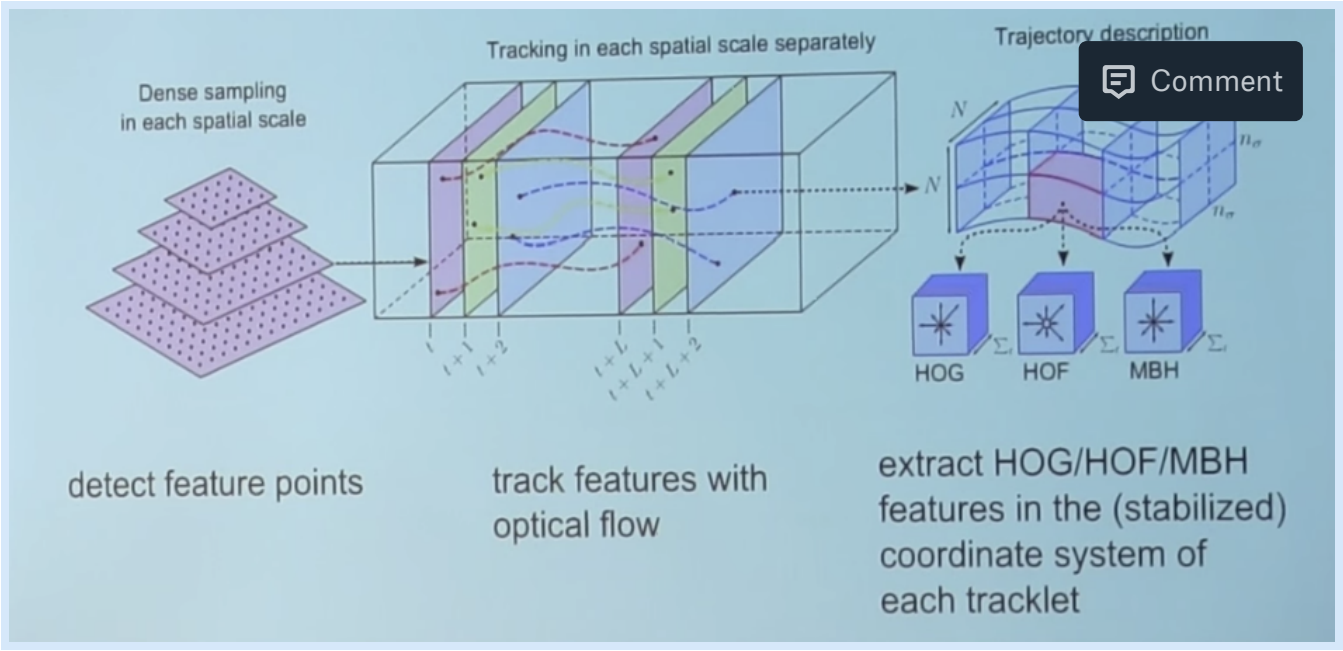
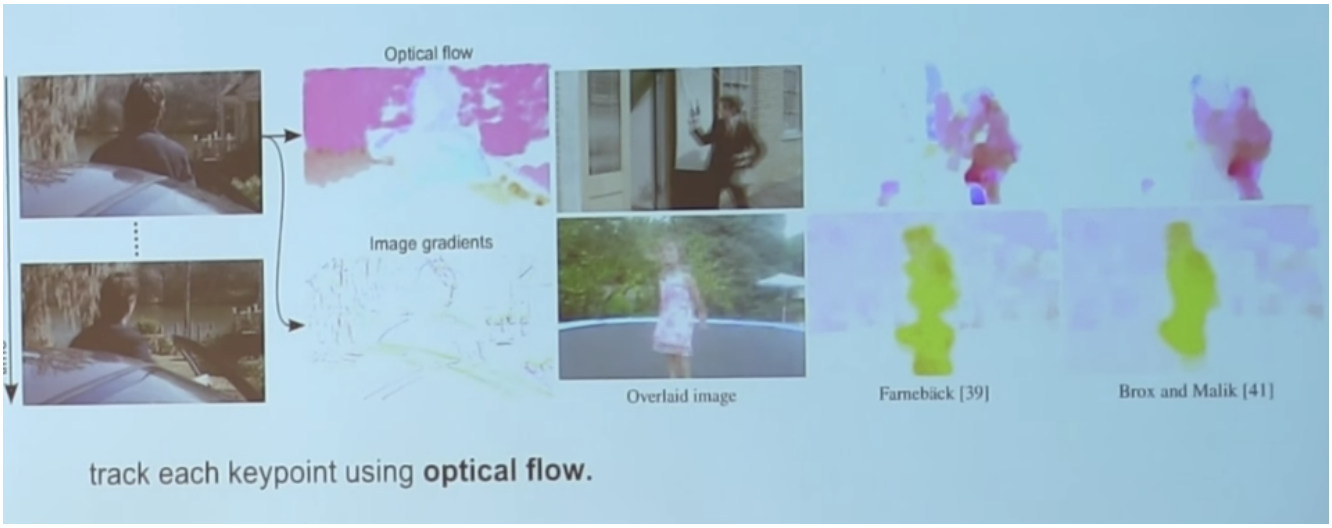
