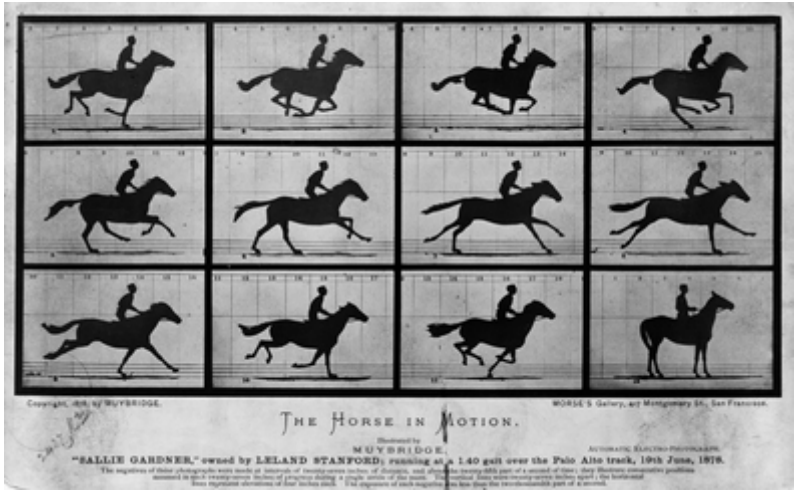
Temporal Models for Video Processing

Tasks

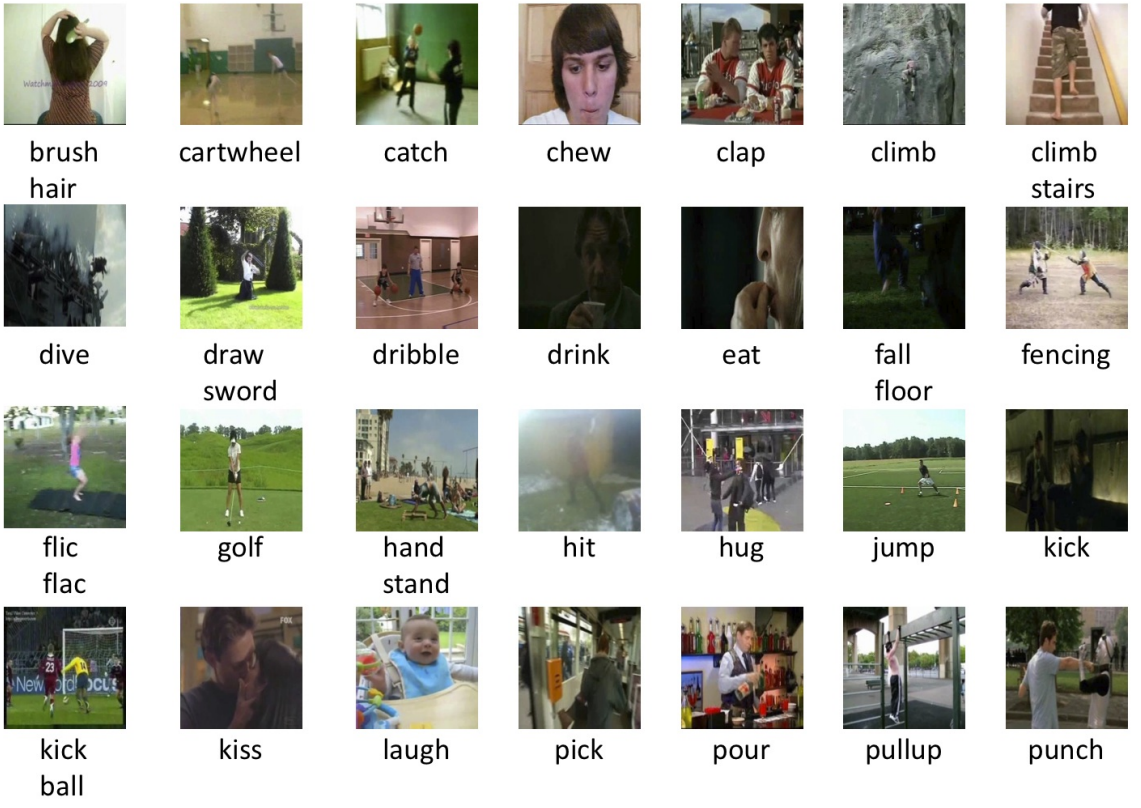
- Action recognition
- Action prediction
- Tracking



