Large-scale Video Classification with Convolutional Neural Networks

Andrej Karpathy, George Toderici, Sanketh Shetty,
Thomas Leung, Rahul Sukthankar, Li Fei-Fei

Note: Slide content mostly from : Bay Area Multimedia Forum - 20 June 2014 - Andrej Karpathy - Large-scale Video Classification with Convolutional Neural Networks

16-824 Spring 2015
Presenter : Esha Uboweja

Problem

Classification of videos in sports datasets

Standard approach to video classification

Bag of Words (BoW) approach:

1. Extraction of local visual features (dense/sparse)

2. Visual word encoding of features

3. Training a classifier (e.g. SVM)

Convolutional Neural Networks (CNNs) emulate all these stages in a single neural network

Motivations for using CNNs for video classification

1. CNNs outperform other approaches in image classification tasks (e.g. ImageNet challenge)

2. Features learned in CNNs transfer well to other datasets (e.g. fine-tuning top layers of a network trained using ImageNet for food recognition)

Dataset

Current video datasets lack variety and number of videos to train a CNN:

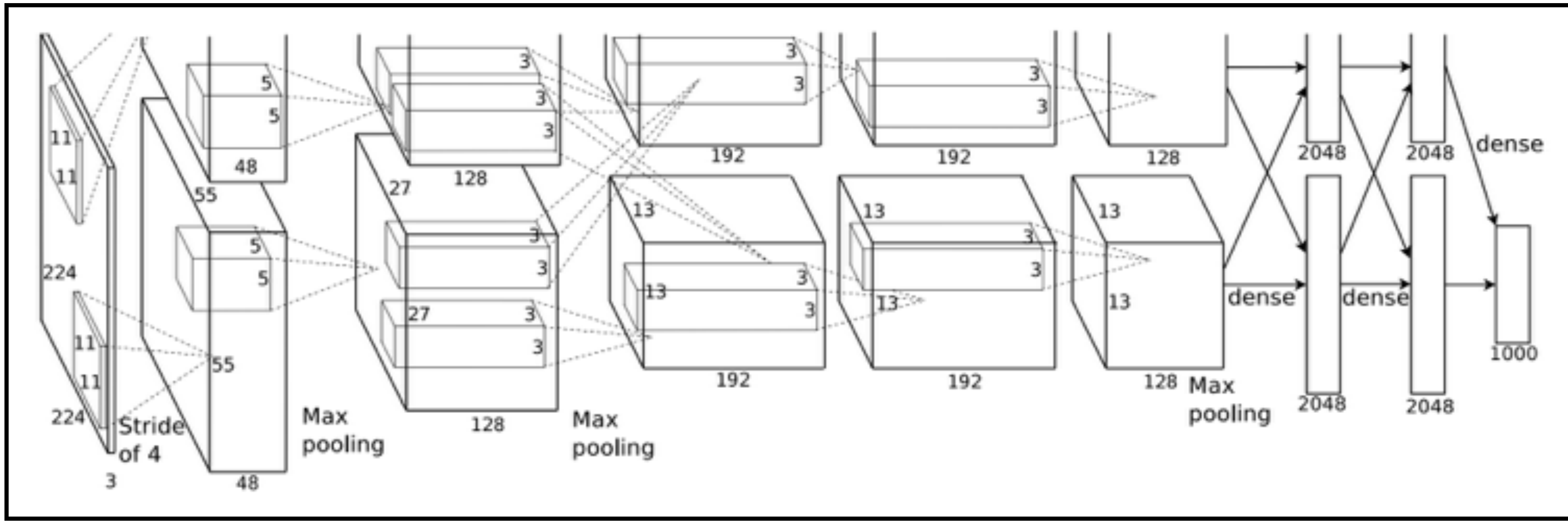
UCF 101 dataset : 13,320 videos, 101 classes

KTH (human action) : 2391 videos, 6 classes

Sports-1M dataset : 1.1 million videos, 487 classes (new!)