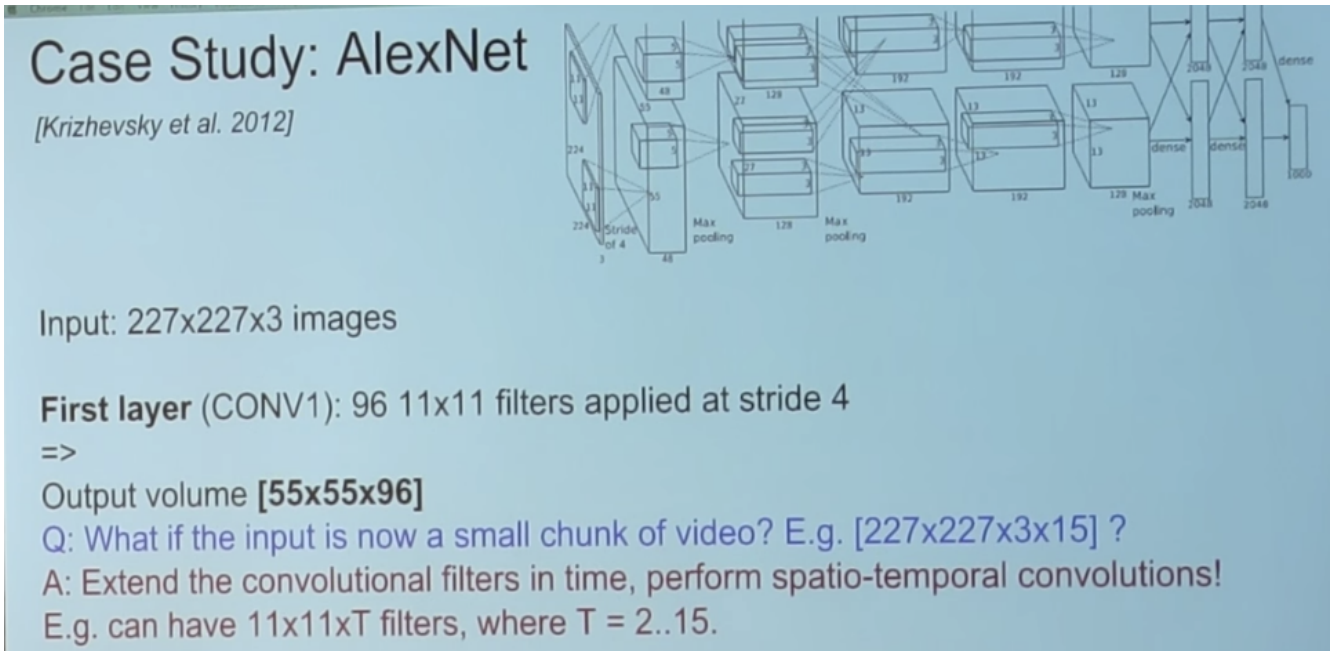
video analysis

