CS698U: Topics in Computer Vision

Jan—May 2017

Lecture 1

Gaurav Sharma

Indian Institute of Technology Kanpur

www.grvsharma.com

Welcome to CS698U

- CS698U — Topics in Computer Vision
- Instructor: Gaurav Sharma, CSE, IITK
  - grv@cse.iitk.ac.in **[CS698U] in email subject**
- Lectures: Mon, Wed @ 17:00 — 18:30, RM101
- Office hours: TBD, send mail for appointment
- Webpage: www.grvsharma.com/tcv_y1617s2.html

Grading (tentative)

- 40% Assignments
- 20% Quizzes (to be announced a day before)
- 15% Mid-sem exam
- 25% End-sem exam

- (Class project by invitation)

Course goals

- Study, understand and implement
  - Learning based methods
  - Geometry based methods
  - … for Computer Vision
- Hopefully, be excited to work in this area beyond this course

Tentative Topics 1/2

- Learning for Computer Vision
  - Convolutional Neural Networks
  - Recurrent Neural Networks
  - AutoEncoders
  - Generative Adversarial Networks
  - … (any other requests, if time permits ?)

Tentative Topics 2/2

- Geometry for Computer Vision
  - Pin-hole camera model
    - Projection matrix
    - Calibration etc
  - Two-view geometry
    - Epipolar geometry
    - Mosaics
    - 3D reconstruction

Introduction 1

Learning based CV

Disclaimer: The images used have been sourced from the internet and are used for no-profit. If you own any of the images and would like to have them removed, please let me know.

Image Classification

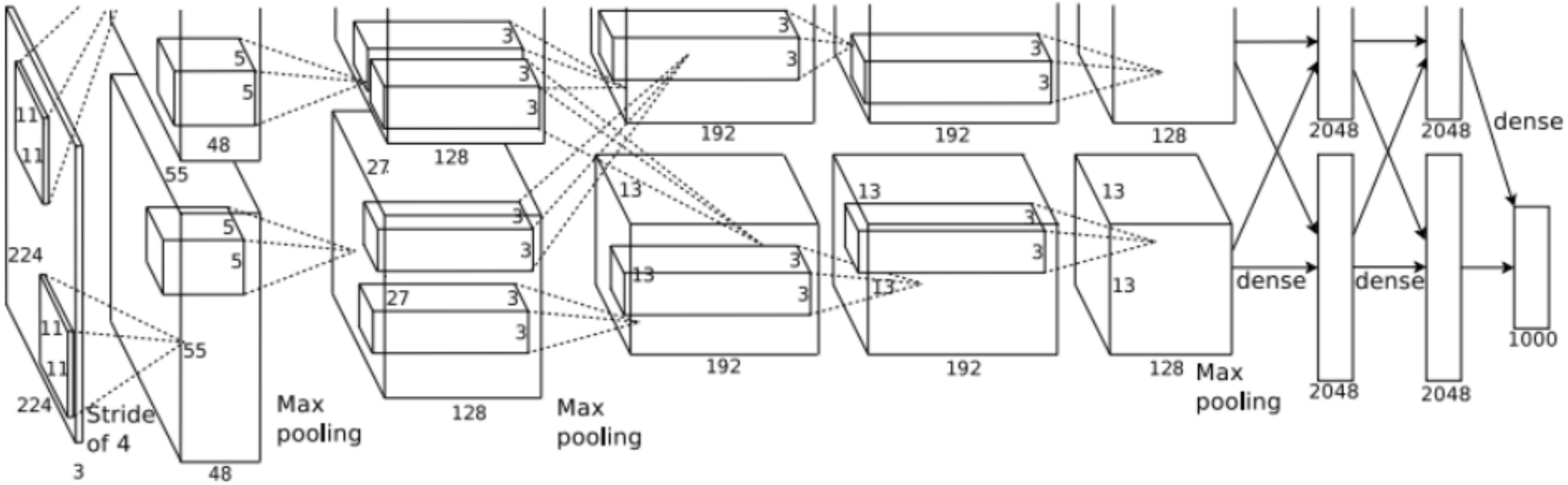
Challenges

- Pose
- Illumination
- Intra class variability
- Inter class similarity
- …





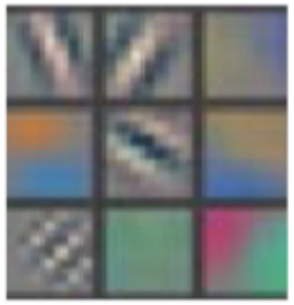

AlexNet — Avalanche



input	conv1	conv2	conv3	conv4	conv5	fc6	fc7	fc8
	11x11x3	5x5x48	3x3x128	3x3x192	3x3x192		Filter sizes	
	48x2	128x2	192x2	192x2	128x2		Num filters	

