Poselets: Body part detectors trained using 3d human pose annotations (ICCV 2009)

Lubomir Bourdev and Jitendra Malik

Dinesh Jayaraman

Person detection results on the PASCAL challenge data

	Poselets	Second-highest score
VOC 2010	48.5	47.5 ***
VOC 2009	48.6	47.9 ***
VOC 2008	54.1	43.1 **
VOC 2007	46.9	43.2 *

* P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan Object Detection with Discriminatively Trained Part Based Models, (Release 4, 2010)

** P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan Object Detection with Discriminatively Trained Part Based Models, PAMI (preprint, 2009)

*** P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, PASCAL VOC 2010 competition

What are poselets?

- Poselets are discriminative parts - not necessarily semantic.

- Requirements:
  - Should tell us about the 3D pose
  - Should be easy to find from a 2D input image

What poselets look like

Slide credit: [Bourdev & Malik, ICCV09]

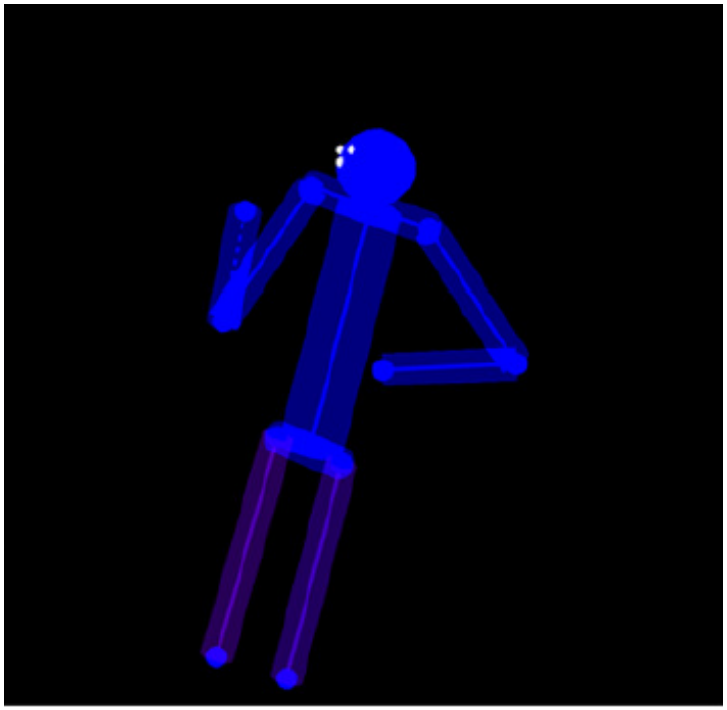
What are poselets?

- Poselets are discriminative “parts”, not necessarily semantic that describes parts of a pose.
- Requirements:
  - Should tell us about the 3D pose
  - Should be easy to find from a 2D input image

How do we enforce these requirements?

Configuration space

- Lost in transfer to 2D
- Fixed number of degrees of freedom
- Specified completely by positions of joints in a 3D coordinate space registered to the camera



"*Should tell us about the 3D pose*" = Tightly clustered in configuration space

Appearance space

- Pixel values i.e. the image itself
- Clothing, illumination, occlusion, background clutter etc.

"Should be easy to find from a 2D input image" =
Should be tightly clustered in appearance space

Poselets

Poselets capture part of the pose from a given viewpoint

Slide credit: [Bourdev & Malik, ICCV09]

Poselets

Examples may differ visually but have common semantics

Slide credit: [Bourdev & Malik, ICCV09]

Poselets

But how are we going to create training examples of poselets?

Slide credit: [Bourdev & Malik, ICCV09]