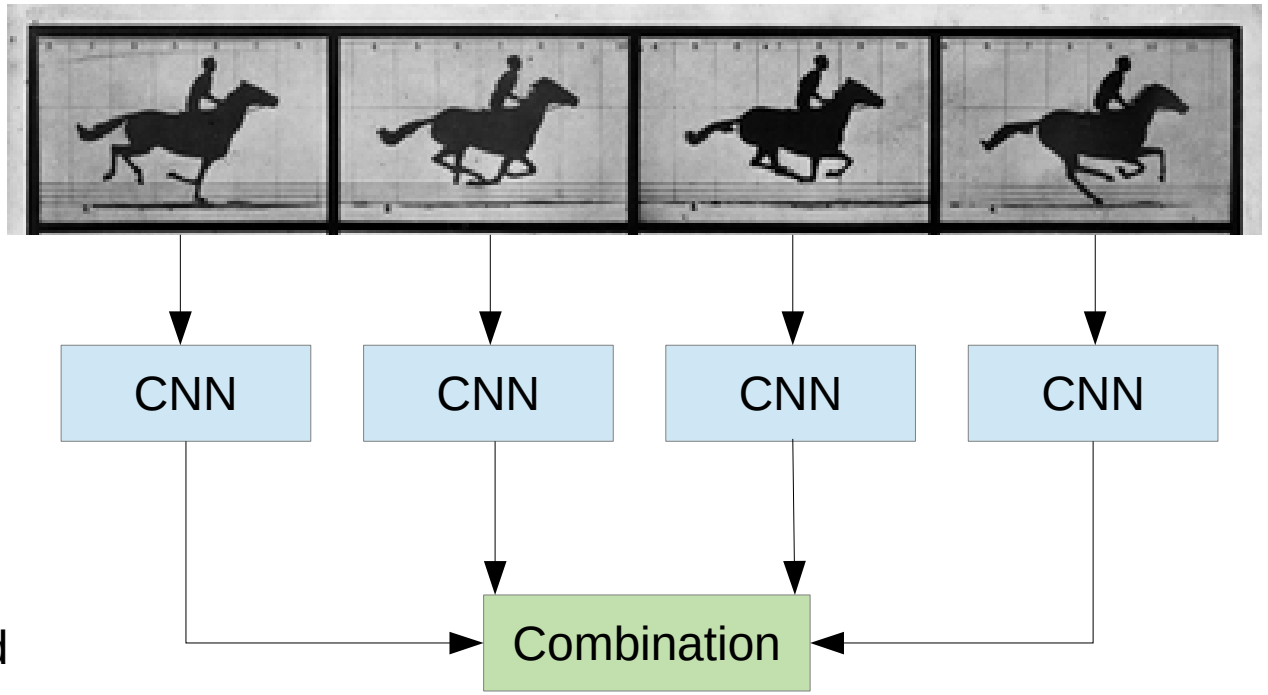
- Monocular 3D estimation (active sensing)

Datasets

- HMDB-51: 7000 videos of 51 actions
- UFC 101: 13,320 videos of 101 actions
- Kinetics: Up to 650,000 videos, up to 700 actions