Krizhevsky et al., ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012

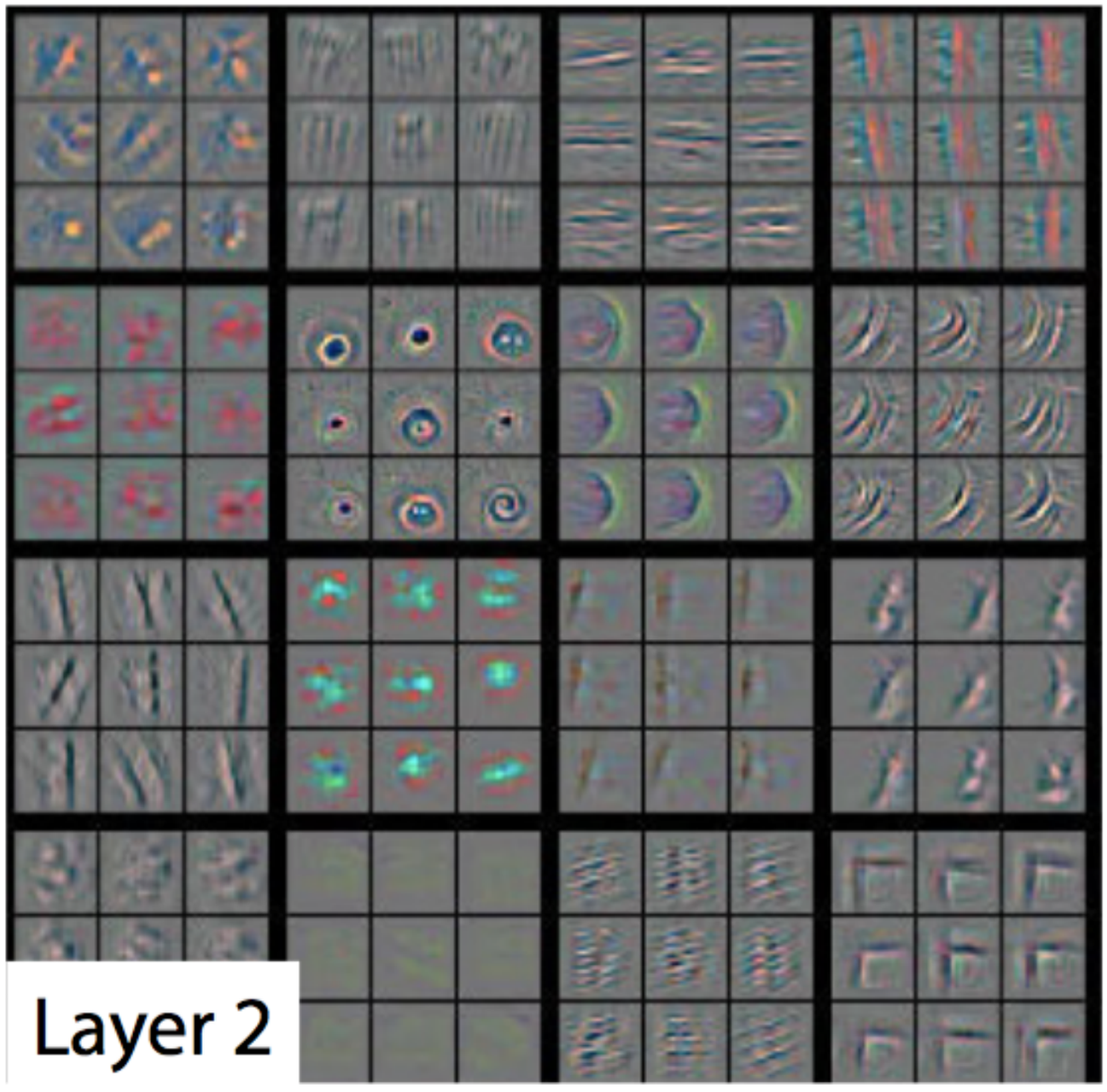
Visualization of CNN Layers

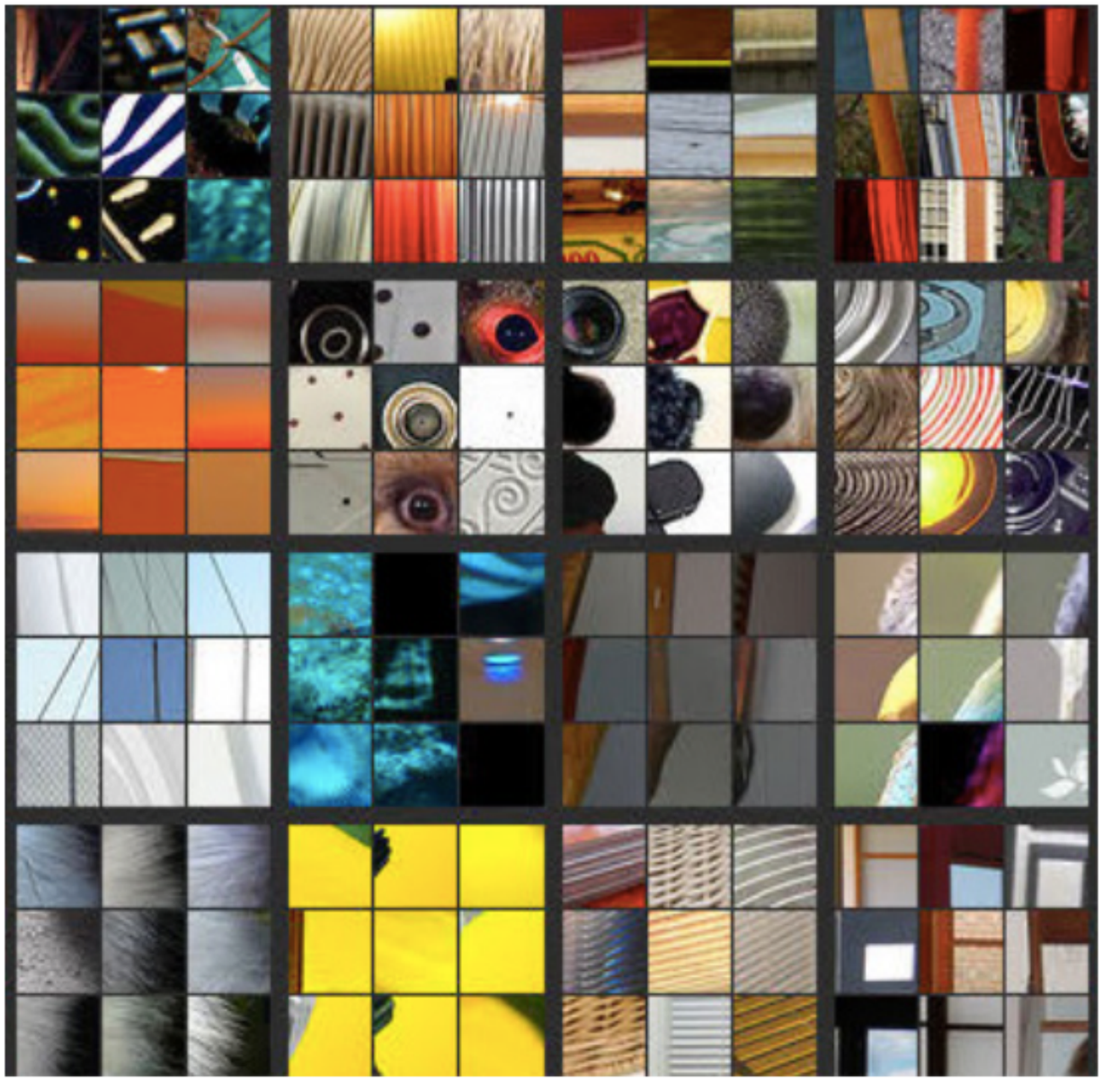


Layer 1

Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, ECCV 2014

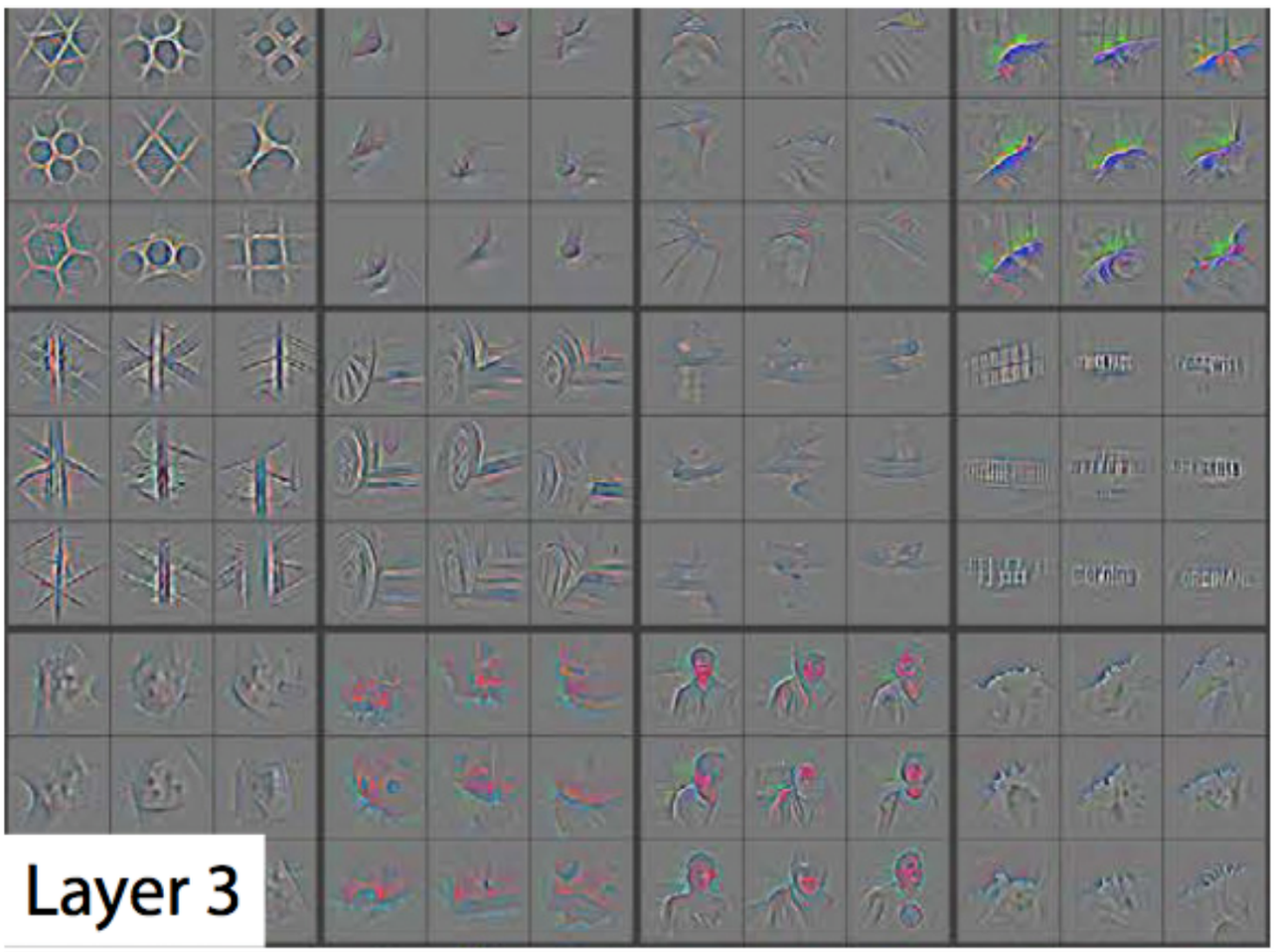
Visualization of CNN Layers





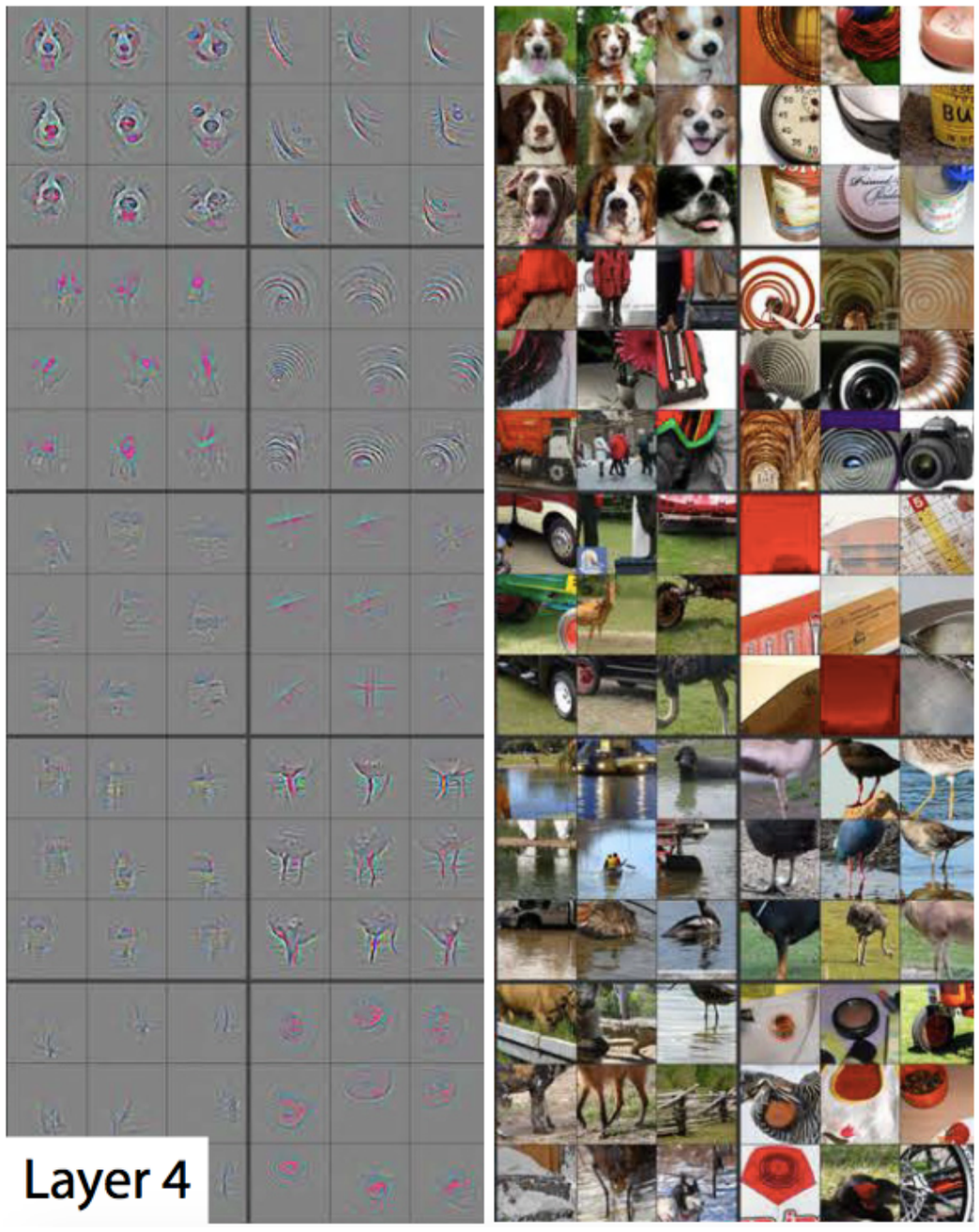
Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, ECCV 2014

