40
hardwired
7x7x3 3D
convolution
H1:
33@60x40
2x2
subsampling
C2:
23*2@54x34
7x6x3 3D
convolution
S3:
23*2@27x17
3x3
subsampling
C4:
13*6@21x12
7x4
convolution
S5:
13*6@7x4
full
connnection
C6:
128@1x1

Benefit & Issue

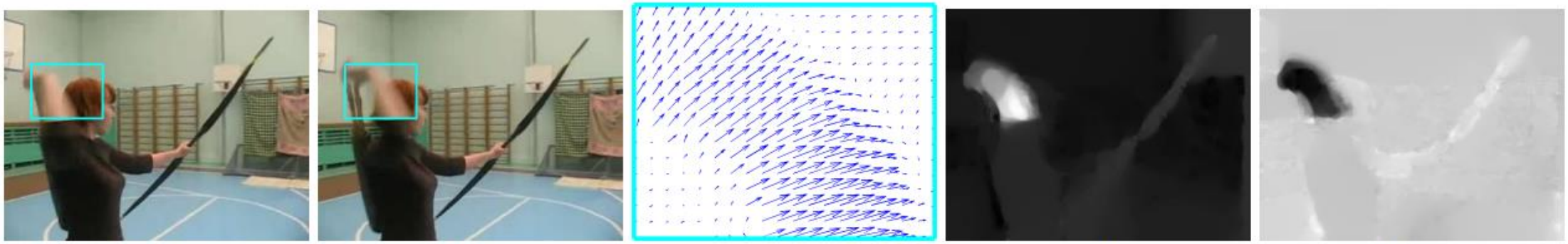
Hierarchical representation of spatio-temporal information

Many more parameters than 2D CNNs

Choice #4: Two-stream nets

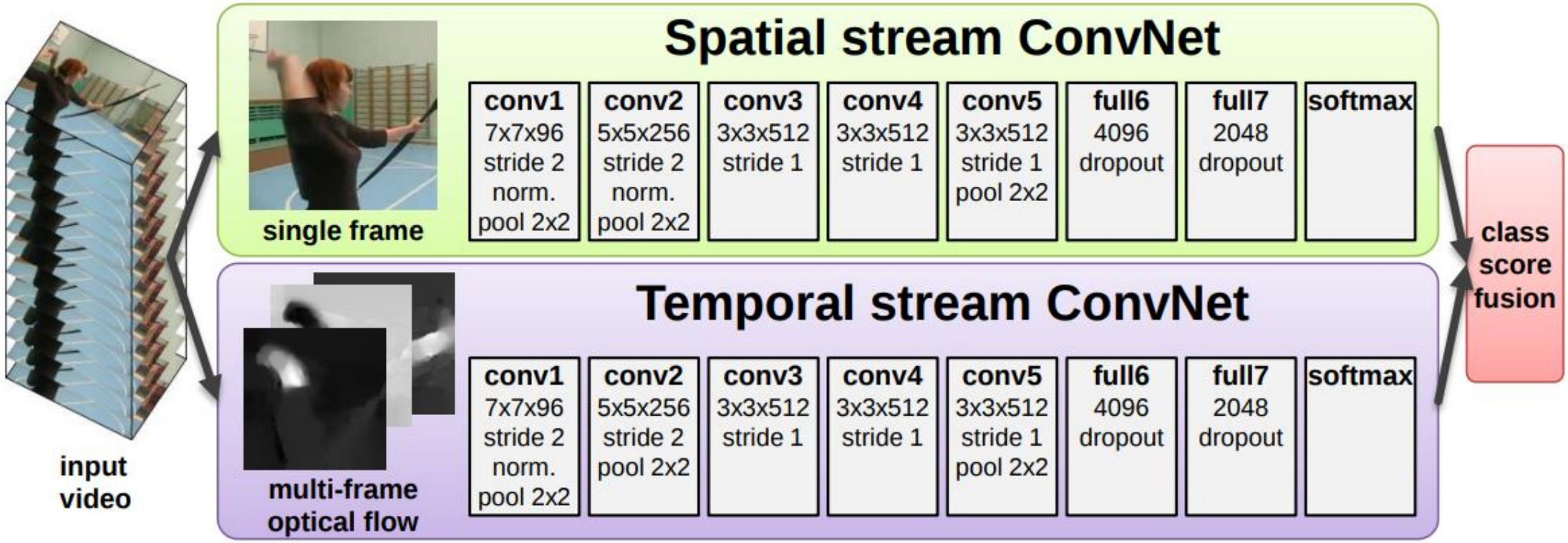
Two-stream convolutional networks for action recognition in videos (NIPS 2014)
Convolutional Two-Stream Network Fusion for Video Action Recognition (CVPR 2016)

