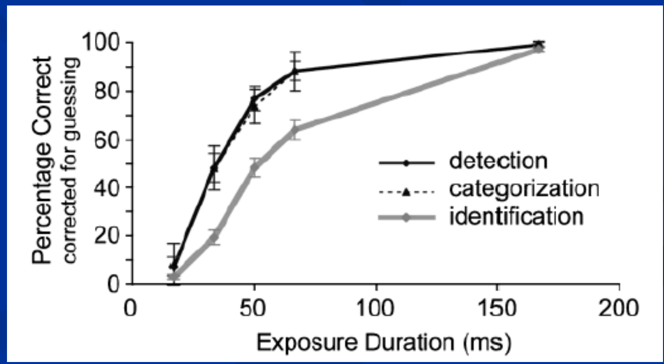
Rich representations for learning visual recognition

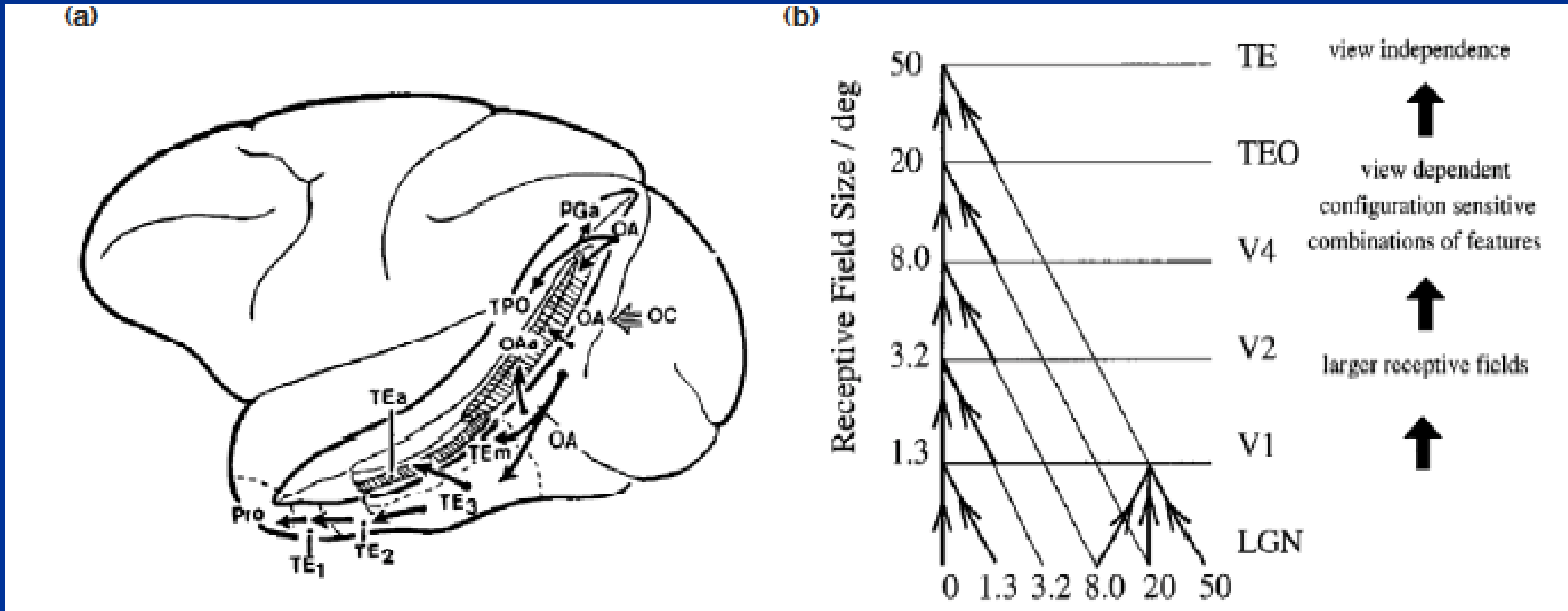
Jitendra Malik
University of California at Berkeley

Detection can be very fast

- On a task of judging animal vs no animal, humans can make mostly correct saccades in 150 ms (Kirchner & Thorpe, 2006)
  - Comparable to synaptic delay in the retina, LGN, V1, V2, V4, IT pathway.
  - Doesn't rule out feed back but shows feed forward only is very powerful
- Detection and categorization are practically simultaneous (Grill-Spector & Kanwisher, 2005)



Rolls et al (2000)



Some opinions…

- A hierarchical, mostly feedforward network is the right model, the question is how to train it
- Unsupervised, sparsity encouraging techniques are promising for lower layers
- But so far the success of this approach at the higher stages has not yet been demonstrated

Insights from child development

•Trying to learn object recognition from bounding boxes is like trying to learn language from a list of sentences.

• The development of visual recognition, like language acquisition, benefits from supportive “scaffolding”

✓ Grouping and tracking can play an important role by helping solve the correspondence problem. In a machine vision system, we can “cheat” by supplying keypoint correspondences

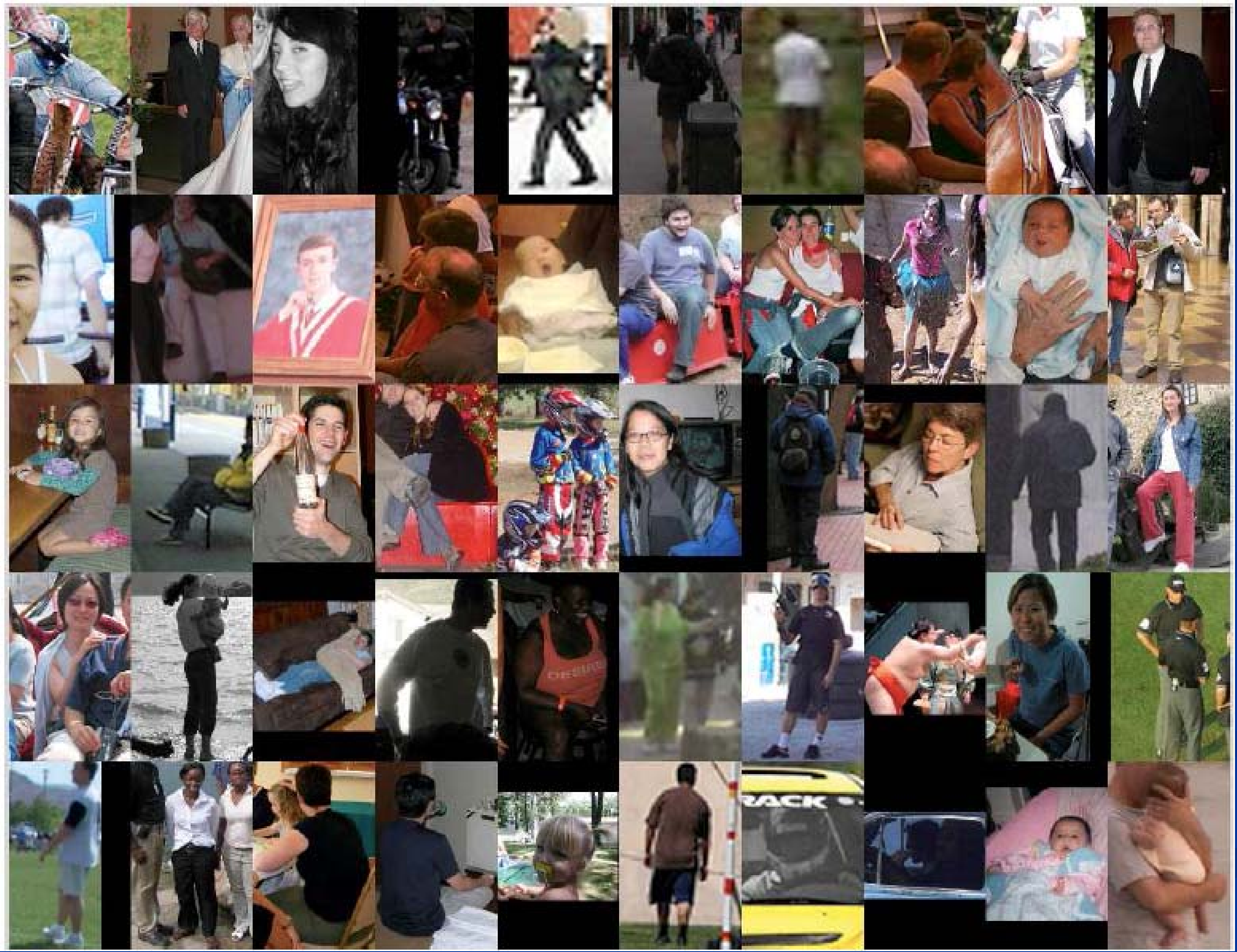
Detecting and Segmenting People

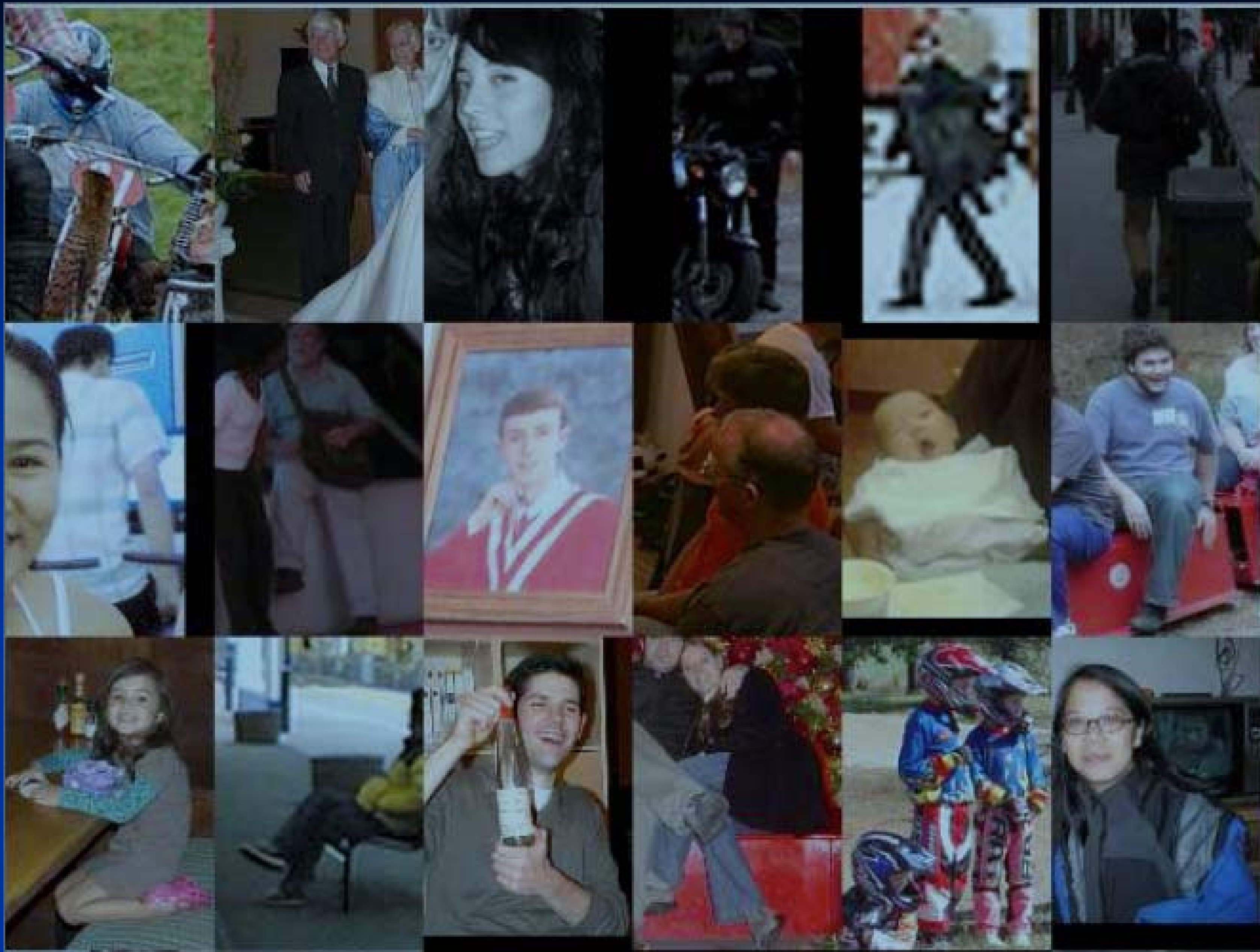
Where are they? What are they wearing? What are they doing?

Jitendra Malik

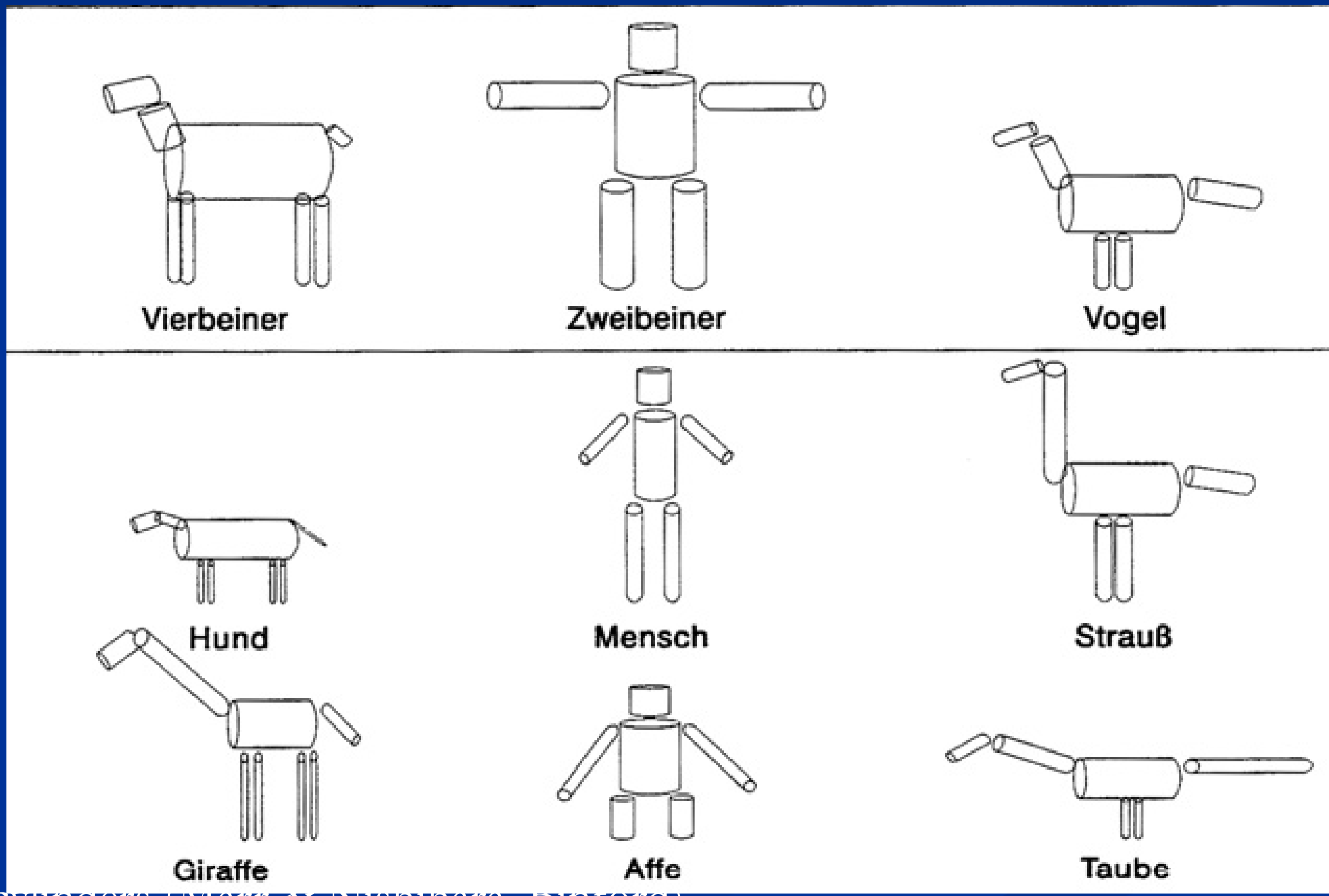
UC Berkeley

This is joint work with L. Bourdev, S. Maji and T. Brox.



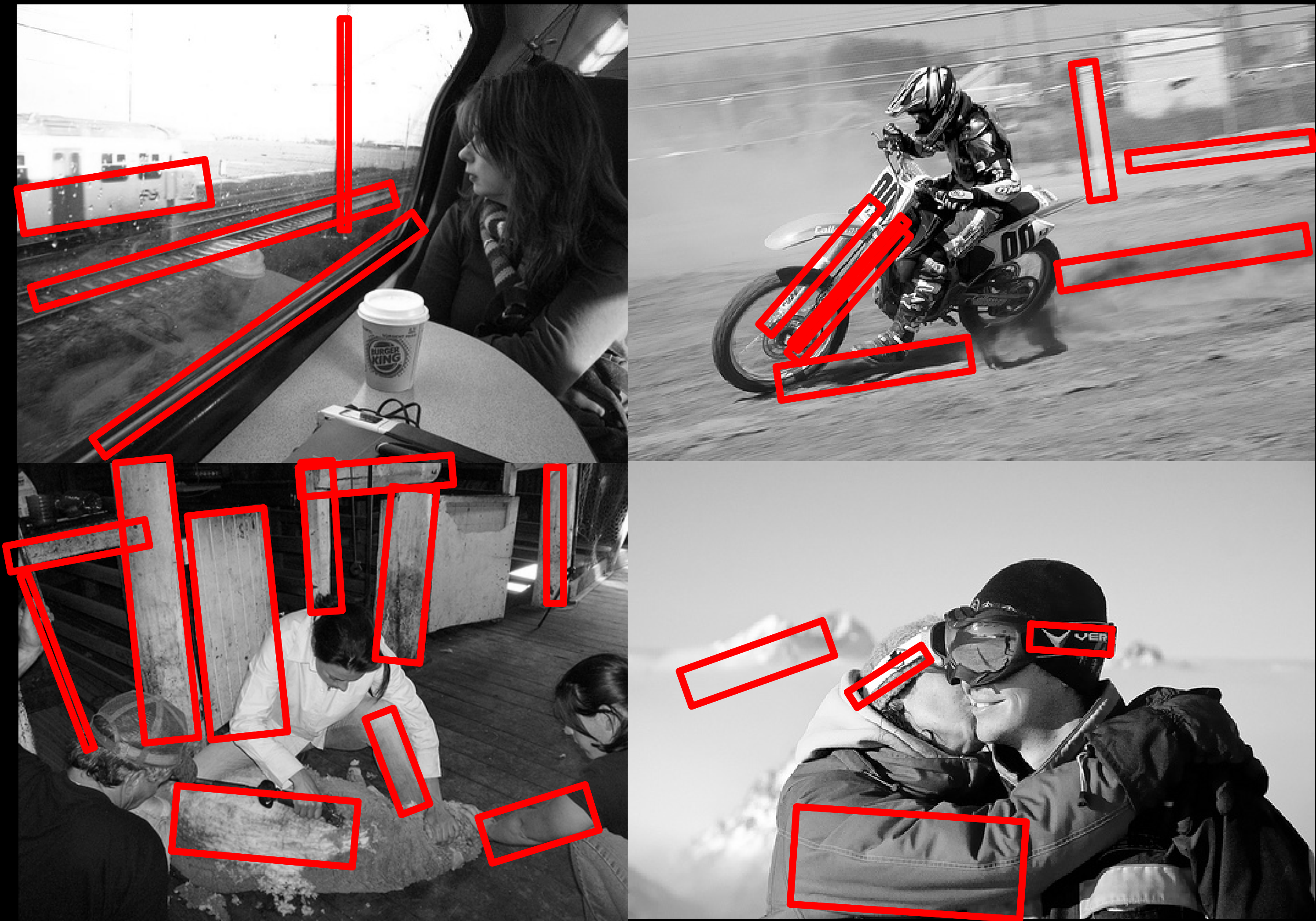


Trying to extract stick figures is hard (and unnecessary!)