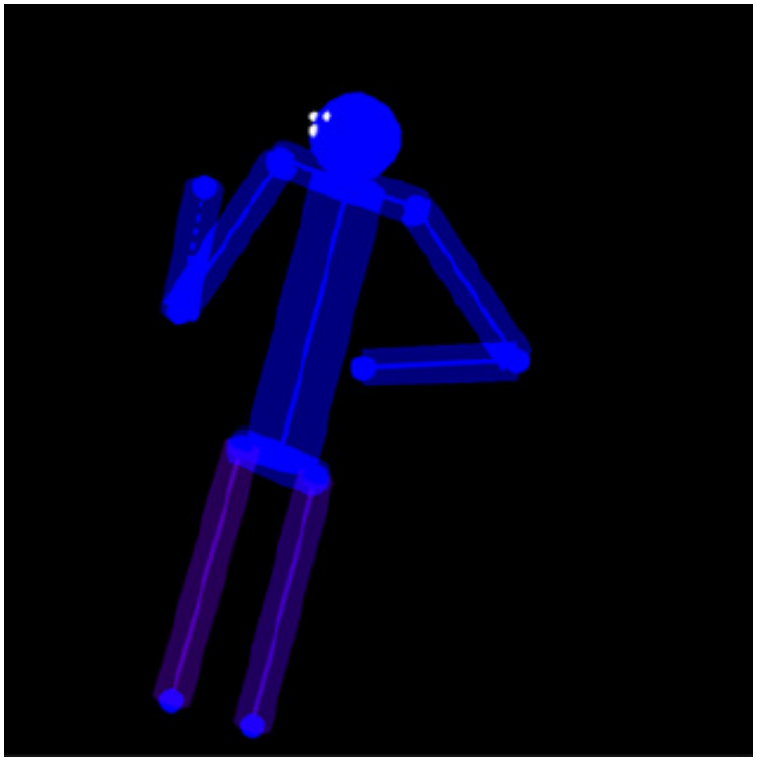
H3D dataset – humans in 3D

- 2000 annotated people
- 19 keypoint annotations

Appearance space annotation

H3D dataset – humans in 3D

- 2000 annotated people
- 19 keypoint annotations



Configuration space annotation

H3D dataset – humans in 3D

- 2000 annotated people
- 19 keypoint annotations
- 15 regions – “face”, “hair” etc.

Region label annotation

How do we train a poselet for a given ?

Slide credit: [Bourdev & Malik, ICCV09]

nces at training time

Given part of a human pose

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

Slide credit: [Bourdev & Malik, ICCV09]

Finding correspondences at training time

Slide credit: [Bourdev & Malik, ICCV09]

Finding correspondences at training time

Slide credit: [Bourdev & Malik, ICCV09]

Finding correspondences at training time

Residual Error

Sum over keypoints

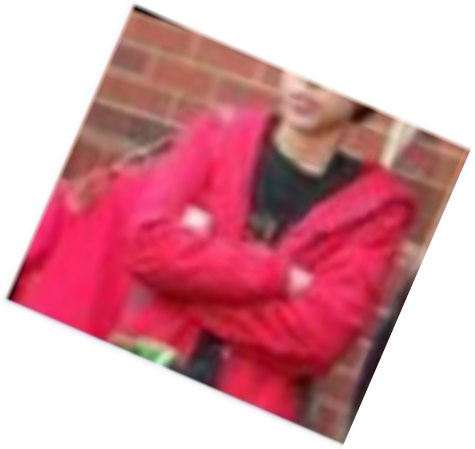
$$d_s(r) = \sum_i w_s(i)||\mathbf{x_s}(i) - \mathbf{x_r}(i)||_2^2(1 + h_{s,r}(i))$$

$$exp(-\mathbf{x_s}(i)^2/(2\sigma^2))$$

0 if visible or invisible in both

a otherwise

Registering candidate matches



Similarity transformation (4DOF):

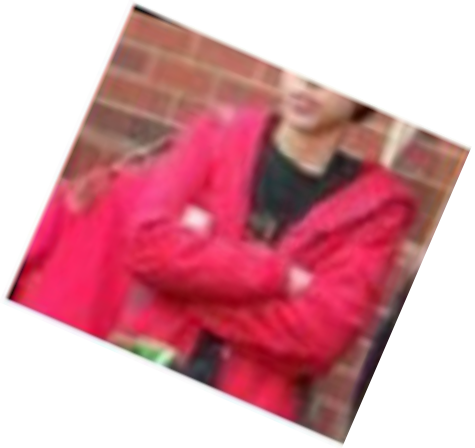
- X translation: t_x
- Y translation: t_y
- Rotation: α
- Scaling: s

Registering candidate matches

Seed s

$$d_s{}^*(r) = \min_{t_x,t_y,\alpha,s} d_s\left(r_{t_x,t_y,\alpha,s}\right)$$

$$x_{r_{t_x,t_y,\alpha,s}} = transform(x|t_x, t_y, \alpha, s)$$



Example r

$$d_s(r) = \sum_i w_s(i)||\mathbf{x_s}(i) - \mathbf{x_r}(i)||_2^2(1 + h_{s,r}(i))$$

Matching done in 3D => clustered in configuration space

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

Slide credit: [Bourdev & Malik, ICCV09]

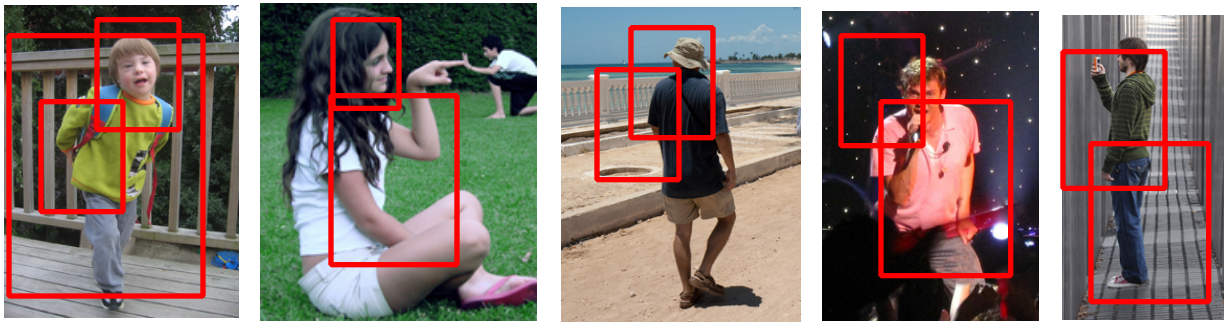
Training poselet classifiers

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 to train a **linear SVM with HOG features**

