Detecting Actions, Poses, and Objects with Relational Phraselets

Chaitanya Desai Deva Ramanan
UC Irvine

Detecting Actions, Poses, and Objects with Relational Phraselets

Chaitanya Desai Deva Ramanan
UC Irvine

K-way action classification

walking
riding-bike
jumping
phoning
taking-picture
using computer

K-way action classification

walking
riding-bike
jumping
phoning
taking-picture
using computer

Requires bounding-box annotation on test images

What's wrong with K-way classification?

Detecting and localizing people performing actions is challenging

What’s wrong with K-way classification?

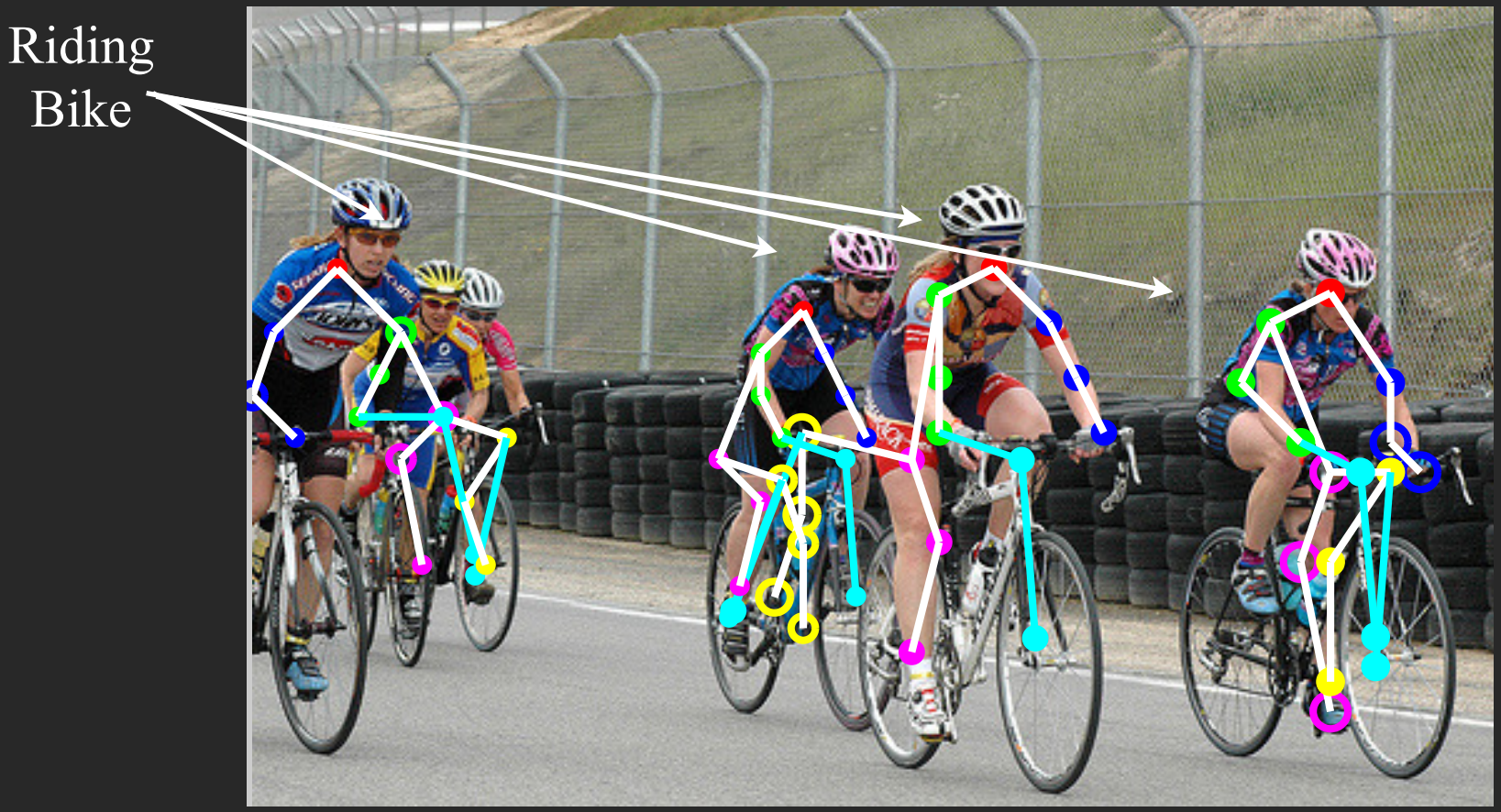
Ignores the complexity of the ways in which people + objects can interact

Our goal: detailed action understanding

Our goal: detailed action understanding

1. Localize person (+ interacting object)

Our goal: detailed action understanding



1. Localize person (+ interacting object)
2. Classify action of each detected instance

Our goal: detailed action understanding

1. Localize person (+ interacting object)
2. Classify action of each detected instance
3. Estimate pose of person (+ interacting object)