Optic flow (precomputed, dense)

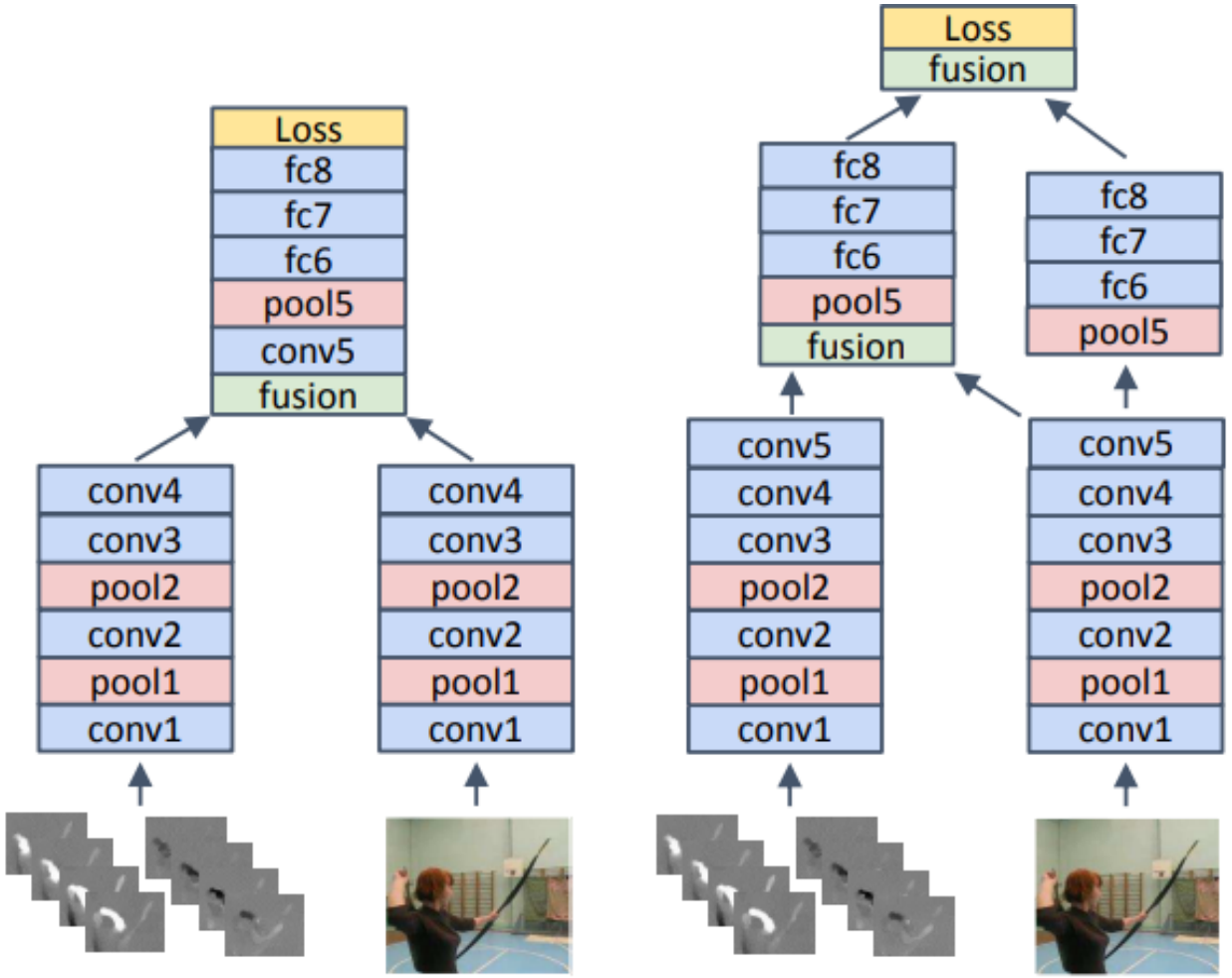


Not "end-to-end" (OF precomputed)