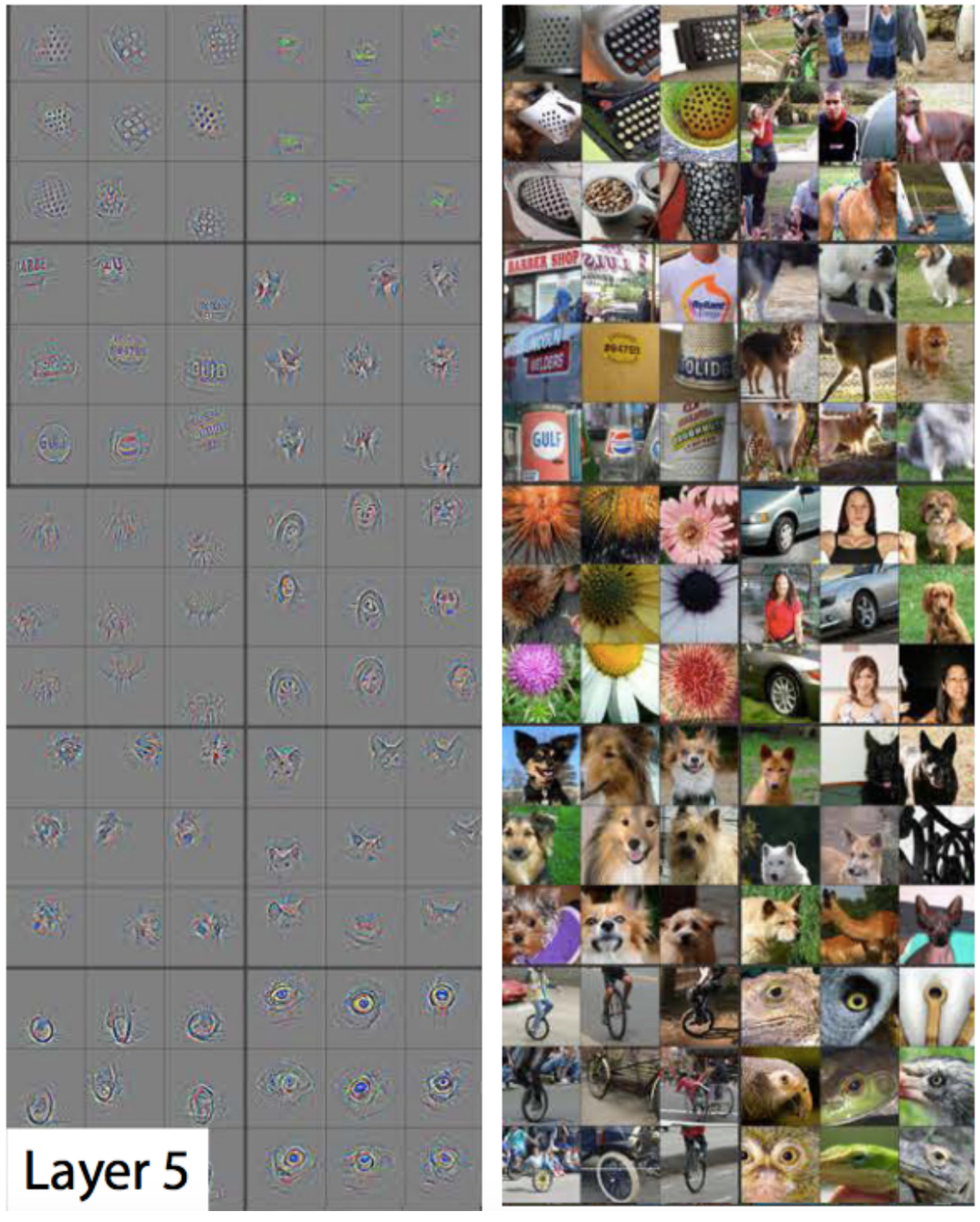
Visualization of CNN Layers



Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, ECCV 2014

Visualization of CNN Layers





Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, ECCV 2014

VGG nets

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**

Cf. 8 Layer AlexNet

Common

maxpool
FC-4096
FC-4096
FC-1000
soft-max

GoogLeNet

22 Layers

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Residual Networks (ResNets)

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

He et al., Deep residual learning for image recognition, CVPR 2016

Significant Progress

method	top-5 err. (**test**)
VGG [41] (ILSVRC'14)	7.32
GoogLeNet [44] (ILSVRC'14)	6.66
VGG [41] (v5)	6.8
PReLU-net [13]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	**3.57**

With "big data" and better methods, close to human performance (on this 'narrow' task)

Error rates just before were ~25%

He et al., Deep residual learning for image recognition, CVPR 2016

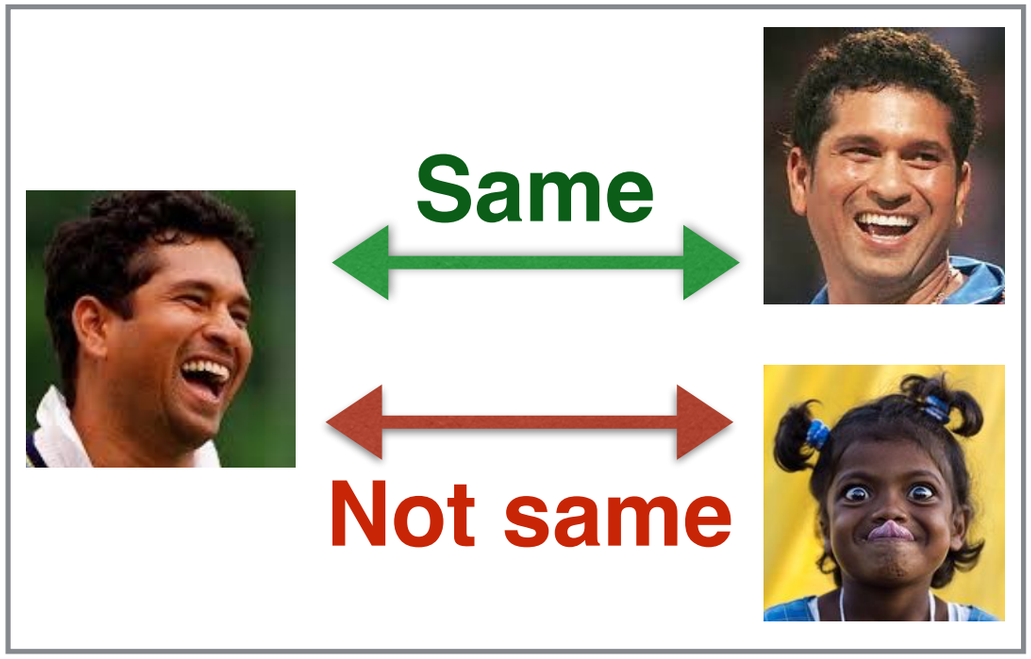
Faces

Faces — Identity

Efficient retrieval in large (~1 billion) face datasets

Face verification

- Given two faces say same or not-same



vs.