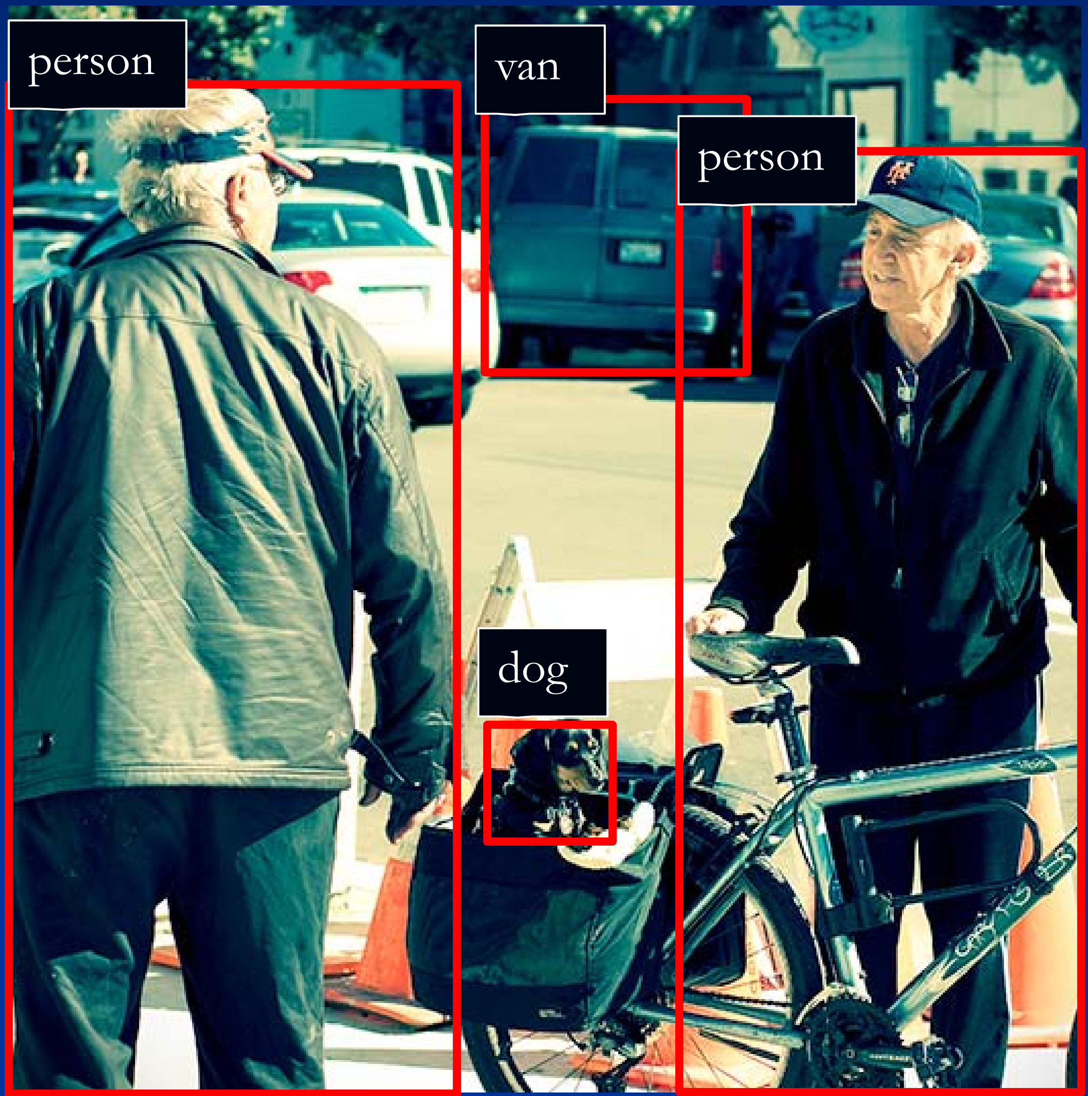
Generalized cylinders (Marr & Nishihara, Binford)
Pictorial Structures (Felszenswalb & Huttenlocher)

All the wrong limbs…



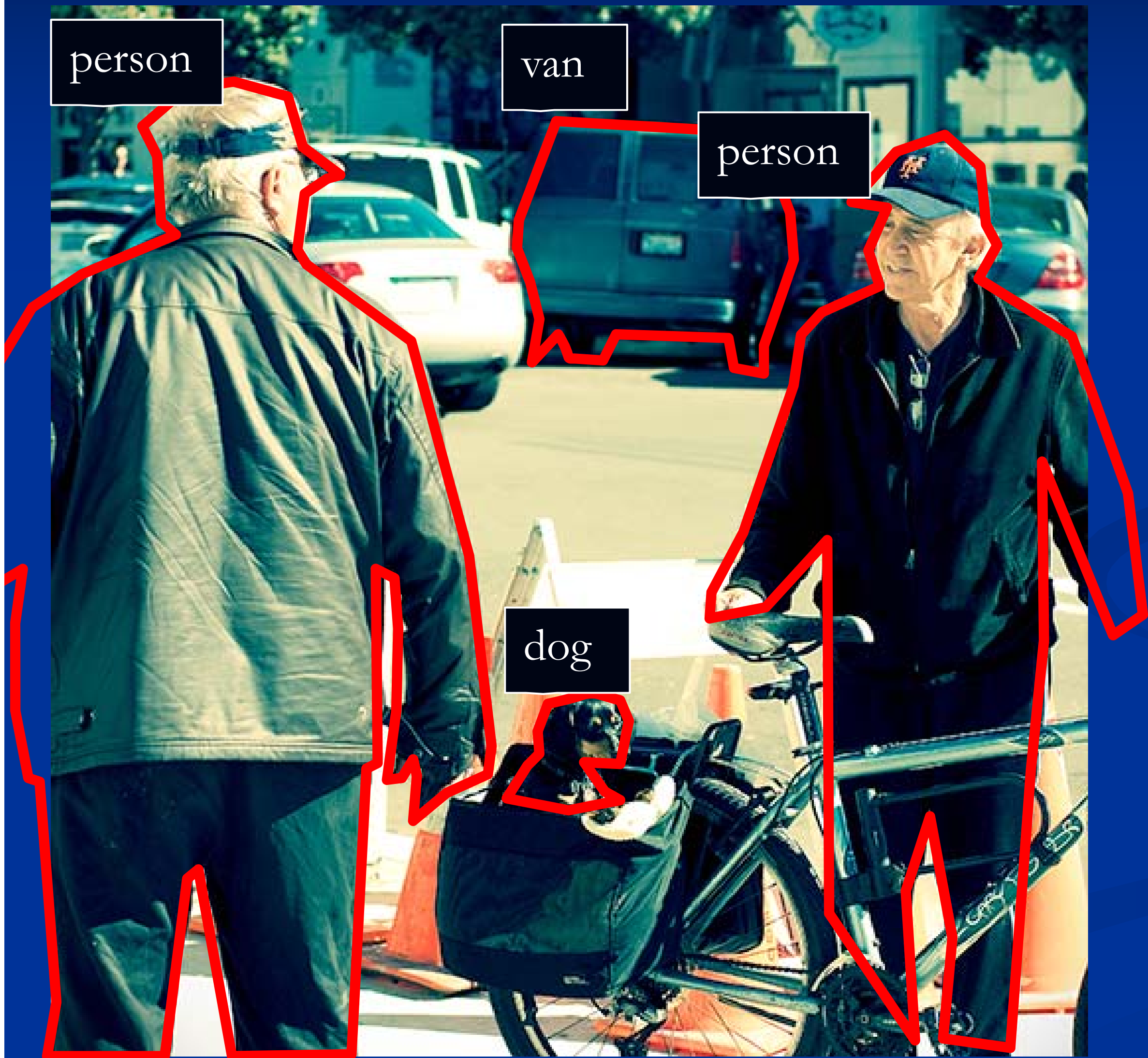
High-Level Computer Vision

High-Level Computer Vision



Object Recognition

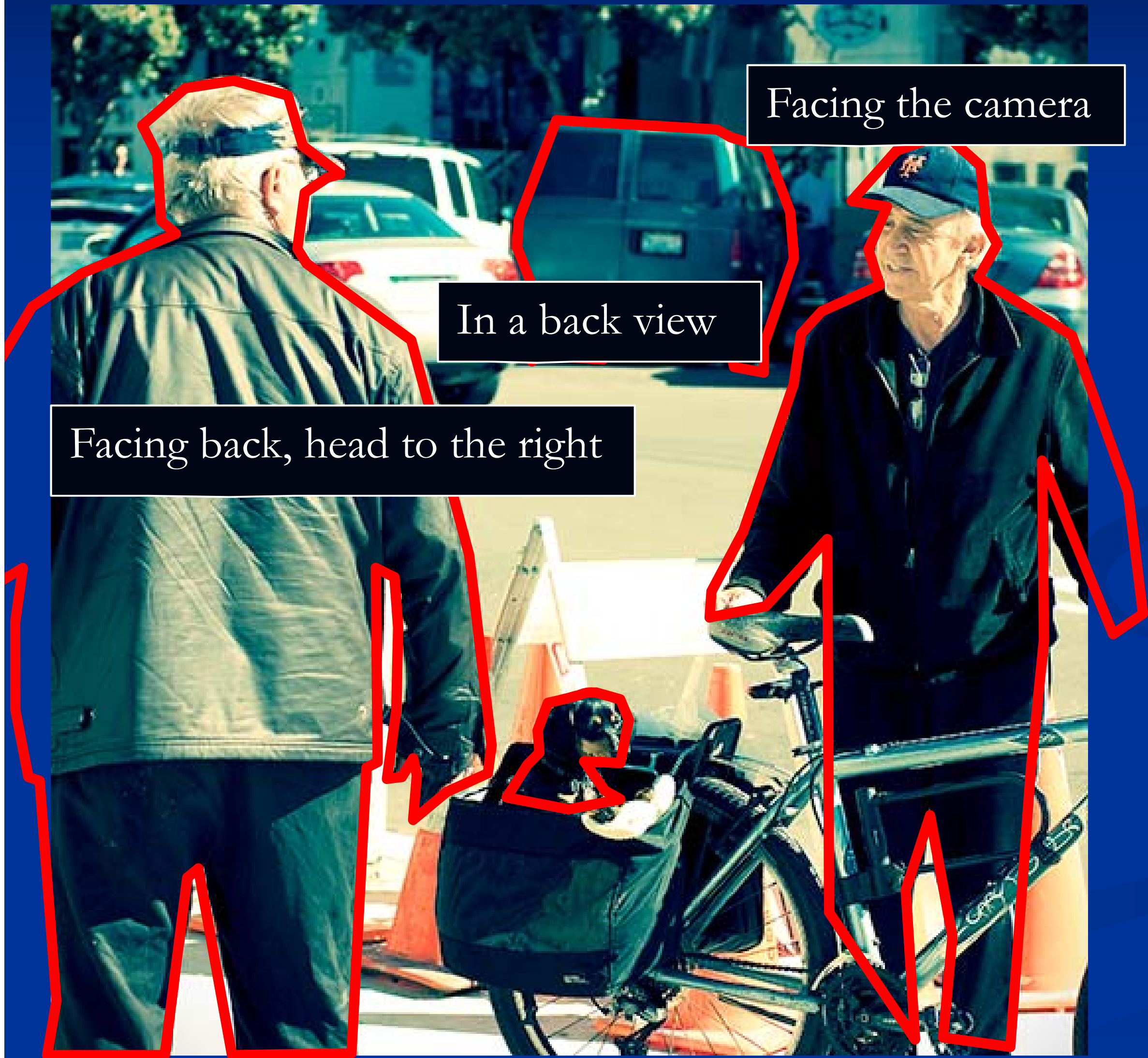
High-Level Computer Vision



Object Recognition

Semantic Segmentation

High-Level Computer Vision



Object Recognition
Semantic Segmentation
Pose Estimation

High-Level Computer Vision

Object Recognition
Semantic Segmentation
Pose Estimation
Action Recognition

High-Level Computer Vision

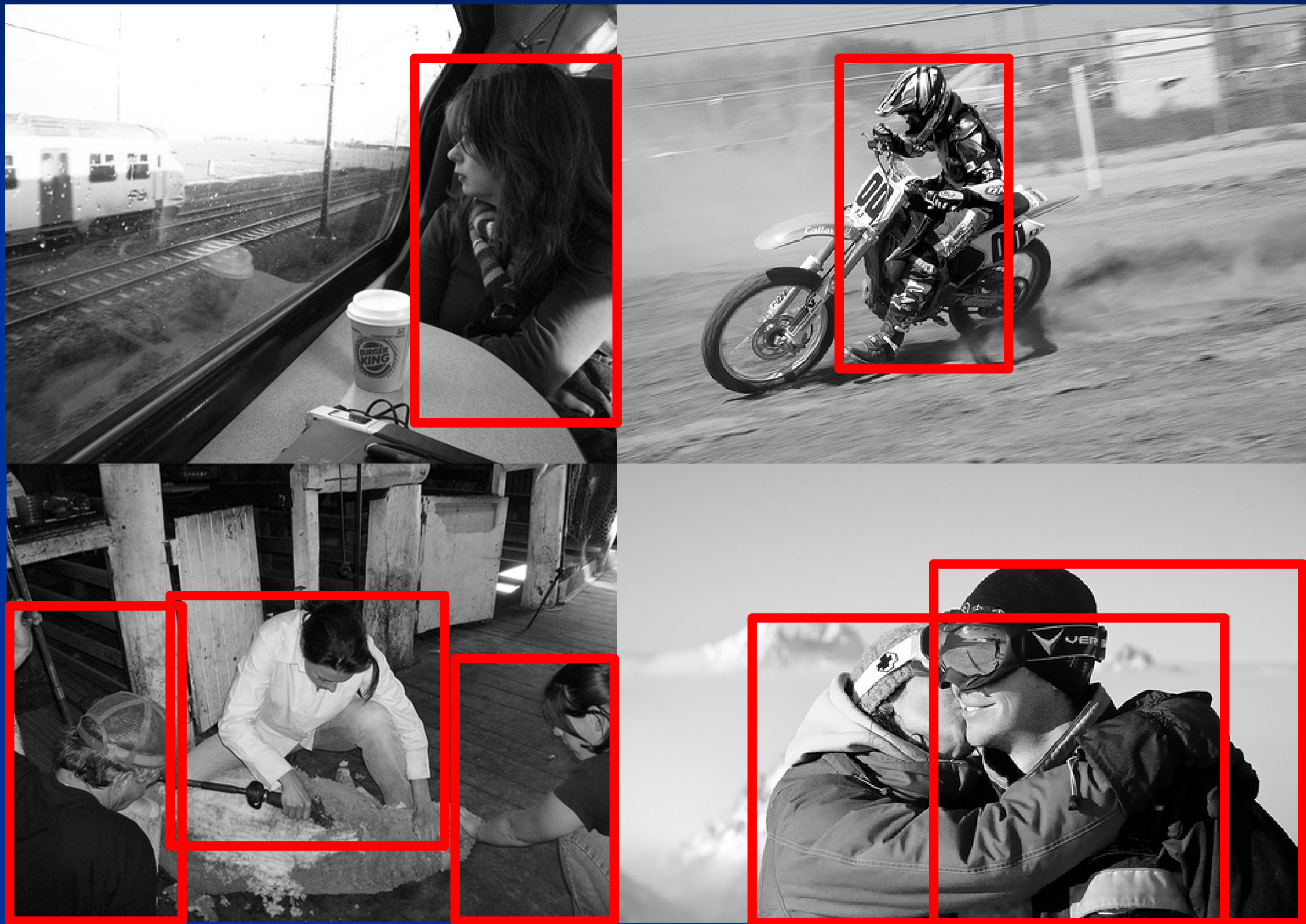
Object Recognition
Semantic Segmentation
Pose Estimation
Action Recognition
Attribute Classification

High-Level Computer Vision

Object Recognition
Semantic Segmentation
Pose Estimation
Action Recognition
Attribute Classification