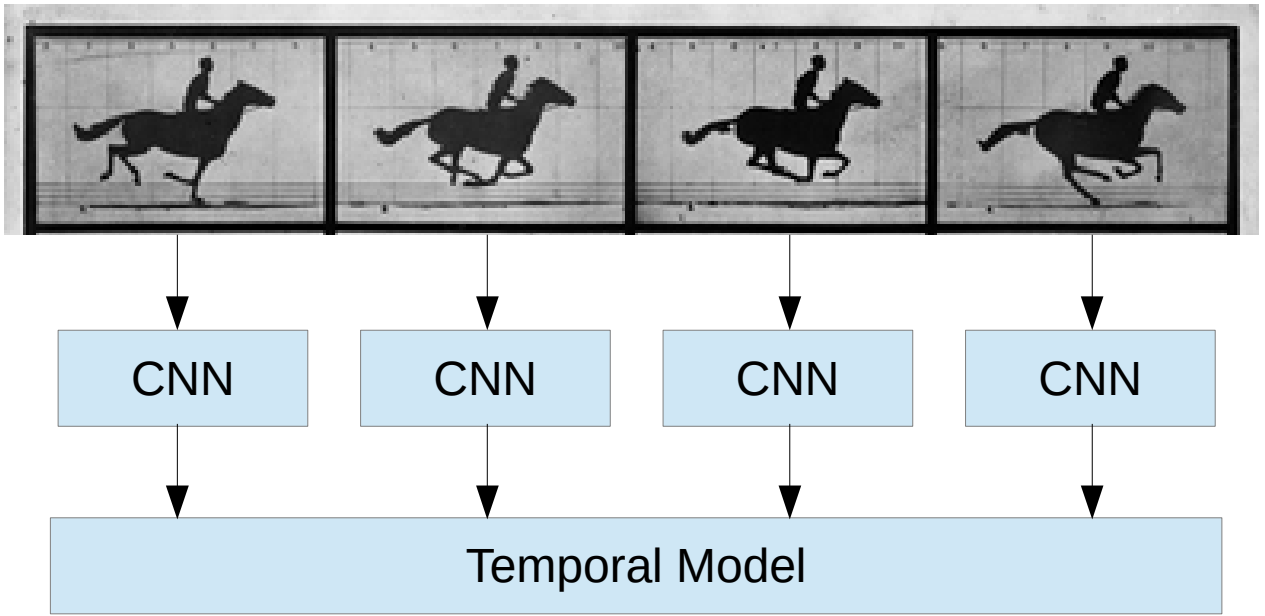


Approach 1: Unordered Frames



Pretty good baseline

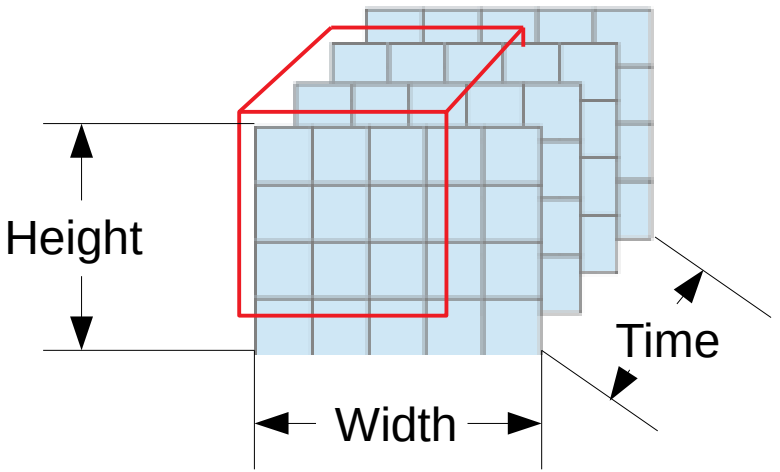
Approach 2: Frames + Global Model



Tends to overfit, hard to train

3D Convolutions

Convolution across both space and time



3D Convolutions

Input $\mathbf{X} \in \mathbb{R}^{C_i \times T \times H \times W}$

Kernel $\mathbf{W} \in \mathbb{R}^{C_o \times C_i \times t \times h \times w}$