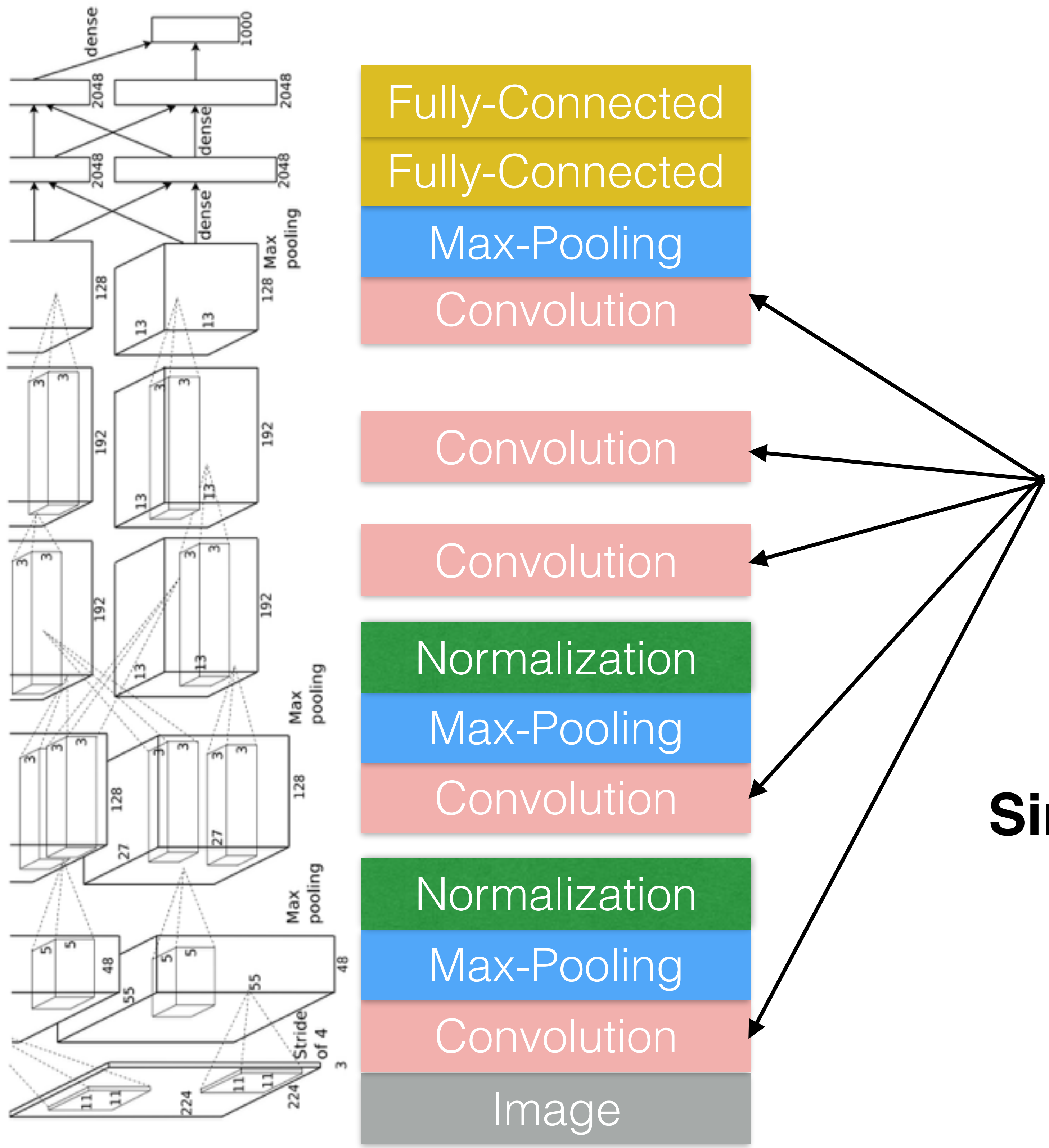
Models

Baseline CNN



Krizhevsky et al. ’12

Baseline CNN



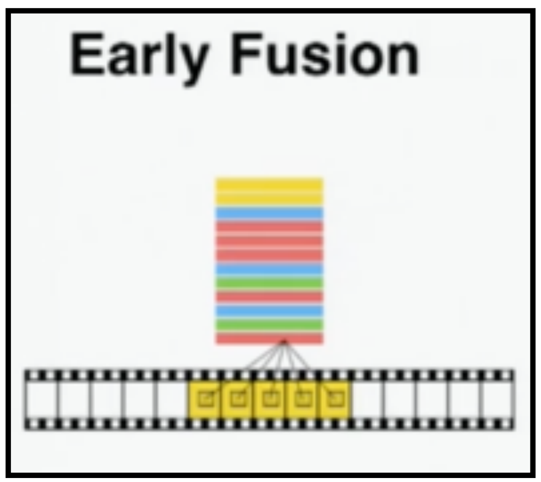
Goal: Extend 2D convolutional layers to 3D to learn spatio-temporal filters

Call this
Single-frame baseline

Temporal Fusion in CNNs

Modify 1st convolutional layer to be of size 11 x 11 x 3 x T pixels

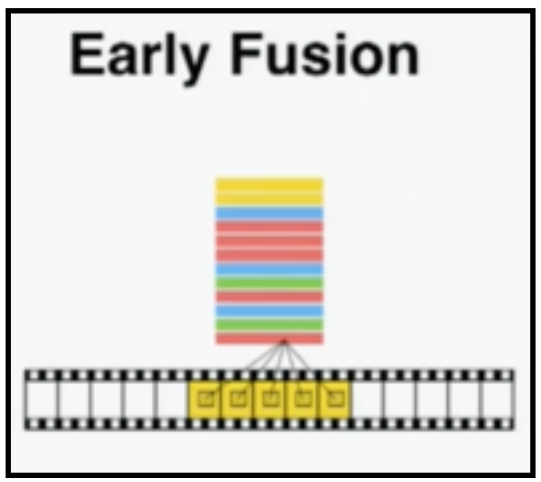
T = # frames (authors use 10)