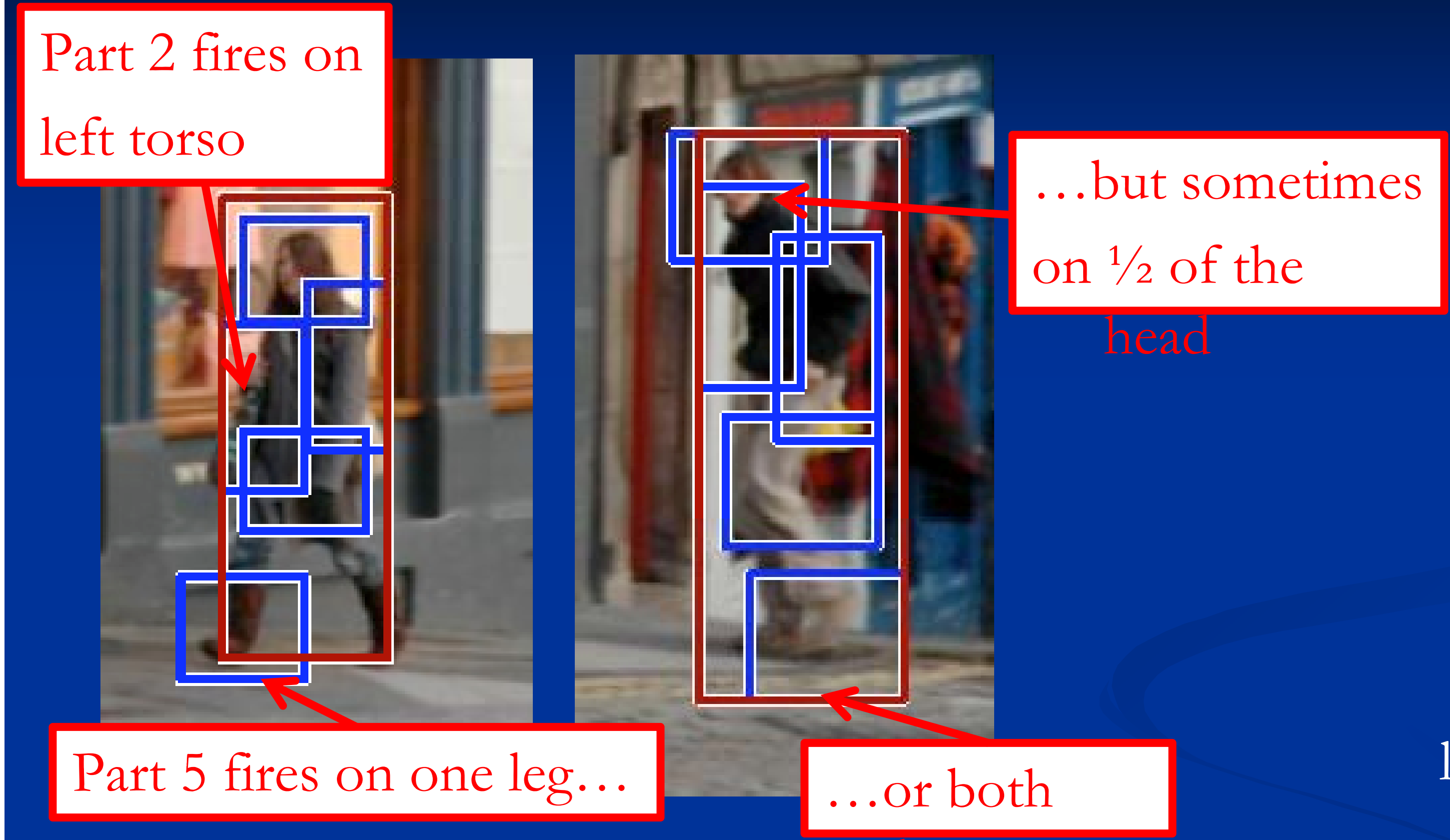
Person Detection is Challenging

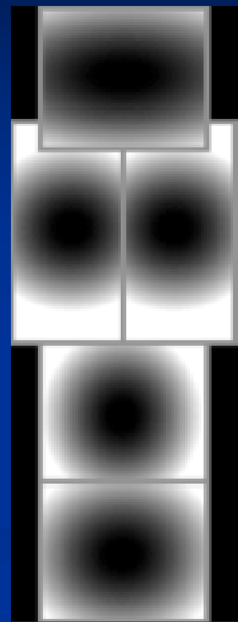
How can we make the problem harder?



- Solution: Severely limit the supervision

The best approach in such setup?





Learned part location penalty

- Divide-and-conquer: One global template + five parts
- Positions and appearance of parts trained jointly (Latent SVM)
- Mixture of models for various poses (standing, sitting, etc)
- Parts are not well localized and have large appearance variations

[Felzenszwalb et al. PAMI 2010]

Radical idea: What if, instead, we try to make the problem easier?

Nose

Right Shoulder

Left Shoulder

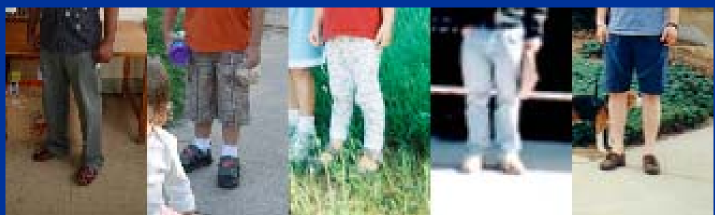
Right Elbow

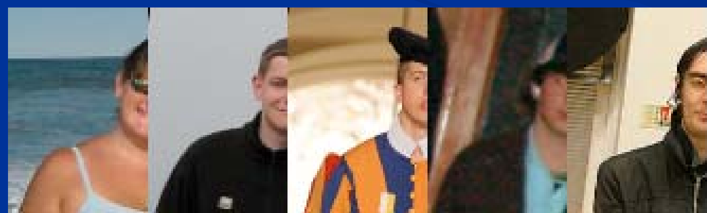
Left Elbow

[Bourdev and Malik, ICCV 2009]

Can we build upon the success of faces and pedestrians?

- Both do template matching
- Capture salient and common patterns
- Are these the only two salient & common patterns?





- But how are we going to create the training set?

Agenda

- Poselets
  - Training a poselet
  - Selecting a good set of poselets
  - Improving poselets with context
  - Detection with poselets
- Segmentation
- Attributes
- Action Recognition

Agenda

- **Poselets**
  - Training a poselet
  - Selecting a good set of poselets
  - Improving poselets with context
  - Detection with poselets
- Segmentation
- Attributes
- Action Recognition

Examples of poselets

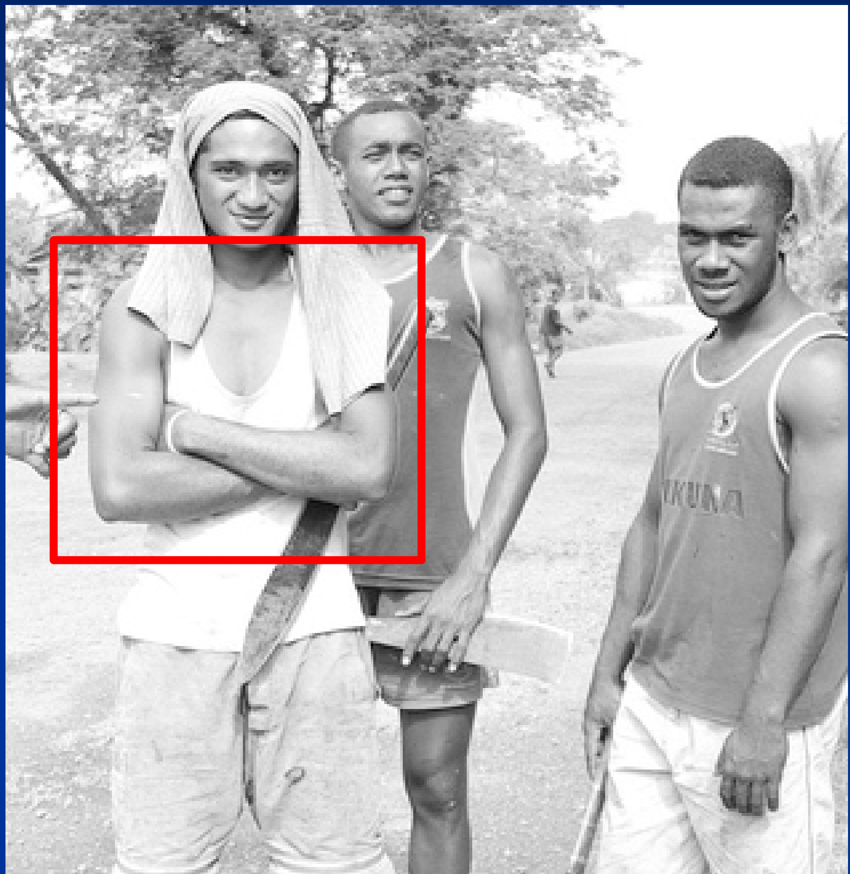
Patches are often far **visually**, but they are close **semantically**

Agenda

- **Poselets**
  - **Training a poselet**
  - Selecting a good set of poselets
  - Improving poselets with context
  - Detection with poselets
- Segmentation
- Attributes
- Action Recognition

How do we train a poselet for a given pose configuration?

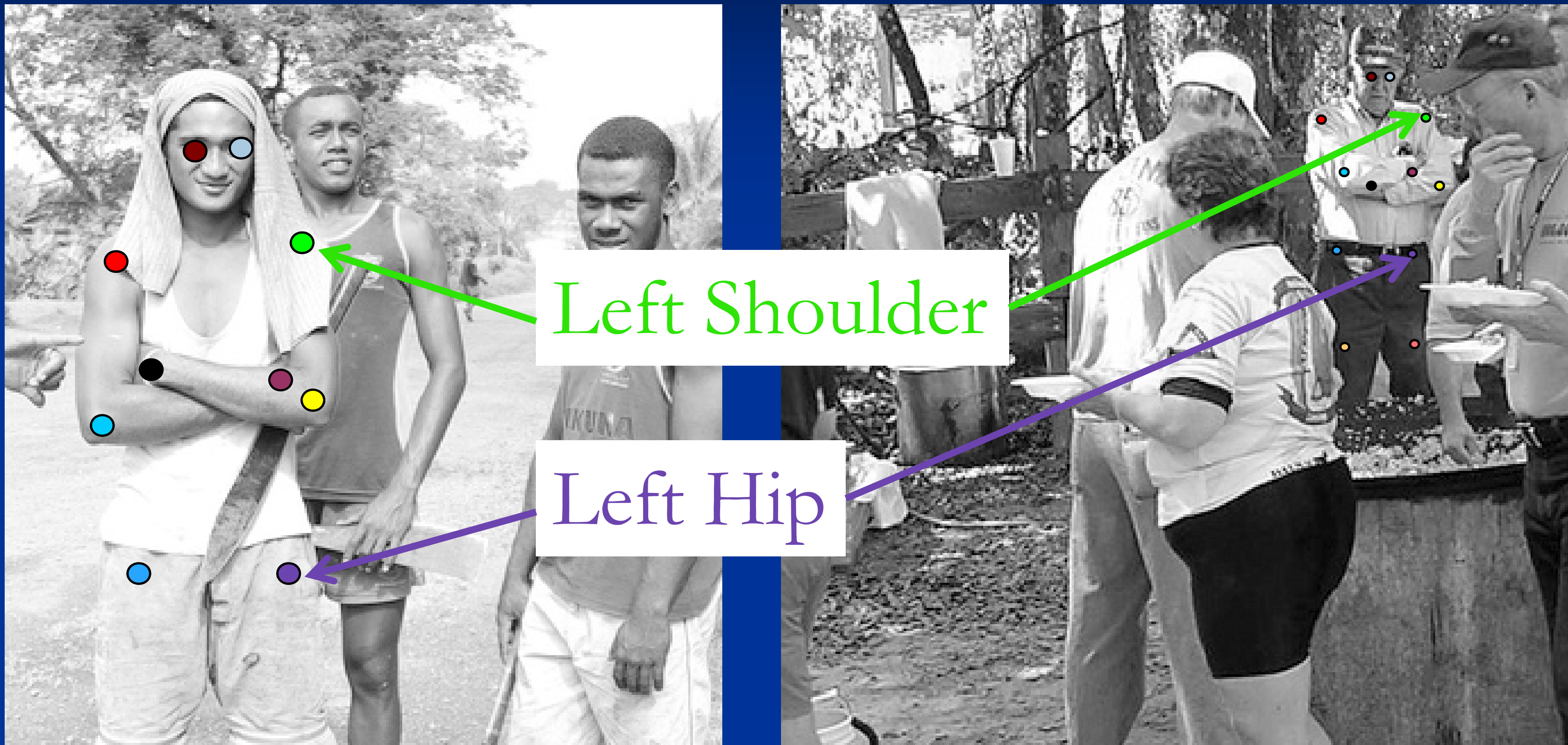
Finding Correspondences



Given part of a human pose

How do we find a similar pose configuration in the training set?

Finding Correspondences



We use keypoints to annotate the joints, eyes, nose, etc. of people

Finding Correspondences

Residual Error

Training poselet classifiers

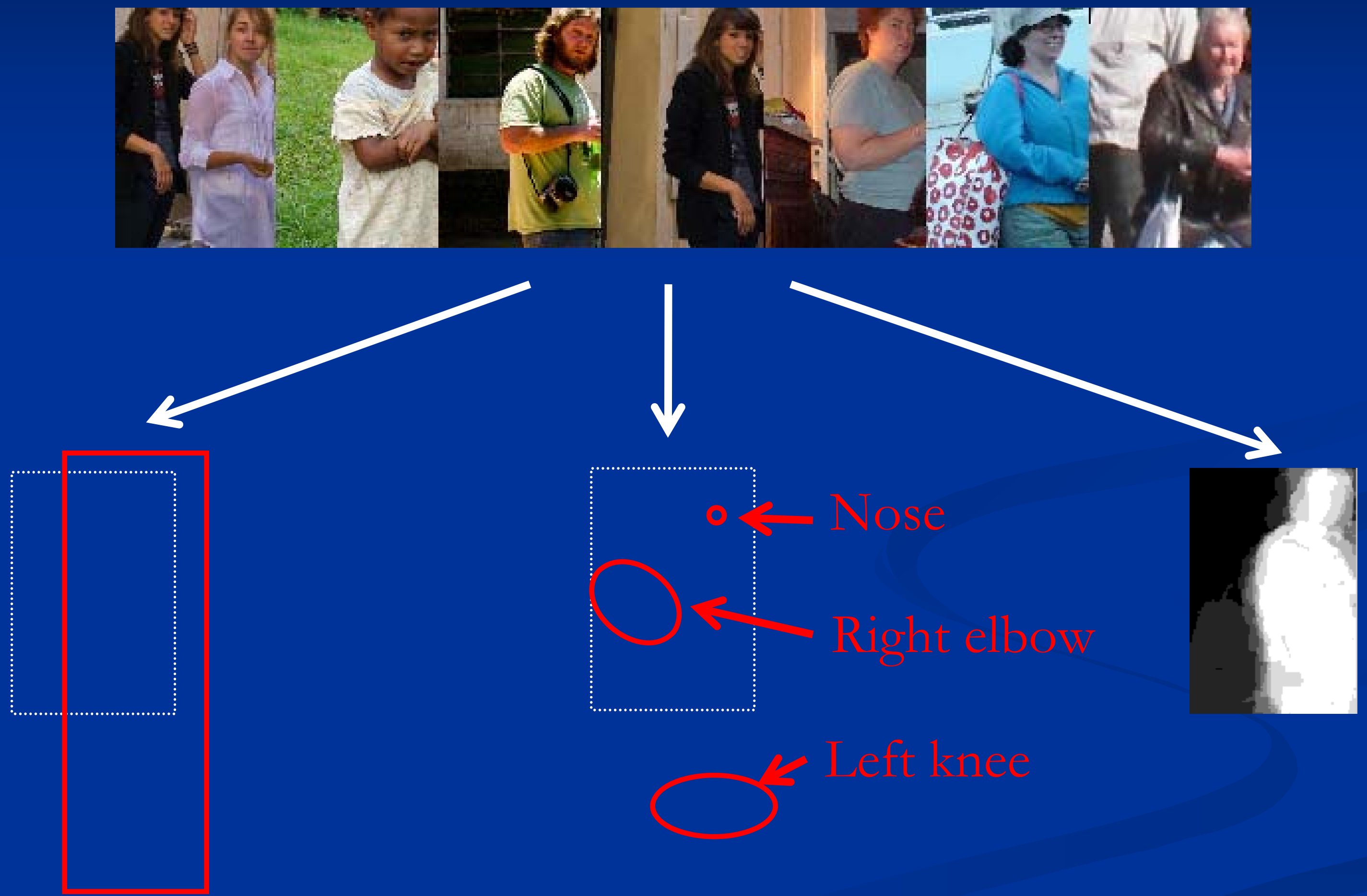
Residual Error: 0.15 0.20 0.10 0.85 0.15 0.35

1. Given a seed patch
2. Find the closest patch for every other person
3. Sort them by residual error
4. Threshold them

Training poselet classifiers

1. Given a seed patch
2. Find the closest patch for every other person
3. Sort them by residual error
4. Threshold them
5. Use them as positive training examples for a classifier (HOG features, linear SVM)

For a trained poselet we fit:



Expected person bounds

Keypoint predictions

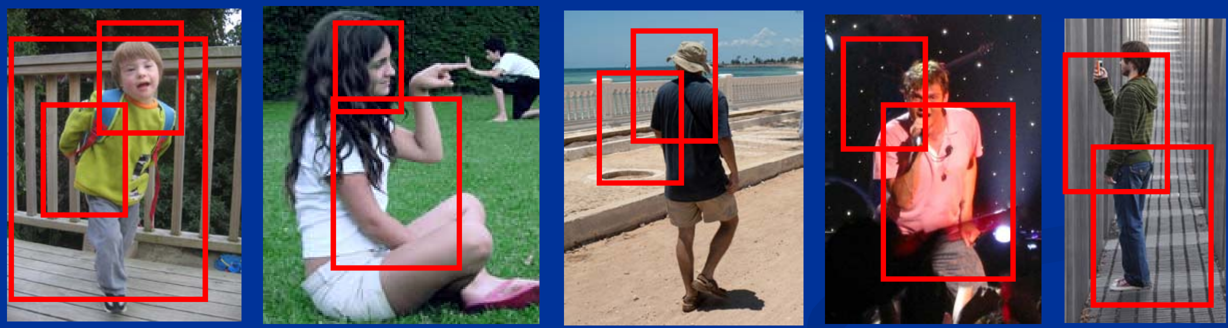
Foreground probability mask

Agenda

- Poselets
  - Training a poselet
  - **Selecting a good set of poselets**
  - Improving poselets with context
  - Detection with poselets
- Segmentation
- Attributes
- Action Recognition