Spatial + Temporal streams



Next version: fusing the streams



Benefit

Although 3D CNNs can capture temporal information, the two-stream architecture improves performance by leveraging on optic flow

Choice #5: Two-stream inflated 3D CNNs

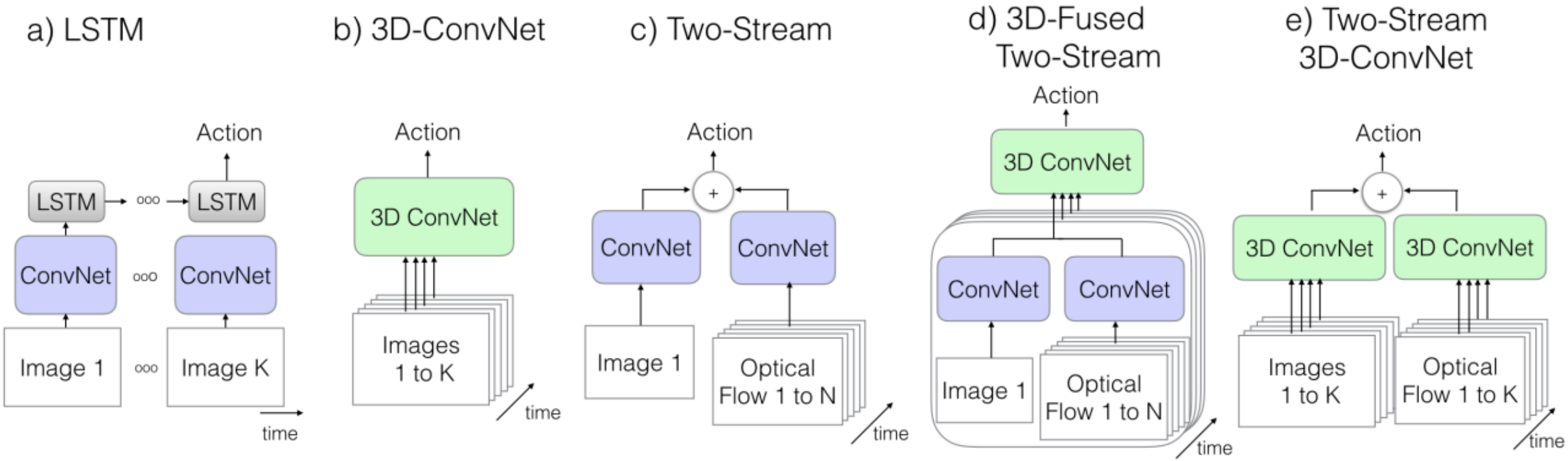
Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (CVPR 2017)

Turn successful 2D CNNs into 3D CNNs

How? By turning NxN filters into NxNxN ones

+ some other modifications

Comparison study



Method	#Params	Training		Testing	
		# Input Frames	Temporal Footprint	# Input Frames	Temporal Footprint
ConvNet+LSTM	9M	25 rgb	5s	50 rgb	10s
3D-ConvNet	79M	16 rgb	0.64s	240 rgb	9.6s
Two-Stream	12M	1 rgb, 10 flow	0.4s	25 rgb, 250 flow	10s
3D-Fused	39M	5 rgb, 50 flow	2s	25 rgb, 250 flow	10s
Two-Stream I3D	25M	64 rgb, 64 flow	2.56s	250 rgb, 250 flow	10s

Architecture	UCF-101			HMDB-51			miniKinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	69.9	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	60.0	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	70.1	58.4	72.9
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	71.4	61.0	74.0
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**74.1**	**69.6**	**78.7**