Temporal Fusion in CNNs

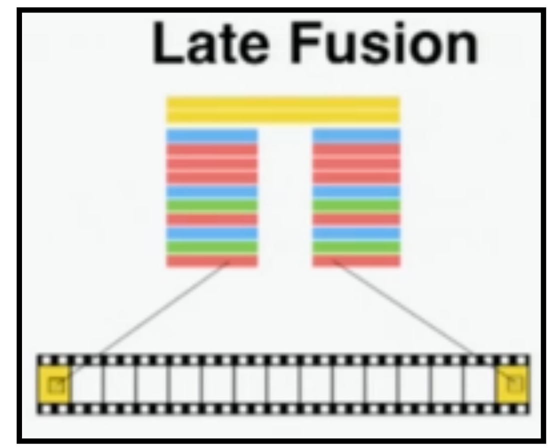
Modify 1st convolutional layer to be of size 11 x 11 x 3 x T pixels

T = # frames (authors use 10)



2 single-frame networks 15 frames apart merge in 1st fully connected layer

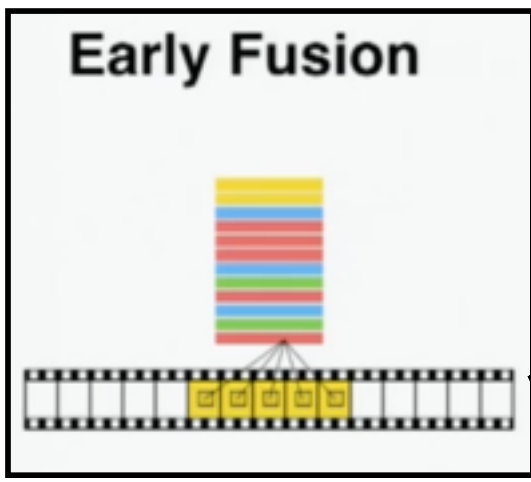
The fully connected layer can compute global motion characteristics



Temporal Fusion in CNNs

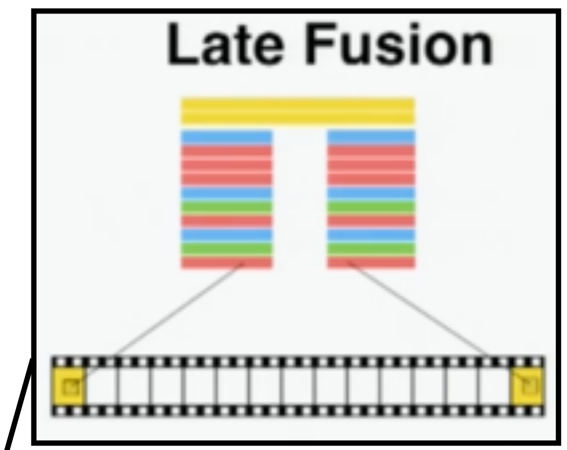
Modify 1st convolutional layer to be of size 11 x 11 x 3 x T pixels

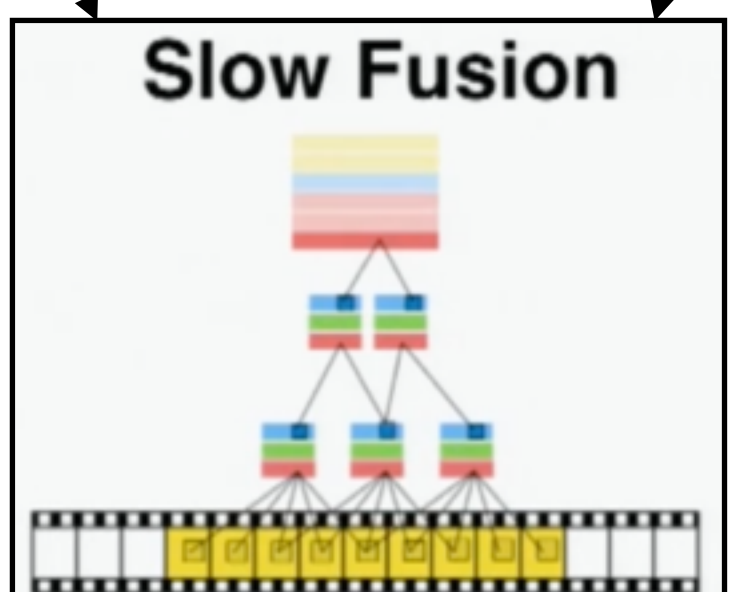
T = # frames (authors use 10)