- Contrast with **recognition** i.e. store and predict identity

Applications

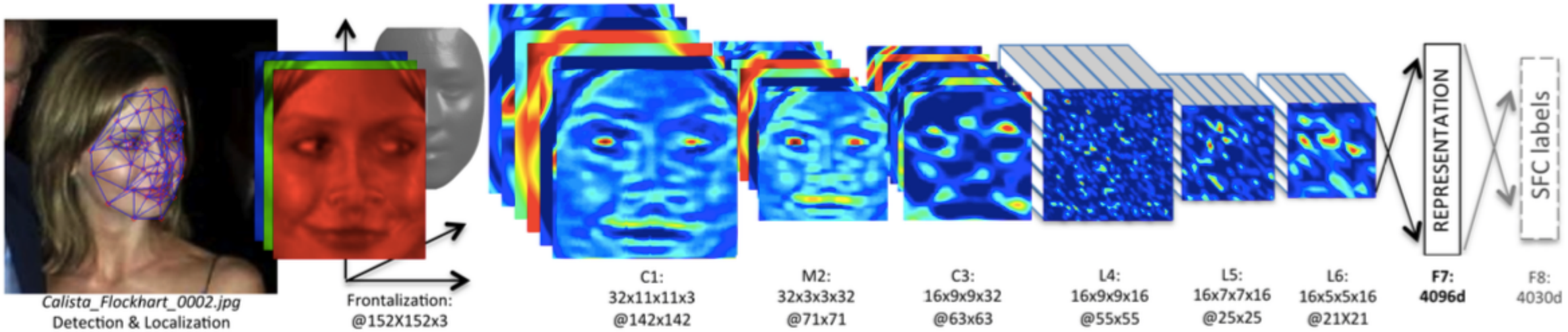
- Indexing large (human centered) video collections
  - Retrieval — personal data, journalism, surveillance
  - Clustering with learnt similarity
  - Visualization, compression etc
- Access control — current person allowed or not ?
  - Maintains privacy — no person specific id kept

Challenges at Large Scale

- Large amount of data
- World scale = ~6 billion people

- Typical face features $O(10^4)$ real numbers
- $\Rightarrow O(10^9)$ faces = $O(10^{13})$ reals $\approx$ 40 Tb

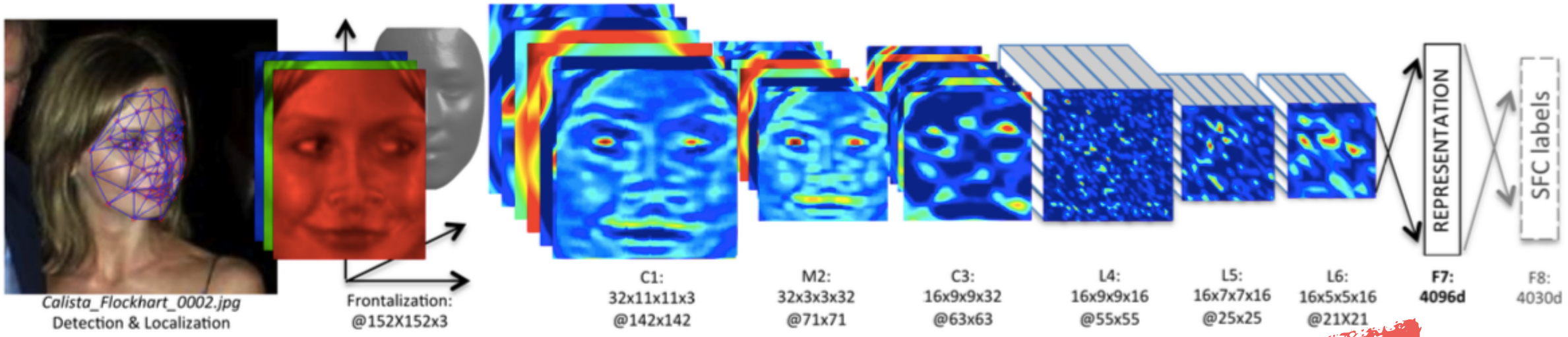
Nonlinear Embeddings



- Deep CNNs as an embedding
- Learnt using Social Face Classification dataset (Facebook proprietary) — 4.4M images, 4k identities

Taigman et al., DeepFace: Closing the Gap to Human-Level Performance in Face Verification, CVPR 2014

Nonlinear Embeddings



- Once learned
  - Chop off final classification layer
  - Use last FC layer as features (with normalization)

Taigman et al., DeepFace: Closing the Gap to Human-Level Performance in Face Verification, CVPR 2014

Siamese Network

CNN

CNN

Same

Abs. diff of embeddings

Single Logistic Unit

End to end learning with back propagation

S. Chopra et al., Learning a similarity metric discriminatively, with application to face verification, CVPR 2005

Labeled Faces in the Wild

- LFW dataset (2007)
- 13k Images of faces of 4k people crawled from web
- Train and test pairs — same and not-same

*Image: https://lrs.icg.tugraz.at/research/kissme/images/datasets/lfw_sample_pairs.jpg

VGG Face

No.	Method	Images	Networks	Acc.
1	Fisher Vector Faces [21]	-	-	93.10
2	DeepFace [29]	4M	3	97.35
3	Fusion [30]	500M	5	98.37
4	DeepID-2,3		200	99.47
5	FaceNet [17]	200M	1	98.87
6	FaceNet [17] + Alignment	200M	1	99.63
7	Ours	2.6M	1	98.95

No big-data

Web-scale data

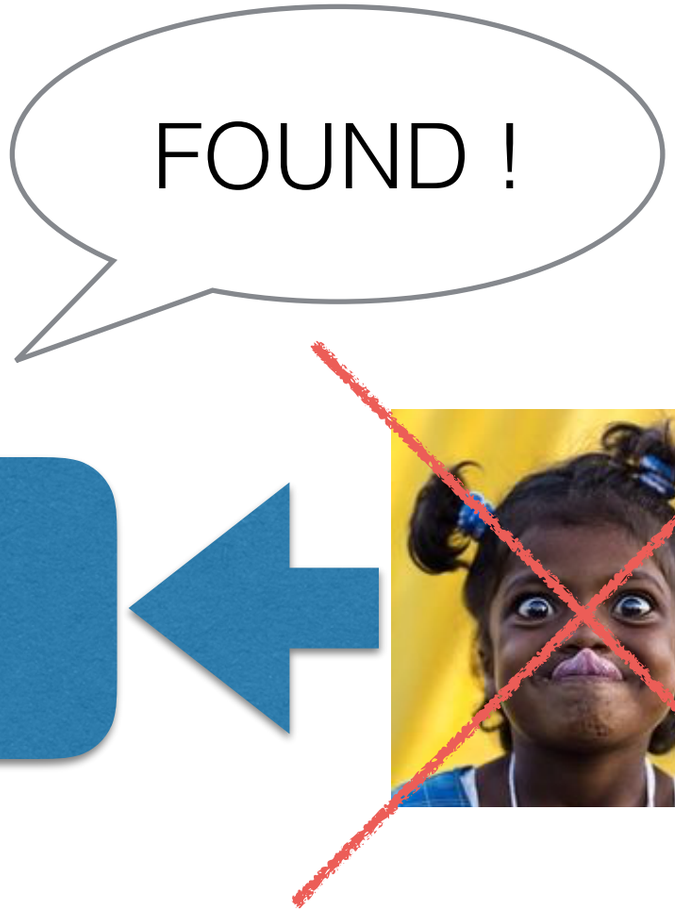
Adapted Objective

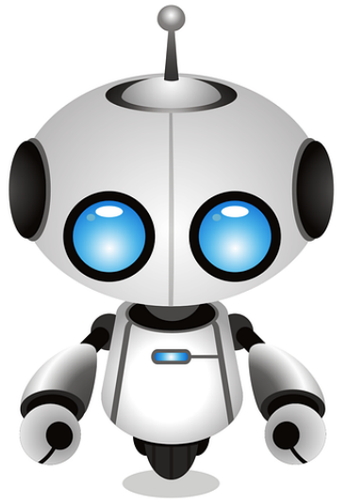
Comparable results on LFW dataset

*Ref. numbers as in Parkhi et al. BMVC 2015

WIN FOR COMPUTER VISION ?