- First, see if you can try a technique from image analysis on your video task. do you really need to use spatio temporal information for you video task? Maybe not!
- before CNNs → they would identify these key feature points to track based on gradients and frequency and other features. they would create frames with the origin at these points and track the points and use that to compute optical flow which is basically a displacement vector of given point across frames.

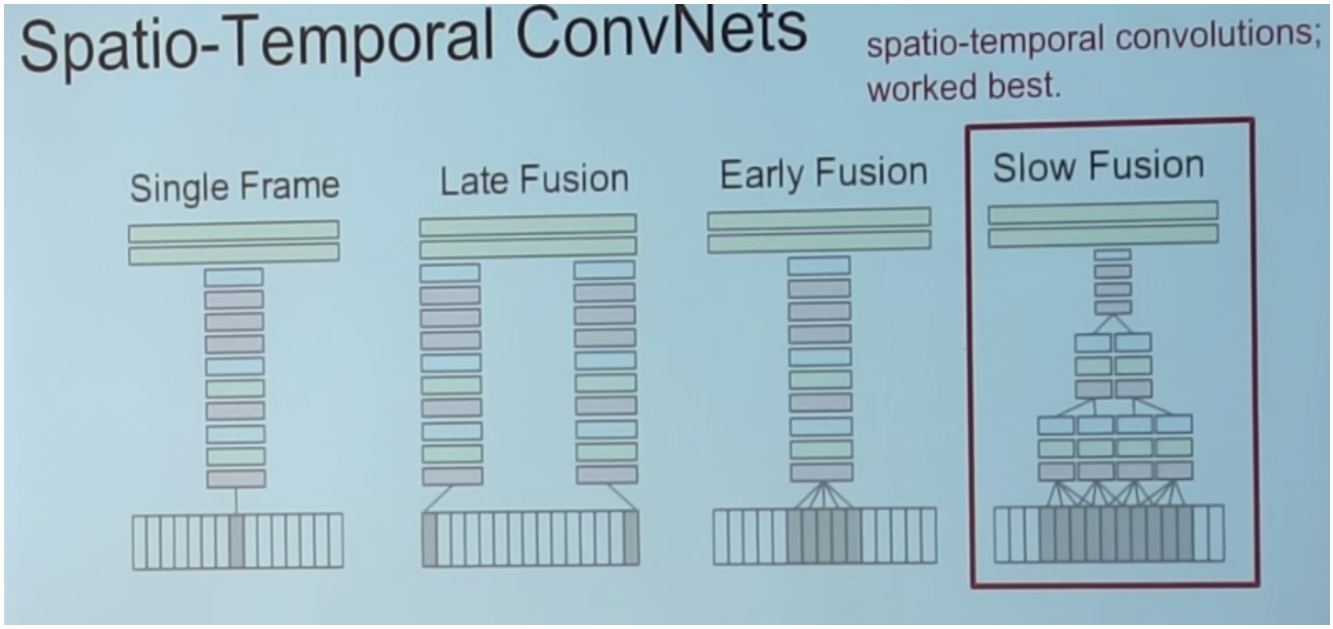




- then came CNNs



- the method in the black square worked best