2 single-frame networks 15 frames apart merge in 1st fully connected layer

The fully connected layer can compute global motion characteristics



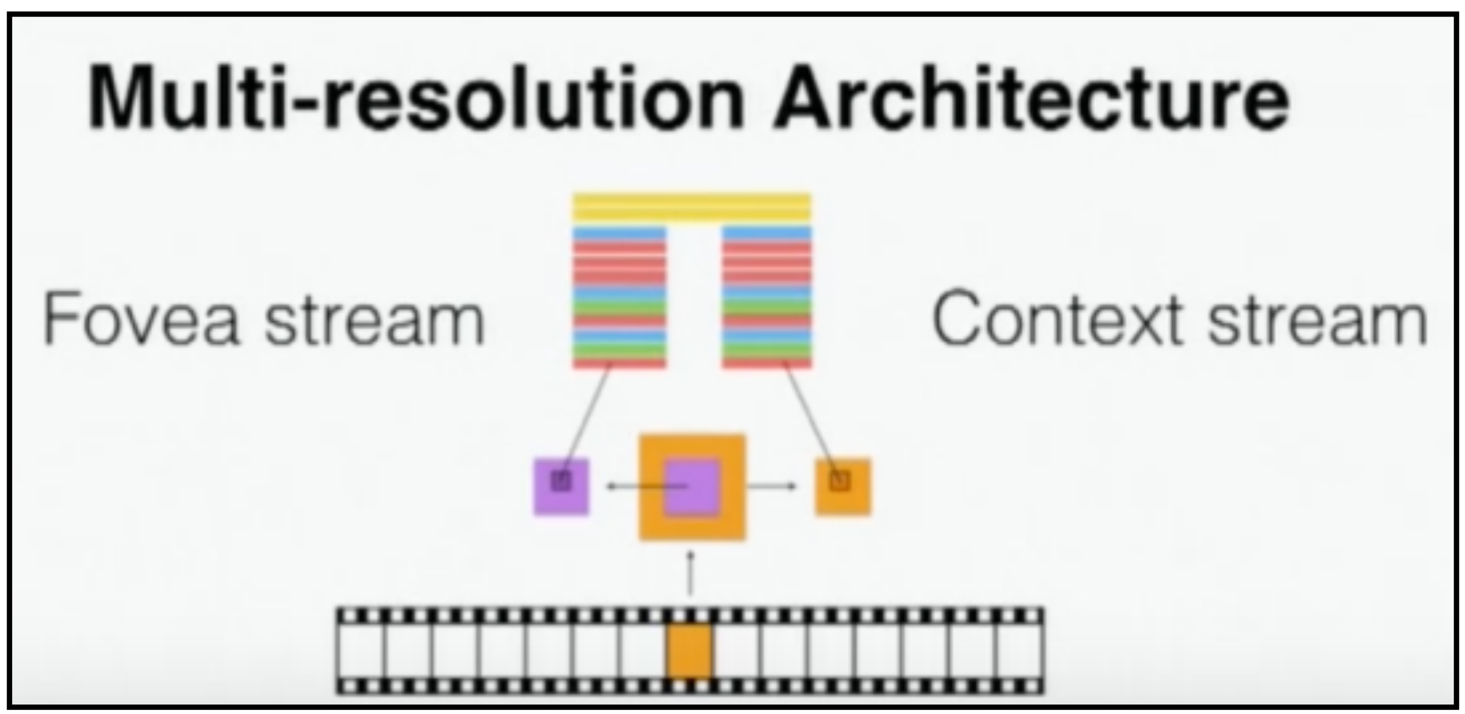


Spatial + temporal convolutions, and higher layers get more global information

Multiresolution CNNs