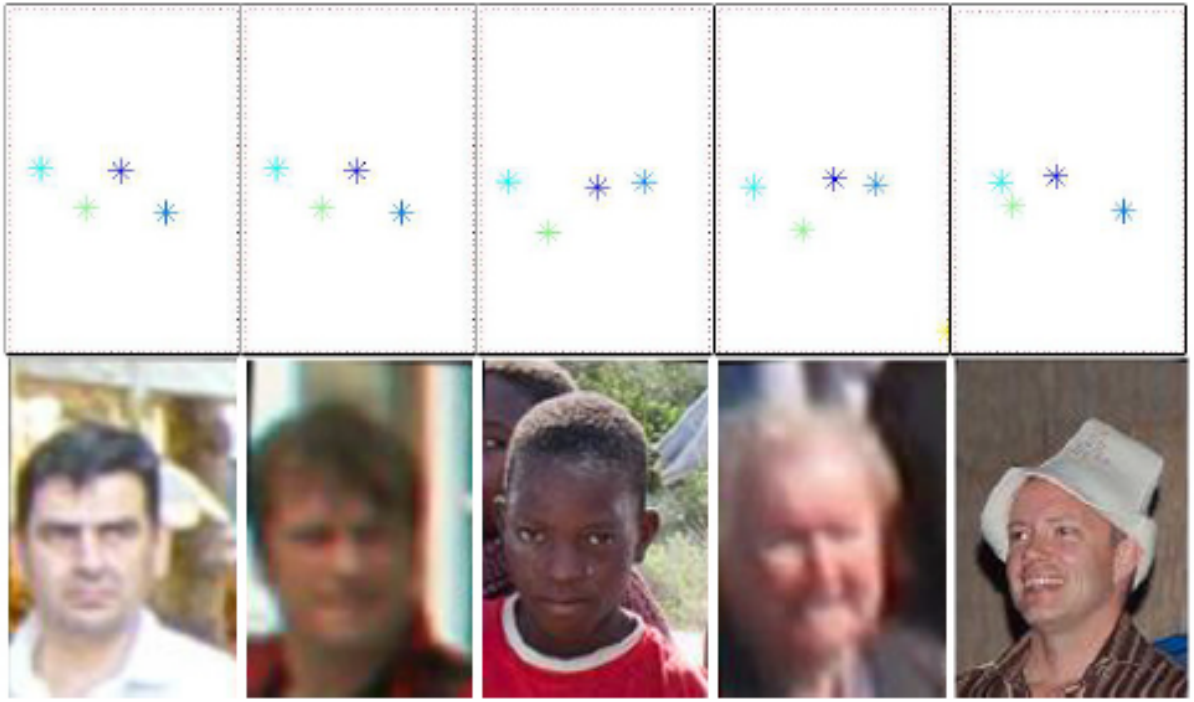
Slide credit: [Bourdev & Malik, ICCV09]

Which poselets should we train?

- 96x64 scanning window over all scales and orientations on 1500 training images.
- 120K poselet candidates generated.
- Strict 2-stage pruning (300 poselets):
  - Individually effective (good cross-validation)
  - Complementary (large pairwise distances)