Kill or Spare

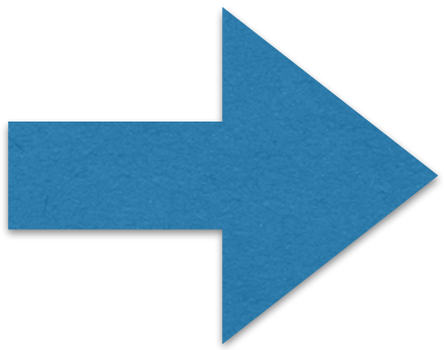




Automatic threat neutralization

In railway stations, airports, movies …

Compression Loss



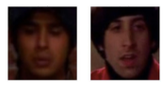
100 x 100 x 8 = 8e4

64 x 32 = 2048

40X compression

Same or not same

1



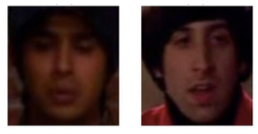
2

3

4

Same or not same

1



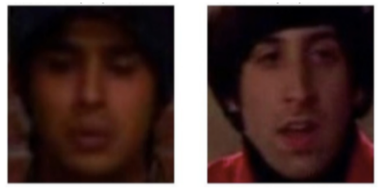
2

3

4

Same or not same

1



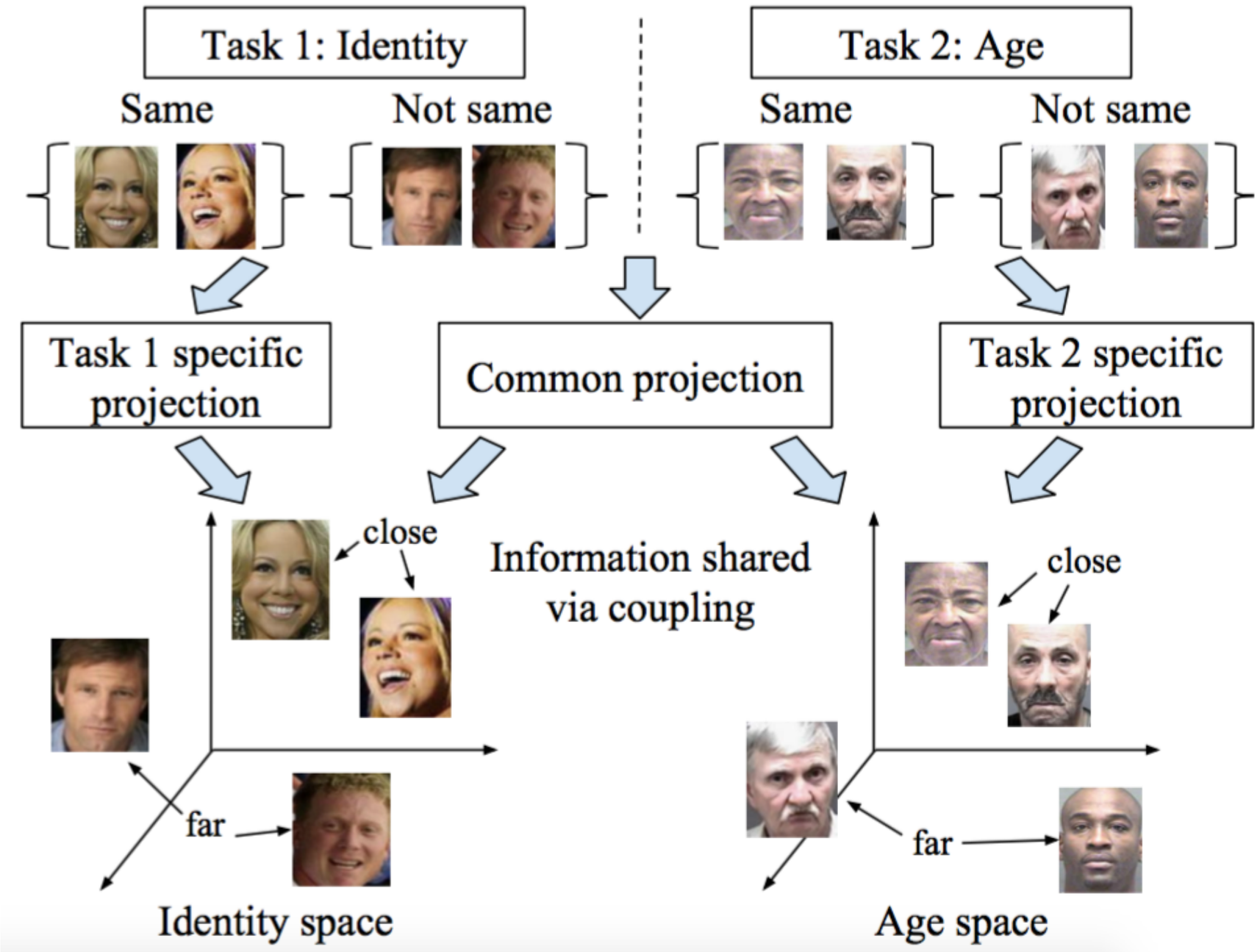
2

3

4

Open problem — Large Scale

Identity, Age, Expressions…



Bhattarai et al., CPmtML: Coupled Projection multi-task Metric Learning for Large Scale Face Retrieval, CVPR 2016

‘Large’ scale

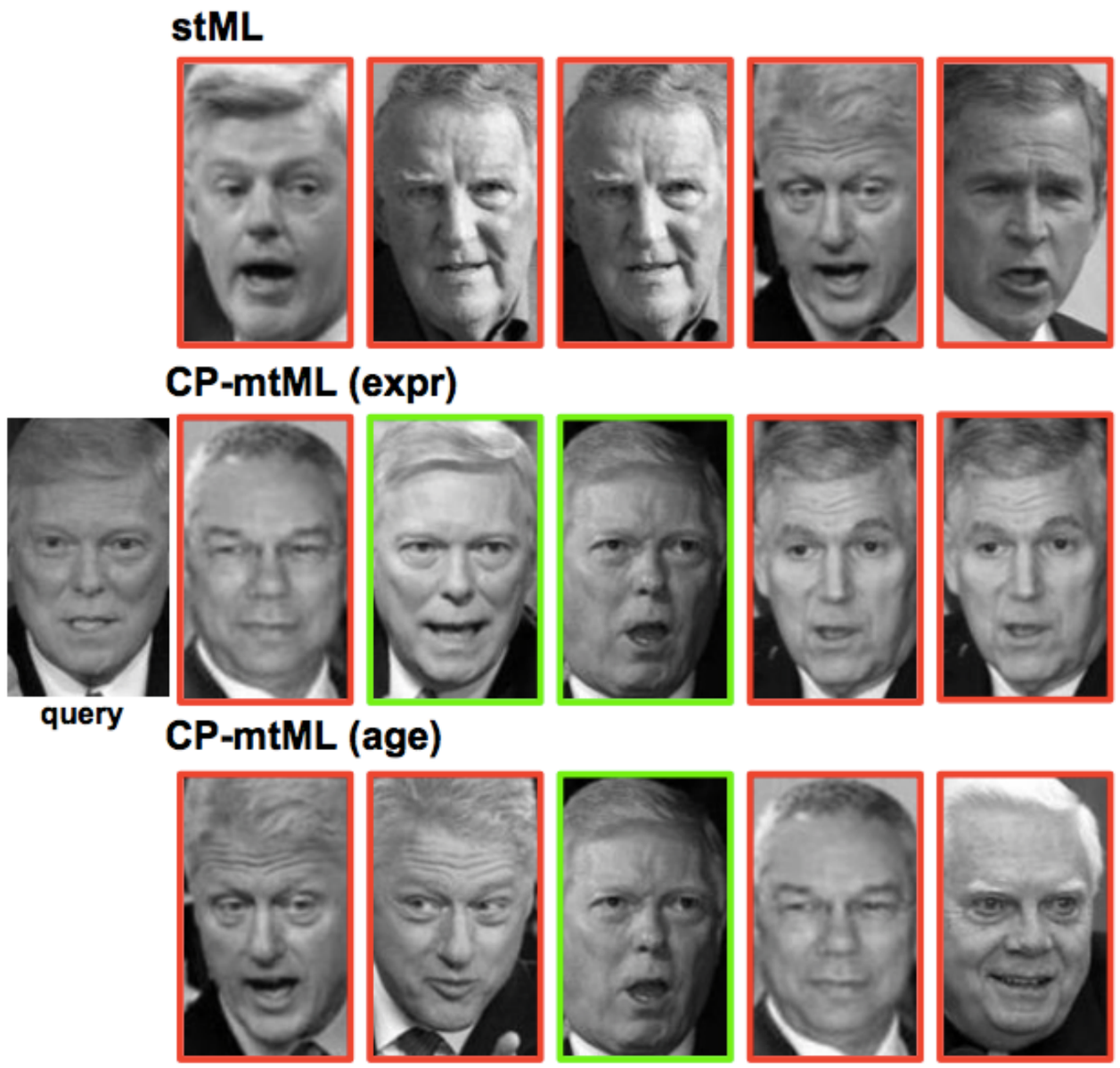
Million face images as distractors

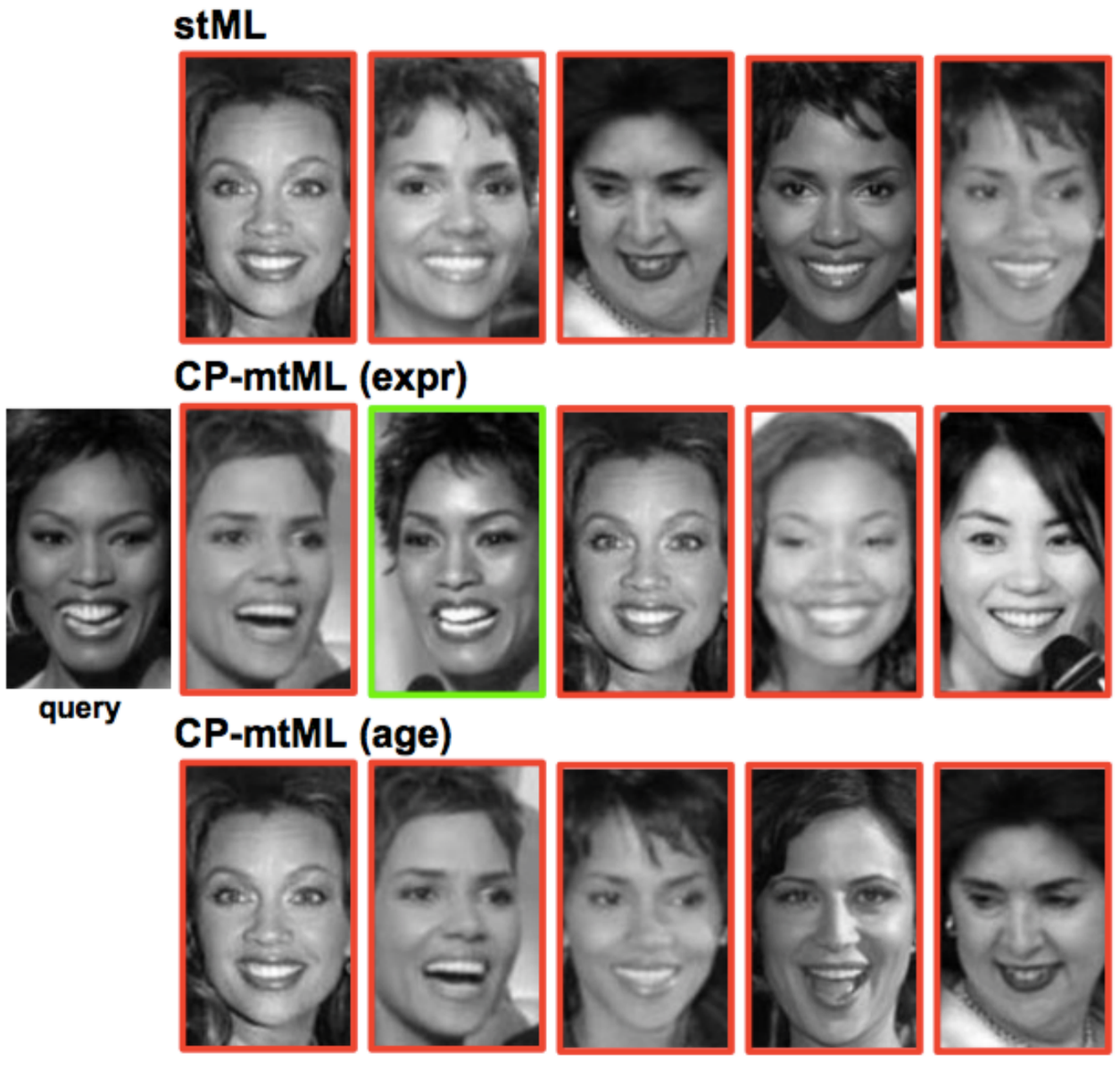
		No distractors				1M distractors			
Method	Aux	$K = 2$	5	10	20	$K = 2$	5	10	20
WPCA	n/a	72.1	80.4	83.7	89.1	65.2	72.1	75.9	78.7
stML	n/a	**76.8**	85.1	89.6	92.0	70.7	78.0	82.0	84.2
utML	expr	73.5	82.3	87.2	90.3	67.1	76.8	79.0	82.0
CP-mtML	expr	**76.8**	**86.5**	**90.3**	**93.4**	**71.2**	**79.7**	**83.2**	**85.3**
utML	age	73.0	82.0	88.2	91.0	68.1	76.1	81.1	82.7
CP-mtML	age	**76.8**	**85.8**	**90.3**	**93.6**	**71.2**	**79.0**	**83.0**	**85.1**