To improve runtime performance:
Input = 178 x 178 frame video clip

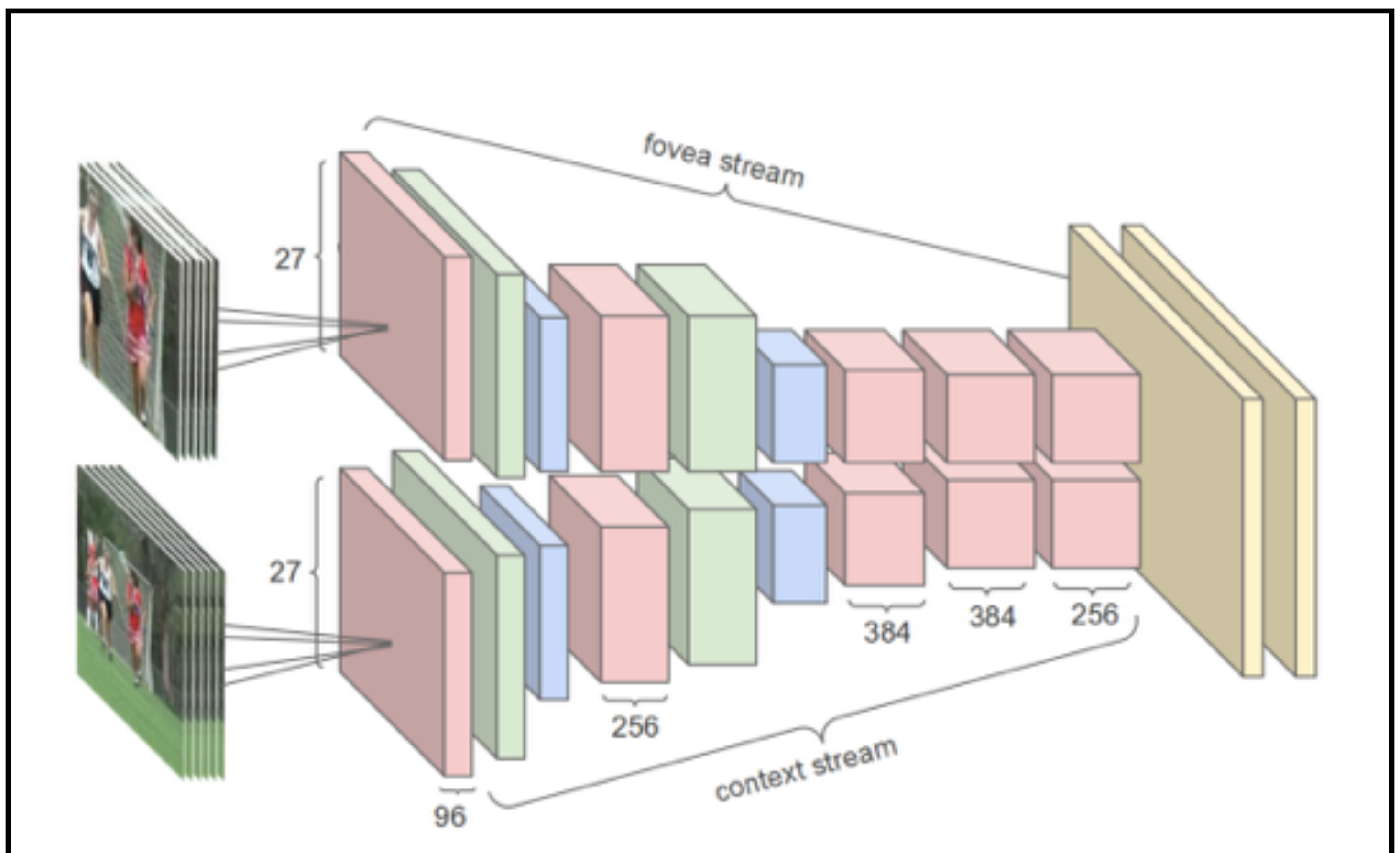
Low-Res Context stream gets down sampled 89 x 89 (entire frame)
High-Res Fovea stream gets cropped center 89 x 89 patch
Both streams merge in 1st fully connected layer