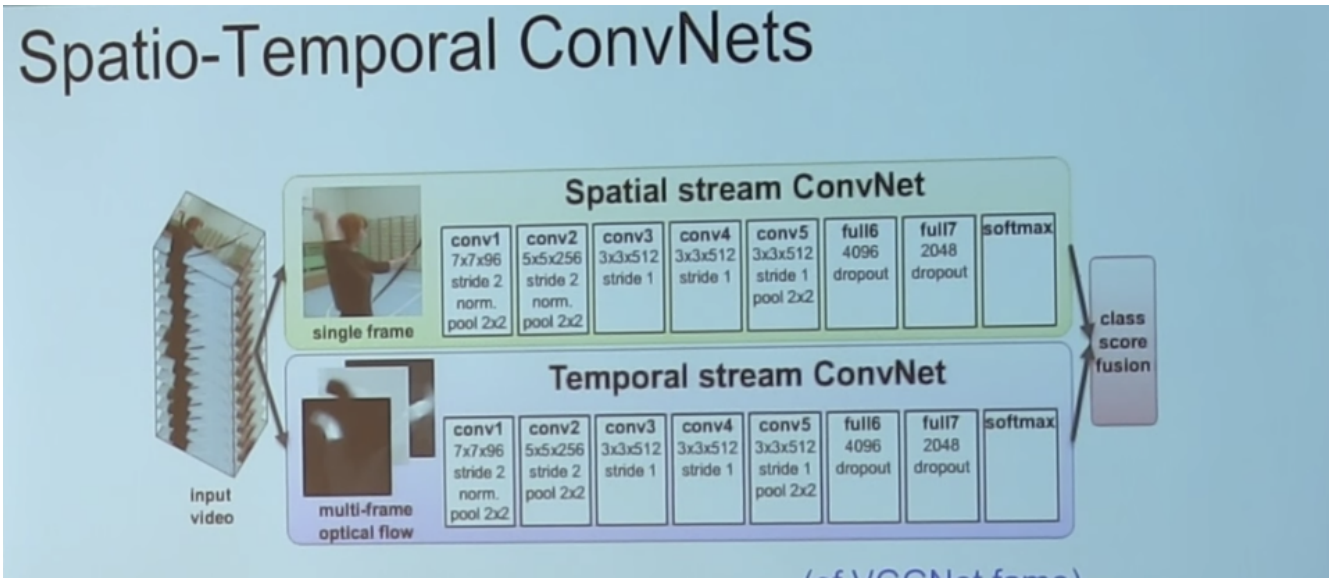
- maybe you don’t need to take advatantage of spatio temporal information and can just use vanilla alexnet etc frame by frame

Spatio-Temporal ConvNets

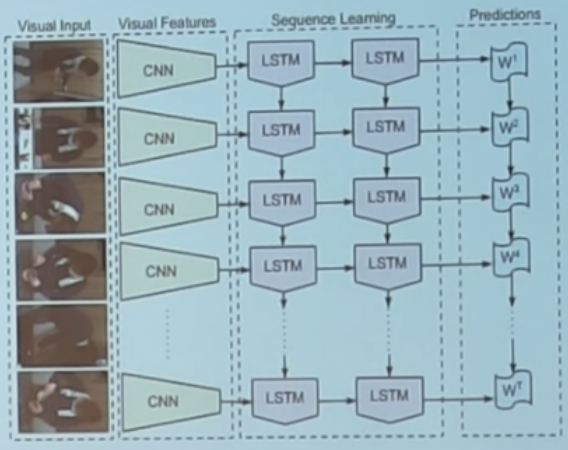
Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Single-Frame + Multires	**42.4**	**60.0**	**78.5**
Single-Frame Fovea Only	30.0	49.9	72.8
Single-Frame Context Only	38.1	56.0	77.2
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	**41.9**	60.9	**80.2**
CNN Average (Single+Early+Late+Slow)	41.4	63.9	82.4

The motion information didn't add all that much...