Accuracy when at least one of top K correct

Bhattarai et al., CPmtML: Coupled Projection multi-task Metric Learning for Large Scale Face Retrieval, CVPR 2016

Large scale





Bhattarai et al., CPmtML: Coupled Projection multi-task Metric Learning for Large Scale Face Retrieval, CVPR 2016

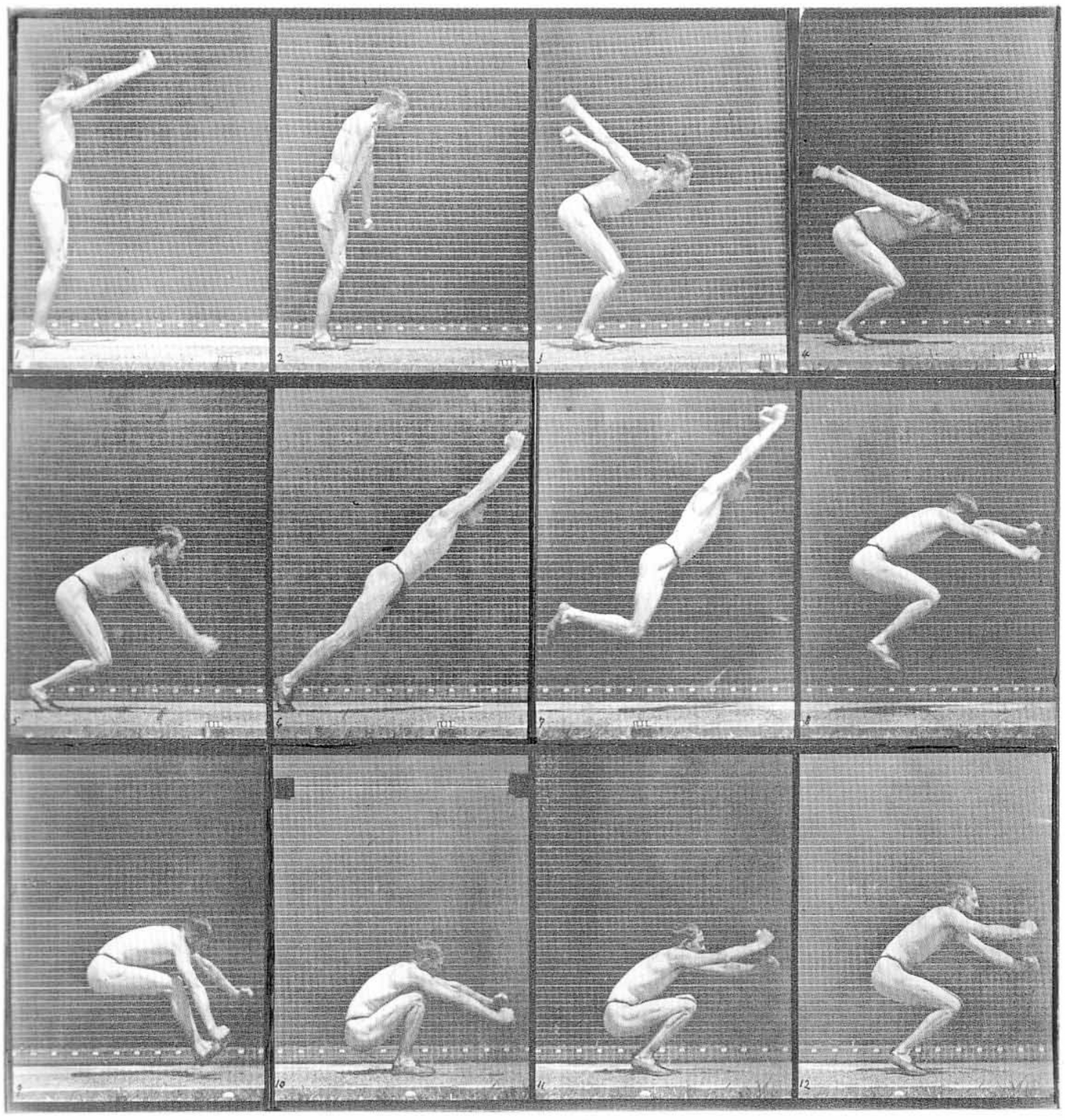


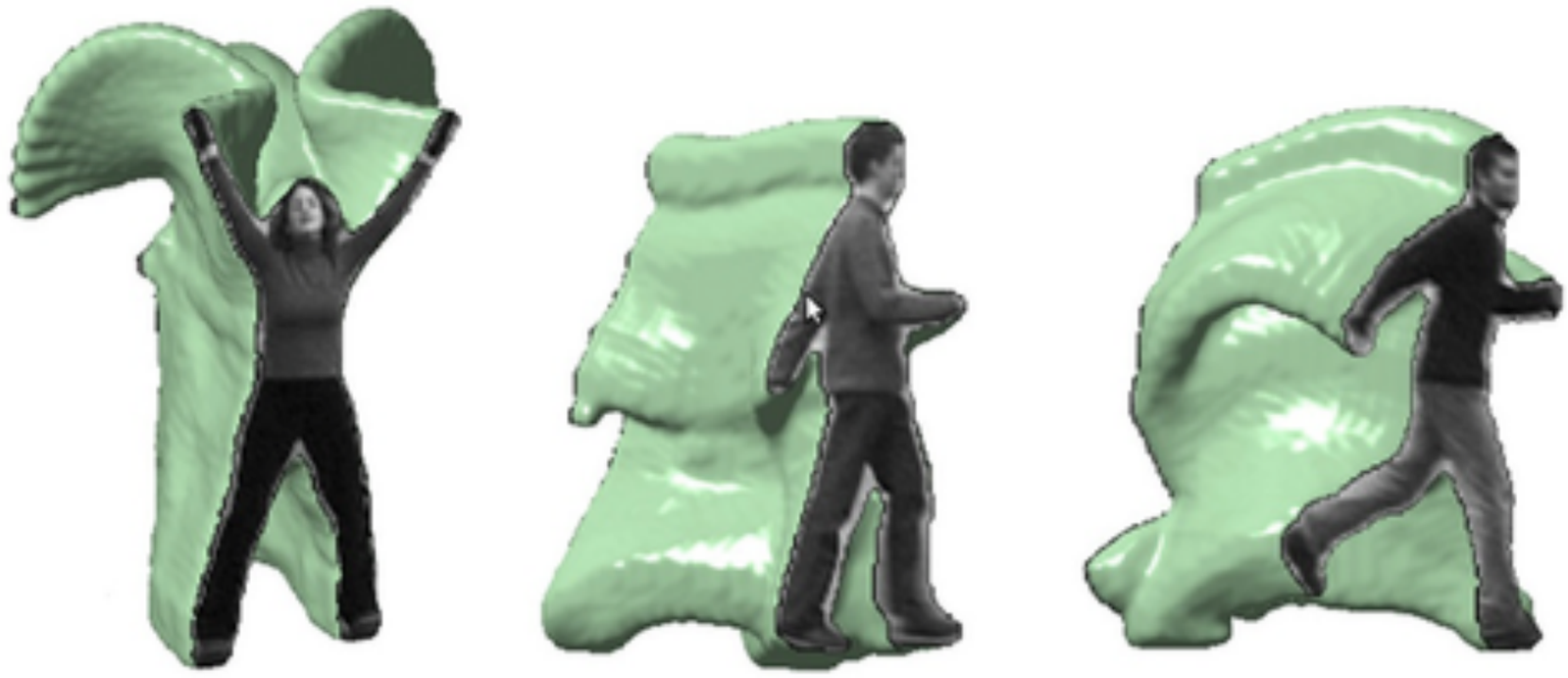
*Image from The Animator’s Survival Kit

Human Actions

Actions in videos

KTH dataset (2004)