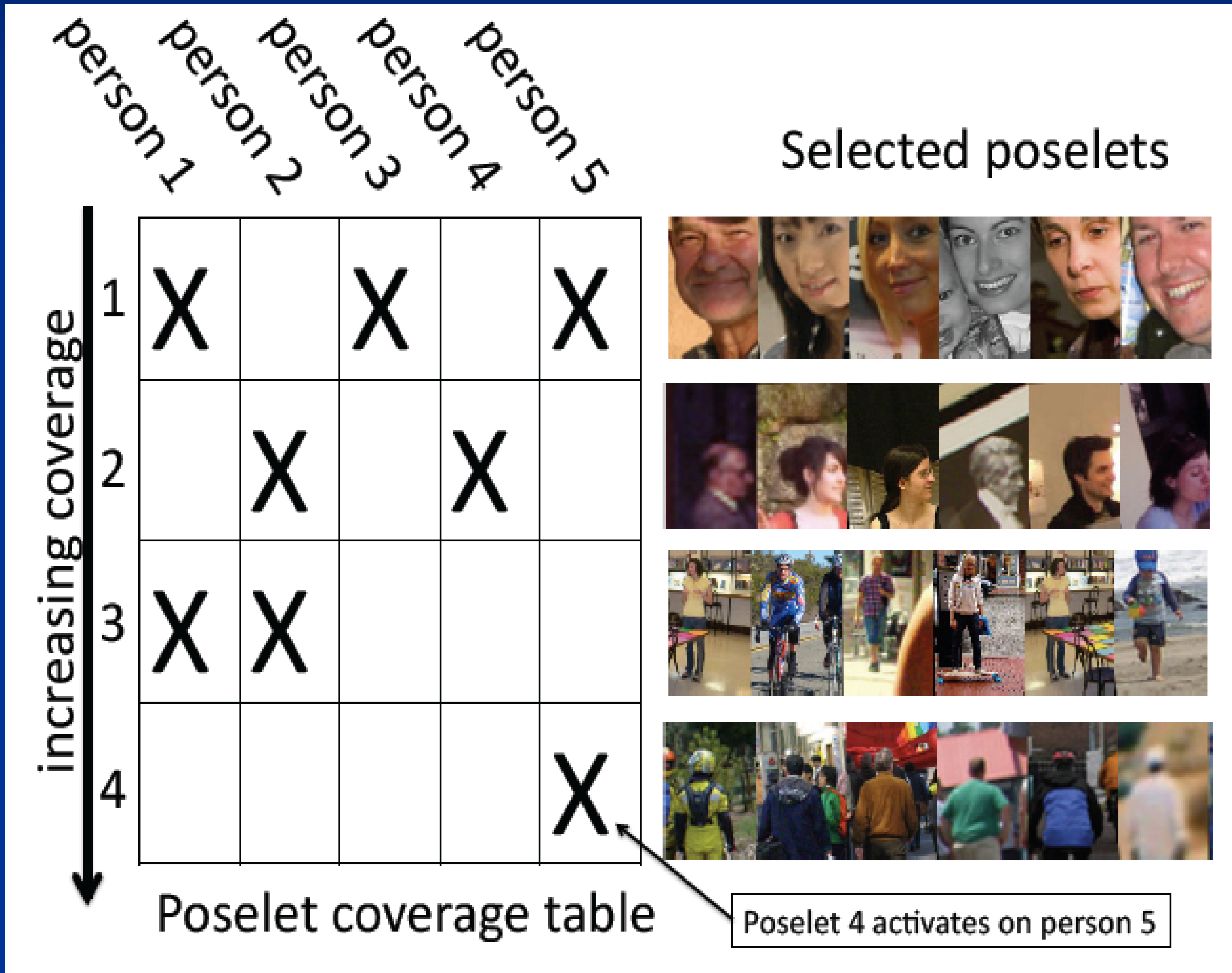
How do we find poselets?

- Choose thousands of random windows, generate poselet candidates, train linear SVMs

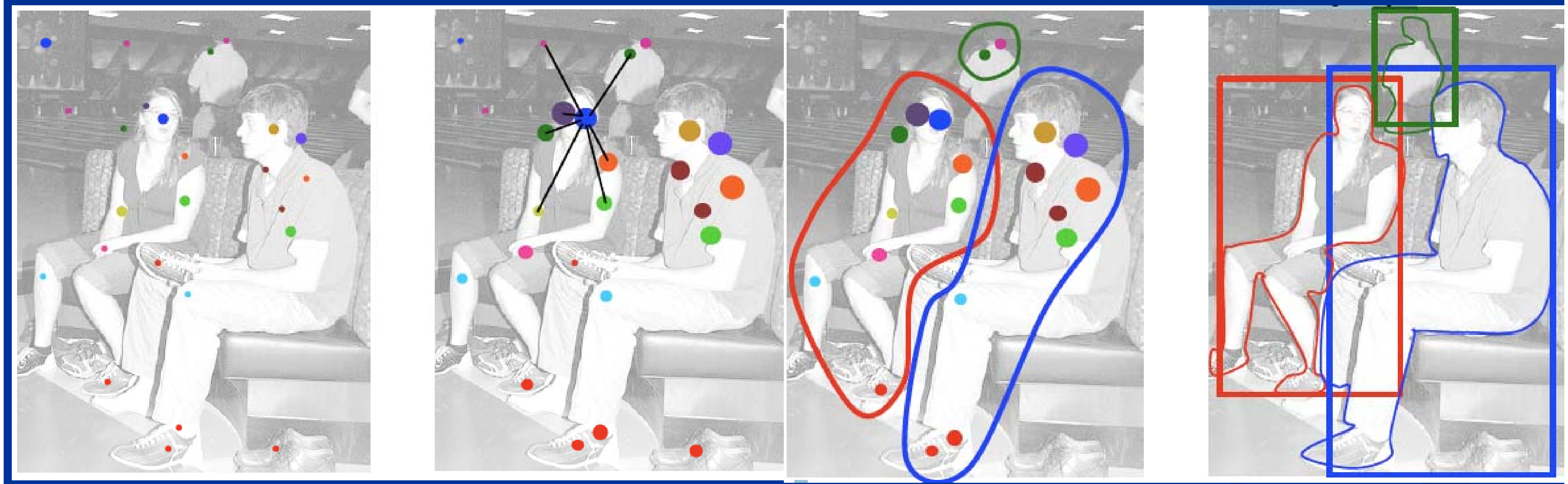


- Select a small set of poselets that are:
  - Individually effective
  - Complementary

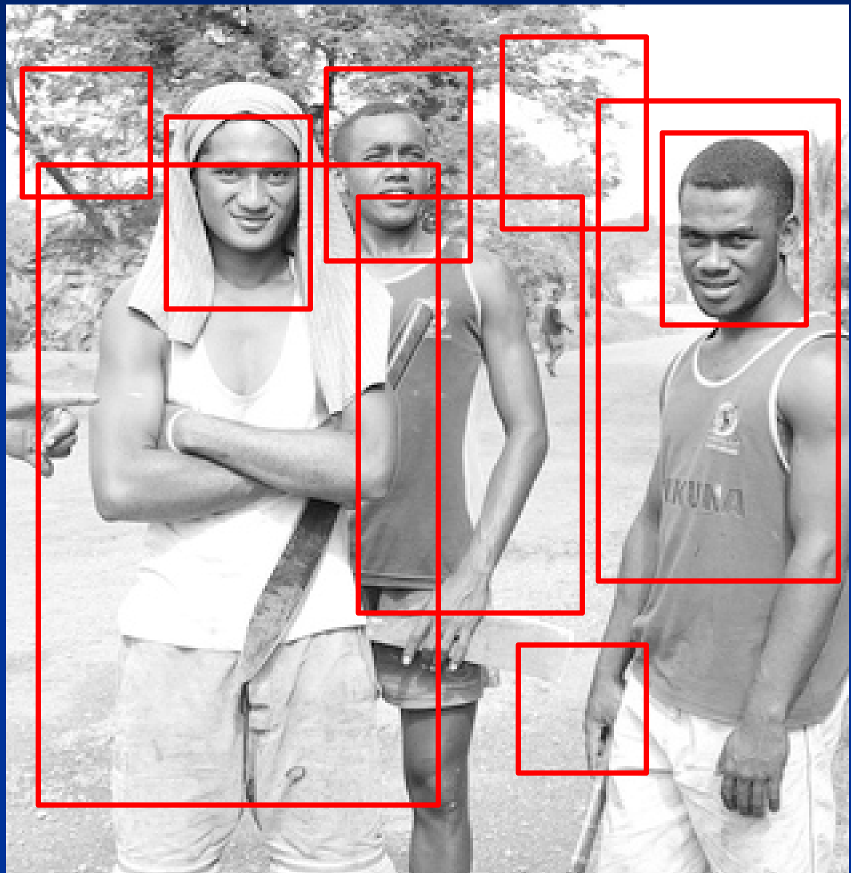
Selecting a small set of complementary poselets



Poselet Activations ➔ Detections & Segmentations

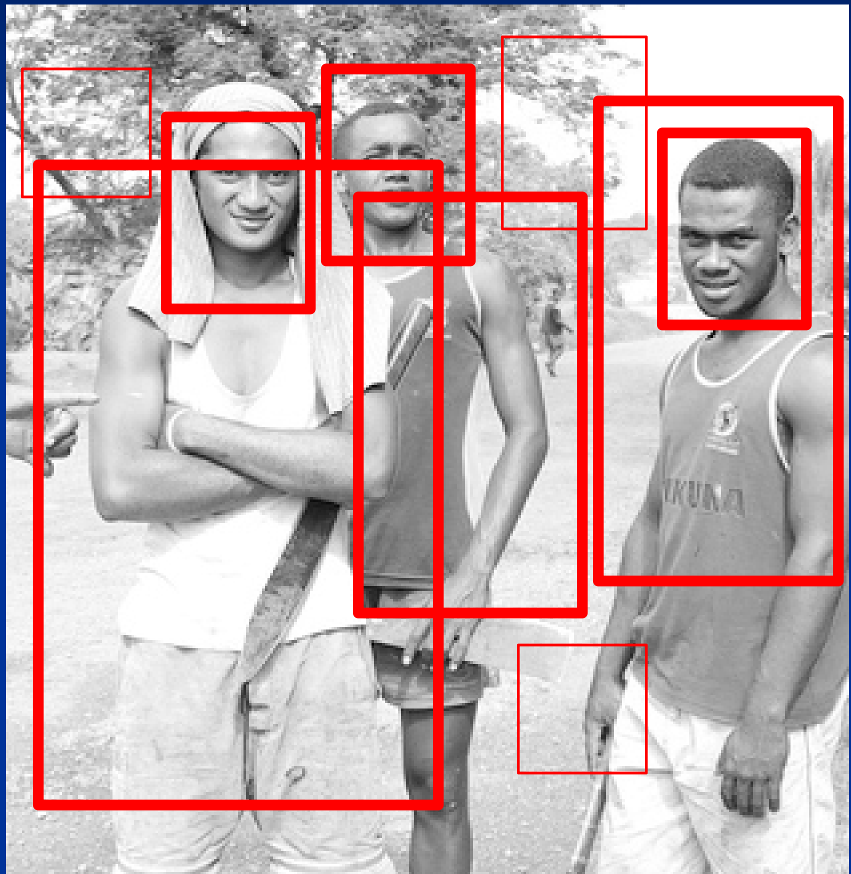


Creating Poselet Activation Vector



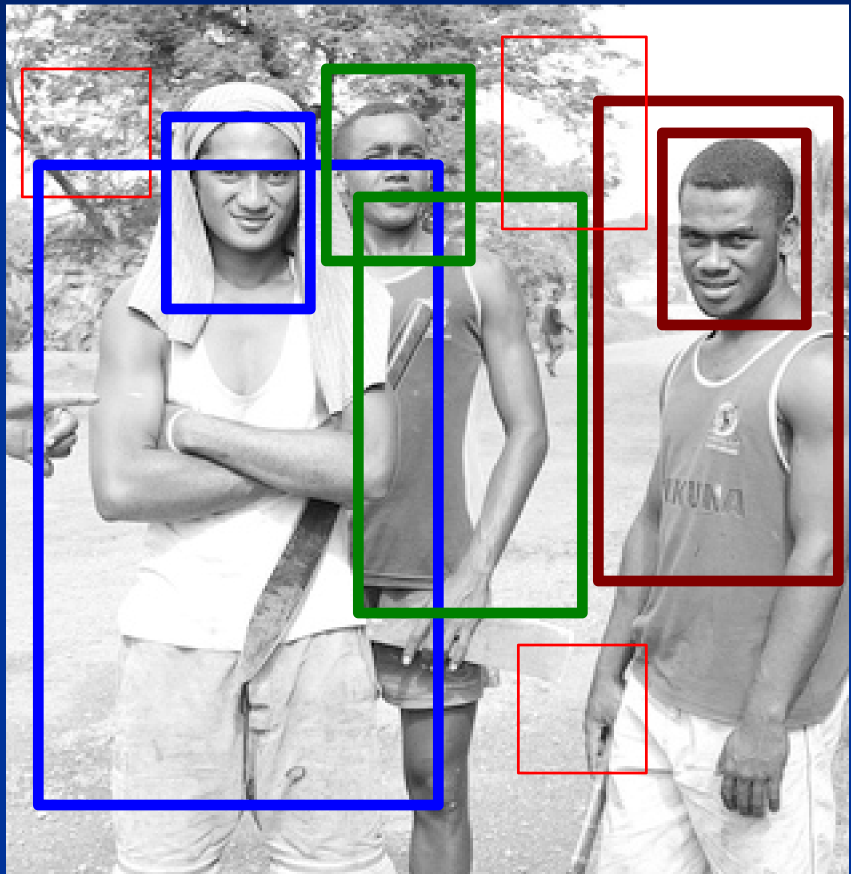
- Step 1: Detect poselets in the image

Creating Poselet Activation Vector



- Step 2: Enhance their scores using context

Creating Poselet Activation Vector

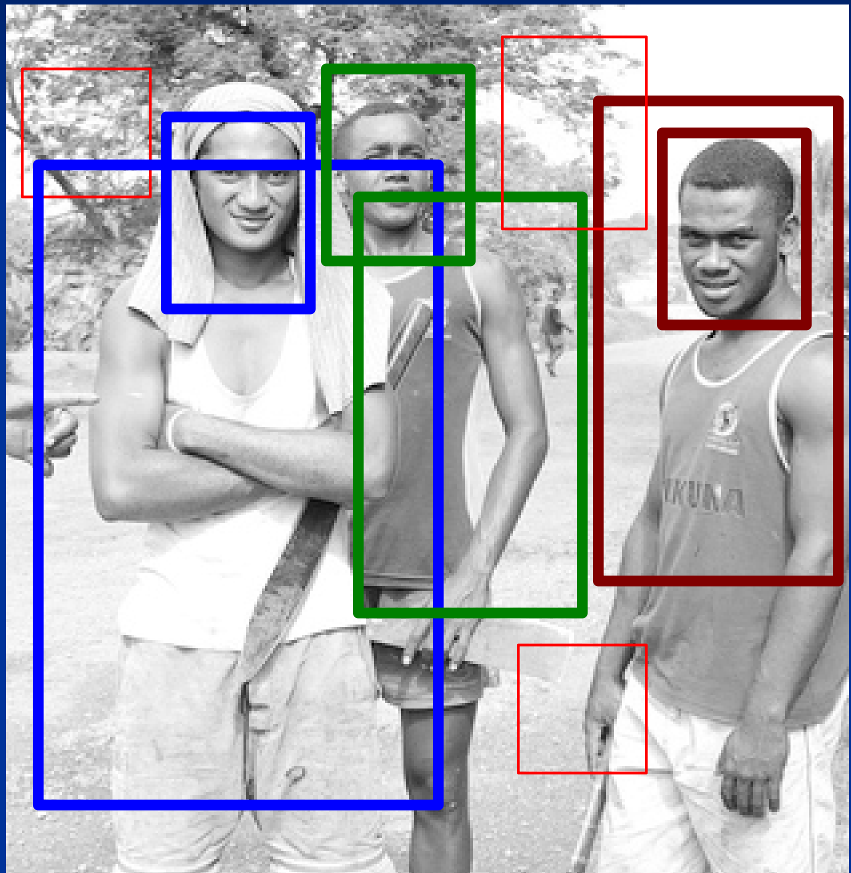


Two poselets refer to the same person if their keypoint predictions are consistent:

Consistent Not consistent

- Step 3: Cluster poselets of the same person together

Creating Poselet Activation Vector



Poselet type	Cluster1	Cluster2	Cluster3
	0.32	0.11	0.95
	0.77	0	0.08
	0	0	0.72
	0.41	0	0

Poselet activation vector

- Step 4: Collect the scores of all poselets in a cluster into a poselet activation vector

Poselet Activation Vector

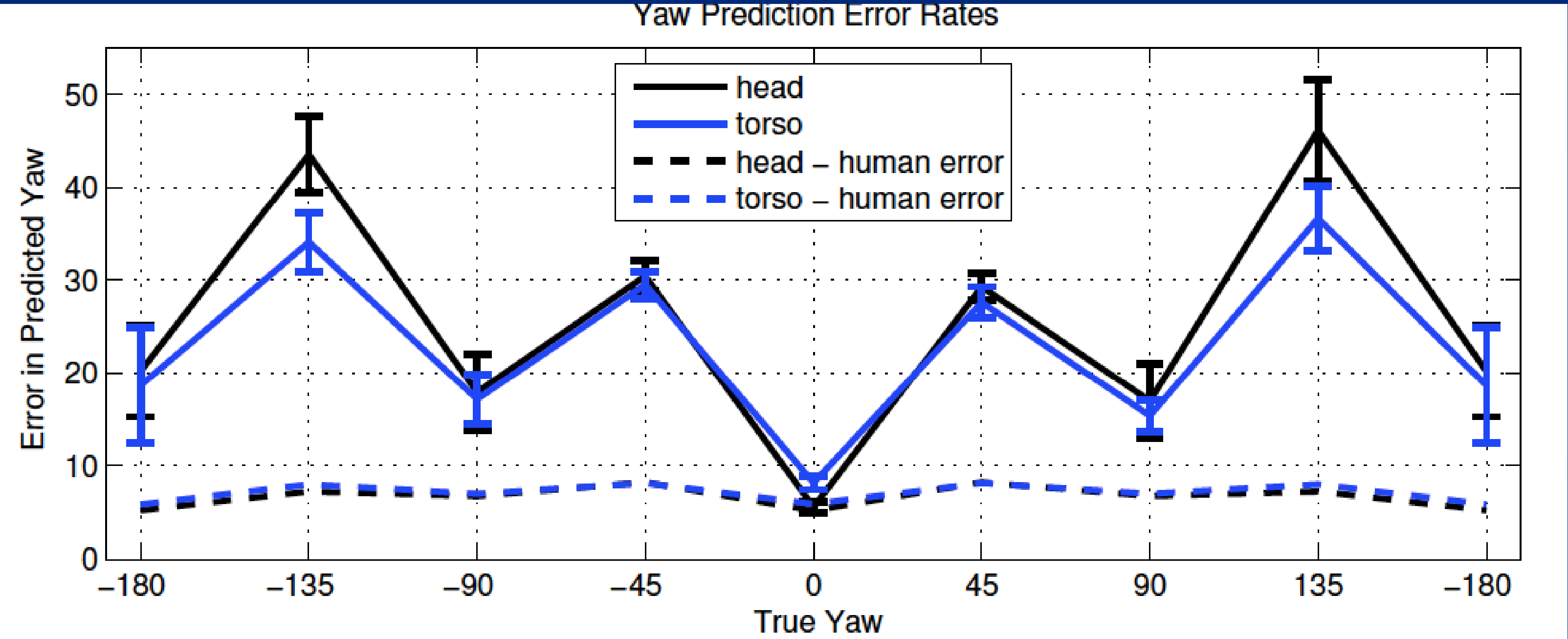
Poselet type	Cluster1	Cluster2	Cluster3
	0.32	0.11	0.95
	0.77	0	0.08
	0	0	0.72
	0.41	0	0

Mostly facing right

Front facing

Likely false positive

- PAV provides a **distributed representation** of the pose and is the basis for poselet-based tasks

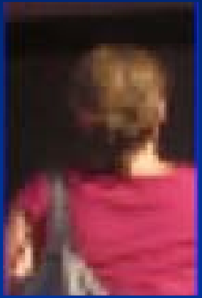
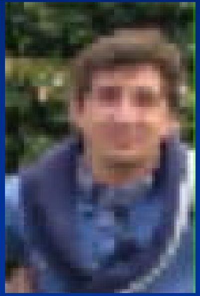


(c) Error in predicting yaw across views

Agenda

- Poselets
  - Training a poselet
  - Selecting a good set of poselets
  - **Improving poselets with context**
  - Detection with poselets
- Segmentation
- Attributes
- Action Recognition

Problem: The patch may have weak signal

Front and back look similar

Face false positive

Left or right leg?