- compute optical flow image and use that as a feature



Long-time Spatio-Temporal ConvNets

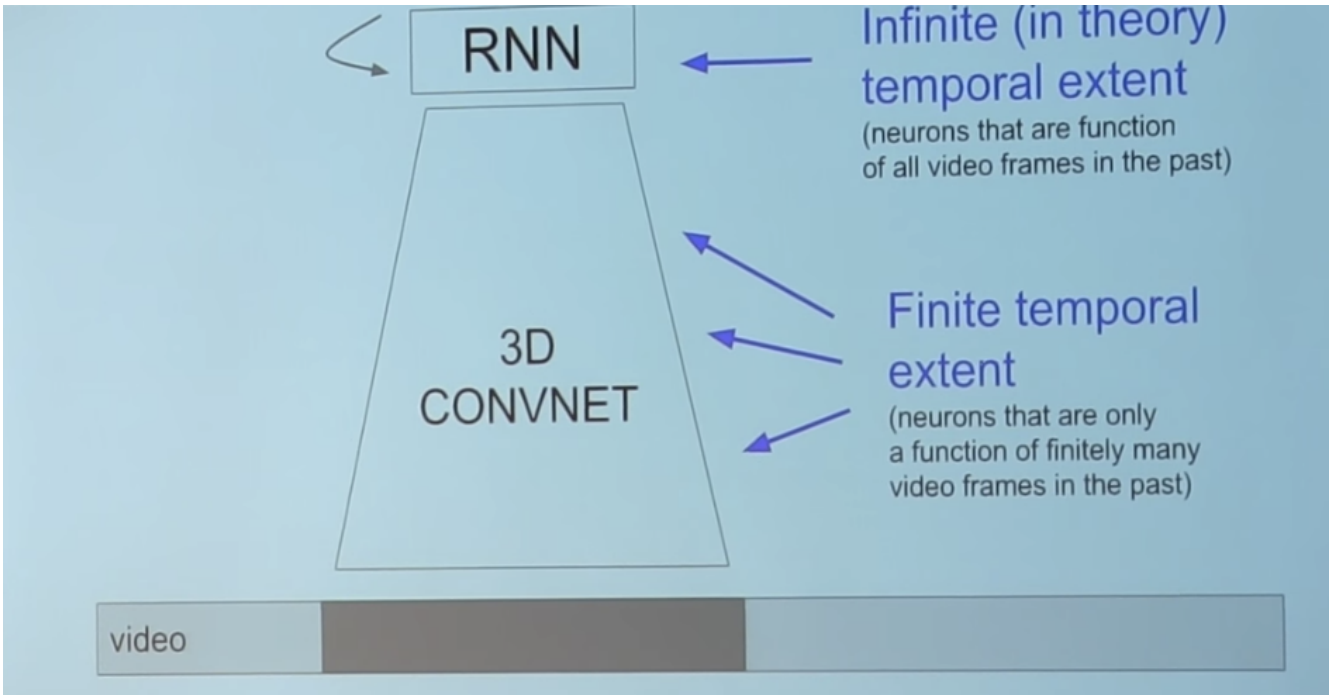


Summary so far

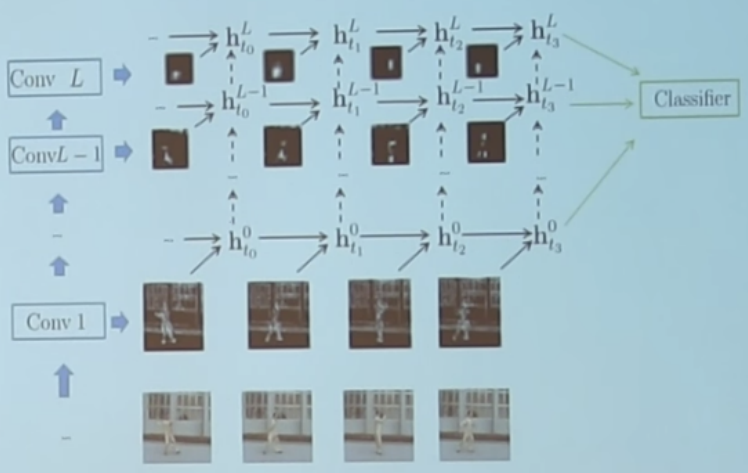
We looked at two types of architectural patterns:

1. Model temporal motion locally (3D CONV)
2. Model temporal motion globally (LSTM / RNN)

\+ Fusions of both approaches at the same time.