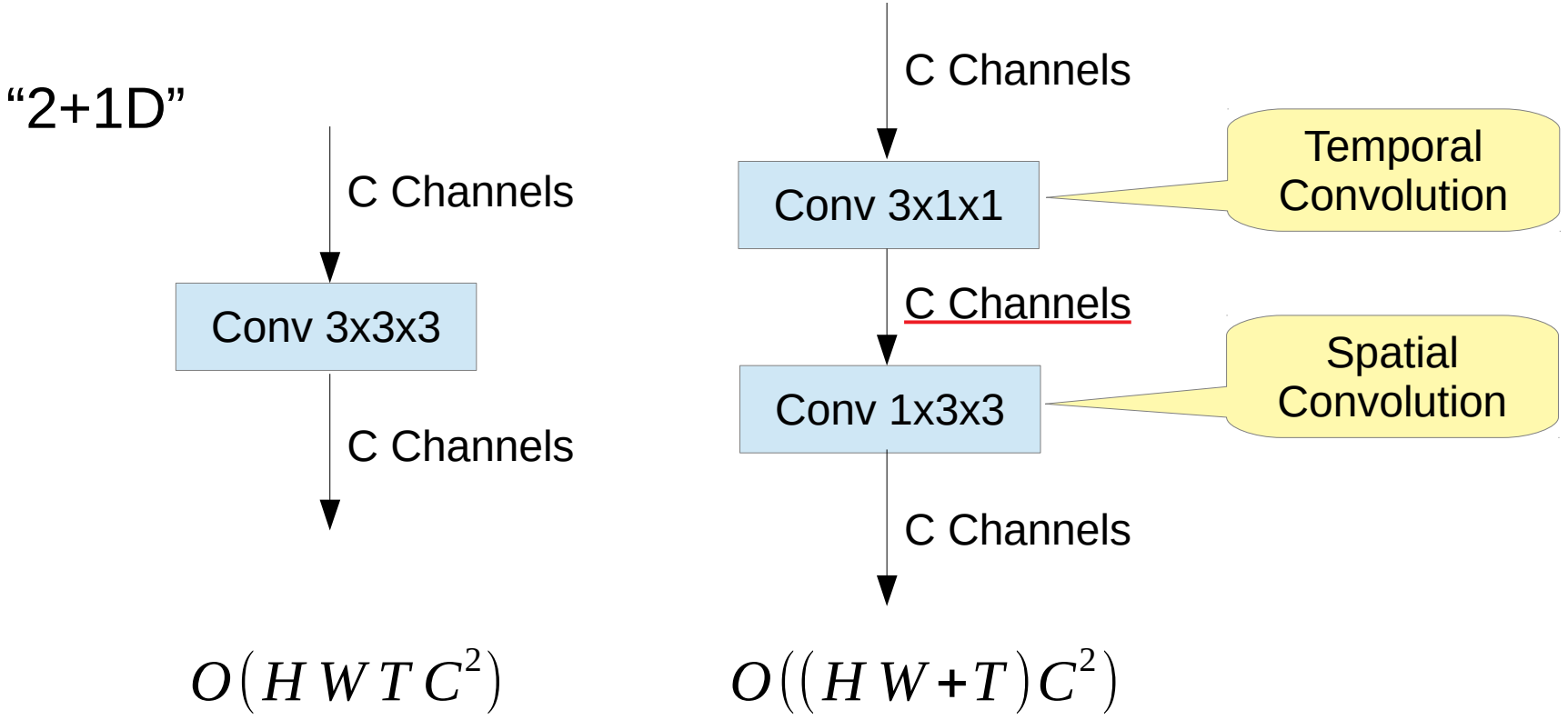
Bias $\mathbf{b} \in \mathbb{R}^{C_o}$

Output $\mathbf{Z} \in \mathbb{R}^{C_o \times \left(\frac{T-t+2p_t}{s_t}+1\right) \times \left(\frac{H-h+2p_h}{s_h}+1\right) \times \left(\frac{W-w+2p_w}{s_w}+1\right)}$

$$\mathbf{Z}_{c,d,a,b} = \mathbf{b}_c + \sum_{l=0}^{C_i} \sum_{i=0}^{t} \sum_{j=0}^{h} \sum_{k=0}^{w} \mathbf{X}_{k,d+i,a+j,b+k} \mathbf{W}_{c,l,i,j,k}$$

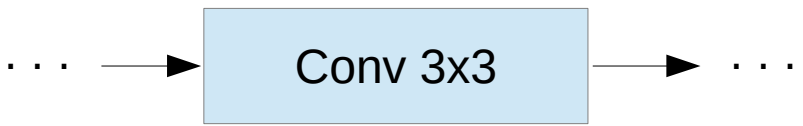
Very slow!