Actions in videos



- Actions = space-time shapes (2005)

Blank et al., Actions as space-time shapes. ICCV, 2005

Actions in videos

UCF Sports (2008)

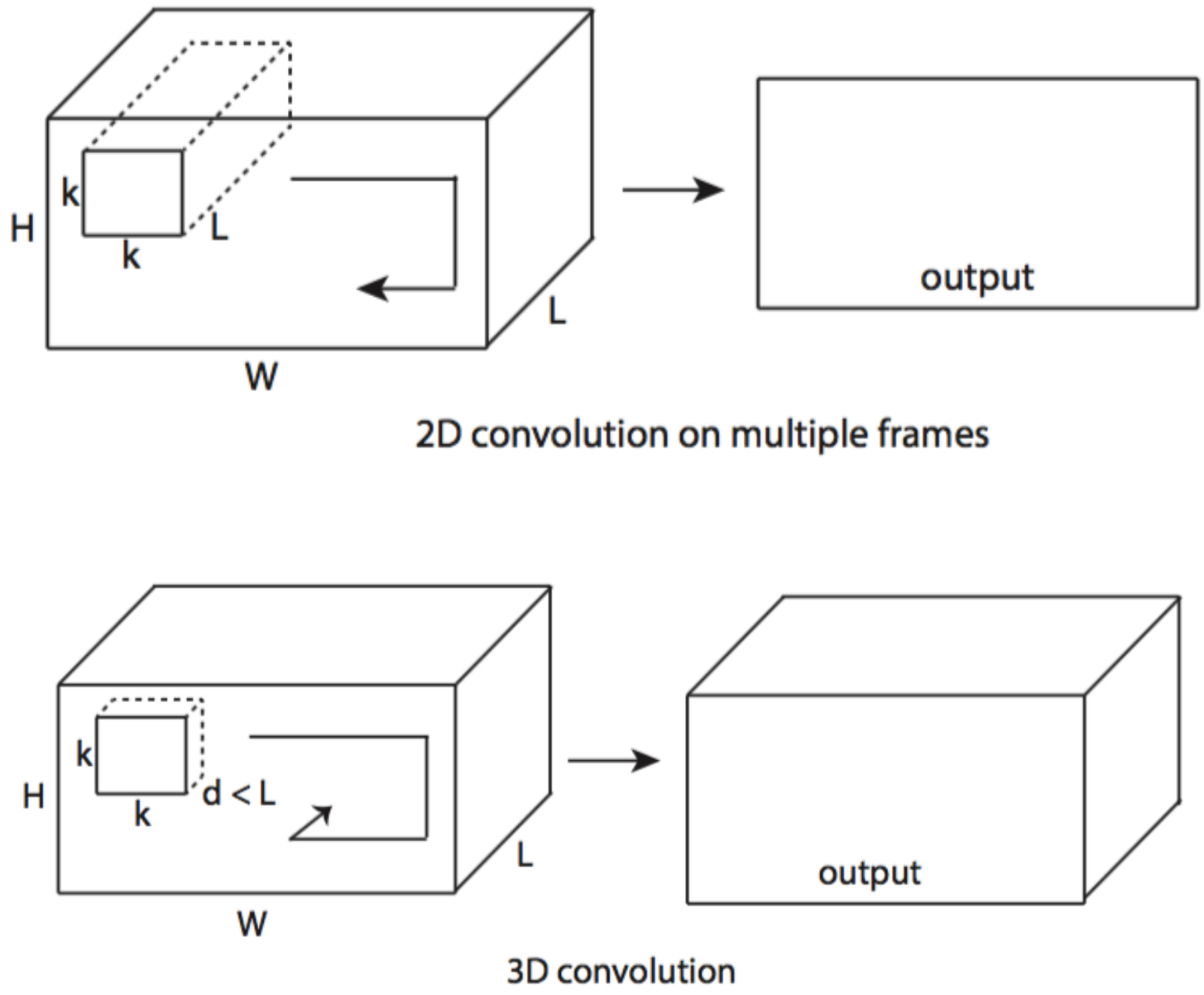
Actions in videos

Youtube Actions (2009)

Actions in videos

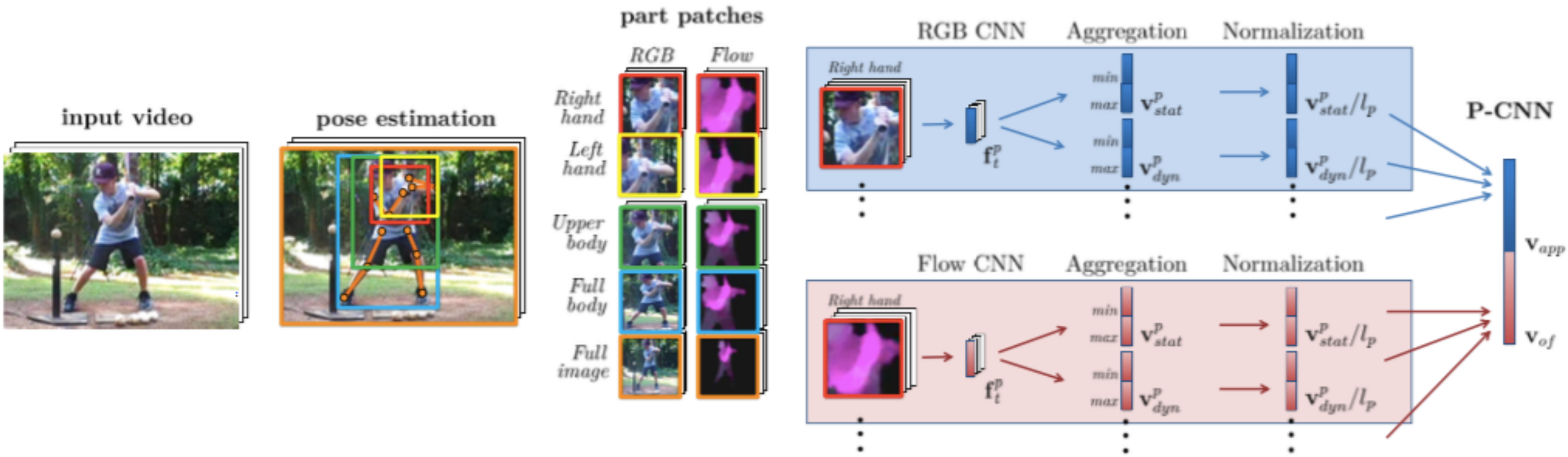
Hollywood (2009)

3D Convnets



Tran et al., Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

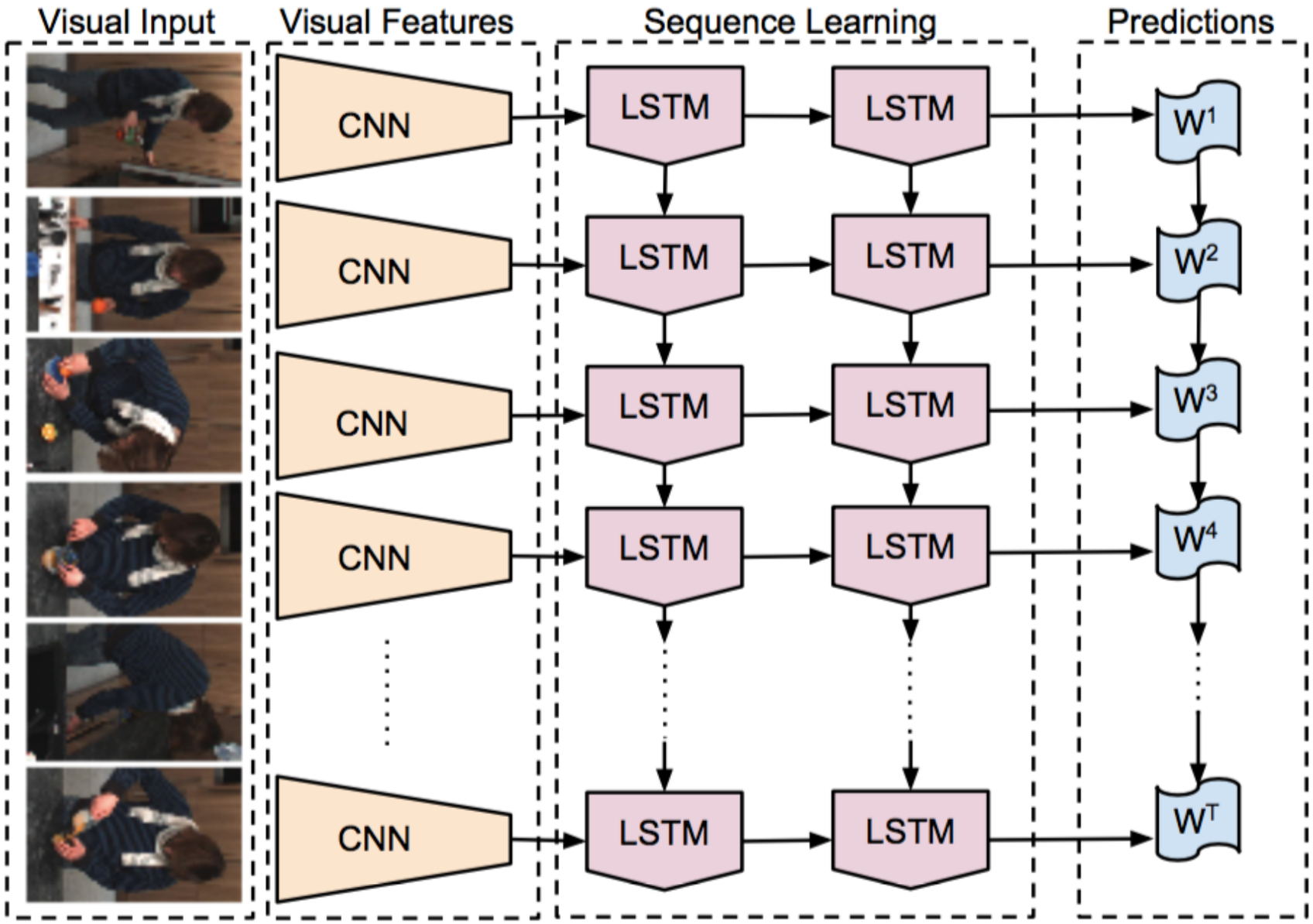
P-CNN



- Parse out the Pose and use it to focus
- Final descriptor is max and min pooled