A front face poselet can disambiguate them

Lack of head-and-shoulders poselet suggests a false positive

Location of pedestrian poselet can disambiguate

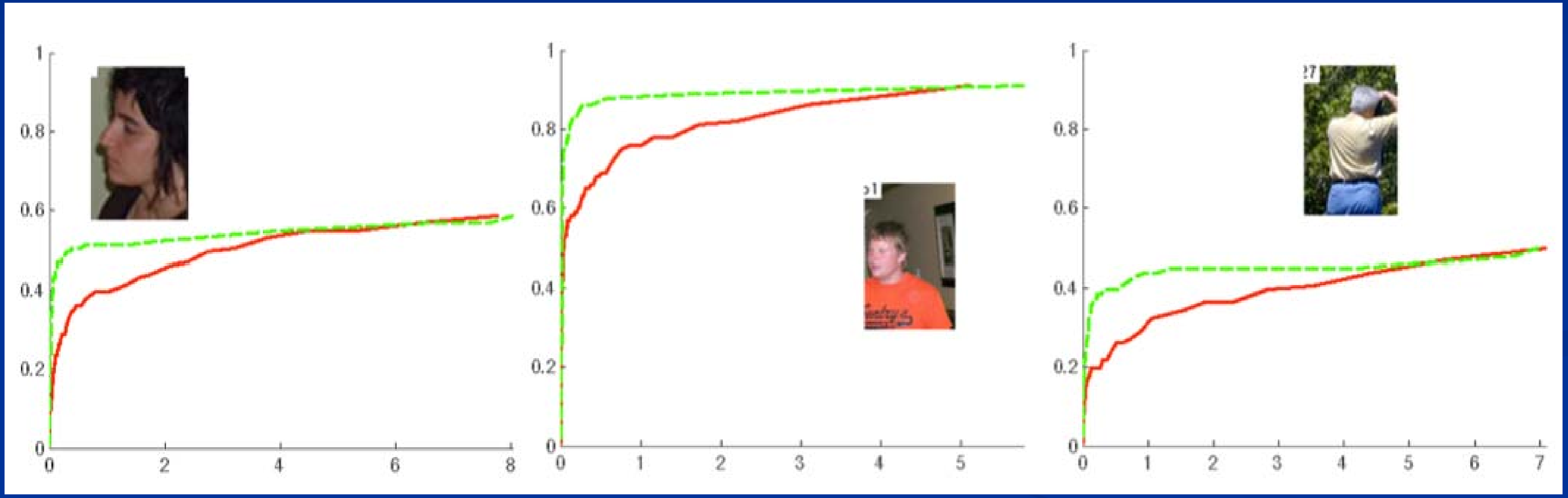
Solution: Enhance the poselet score using other consistent poselets

Using context

1. For each poselet activation on the training set:
   A. Find its label: True positive, False positive, Unknown
   B. Construct a feature vector from activations of other consistent poselets
2. Train a linear classifier for each poselet
3. Convert score to probability via logistic regression

The effect of using context

ROC curves for three random poselets



Green: Context

Red: No context

Agenda

- Poselets
  - Training a poselet
  - Selecting a good set of poselets
  - Improving poselets with context
  - **Detection with poselets**
- Segmentation
- Attributes
- Action Recognition

Object Detection with Poselets

1. Detect poselets in the image
2. Enhance their scores via context
3. Cluster consistent ones into object hypotheses

The most salient poselet creates the first hypothesis

If a poselet is consistent with an existing hypothesis it gets assigned to it

Otherwise it starts a new hypothesis

4. Predict bounding box and score of the cluster

Object Detection with Poselets

1. Detect poselets in the image
2. Enhance their scores via context
3. Cluster consistent ones into object hypotheses

The most salient poselet creates the first hypothesis

If a poselet is consistent with an existing hypothesis it gets assigned to it

Otherwise it starts a new hypothesis

4. Predict bounding box and score of the cluster

Results

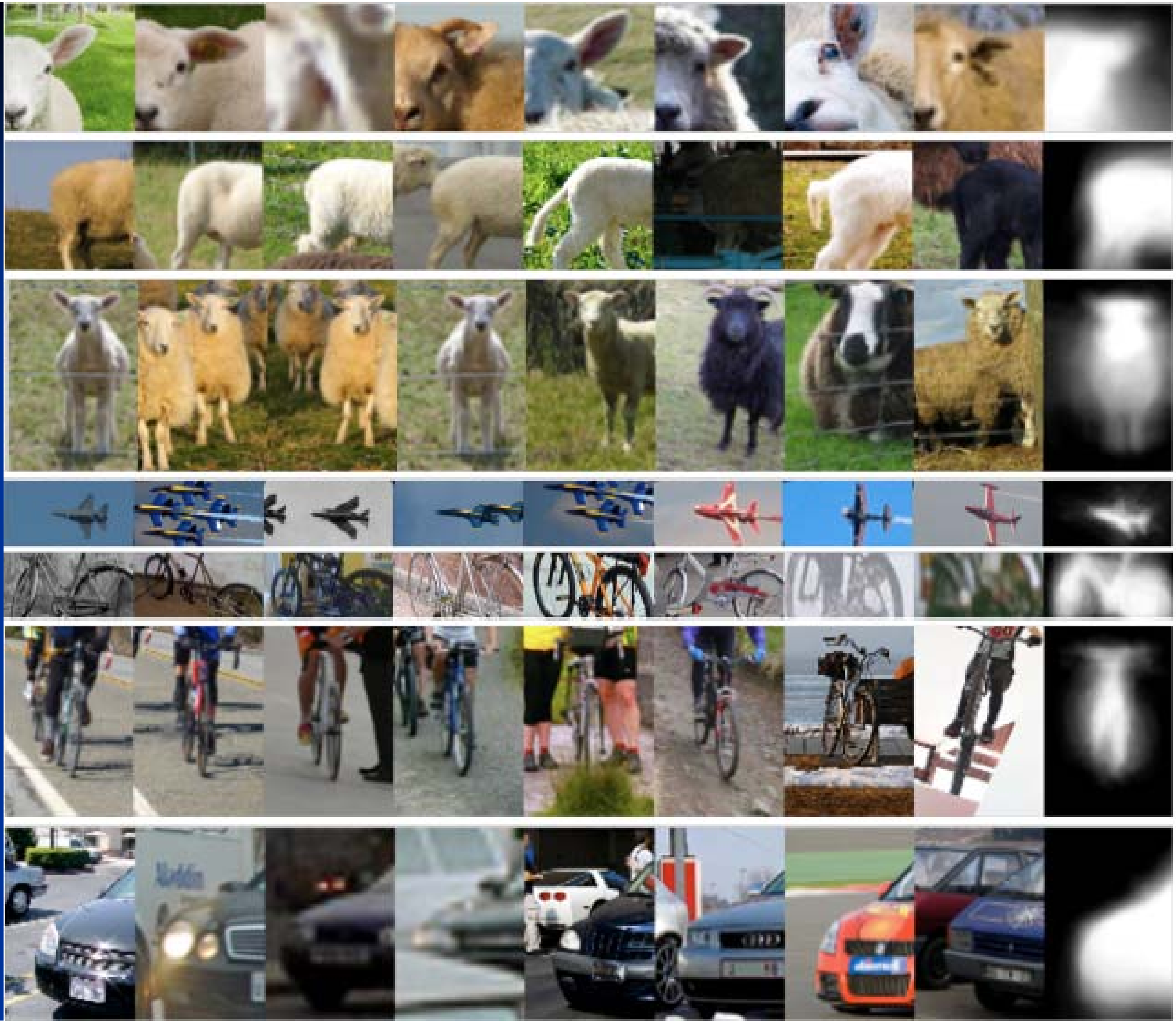
- Best results on all PASCAL person detection competitions

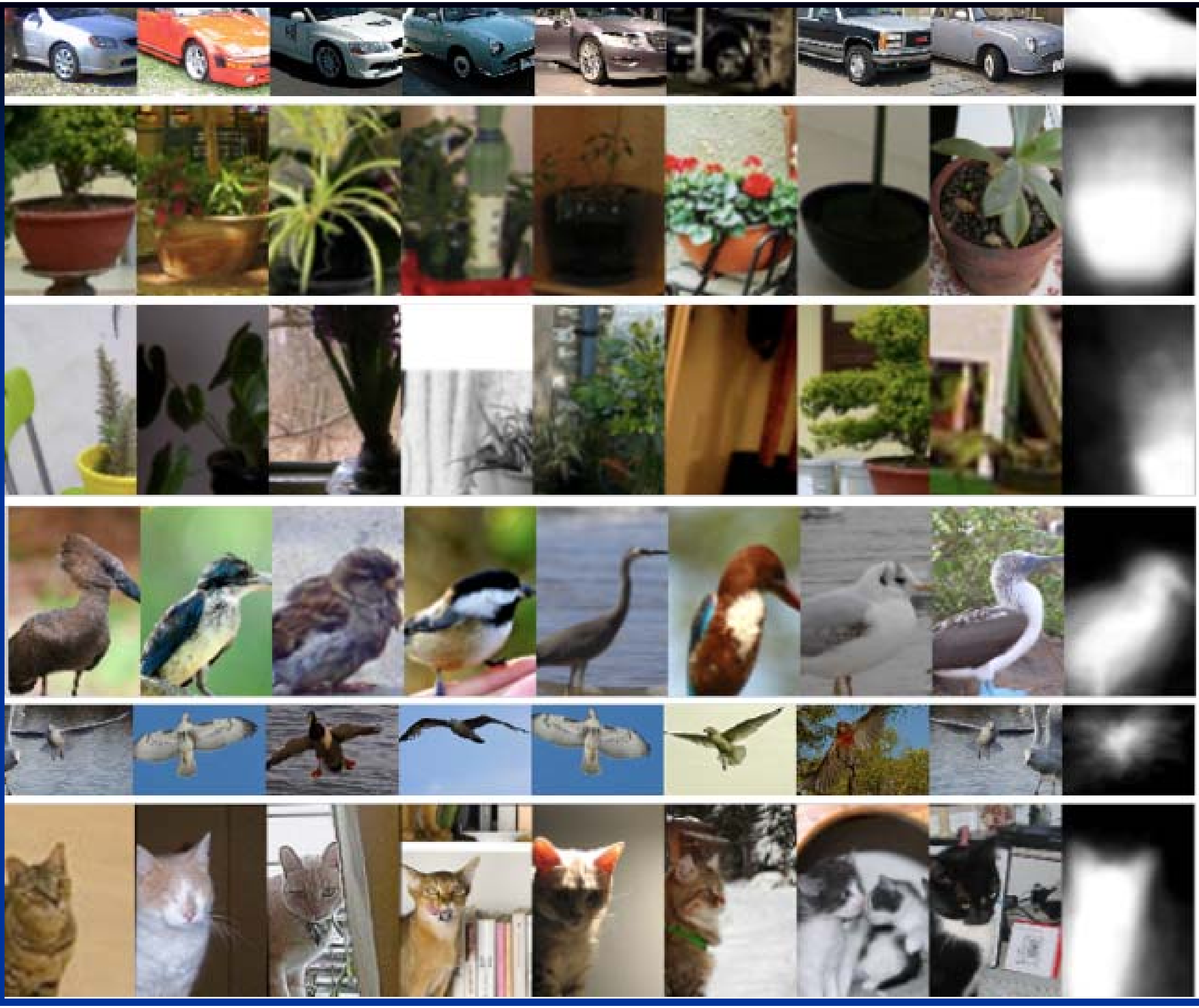
	POSELETS	Felzenszwalb et al.
2010	**48.5%**	47.5%
2009	**48.3%**	47.4%
2008	**54.1%**	43.1%

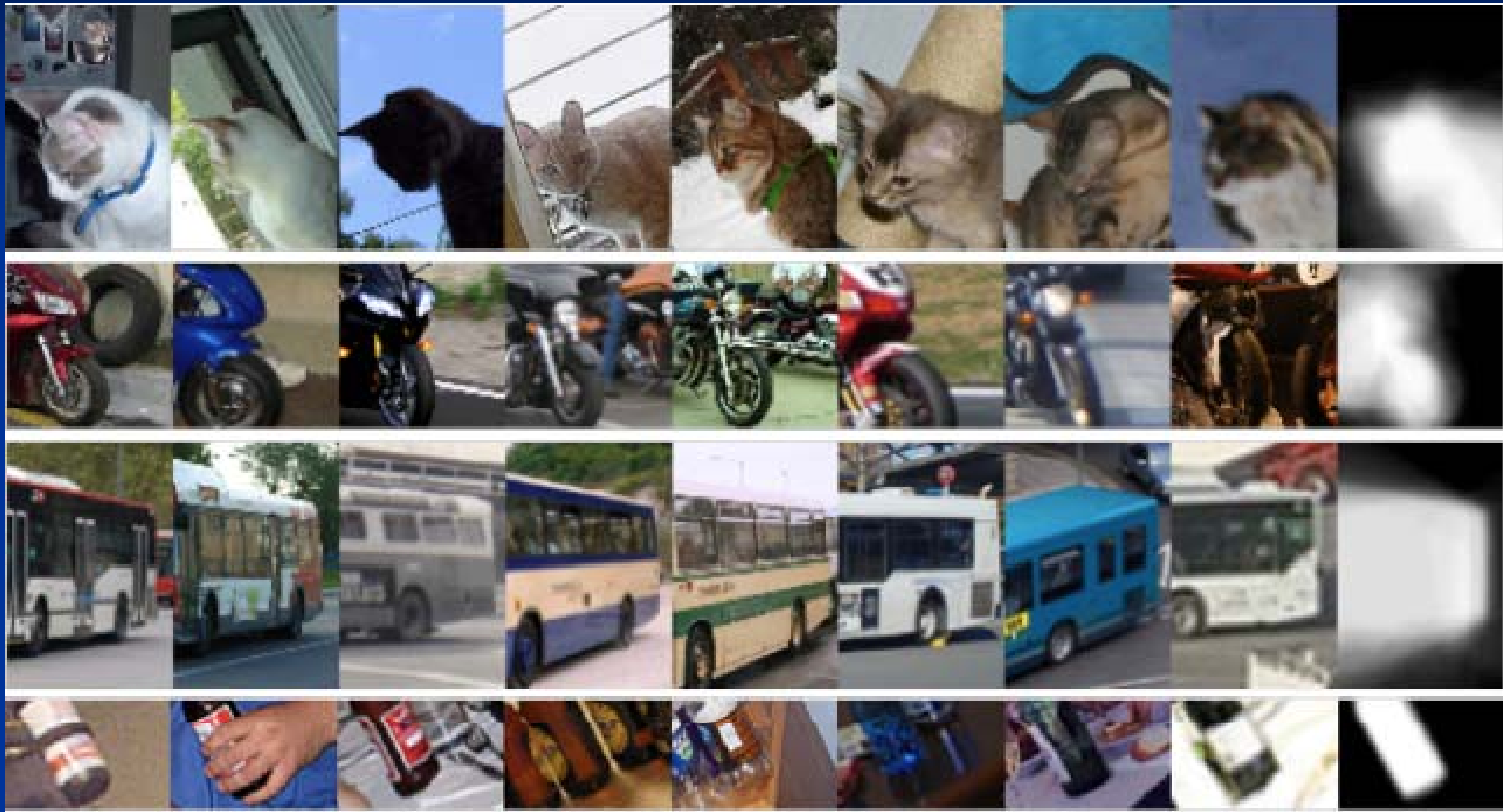
Agenda

- Poselets
  - Training a poselet
  - Selecting a good set of poselets
  - Improving poselets with context
  - Detection with poselets
- **Segmentation**
- Attributes
- Action Recognition

Segmenting people

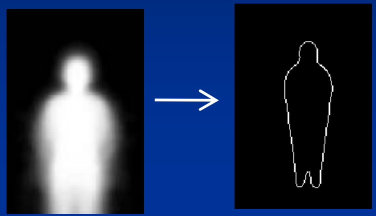


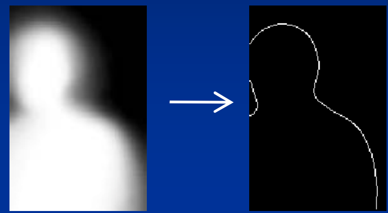




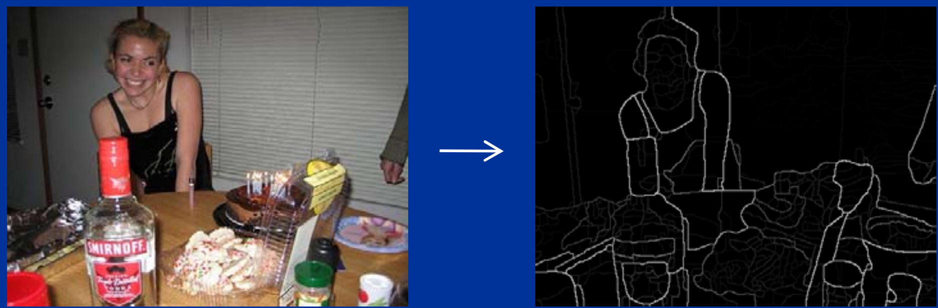
Align poselet activations (1 of 3)

- Threshold the mask of each poselet





- Make boundary map of the image (Arbelaez et al.)