Multiresolution CNNs

To improve runtime performance:
Input = 178 x 178 frame video clip

Low-Res Context stream gets down sampled 89 x 89 (entire frame)
High-Res Fovea stream gets cropped center 89 x 89 patch
Both streams merge in 1st fully connected layer



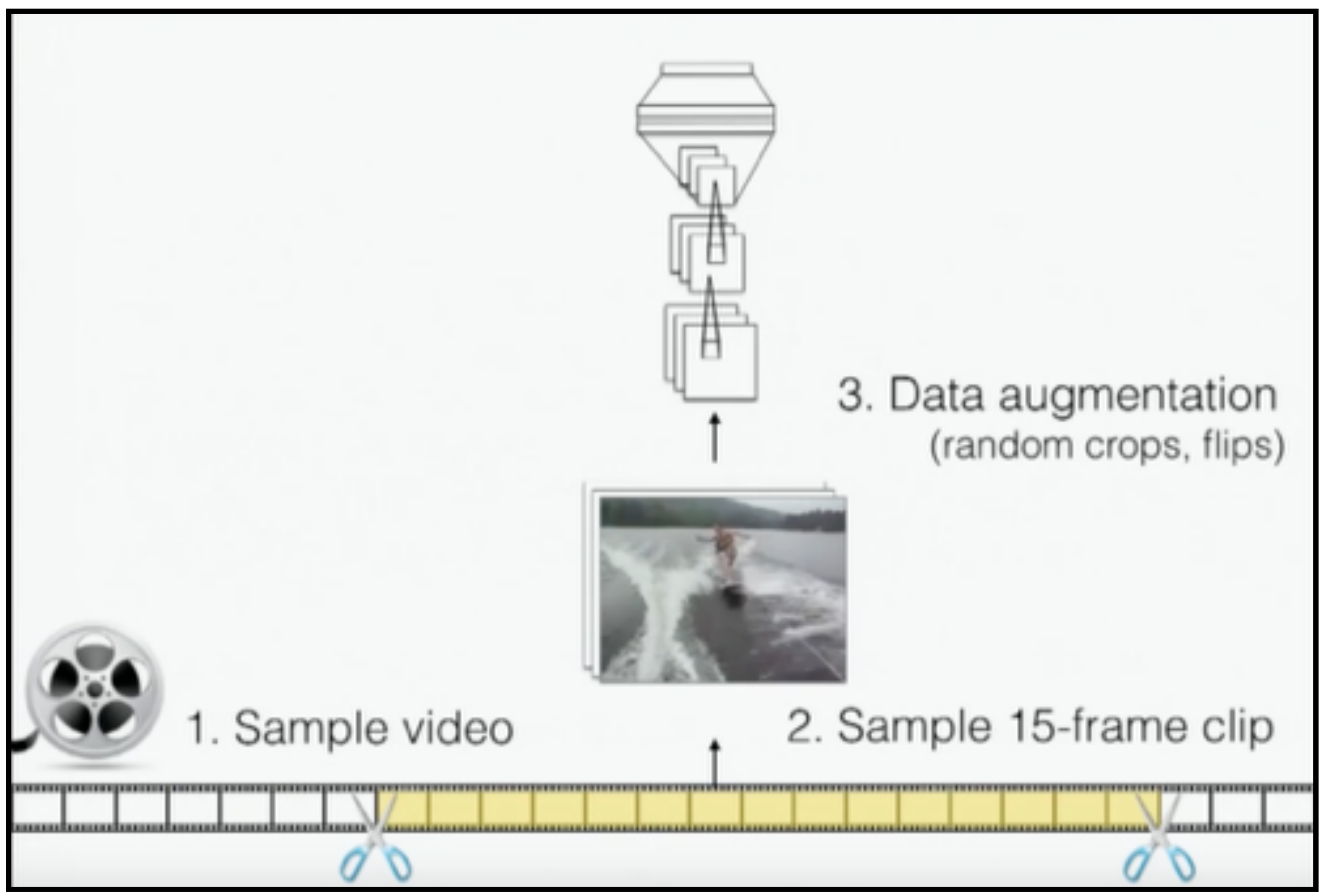
Train Procedure

1. Randomly sample a video

2. Sample a 15 frame (~0.5 secs) clip from (1)

3. Randomly crop, flip frames in clip, subtract mean of all pixels in images (data augmentation + preprocessing)

   Test Procedure is similar



Experiments

Feature Histogram Baseline

1. Extraction of local visual features :
   HoG, Texton, Cuboids, Hue-Saturation, Color moments, #Faces detected

2. Visual word encoding of features:
   Spatial pyramid encoding in histograms after k-means : Finally obtain a 25,000 D feature vector for the entire video