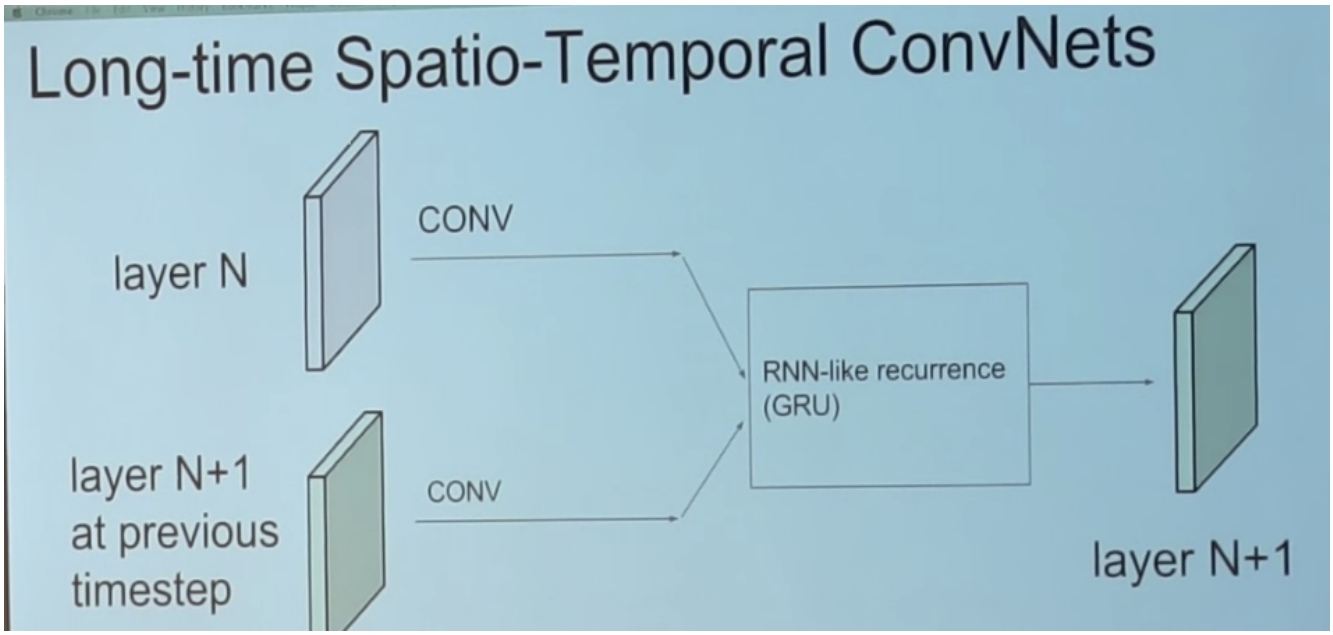
Long-time Spatio-Temporal ConvNets



Beautiful:
All neurons in the ConvNet are recurrent.

$$
\begin{aligned}
\mathbf{z}_t^l &= \sigma(\mathbf{W}_z^l * \mathbf{x}_t^l + \mathbf{U}_z^l * \mathbf{h}_{t-1}^l), \\
\mathbf{r}_t^l &= \sigma(\mathbf{W}_r^l * \mathbf{x}_t^l + \mathbf{U}_r^l * \mathbf{h}_{t-1}^l), \\
\bar{\mathbf{h}}_t^l &= \tanh(\mathbf{W}^l * \mathbf{x}_t^l + \mathbf{U} * (\mathbf{r}_t^l \odot \mathbf{h}_{t-1}^l), \\
\mathbf{h}_t^l &= (1 - \mathbf{z}_t^l)\mathbf{h}_{t-1}^l + \mathbf{z}_t^l \bar{\mathbf{h}}_t^l,
\end{aligned}
$$

Only requires (existing) 2D CONV routines. No need for 3D spatio-temporal CONV.


Long-time Spatio-Temporal ConvNets
layer N
CONV
RNN-like recurrence
(GRU)
layer N+1
at previous
timestep
CONV
layer N+1