Challenge 1: human pose estimation

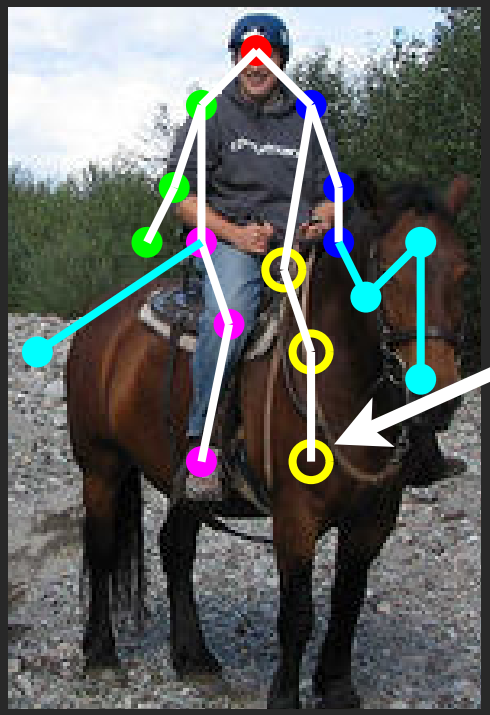
variation in appearance

variation in pose, viewpoint

Challenge 2: person-object occlusions

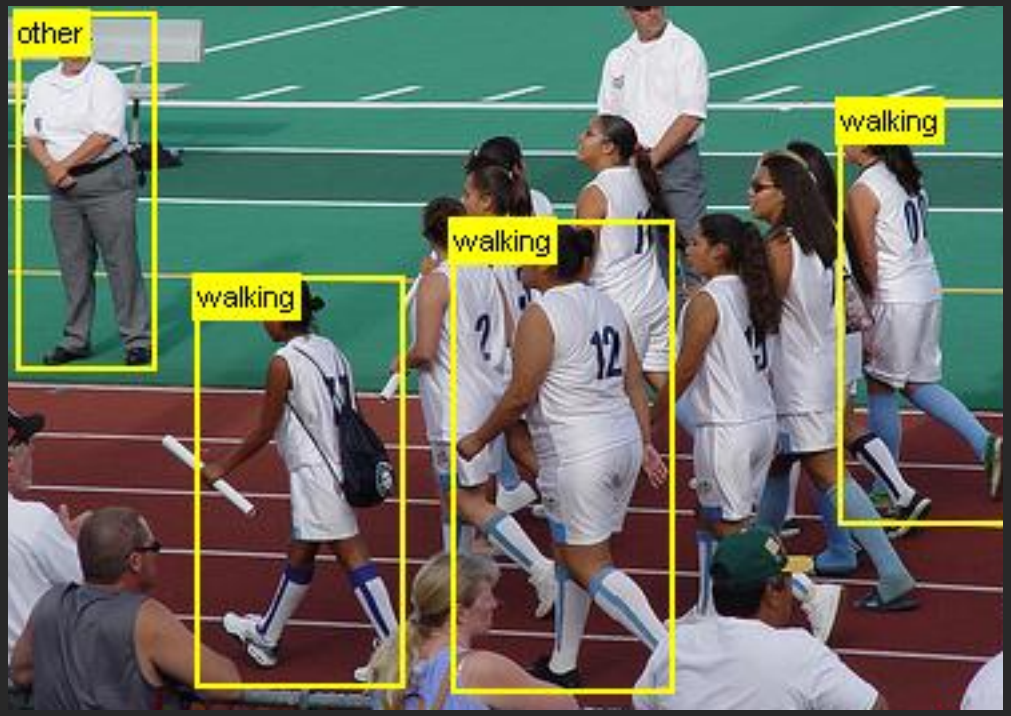
Occluded person leg

(Revised) action understanding



1. Localize
2. Estimate pose
3. Classify action
4. Estimate occlusions

Related work: PASCAL Action Classification Challenge



Everingham et al 2011
Yao et al ICCV 11
Maji et al CVPR11

Few previous entries appear to output an explicit human skeleton

Exceptions: next talk, Yang et al CVPR 10

Our approach

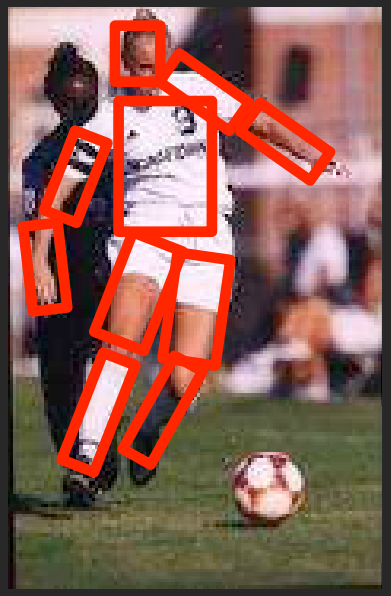
Articulated pose estimation

Visual composites

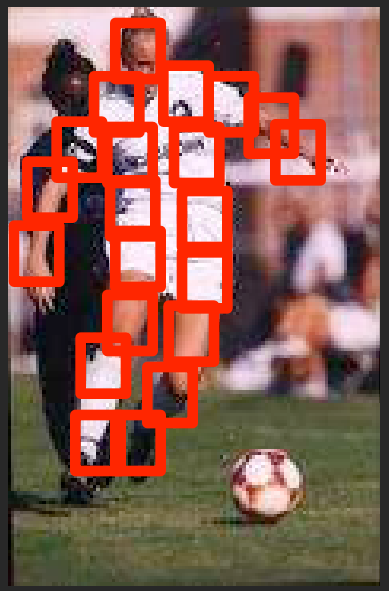
Geometric parts

Articulated pose estimation

Pictorial structures



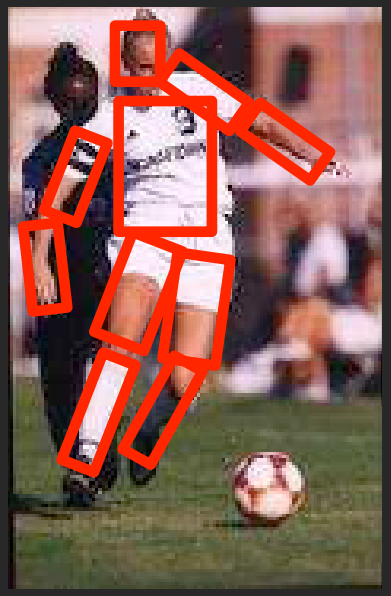
Ioffe & Forsyth 01
Felzenswalb & Huttenlocher 05
Ferrari et al.08
Andruikula et al. 09
Johnson & Everingham 11

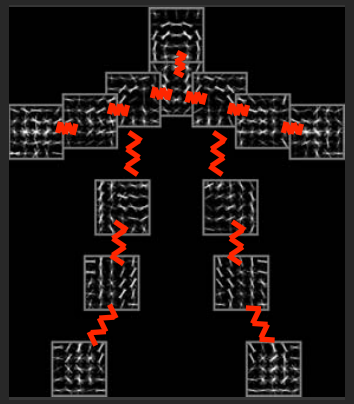


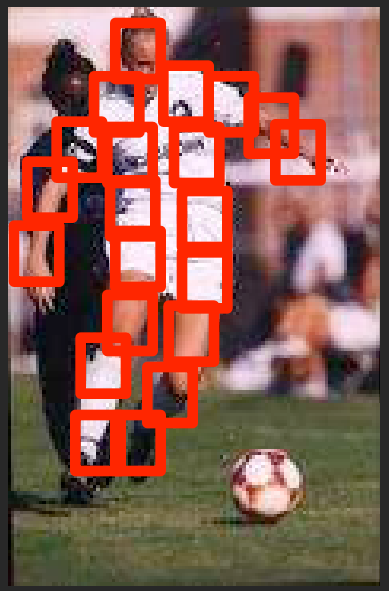
Yang & Ramanan 11

Articulated pose estimation

Pictorial structures



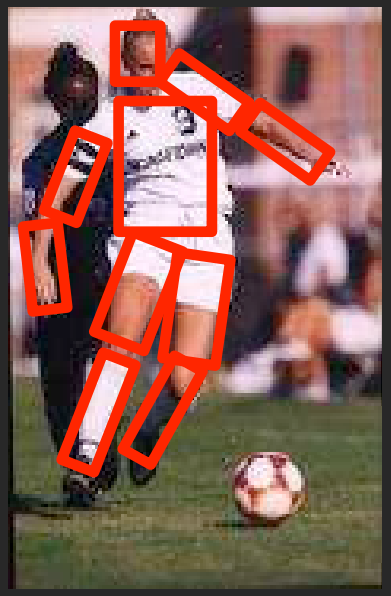


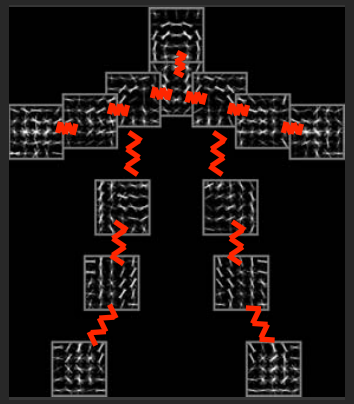


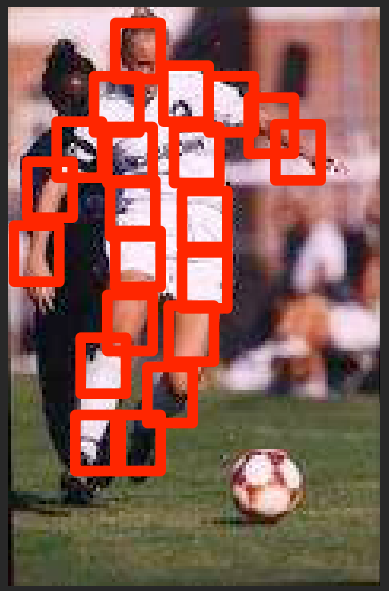
Models assume local appearance of parts is independent of global geometry