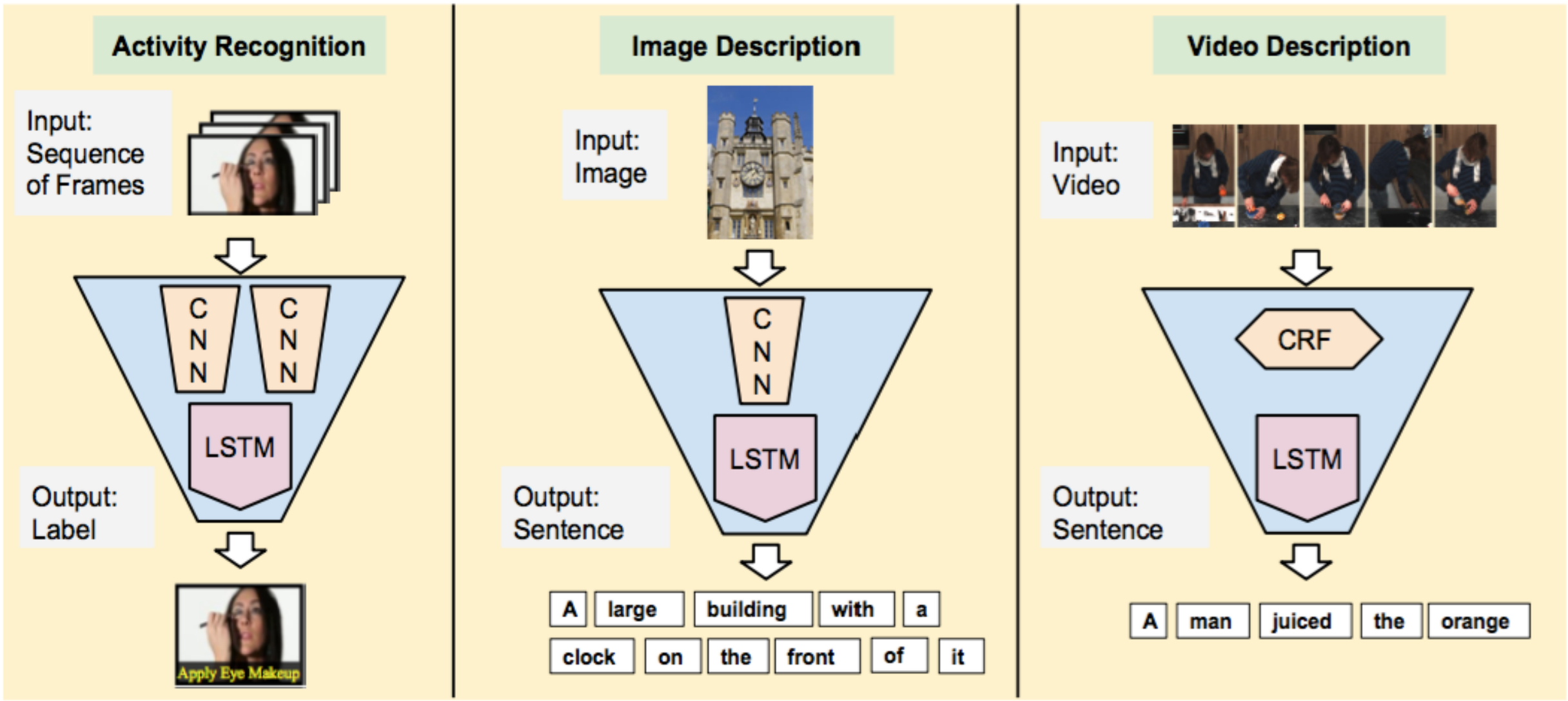
Cheron et al., P-CNN: Pose-based CNN Features for Action Recognition, ICCV 2015

LSTM based



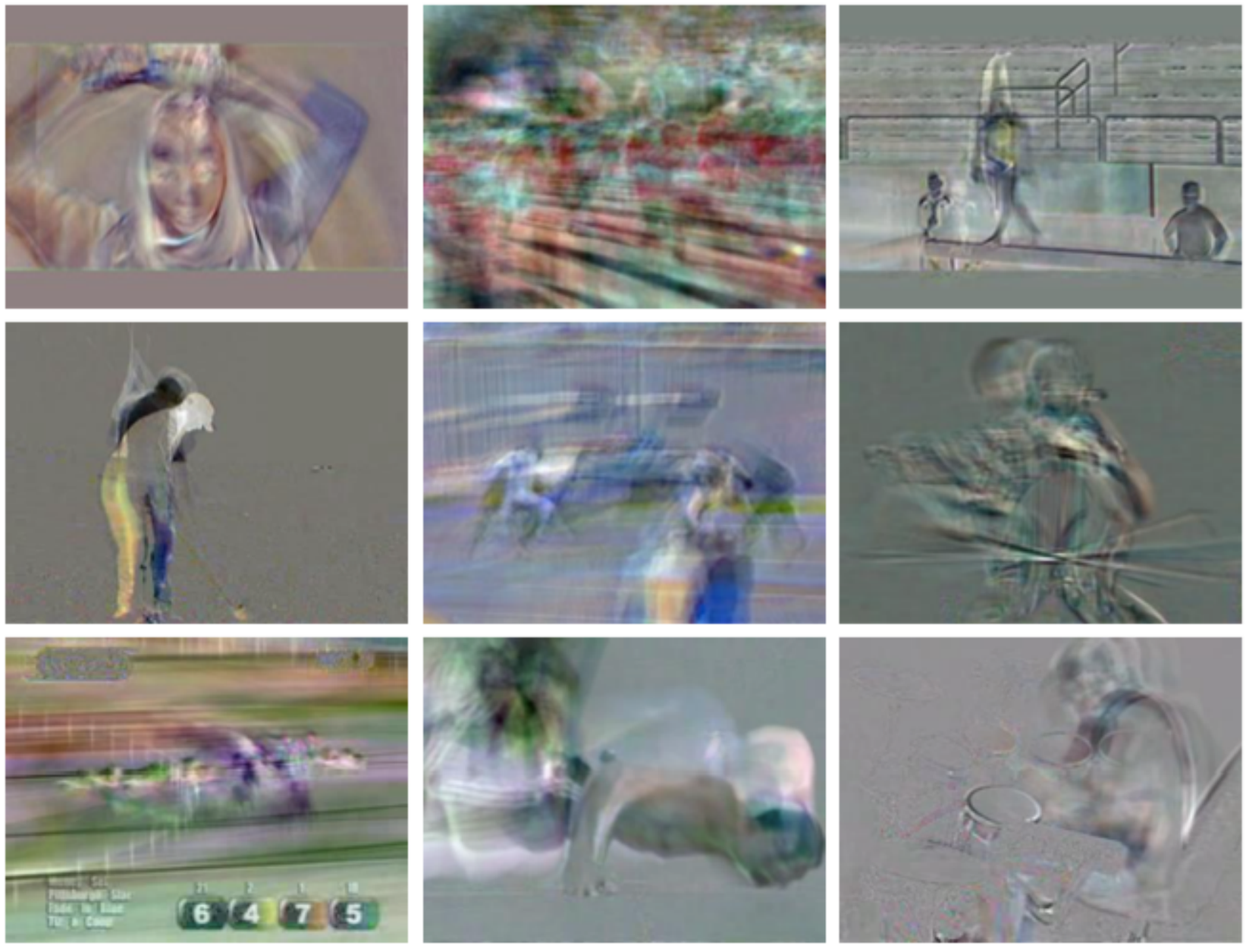
Donahue et al., Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015

Different architectures



Donahue et al., Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015

Dynamic Image Networks



From left to right and top to bottom: “**blowing hair dry**”, “**band march- ing**”, “**balancing on beam**”, “**golf swing**”, “**fencing**”, “**playing the cello**”, “**horse racing**”, “**doing push-ups**”, “**drumming**”.

Bilen et al., Dynamic Image Networks for Action Recognition, CVPR 2016

Dynamic Image Networks

- Dynamic Image = RGB images to summarize vid.
  - Appearance
  - Dynamics

Bilen et al., Dynamic Image Networks for Action Recognition, CVPR 2016

What will we do in this course ?

Start with modest goals

- Learn, thoroughly, the basics of select learning based methods for Computer Vision
  - Potentially, implement them from scratch
    - Also, understand select modern architectures
- Ditto for geometry based Computer Vision
  - (Make a full scale 3D map of IITK ? … Academic area ? … CSE dept ?)