3. Training a classifier:
   Use a 2-hidden layer neural net (worked better than any linear classifier)

Testing Procedure

1. Randomly sample 20 clips for a given test video

2. Present each clip individually to the network (with different crops and flips)

3. Individual clip class predictions are averaged to get a class result for the entire video

Results on Sports-1M dataset

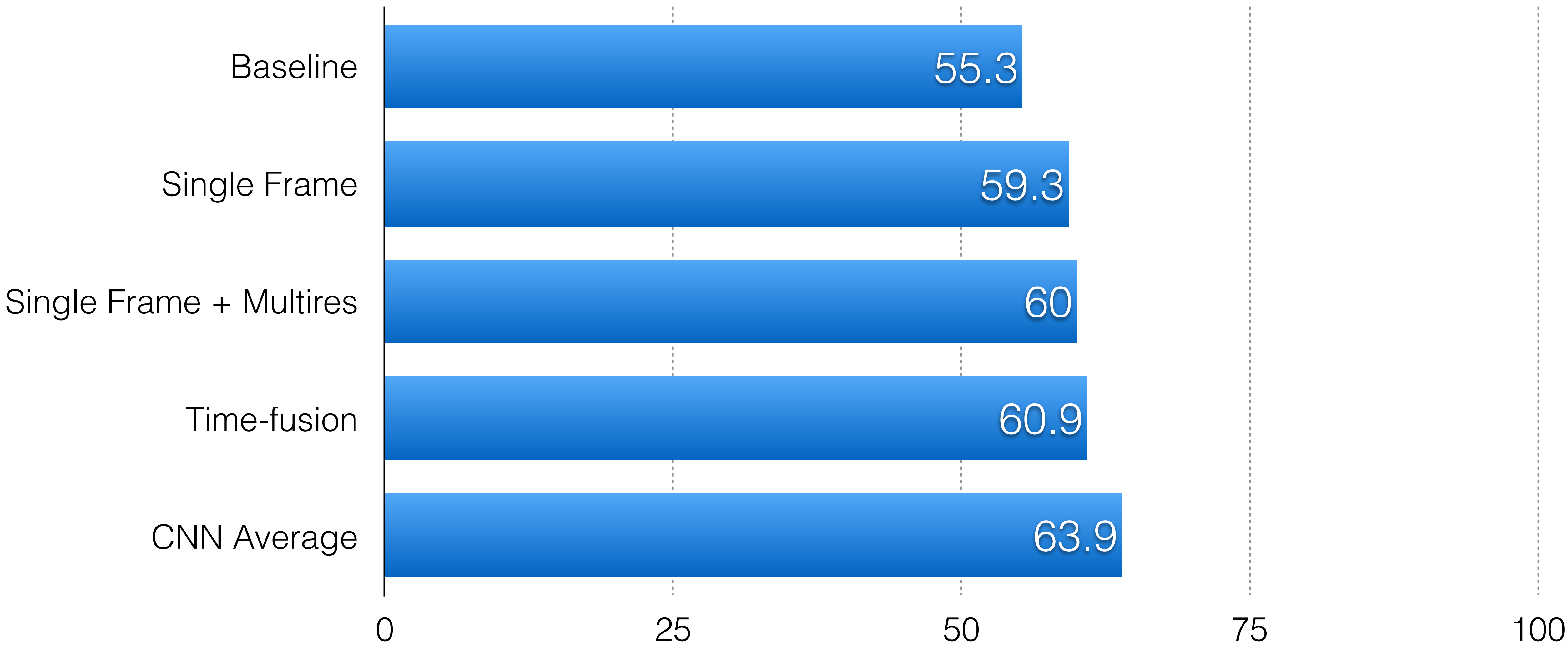
Video Results

https://www.youtube.com/watch?v=qrzQ_AB1DZk

Cycling

Basketball

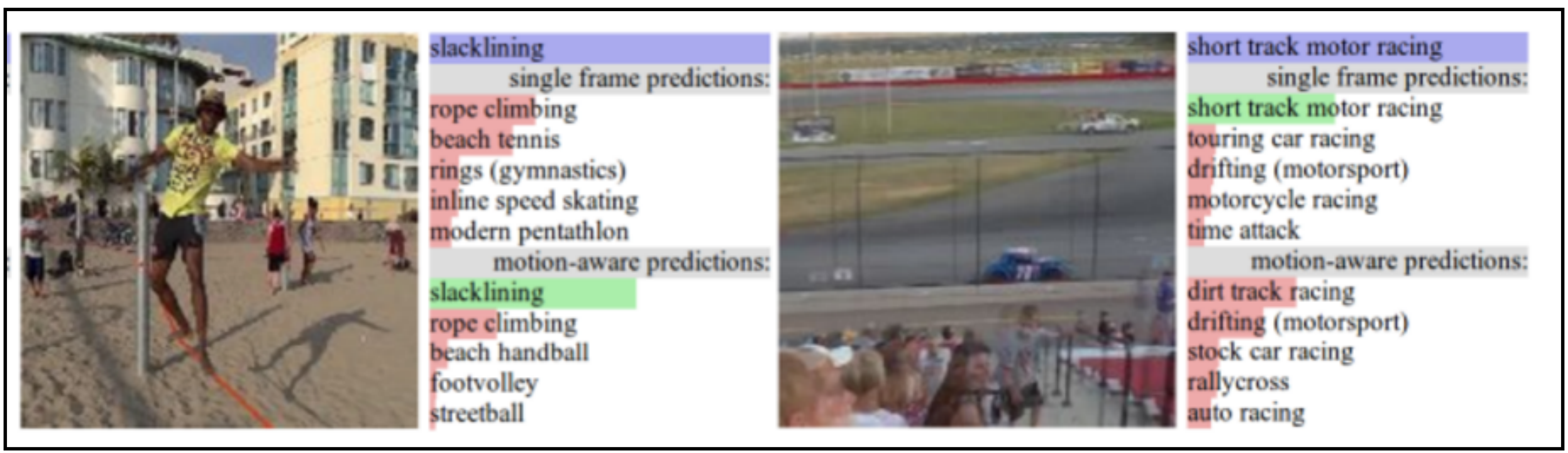
Quantitative Results



Qualitative Results

1. The confusion matrix shows that the network doesn't do well on fine-grained classification

2. Slow-fusion networks are sensitive to small motions, hence "motion-aware", but don't work well with presence of camera translation and zoom