Articulated pose estimation

Pictorial structures

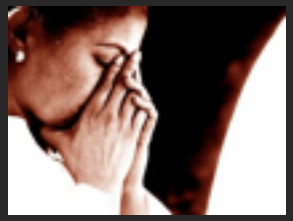






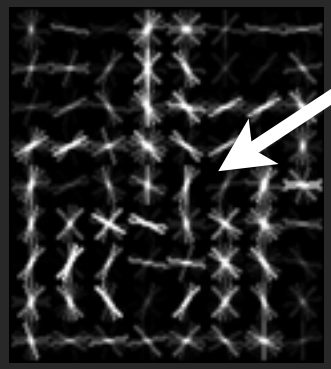
Models assume local appearance of parts is independent of global geometry

Problem: This fails to capture occlusions



Visual Phrases

Sadeghi and Fahardi, CVPR 11

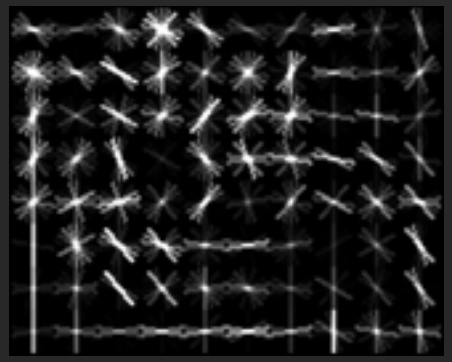


Occluded leg not present in template

Person on horse

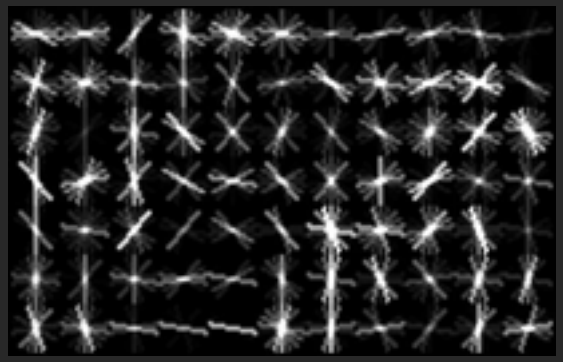
Visual Phrases

Sadeghi and Fahardi, CVPR 11



Person on jumping horse

Person on horse

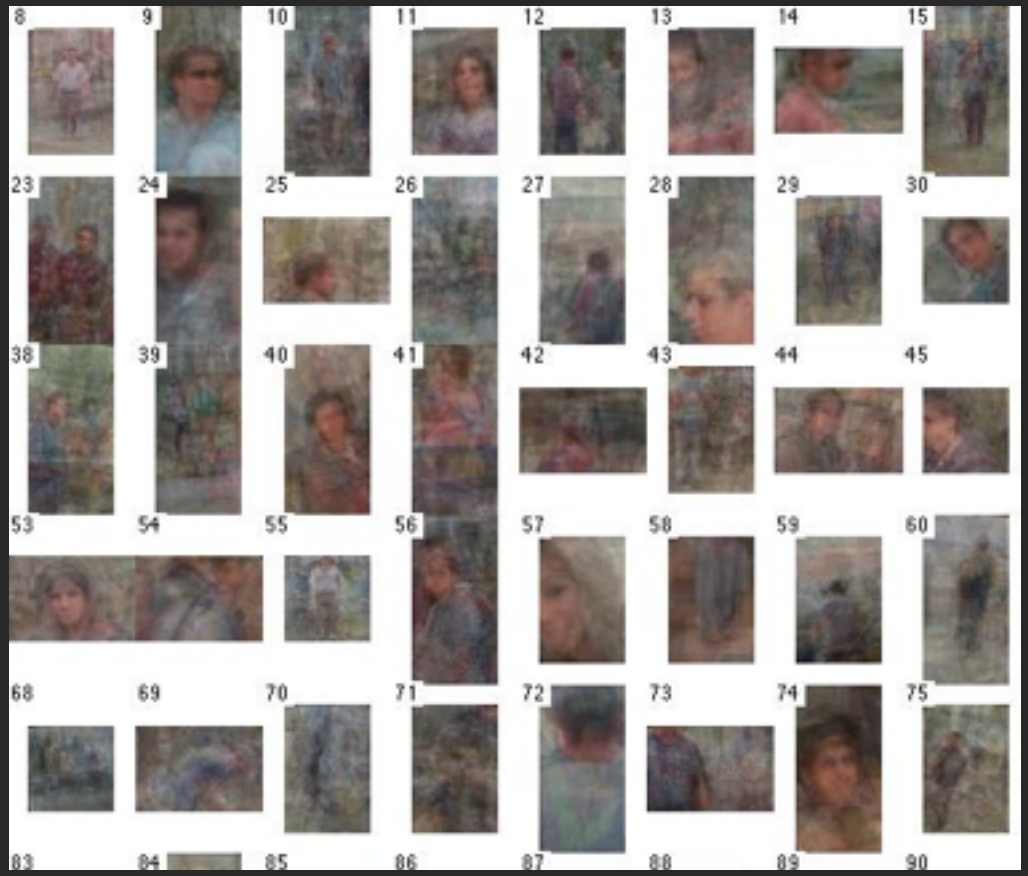


Person standing next to horse

Problem: one may need lots of large composite templates

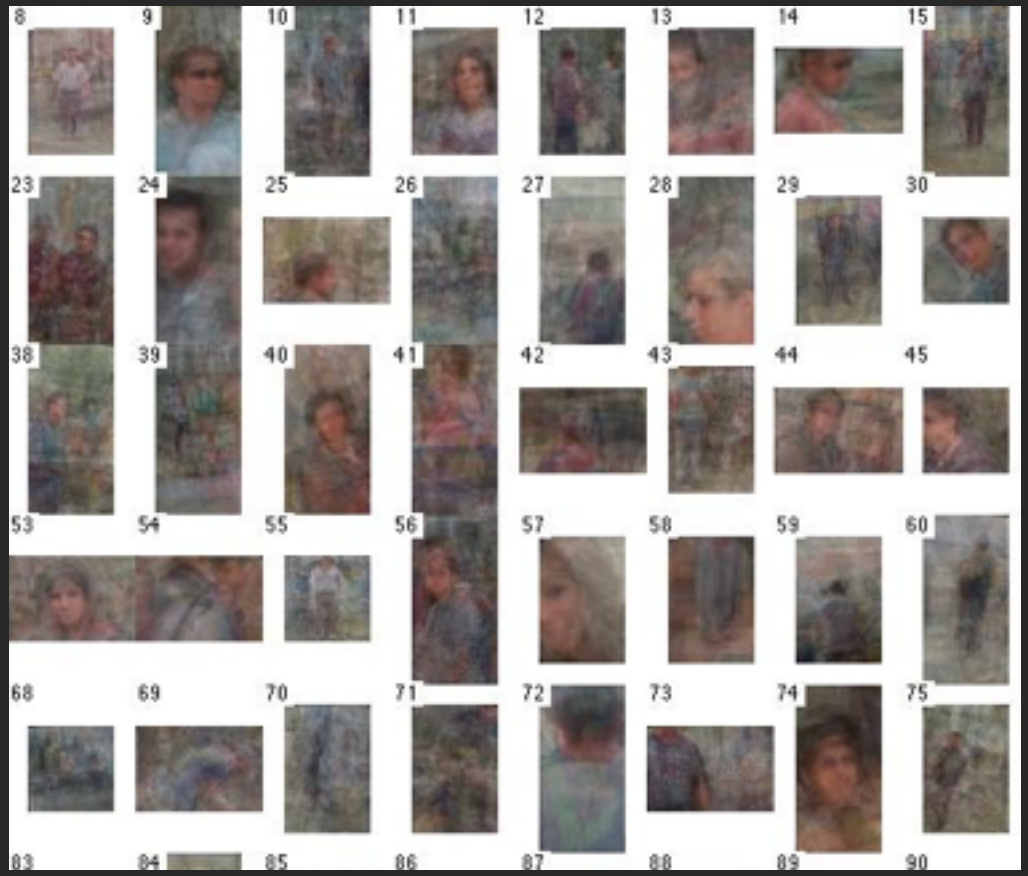
Geometric parts (poselets)