Selected poselets

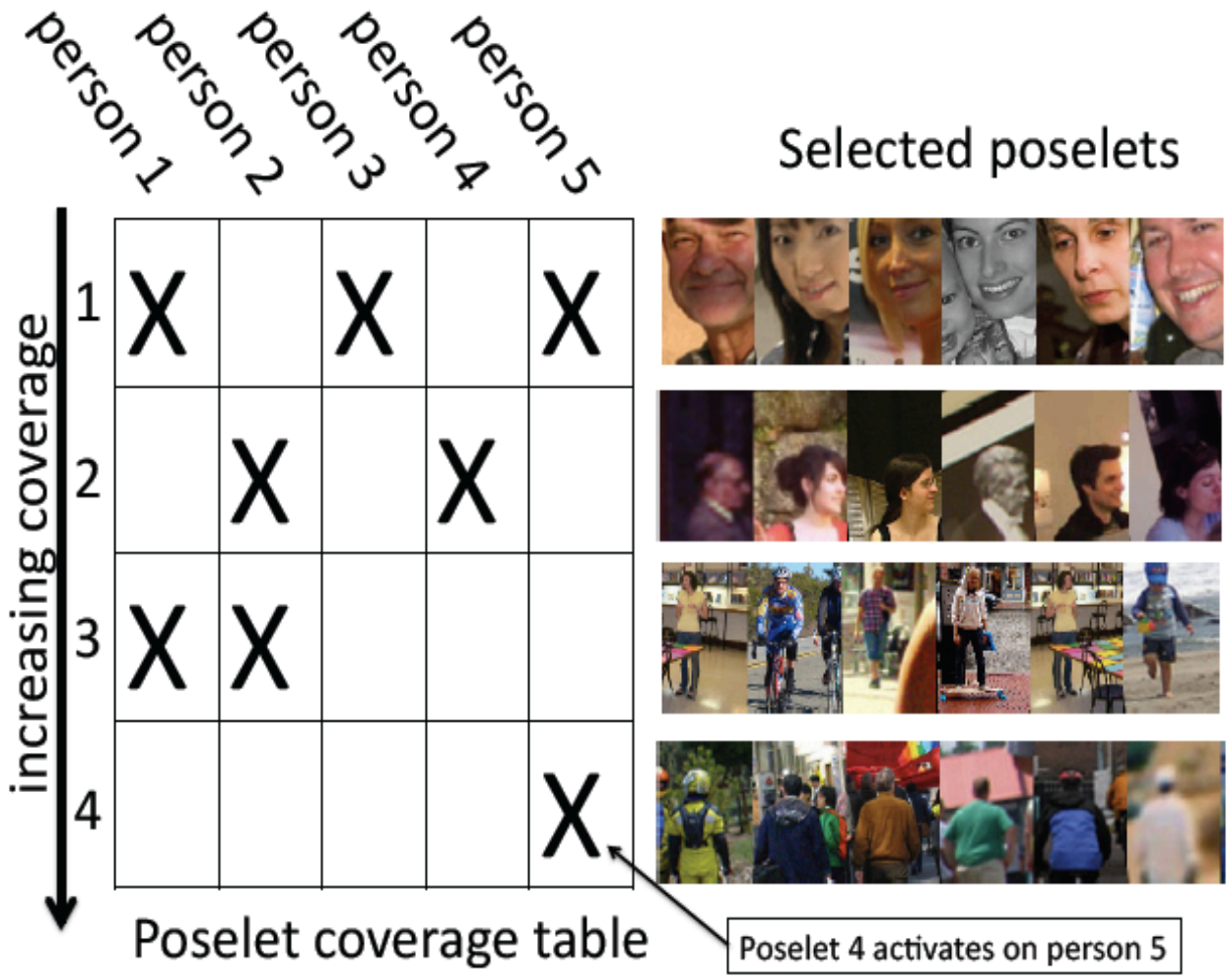


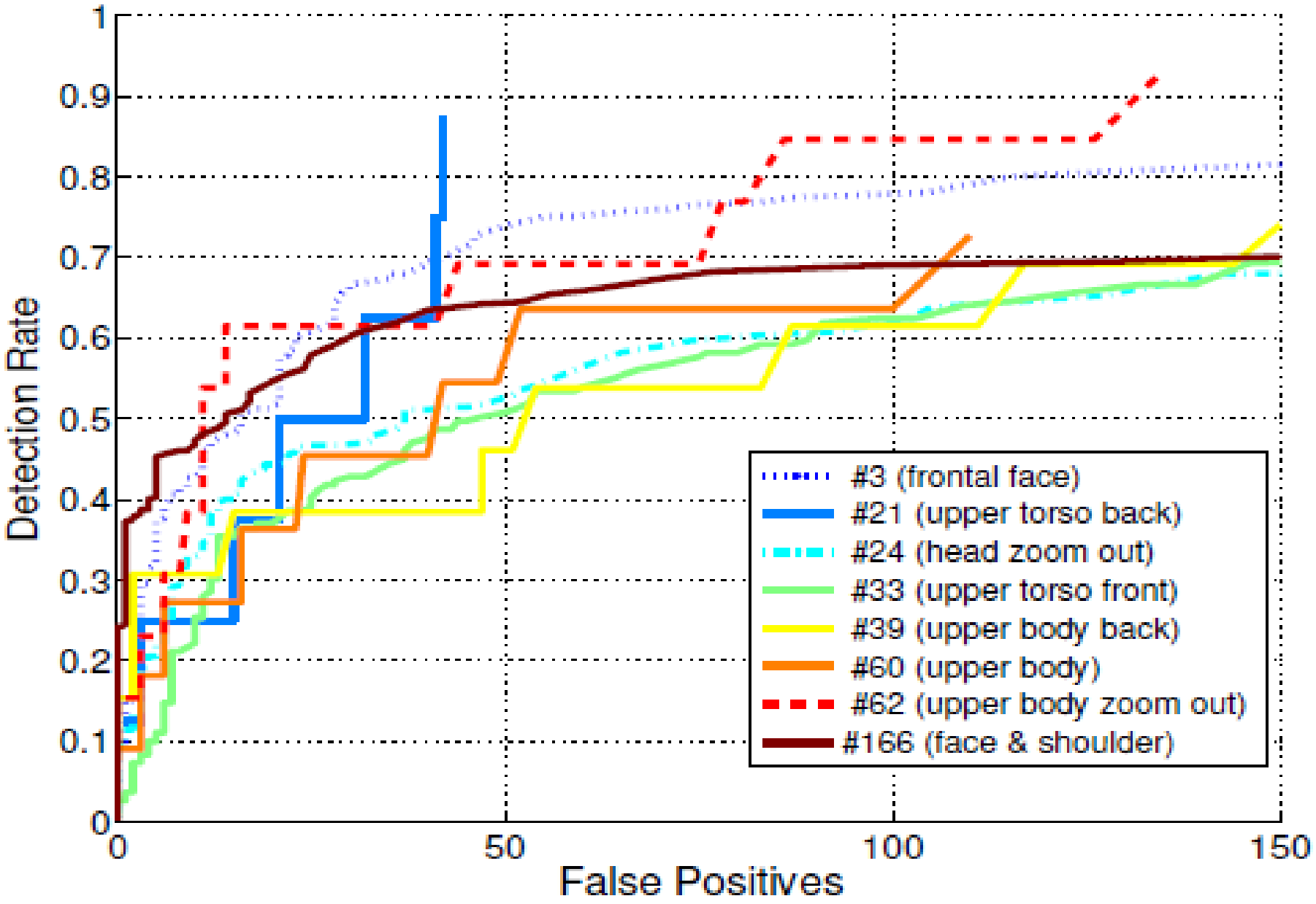
Configuration space

Examples

A frontal face poselet

Selected poselets





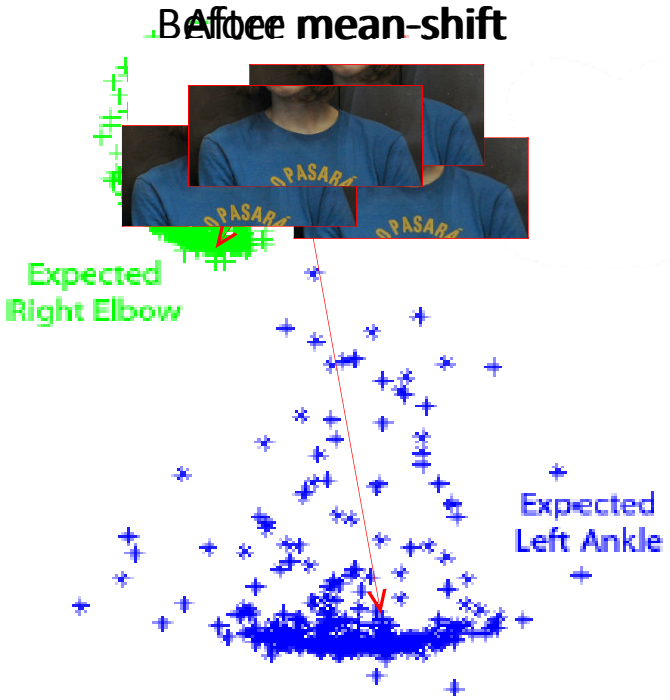
Poselets in isolation – “individually effective”

Discussion

- Why do upper body poselets do better?
- Why do we want complementarity?

Detecting torsos and localizing keypoints – The Hough transform

- H3D can find expected keypoint location given poselet detection.
- Poselet detector in scanning window, mean-shift for non-maximum suppression



Discussion

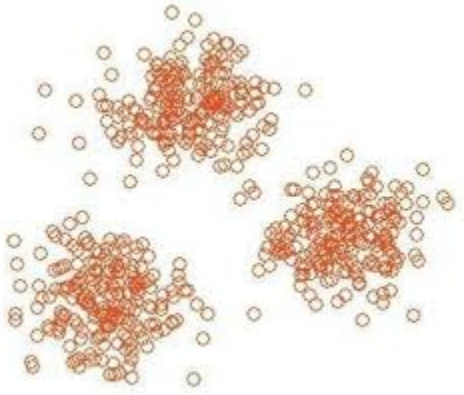
- What is the dimension of the voting space?
- Does a poselet predict *all* keypoints equally well?
- Are all poselets equal?