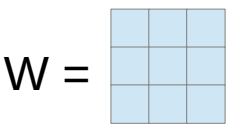
Factorized 3D Convolutions



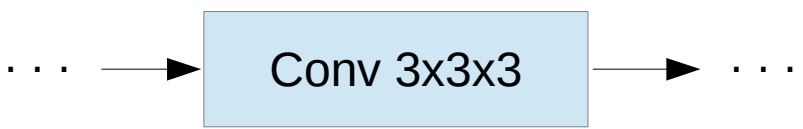
I3D

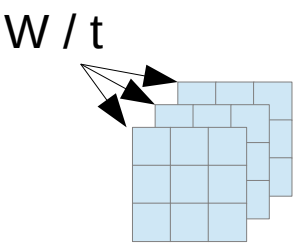
- Pre-train a network on ImageNet





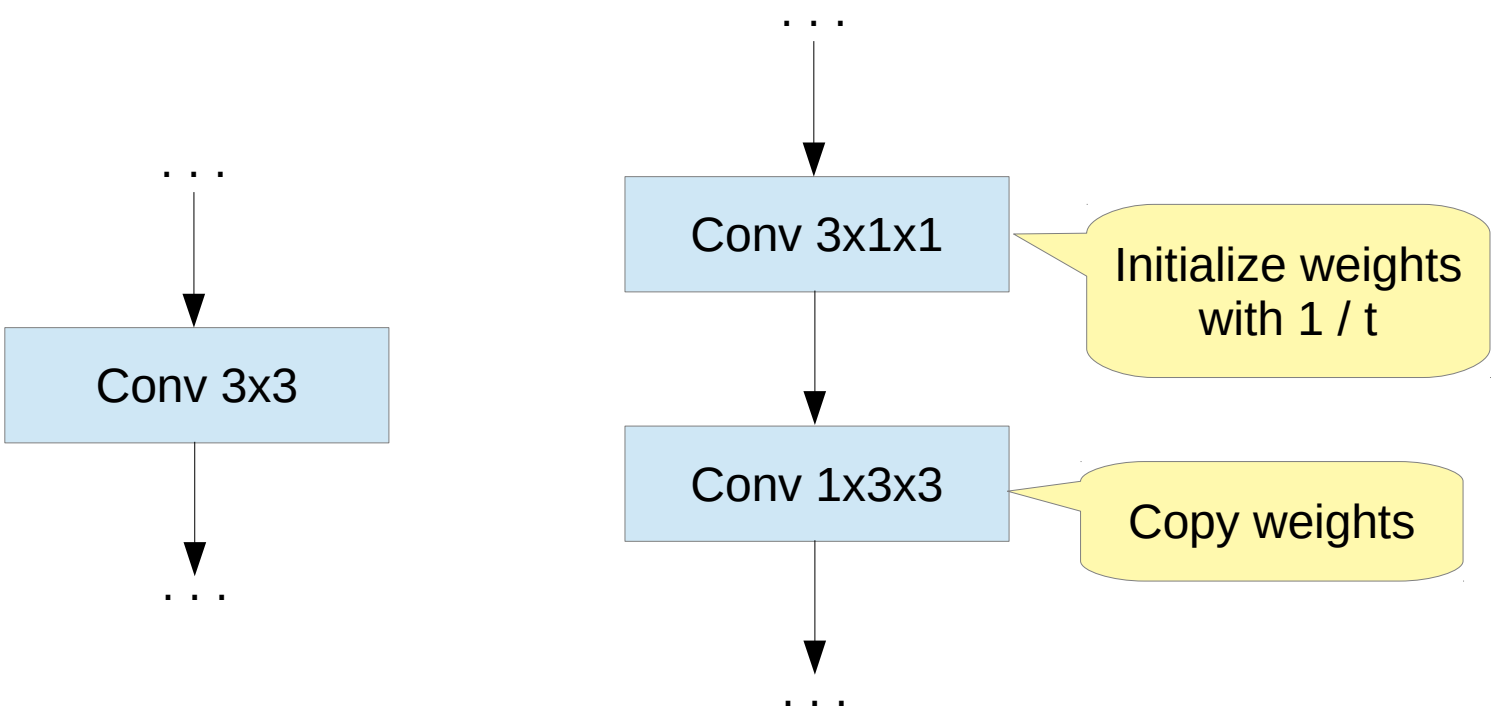
- “Inflate” some 2D convolutions to 3D





João Carreira, Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017.

I2+1D


. . .
Conv 3x3
. . .
. . .
Conv 3x1x1
Initialize weights with 1 / t
Conv 1x3x3
Copy weights
. . .

Open Problem: What Tasks Should We Care About?

- Vision tasks are often proxies or initial steps in other applications
  - But they often don’t really capture the downstream task.
- Vision is good as a test bed.
  - New architectures
  - Pre-training