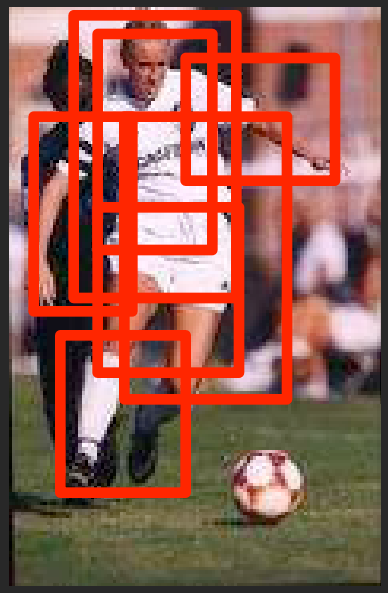
Bourdev & Malik ICCV09
Maji et al CVPR11



Geometric parts (poselets)

Bourdev & Malik ICCV09
Maji et al CVPR11





Problem: difficult to ensure that a globally-consistent arrangement of poselets will fire on a detection

Approach

Articulated pose estimation

Visual composites

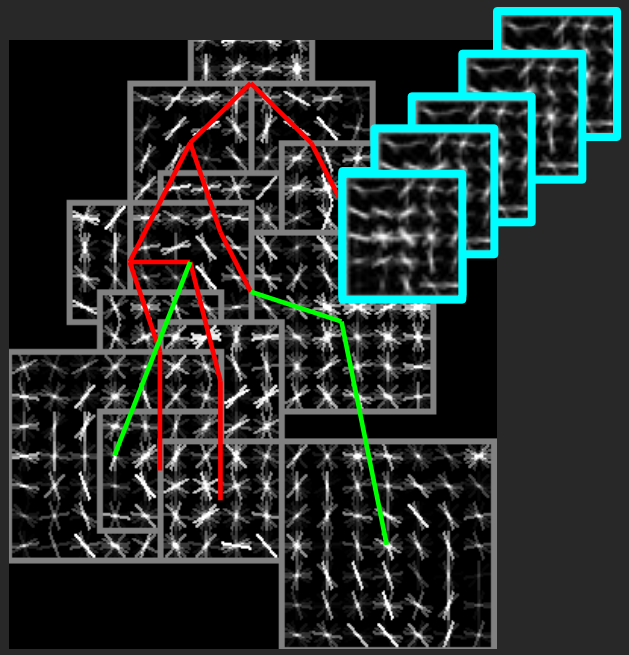
Geometric parts

Articulated models + visual composites

1. Define articulated model for person+object composite

Articulated models + visual composites + geometric parts

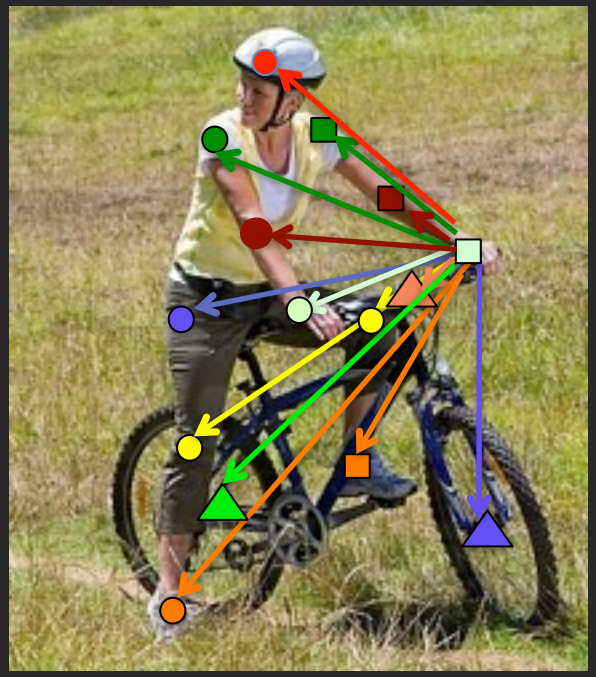
1. Define articulated model for person+object composite
2. Use local part mixtures (“phraselets”) to capture different occlusion states



Learning phraselets

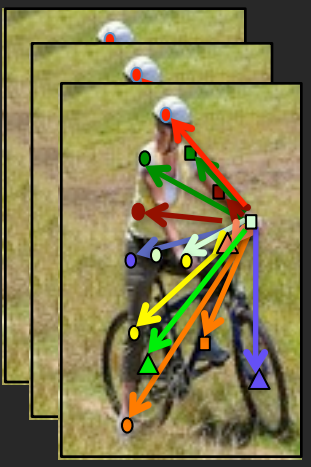
Define phraselets as commonly-occuring geometric configurations

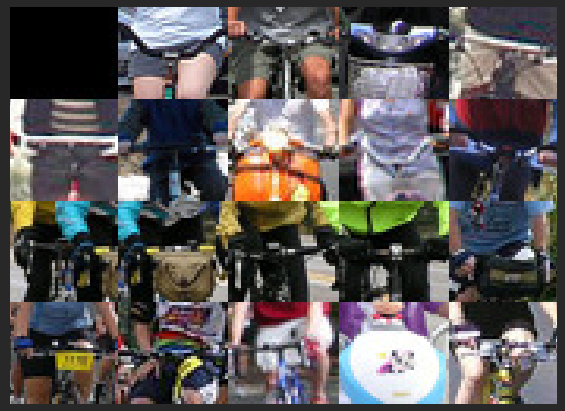
"Poselet-like" clusters



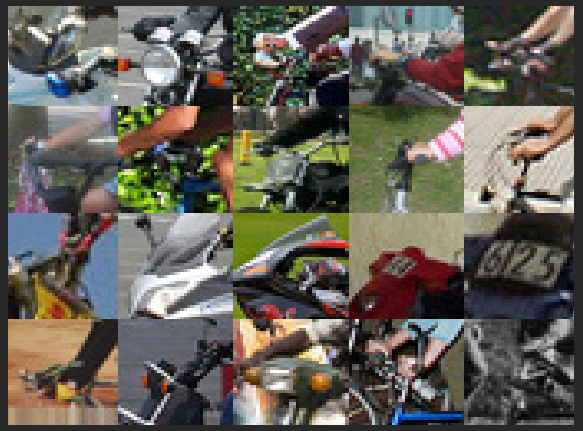
Given training data (with annotated landmarks), find clusters of landmark configurations relative to each joint