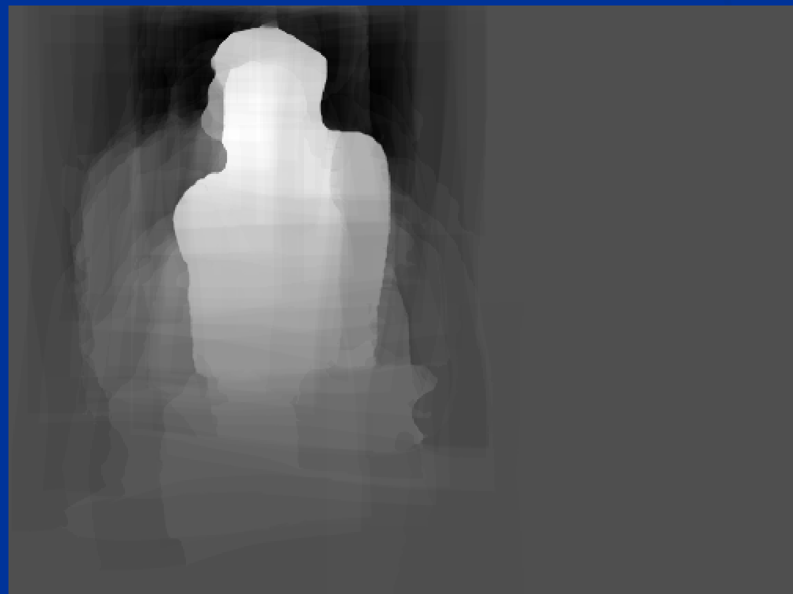
- Align the poselet activations using this non-rigid deformation:

$$E(u,v) = \int_{\mathbb{R}^2} |f(x,y) - g(x+u, y+v)| + \alpha \left(|\nabla u|^2 + |\nabla v|^2\right) dx dy.$$

Variational smoothing (2 of 3)

- The initial object mask $\tilde{M}$ is smoothed by taking into account the predicted object boundary C:

$$E(M) = \int (M - \tilde{M})^2 |\tilde{M}| + \frac{2}{C+1} |\nabla M| \, dx dy$$



Smoothed object mask

Refine via self-similarity (3 of 3)

Before refinement

After refinement

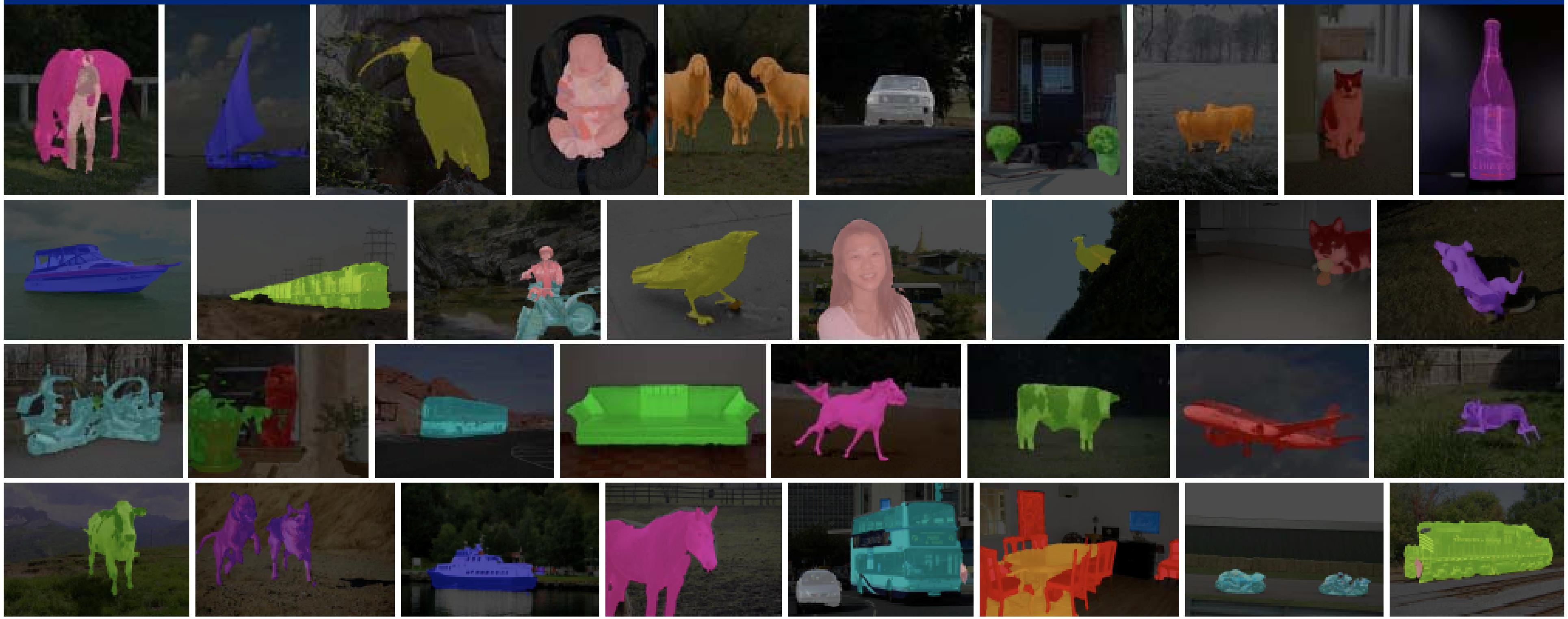
Multi-object segmentation

Person and horse

Multi-object segmentation

Person and bicycle

Some segmentation results…



Categories we are best in

	ours	Barce-lona	Bonn	Chicago/ Irvine	Oxford Brookes
background	82.2	81.1	84.2	80.0	70.1
aeroplane	43.8	58.3	52.5	36.7	31.0
bicycle	23.7	23.1	27.4	23.9	18.8
bird	30.4	39.0	32.3	20.9	19.5
boat	22.2	37.8	34.5	18.8	23.9
bottle	45.7	36.4	47.4	41.0	31.3
bus	56.0	63.2	60.6	62.7	53.5
car	51.9	62.4	54.8	49.0	45.3
cat	30.4	31.9	42.6	21.5	24.4
→ chair	**9.2**	9.1	9.0	8.3	8.2
cow	27.7	36.8	32.9	21.1	31.0
diningtable	6.9	24.6	25.2	7.0	16.4
dog	29.6	29.4	27.1	16.4	16.4
→ horse	**42.8**	37.5	32.4	28.2	27.3
motorbike	37.0	60.6	47.1	42.5	48.1
→ person	**47.1**	44.9	38.3	40.5	31.1
pottedplant	15.1	30.1	36.8	19.6	31.0
sheep	35.1	36.8	50.3	33.6	27.5
→ sofa	**23.0**	19.4	21.9	13.3	19.8
train	37.7	44.1	35.2	34.1	34.8
tvmonitor	36.5	35.9	40.9	48.5	26.4
average	34.9	40.1	39.7	31.8	30.3

Agenda

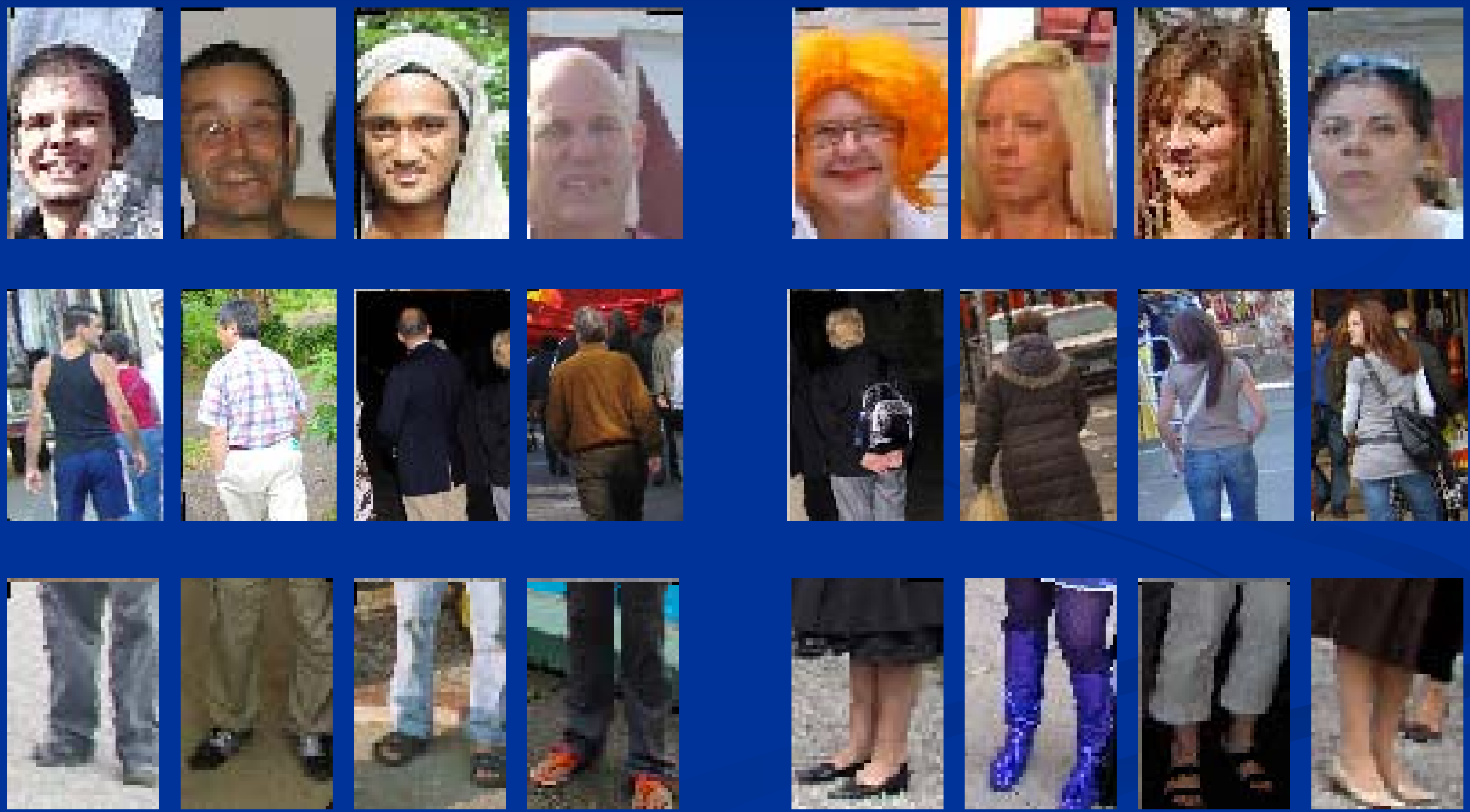
- Poselets
  - Training a poselet
  - Selecting a good set of poselets
  - Improving poselets with context
  - Detection with poselets
- Segmentation
- **Attributes**
- Action Recognition

Male or female?

How do we train attribute classifiers "in the wild"?

- Effective prediction requires inferring the pose and camera view
- Pose reconstruction is itself a hard problem, but we don't need perfect solution.
- We train attribute classifiers for each poselet
- Poselets implicitly decompose the pose

Gender classifier per poselet is much easier to train



Poselets: general-purpose pose decomposition engine. Can be used any time separating pose from appearance is important

Appearance is key for:

- Attribute classification

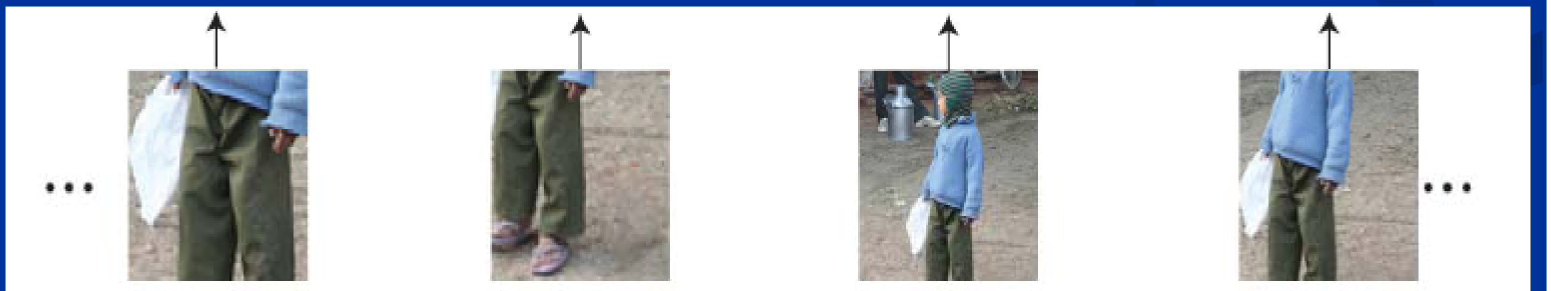
Pose is key for:

- Pose reconstruction
- Action recognition

Attribute Classification Overview

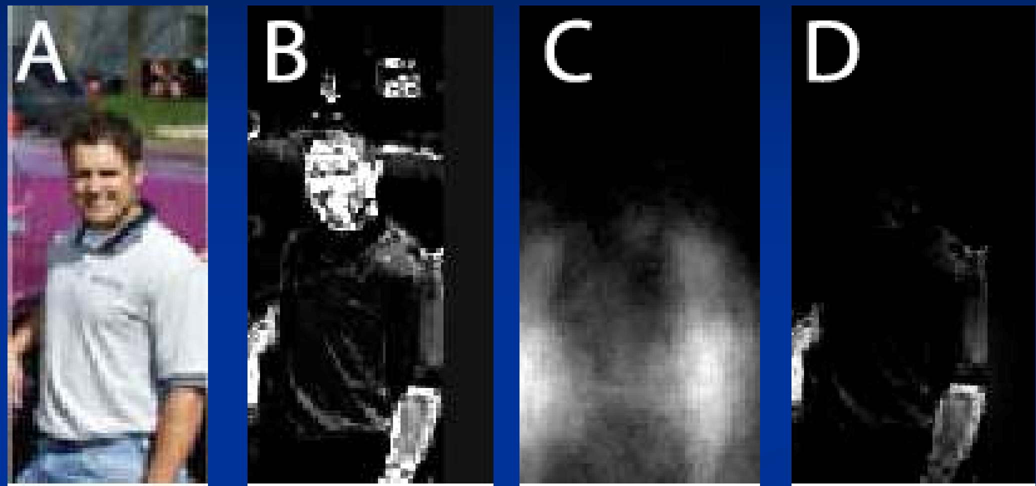
Given a test image

Poselet Activations



Features

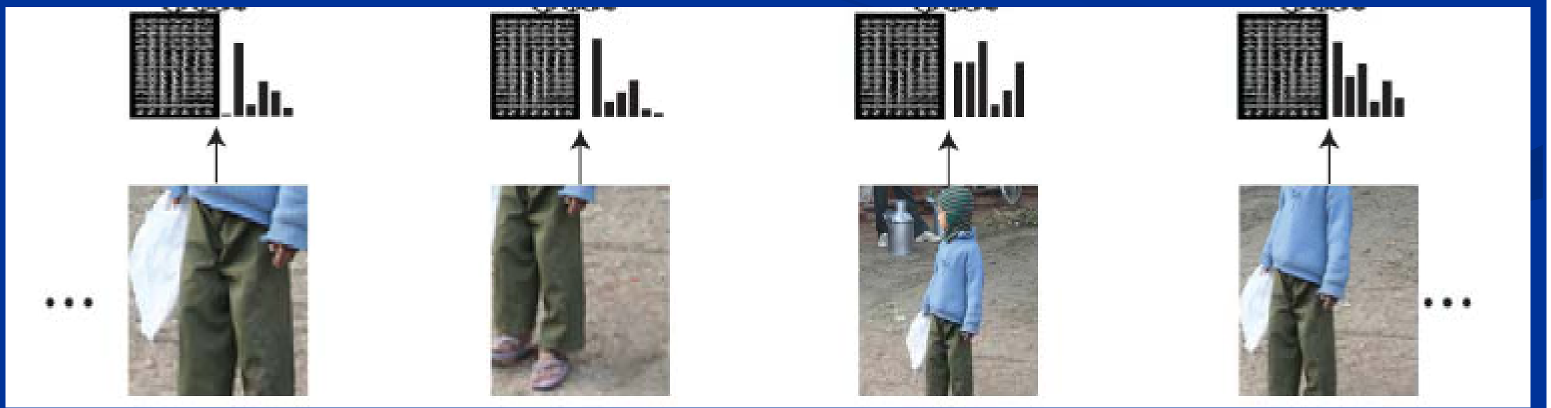
- Pyramid HOG
- LAB histogram
- Skin features
  - Hands-skin
  - Legs-skin