Combining poselet predictions

$$P(O|x) \propto \sum_i w_i a_i(x)$$

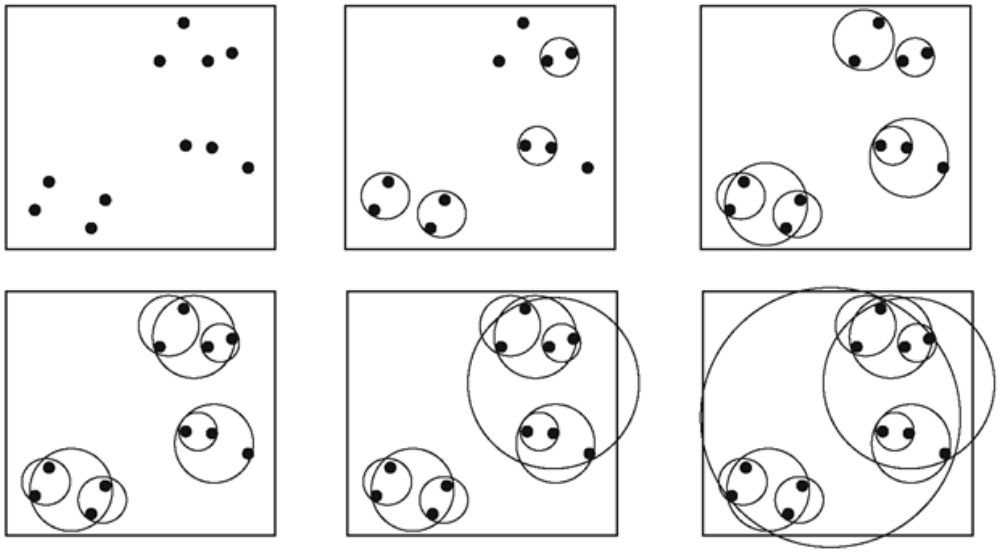
Poselet index

Poselet i vote for O being at x



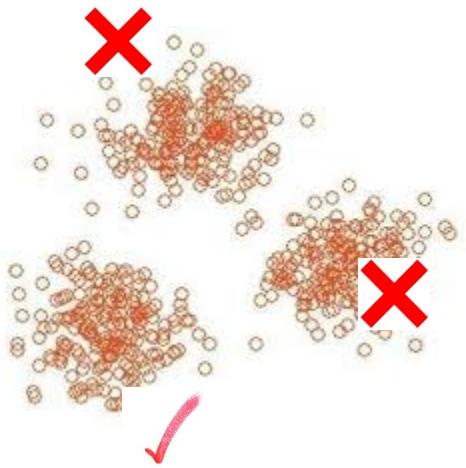
2D Hough space (keypoint)

Agglomerative clustering in Hough space



Learning task-specific weights: Max-Margin Hough transform

- SVM-like formulation that classifies true-positive clusters versus false-positive clusters using poselet votes as descriptors:



$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{T} \xi_i$$

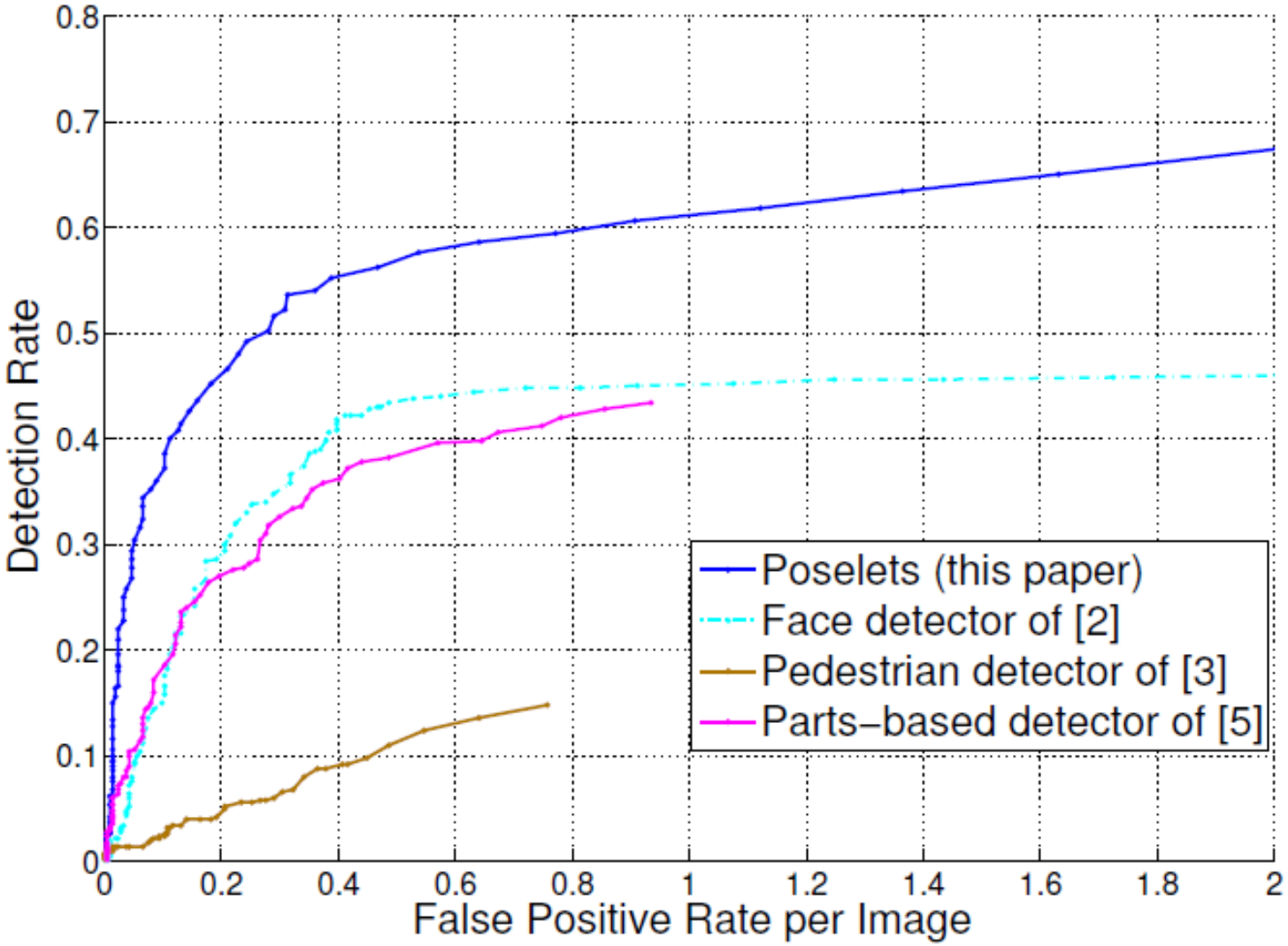
$$\text{s.t. } y_i (w^T A_i + b) \geq 1 - \xi_i$$