Clusters



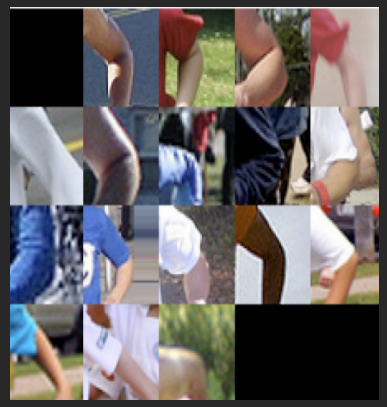


frontal

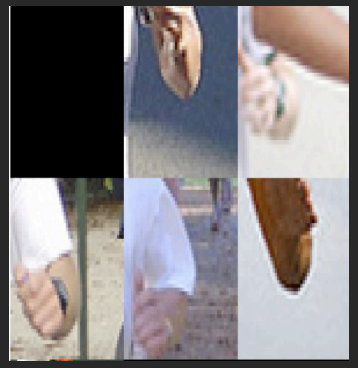


side-view

Model occlusions with separate clusters



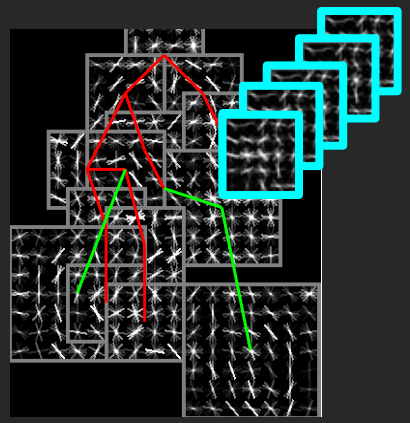
Visible left elbow

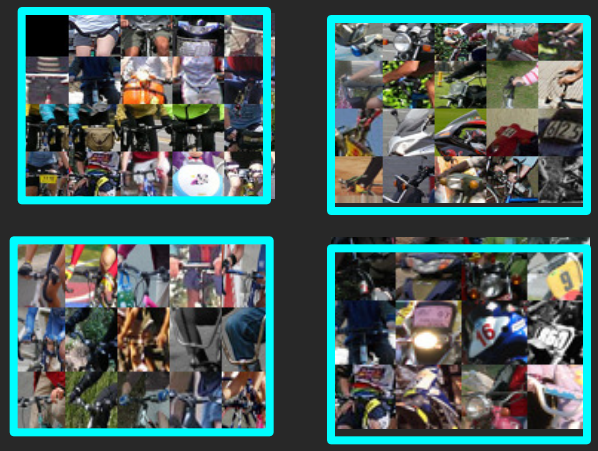


Occluded left elbow

Mixture label corresponds to visible/occlusion state

Local mixtures of phraselets



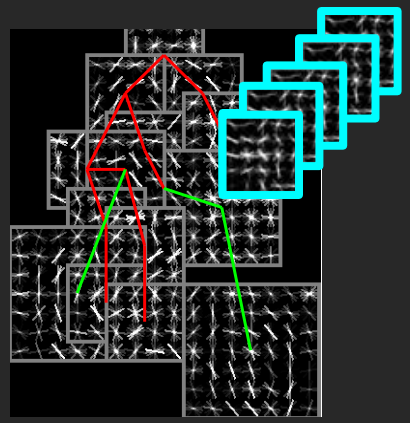


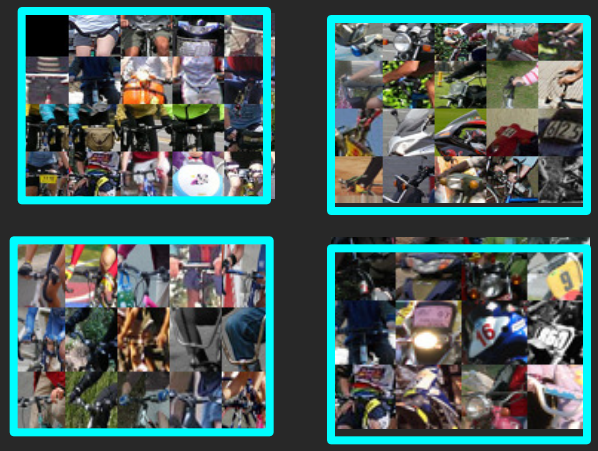
visible occluded

$$S(x, p, t)$$

x:image
p: position of part
t: local mixture type of part

Local mixtures of phraselets



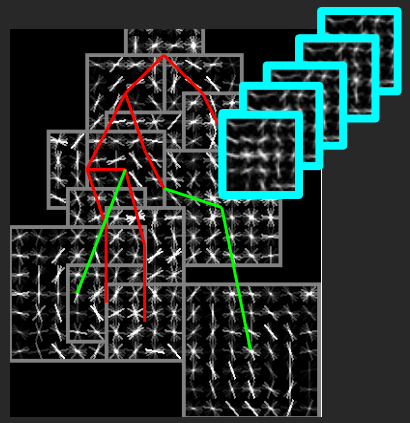


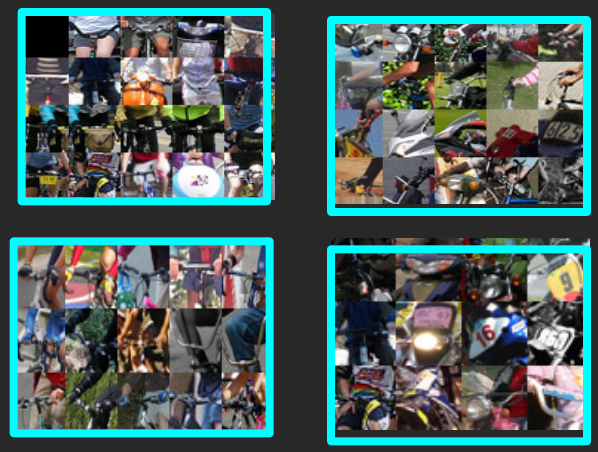
visible occluded

$$S(x,p,t) = \sum_i w_i^{t_i} \cdot \phi(x,p_i)$$

x:image
p:position of part
t: local mixture type of part

Local mixtures of phraselets





visible occluded

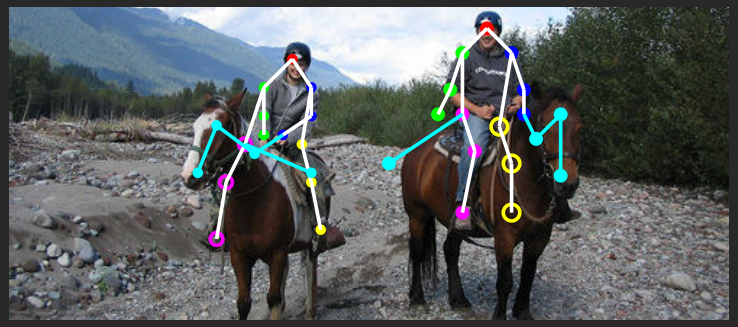
$$S(x,p,t) = \sum_i w_i^{t_i} \cdot \phi(x,p_i) + \sum_{ij \in E} w_{ij}^{t_i,t_j} \cdot \psi(p_i,p_j)$$