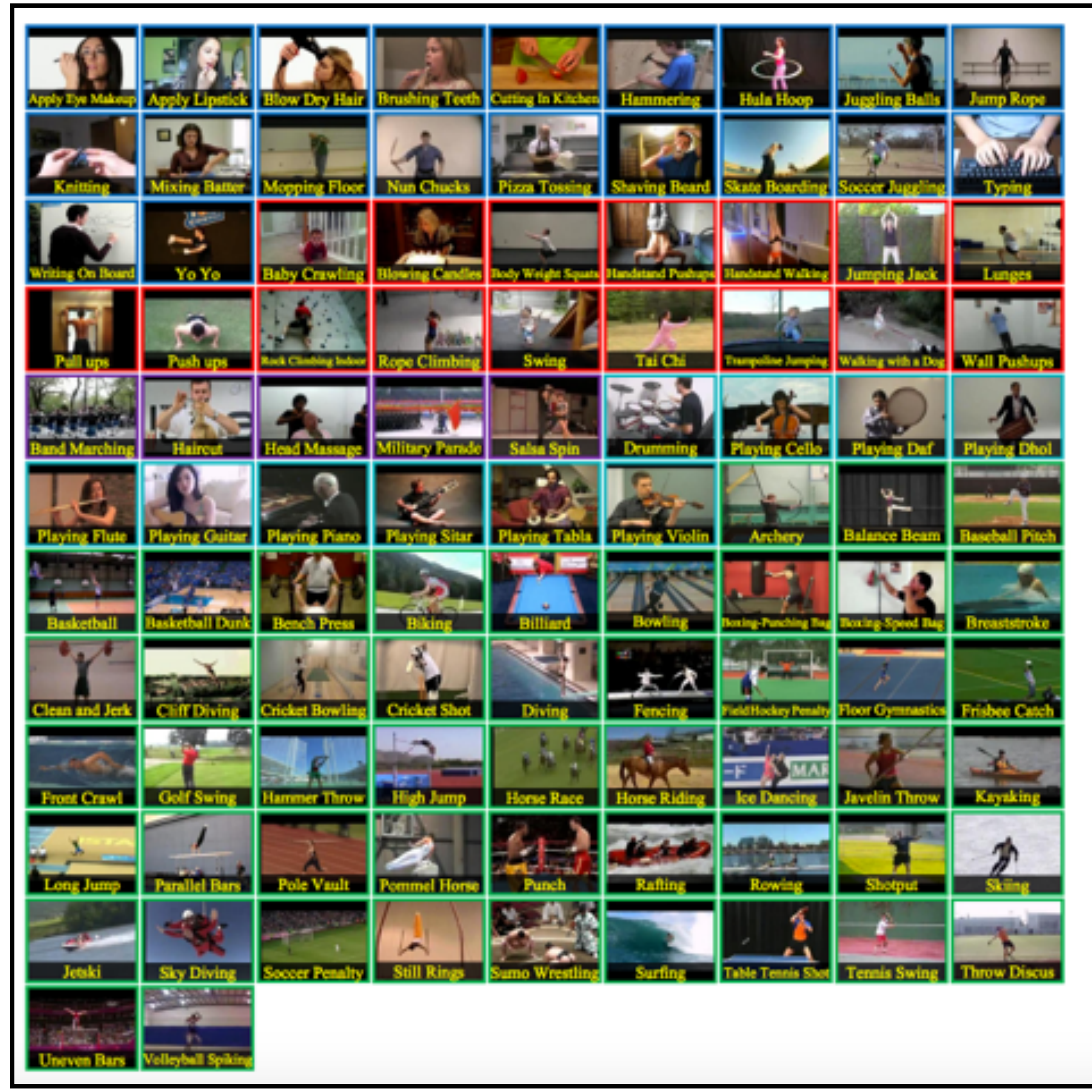


Transfer Learning

UCF-101 dataset

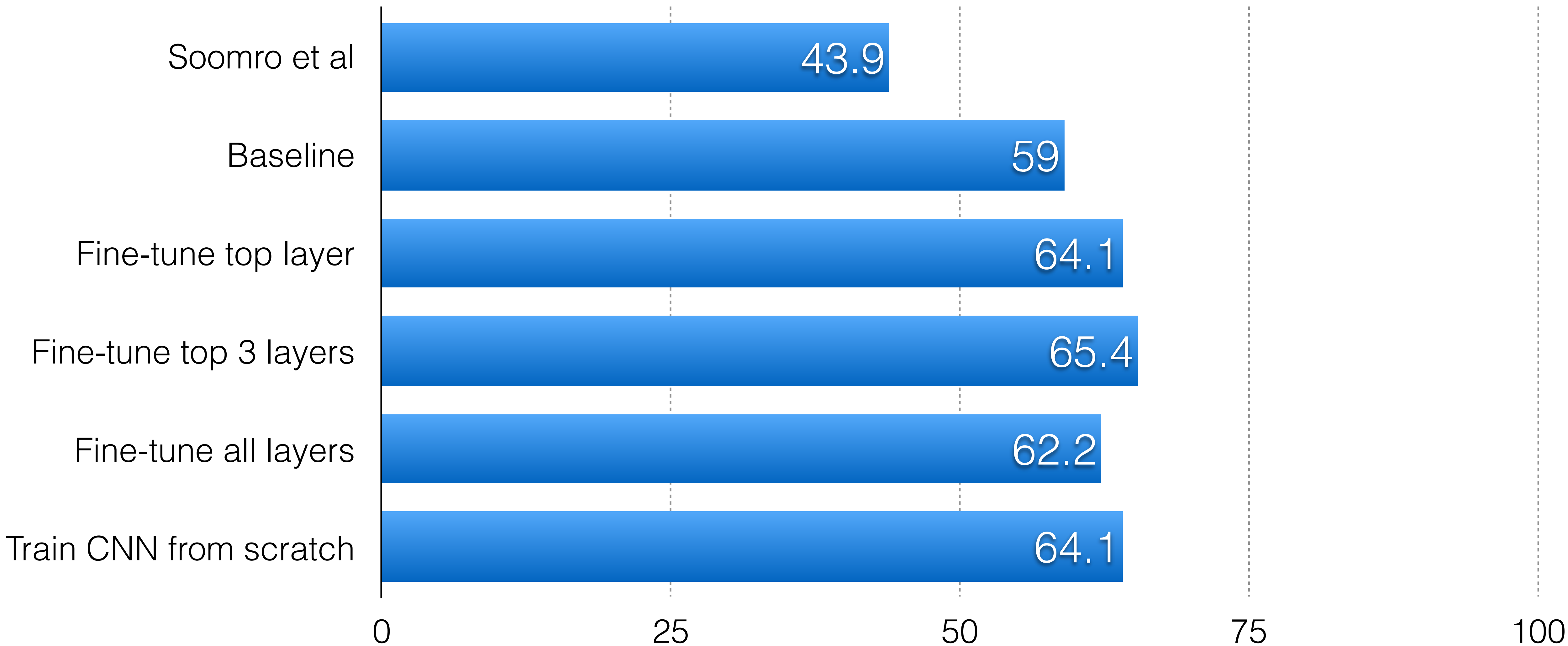
5 main categories of data

1. Human Object Interaction
2. Body-Motion only
3. Human-Human interaction
4. Playing Musical Instruments
5. Sports



Soomro et al. ’12

Transfer learning to Sports data in UCF 101 Results



Discussion