$$w \geq 0, \xi_i \geq 0, \forall i = 1, 2, ..., N$$

1 if true positive peak, 0 else

Score of poselet j in Hough peak i

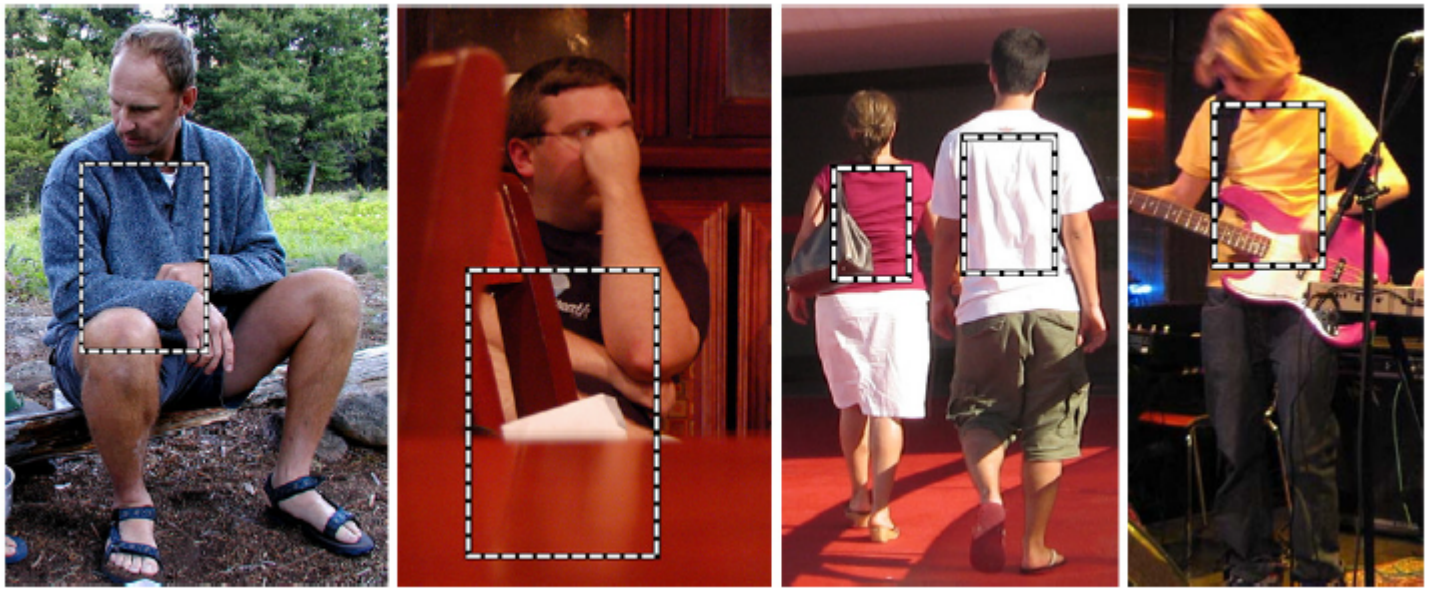
Results – Torso detection



Key strengths

- Discussion: Why does this method do so much better than the next best?
  - Tightly controlled poselet training and selection process enabled by H3D
  - Large number of poselets guaranteed to be semantically same vs. stick-figure approach
  - Lots of occlusions in test set => Hough transform based voting is advantageous