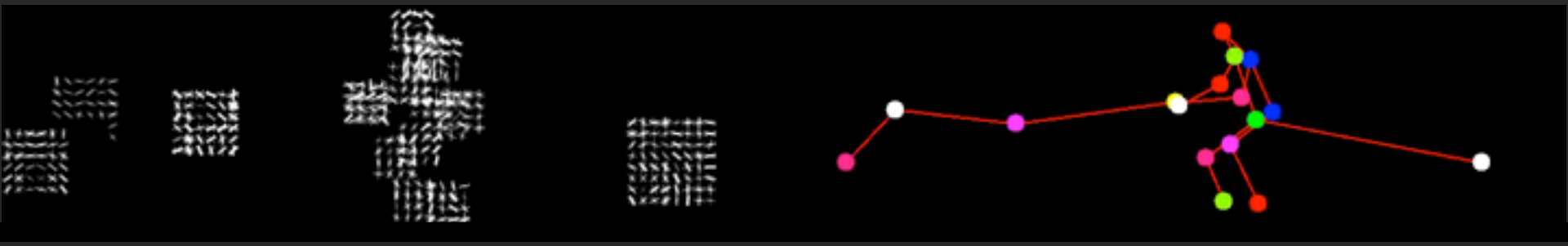
Yang & Ramanan CVPR 11

x:image
p:position of part
t: local mixture type of part

"Switching" spring model tuned for a pair of local mixtures

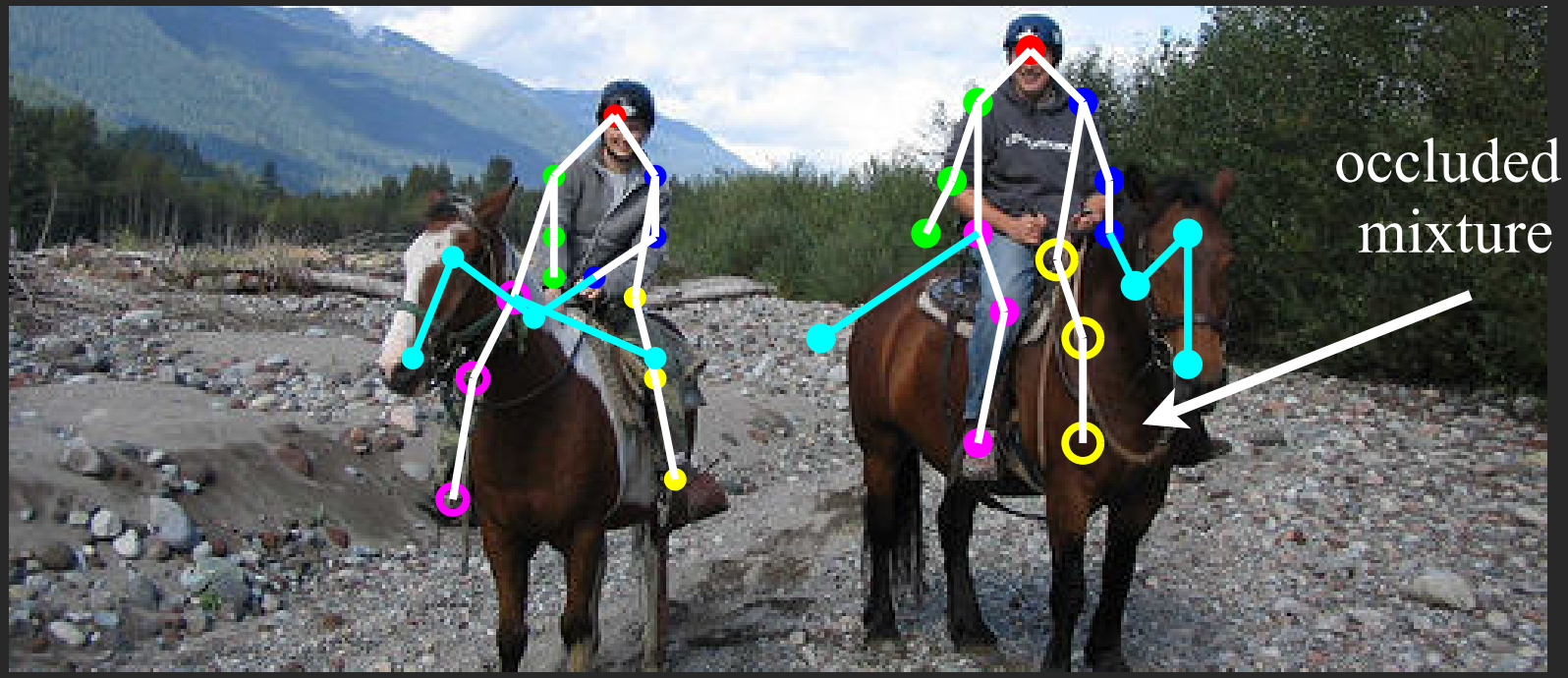
Geometry-dependent parts





Part appearance (local mixture, denoted by color)
depends on the location and appearance of other parts

Inference & Learning

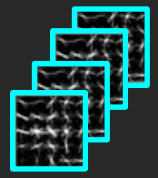


Inference: Infer part locations + mixtures with dynamic programming on trees

Learning: Tune linear parameters (including occlusion constraints) with SVM solver

Possible Criticisms

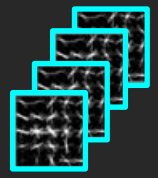
1. One should not score image evidence during occlusions



Possible Criticisms

1. One should not score image evidence during occlusions

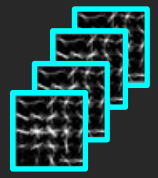
An occluded mixture template may learn all 0 weights; let the learning algorithm decide!



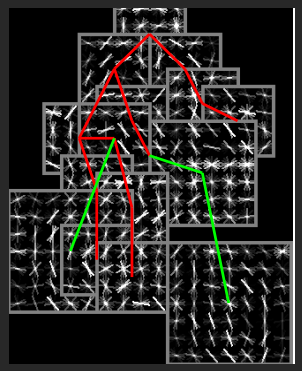
Possible Criticisms

1. One should not score image evidence during occlusions

An occluded mixture template may learn all 0 weights; let the learning algorithm decide!



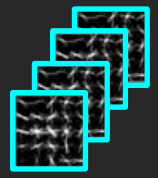
2. Small patches are not as discriminative as larger templates (visual phrases / poselets)



Possible Criticisms

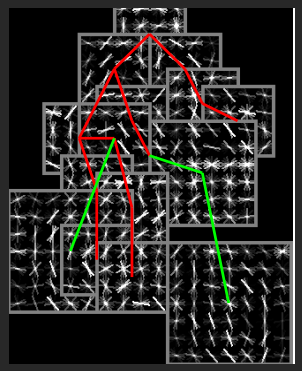
1. One should not score image evidence during occlusions

   An occluded mixture template may learn all 0 weights; let the learning algorithm decide!