Poselet patch | Skin mask | Arms mask | B .* C

Features

Poselet Activations

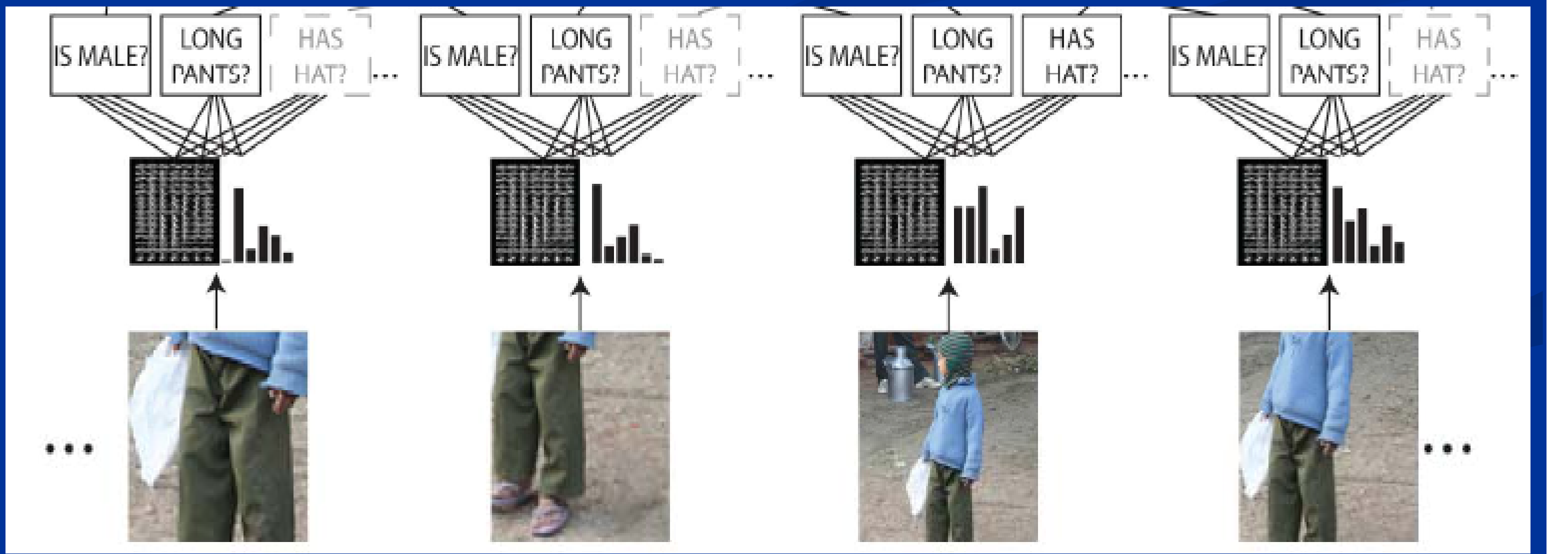


Attribute Classification Overview

Poselet-level Attribute Classifiers

Features

Poselet Activations



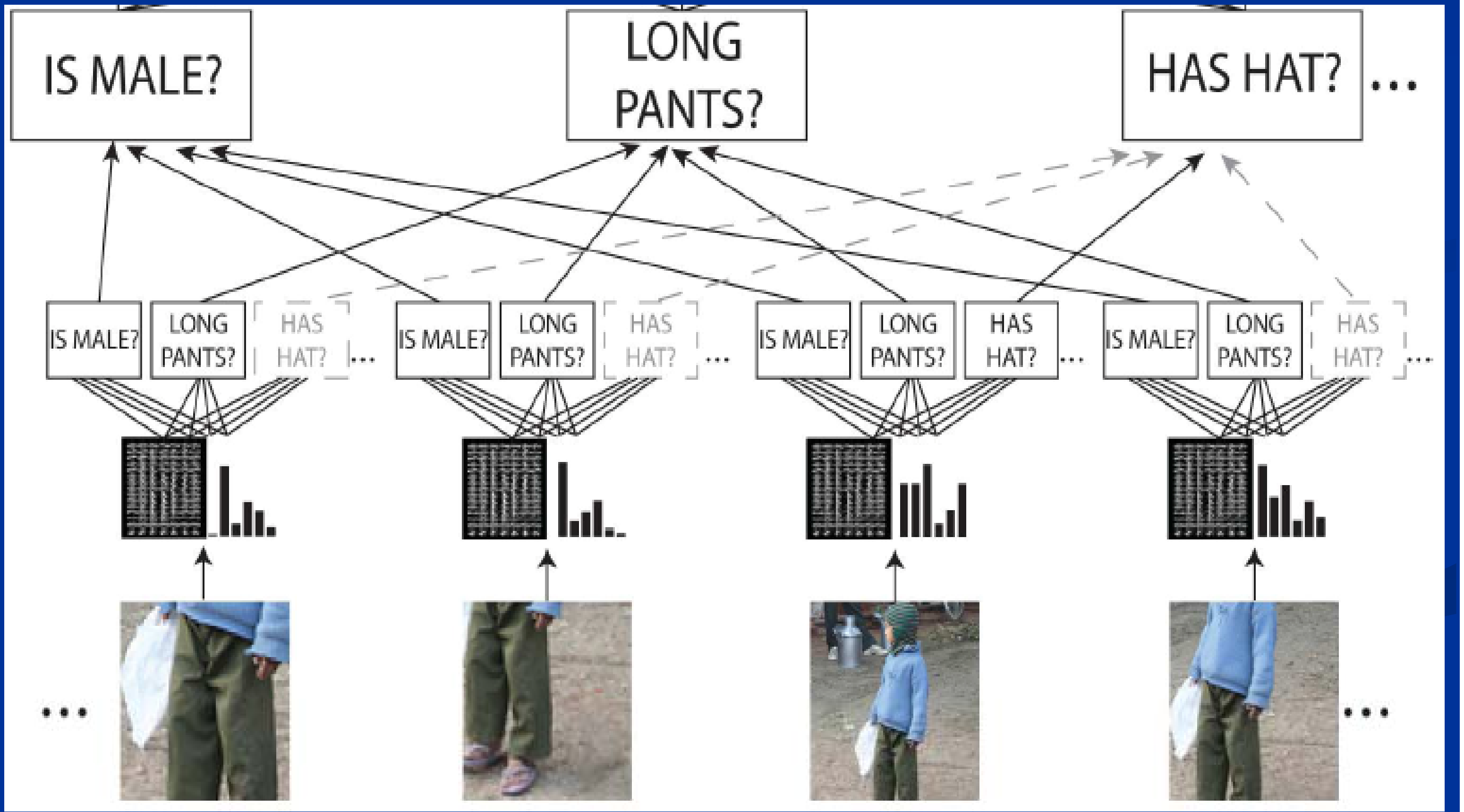
Attribute Classification Overview

Person-level Attribute Classifiers

Poselet-level Attribute Classifiers

Features

Poselet Activations



Attribute Classification Overview

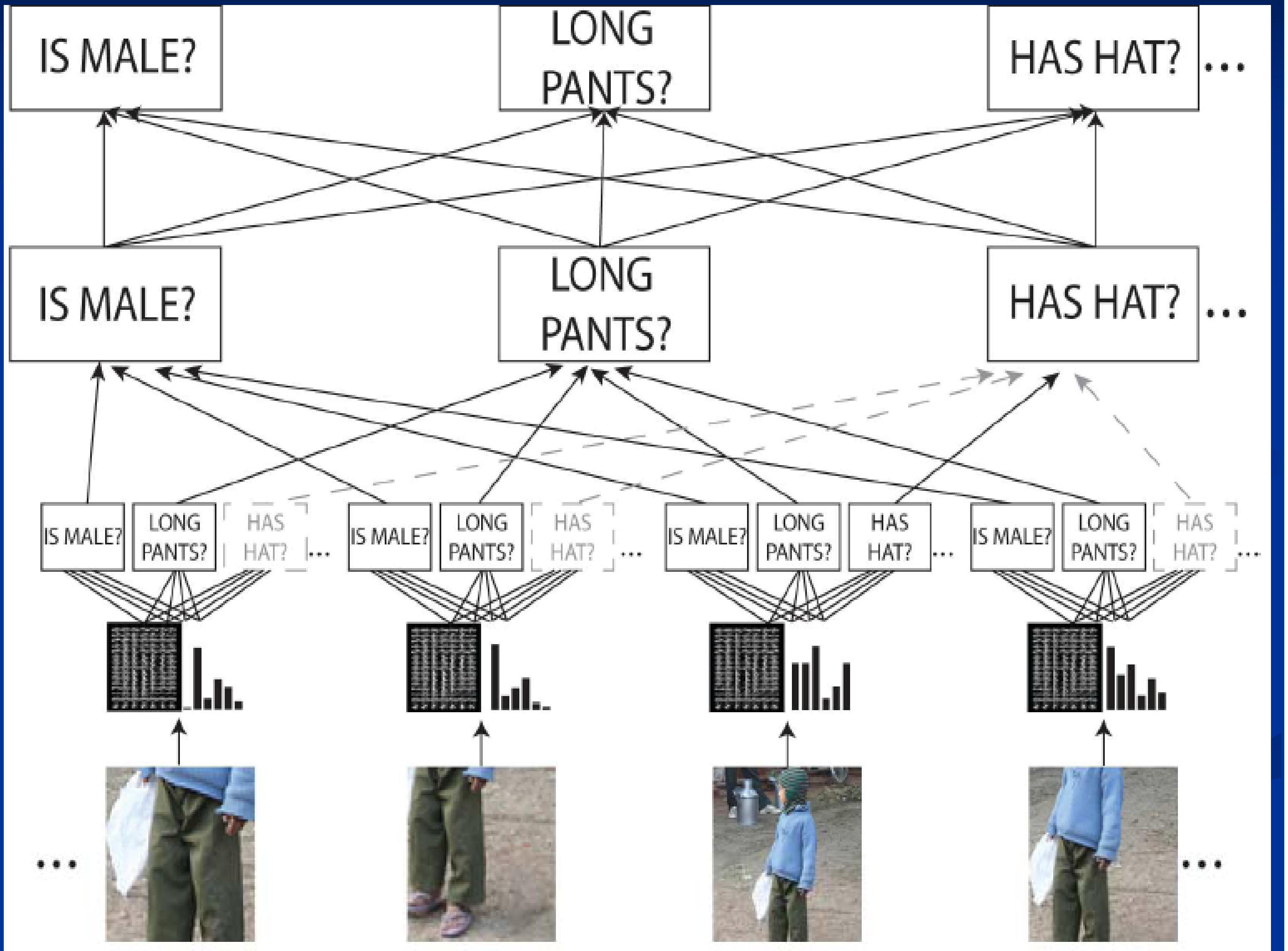
Context-level Attribute Classifiers

Person-level Attribute Classifiers

Poselet-level Attribute Classifiers

Features

Poselet Activations



Is male

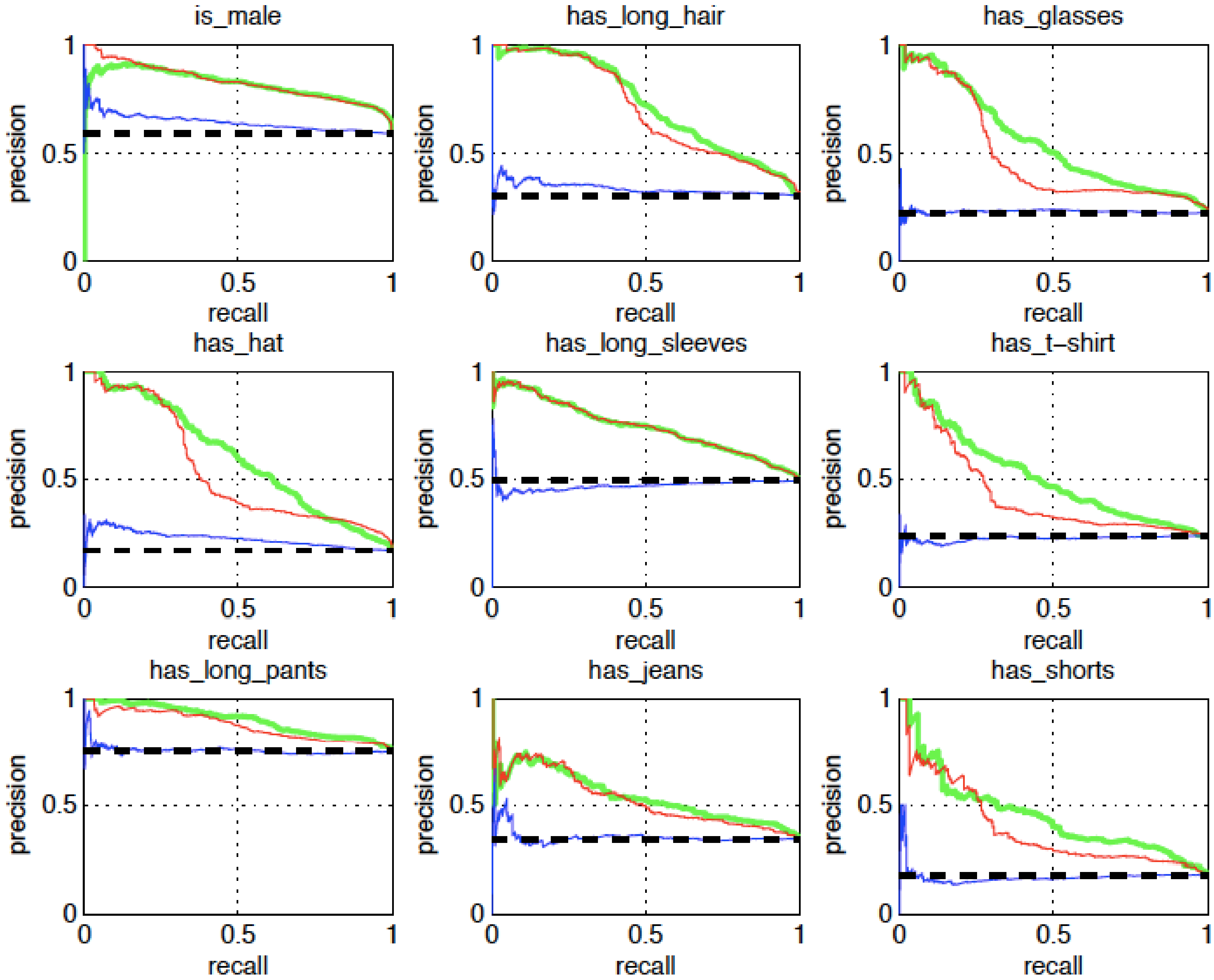
Has long hair

Wears a hat

Wears glasses

Wears long pants

Wears long sleeves

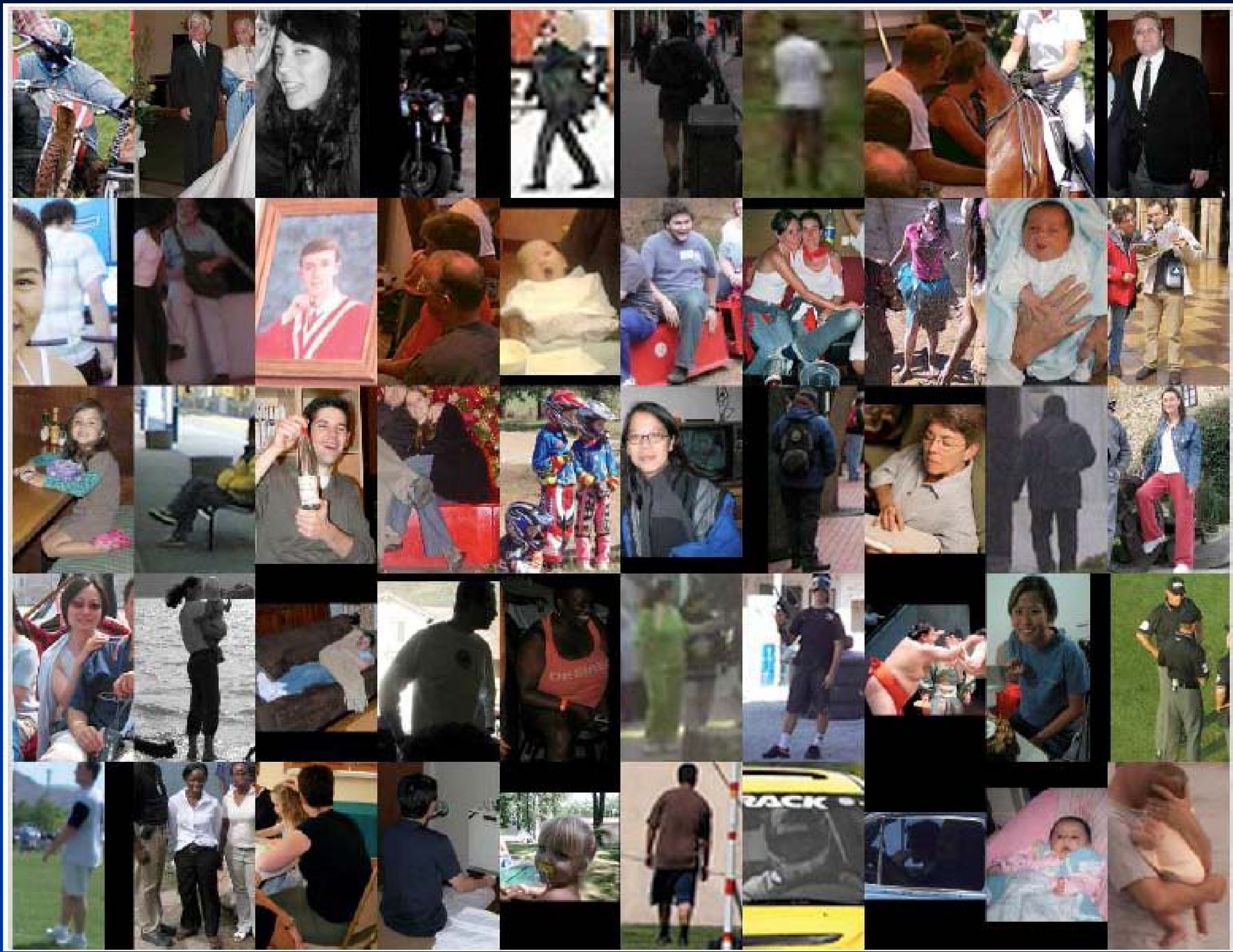


is_male
has_long_hair
has_glasses
has_hat
has_long_sleeves
has_t−shirt
has_long_pants
has_jeans
has_shorts
precision
recall
0
0.5
1

Results – Average Precision

Attribute	Freq	SPM	No cntxt	Cntxt
is male	59.3	64.8	82.9	82.4
has long hair	30.0	34.2	70.0	72.5
has glasses	22.0	23.6	48.9	55.6
has hat	16.6	22.6	53.7	60.1
has long sleeves	49.0	49.4	74.3	74.2
has t-shirt	23.5	23.9	43.0	51.2
has long pants	74.7	76.3	87.8	90.3
has jeans	33.8	36.4	53.3	54.7
has shorts	17.9	19.0	39.2	45.5
Mean AP	36.31	38.90	61.46	65.18

Random 2% of the test set



Agenda

- Poselets
  - Training a poselet
  - Selecting a good set of poselets
  - Improving poselets with context
  - Detection with poselets
- Segmentation
- Attributes
- **Action Recognition**

Actions in still images …

- have characteristic :
  - pose and appearance
  - interaction with objects and agents

PASCAL VOC 2010 Action Classification

- **Action Classification**: Predicting the action(s) being performed by a person in a still image. Bounding boxes are given

9 action classes

Relatively small training data/classes

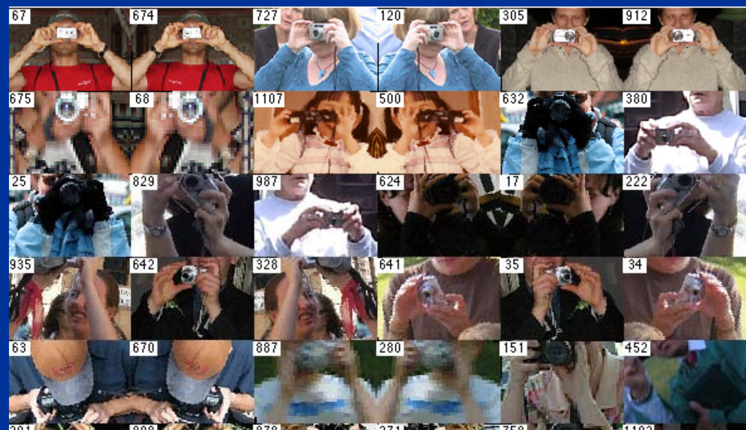
	train		val		trainval	
	Images	Objects	Images	Objects	Images	Objects
Phoning	25	25	25	26	50	51
Playinginstrument	27	38	27	38	54	76
Reading	25	26	26	27	51	53
Ridingbike	25	33	25	33	50	66
Ridinghorse	27	35	26	36	53	71
Running	26	47	25	47	51	94
Takingphoto	25	27	26	28	51	55
Usingcomputer	26	29	26	30	52	59
Walking	25	41	26	42	51	83
Total	226	301	228	307	454	608