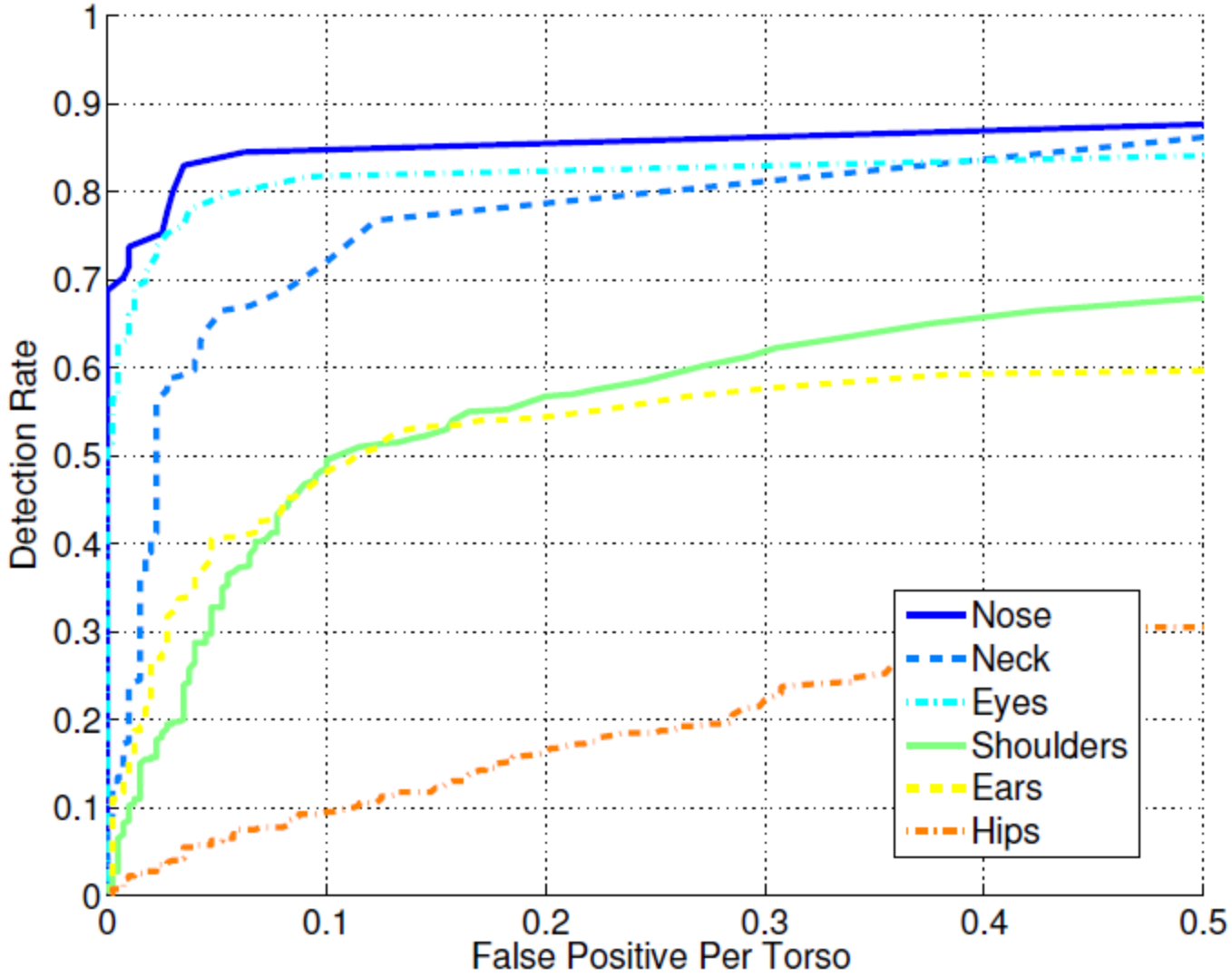
Results - Torso detection



Pascal VOC person detection

	Poselets	Second-highest score
VOC 2010	48.5	47.5 ***
VOC 2009	48.6	47.9 ***
VOC 2008	54.1	43.1 **
VOC 2007	46.9	43.2 *

Keypoint detection



Discussion

- Why are the nose, eyes and neck predicted best?
- A 2D projection can represent more than one 3D configuration - metamers. Are metamers really handled at all by the poselets method?
- Did we really collect 3D data?
- What advantage did 3D really give us?
- Can we do without 3D annotations?

Poselets website

http://eecs.berkeley.edu/~lbourdev/poselets

The set of published poselet papers
H3D data set + Matlab tools
Java3D annotation tool + video tutorial
Matlab code to detect people using poselets
Latest trained poselets: http://www.cs.berkeley.edu/~lbourdev/poselets/poselets_person.html

Slide credit: [Bourdev &