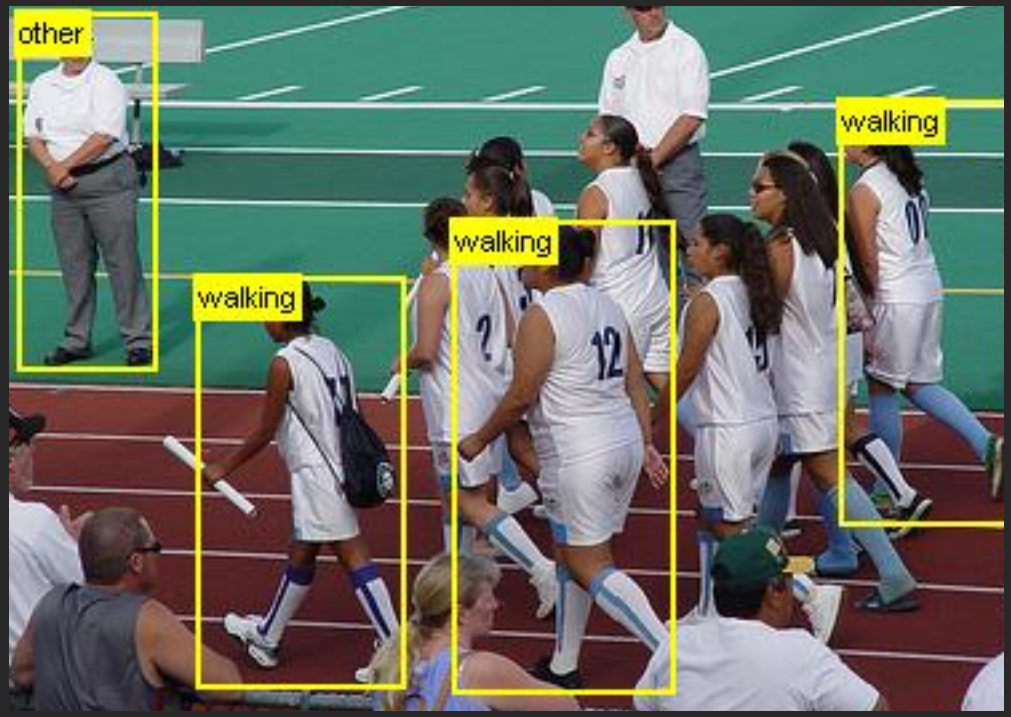


2. Small patches are not as discriminative as larger templates (visual phrases / poselets)

   Any connected set of phraselets can learn to behave like a larger template (rigid springs)



Experimental Results



We use 2010 & 2011 PASCAL Action Recognition benchmark

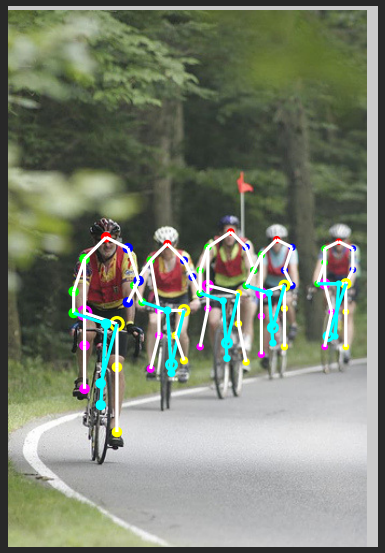
We augment training and validation set with landmark annotations