Poselet selection and training

- Restrict training examples to ones from the category

takingphoto

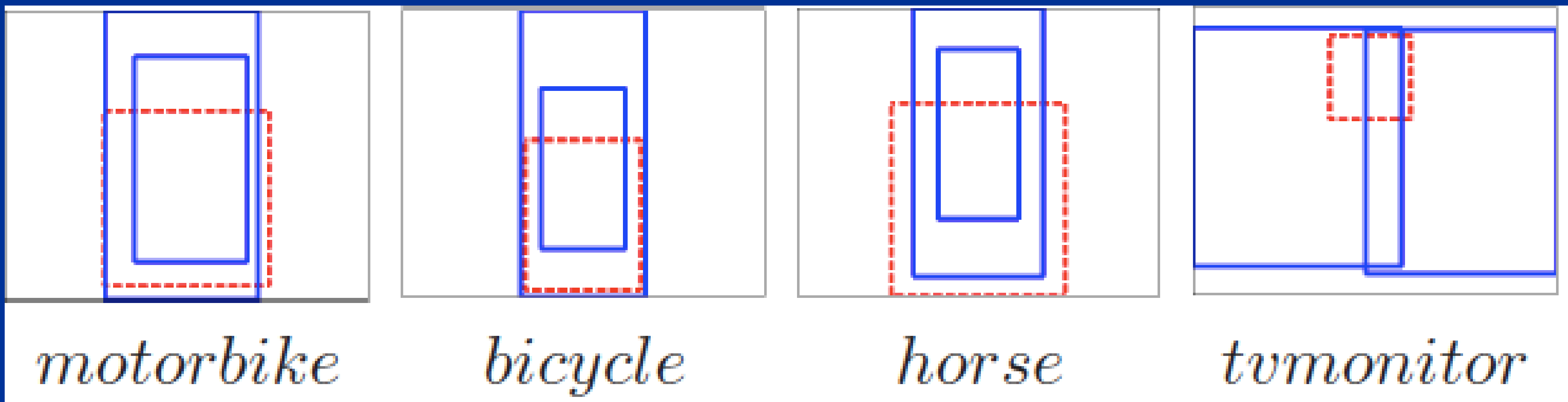
Examples from all actions



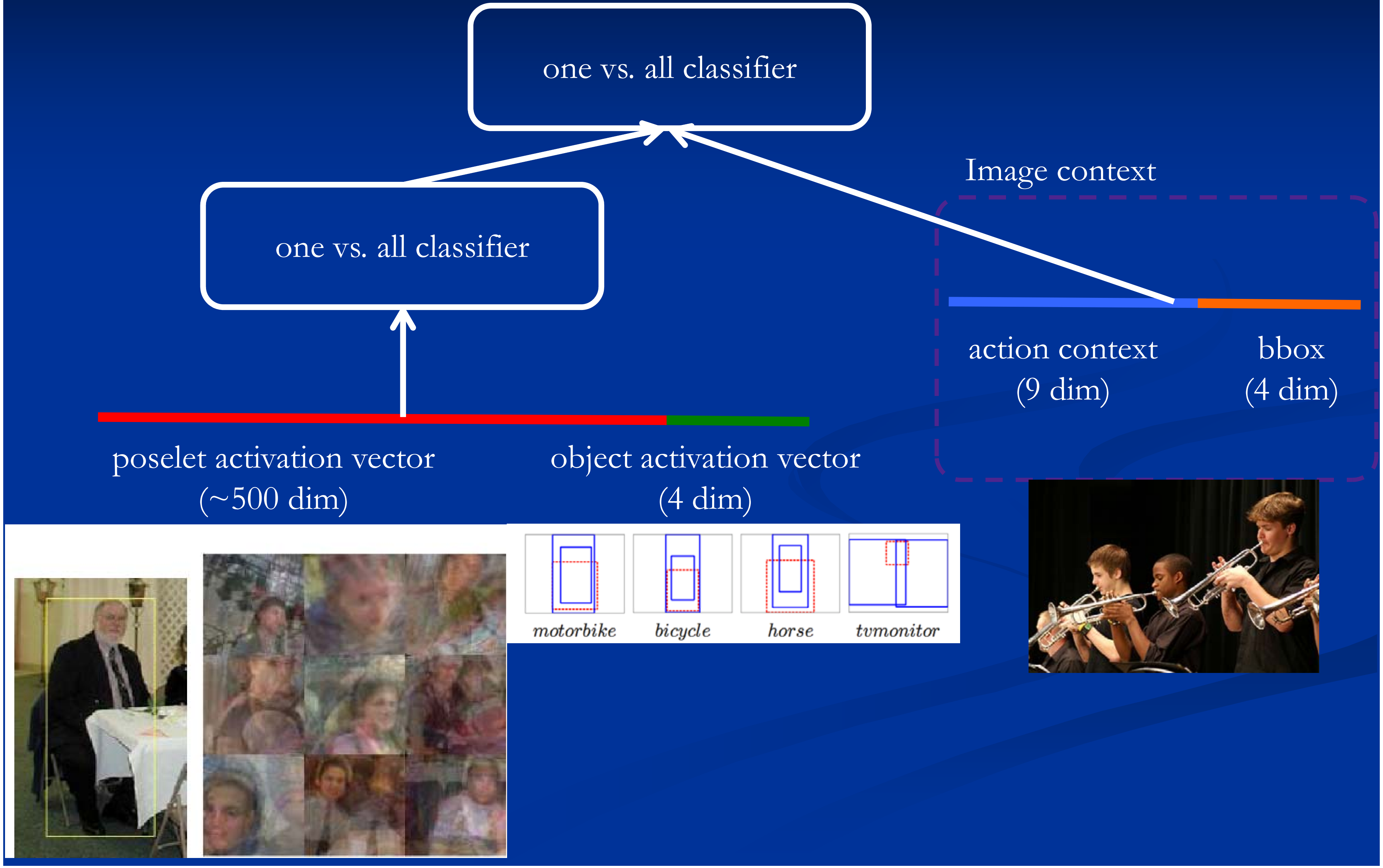
Examples from takingphoto

Some discriminative poselets

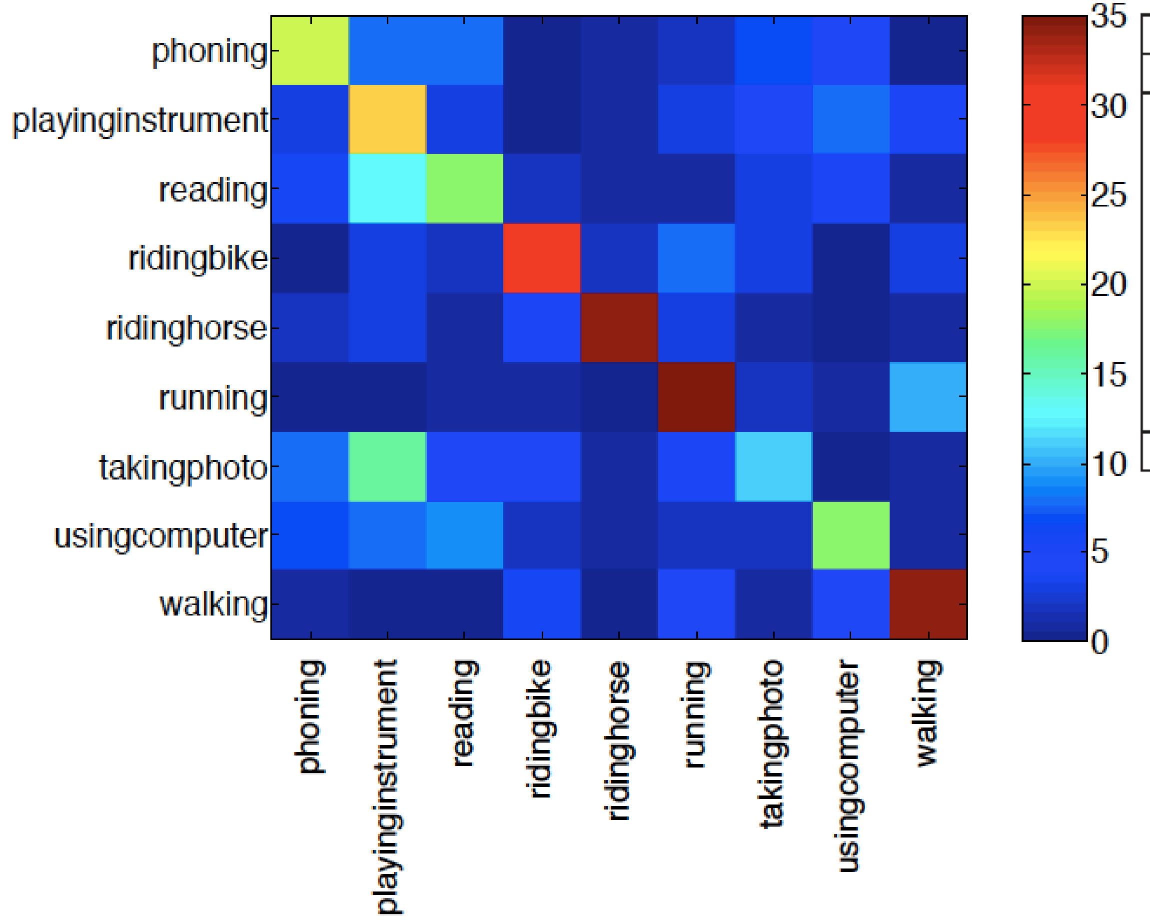
Spatial model of person-object interaction



Action classification

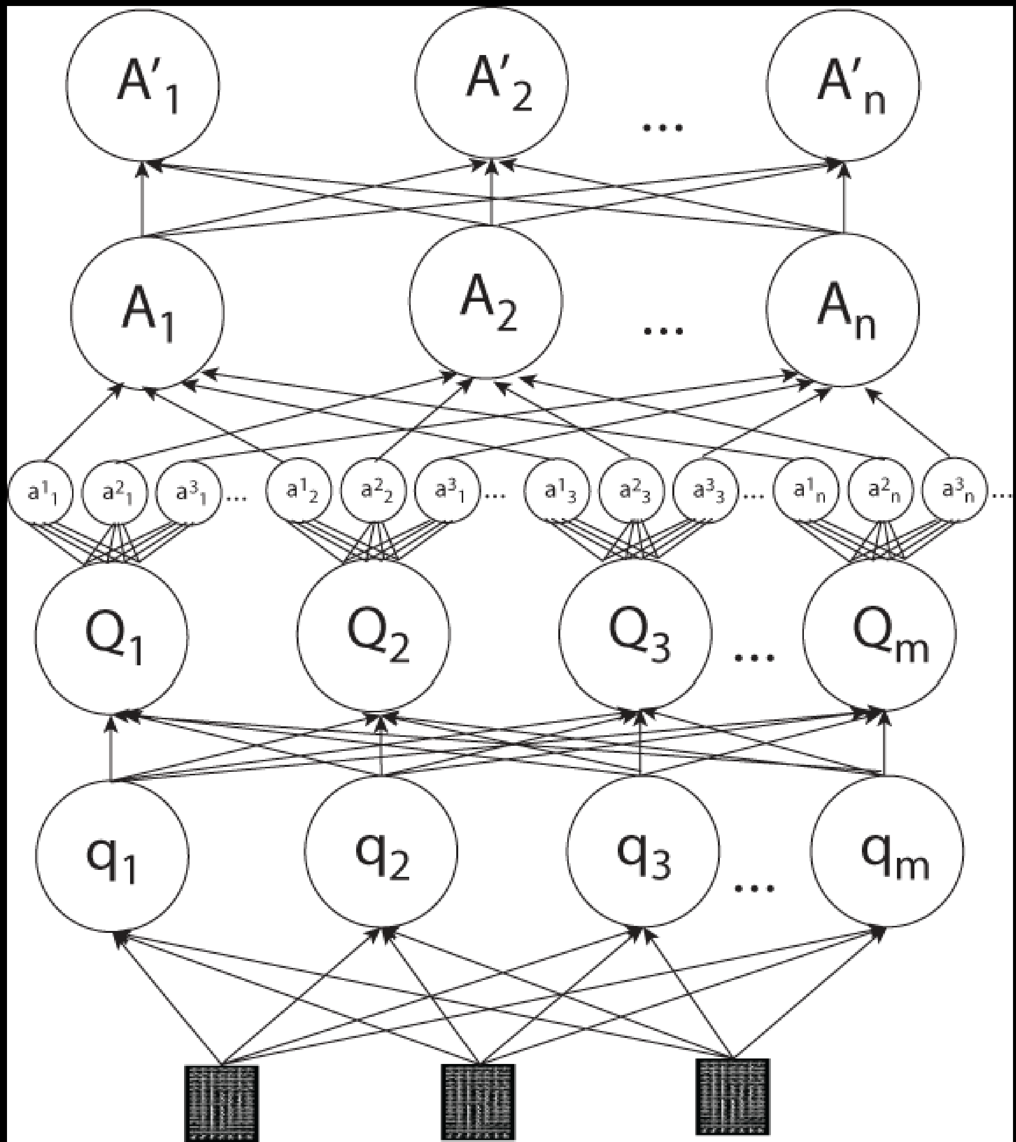


Results on static action classification



	Validation			Test
category	PAV	w/ OAV	w/ C	w/ C
phoning	57.85	57.99	57.63	50.55
playinginstrument	29.32	36.98	51.68	46.41
reading	42.20	45.99	44.55	25.10
ridingbike	61.92	86.46	87.09	83.10
ridinghorse	90.61	94.58	95.81	89.98
running	82.91	82.19	82.90	85.24
takingphoto	20.37	22.42	24.68	30.46
usingcomputer	31.45	53.13	49.63	59.19
walking	84.16	85.13	86.96	70.56
average	55.64	62.76	64.55	60.06

Feed-forward network

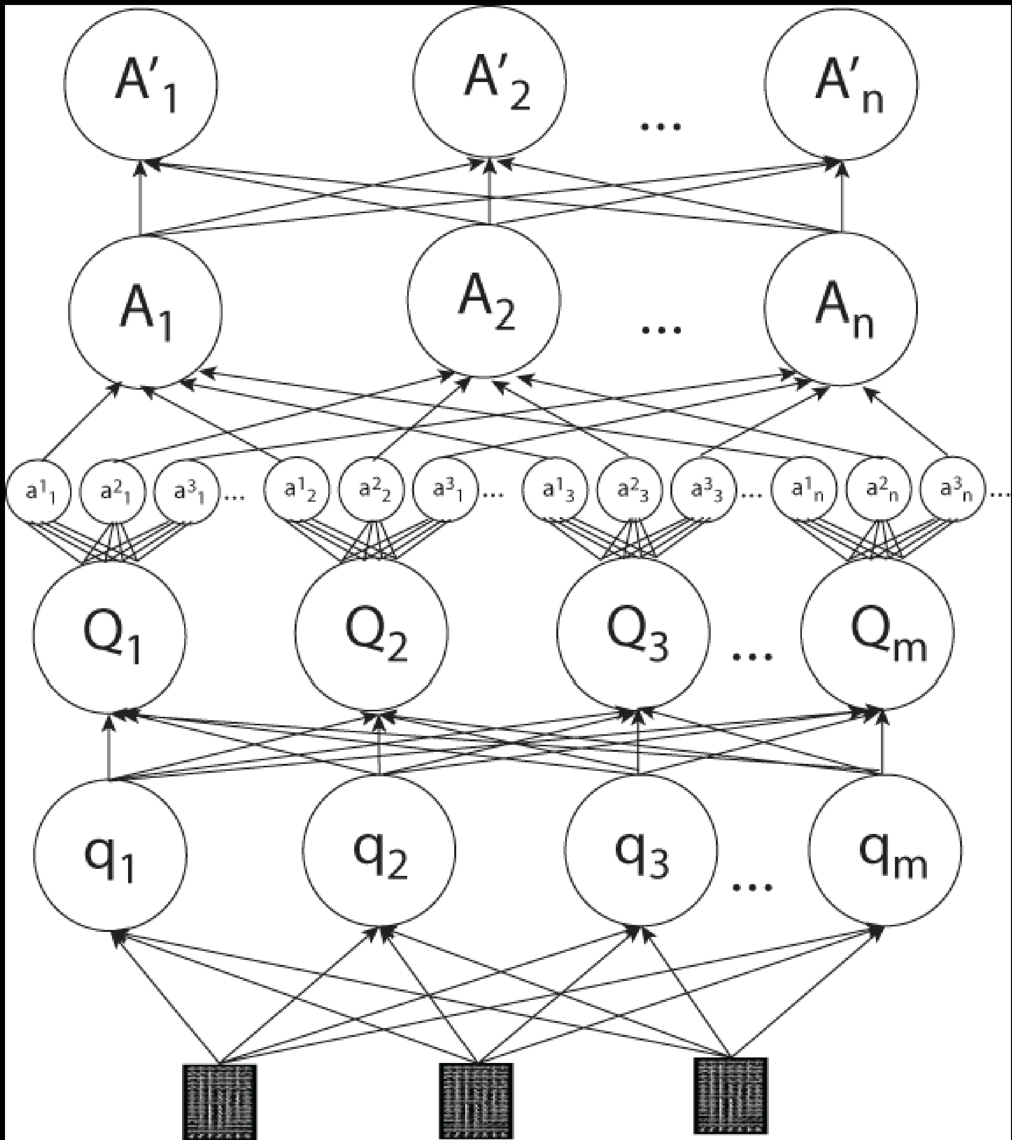


High-level questions:
“is this a woman?”
“is she running?”

Local pattern matching
“left half of head and shoulder”

Oriented gradients

Feed-forward network



Lots of context

No context

View independent

View specific

Poselets: general-purpose pose decomposition engine. Can be used any time separating pose from appearance is important

Appearance is key for:

- Attribute classification

Pose is key for:

- Pose reconstruction
- Action recognition