High-confidence detections

Riding horse

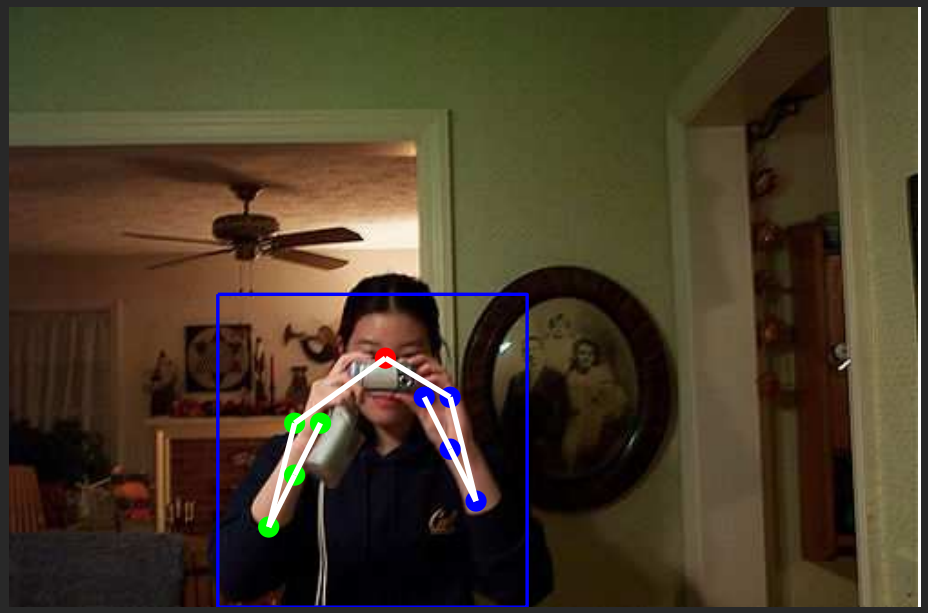


Riding bike


High-scoring false-positives

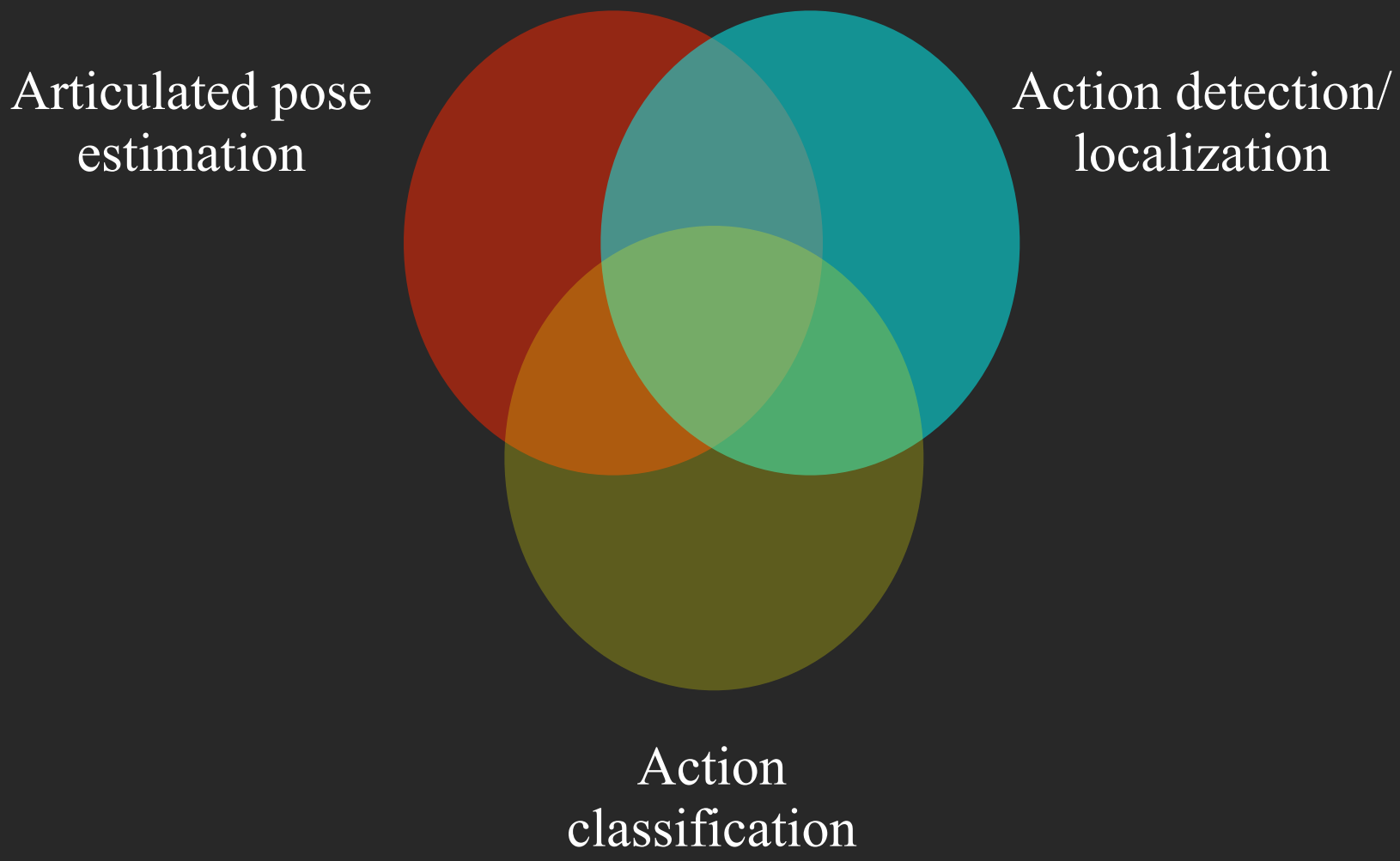
Often produce meaningful poses



False-positive “phoning” detections

(taking a picture and scratching head have similar poses)

Evaluation


Articulated pose estimation
Action detection/ localization
Action classification

Evaluation

Articulated pose estimation

Action detection/ localization

Action classification

On par with Poselets, comparable to state-of-the-art

Evaluation

Articulated pose estimation

Considerably outperform pictorial structures

Action detection/ localization

Action classification

On par with Poselets, comparable to state-of-the-art

Evaluation

Articulated pose estimation

Considerably outperform pictorial structures

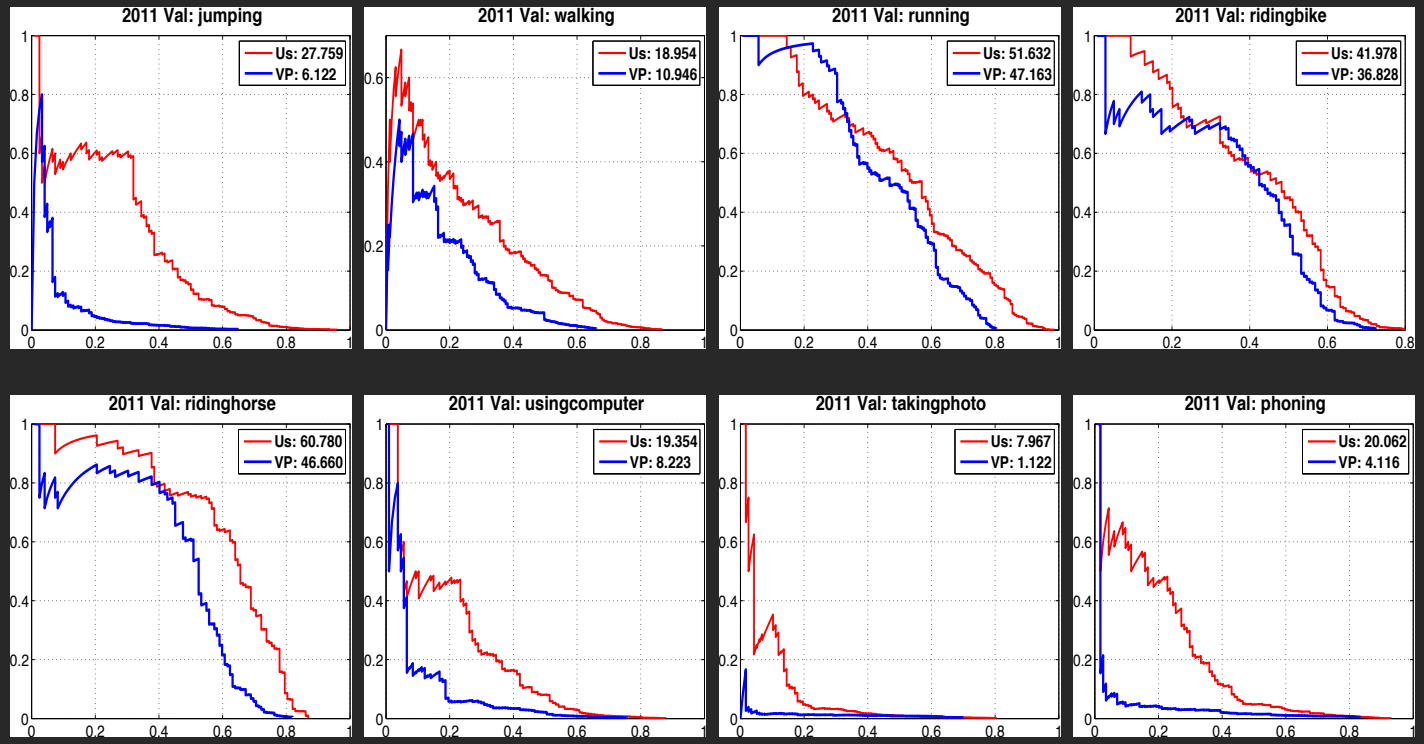
Action detection/ localization

Action classification

On par with Poselets, comparable to state-of-the-art

Action detection

Treat each action as an “object” and evaluate standard criteria (AP)



Blue line: Visual Phrase Red line: Us

Significantly outperform state-of-art for detection

A look back

Action understanding

(detection, classification, & pose estimation)

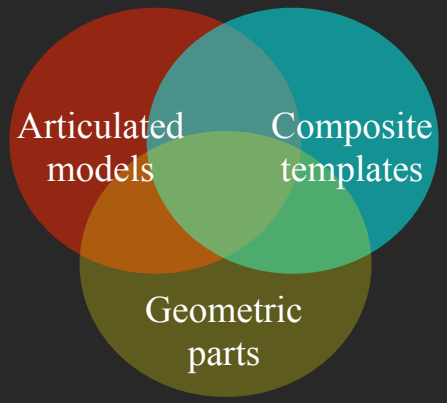
A look back

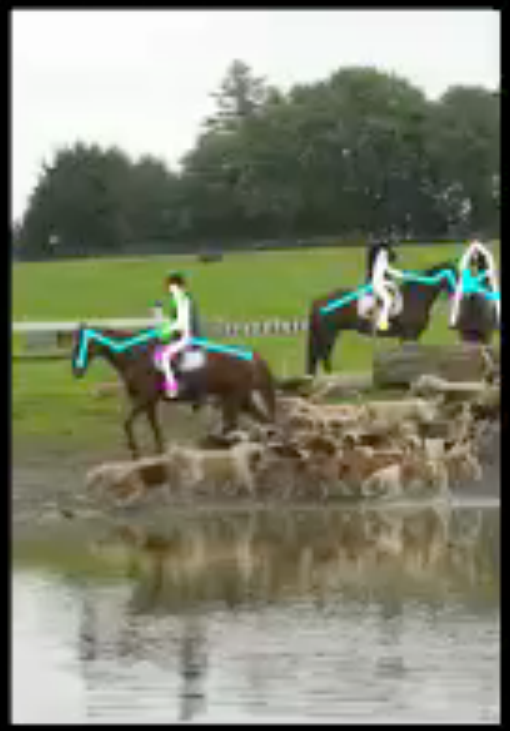
Action understanding

(detection, classification, & pose estimation)

Geometric part models

(interdependence of geometry and appearance)





Riding Horse