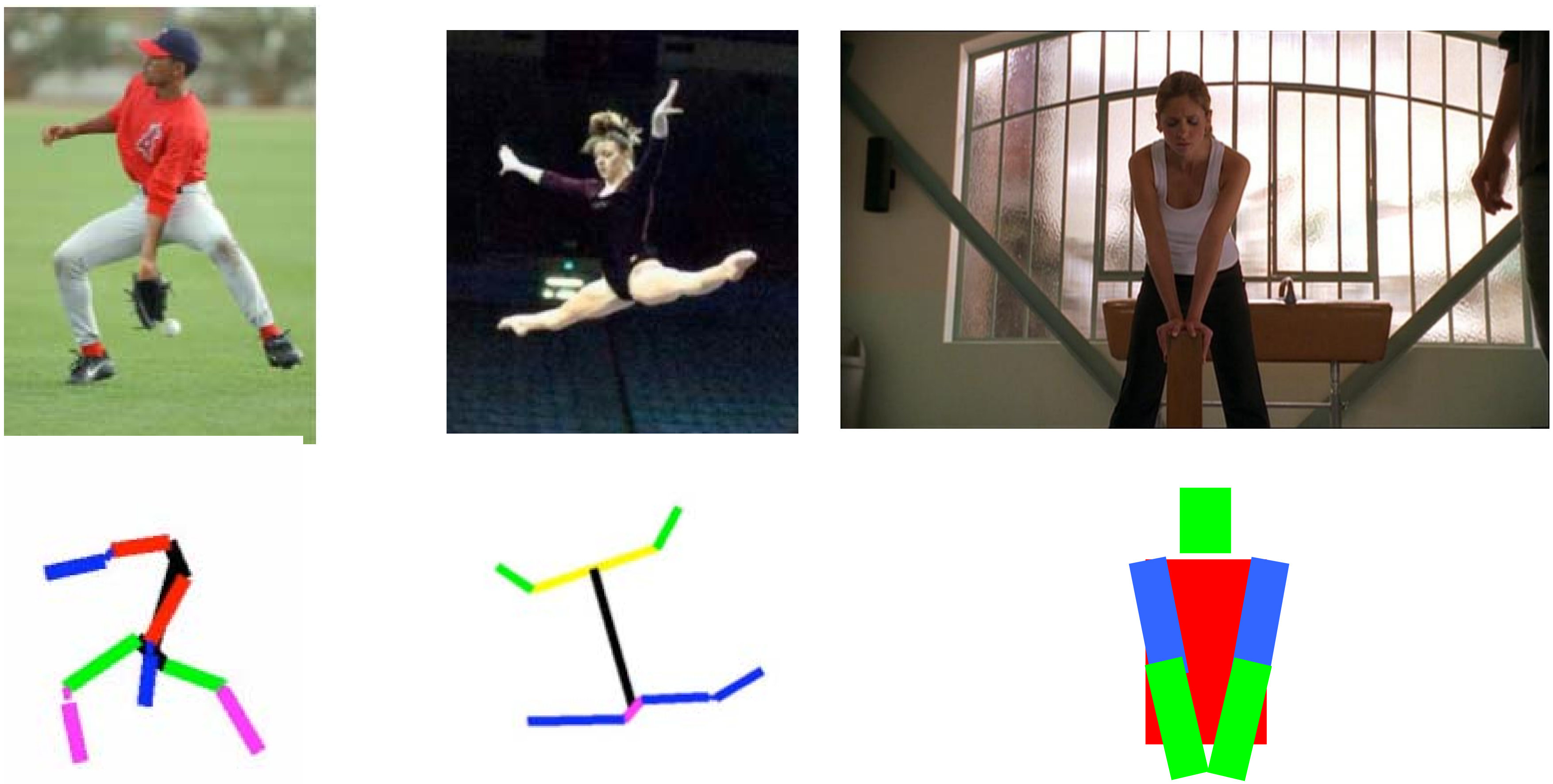
Human Pose Estimation in images and videos

Andrew Zisserman

Department of Engineering Science
University of Oxford, UK
http://www.robots.ox.ac.uk/~vgg/

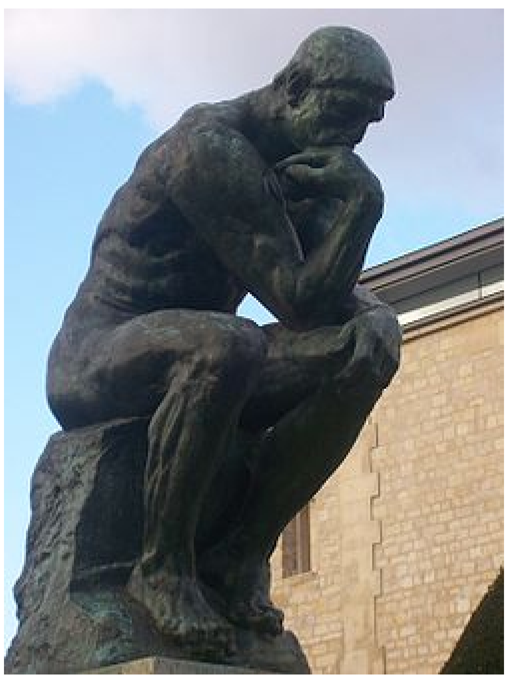
Objective and motivation

Determine human body pose (layout)



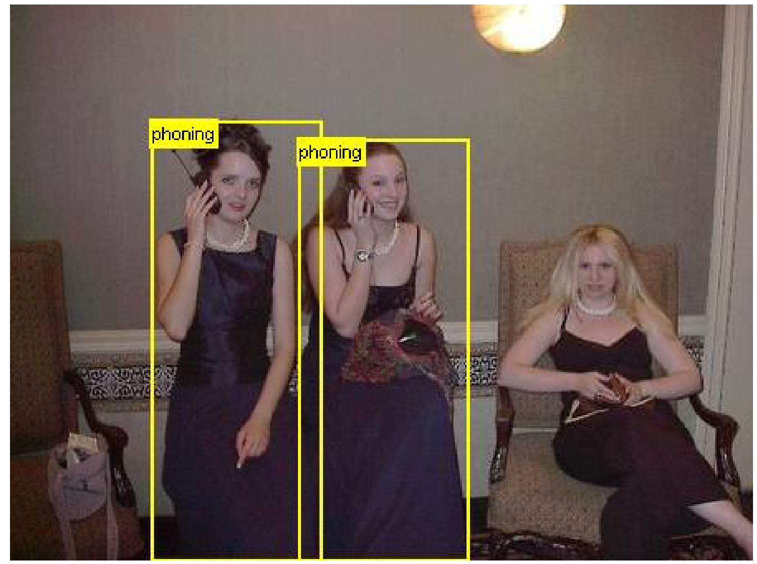
Why? To recognize poses, gestures, actions

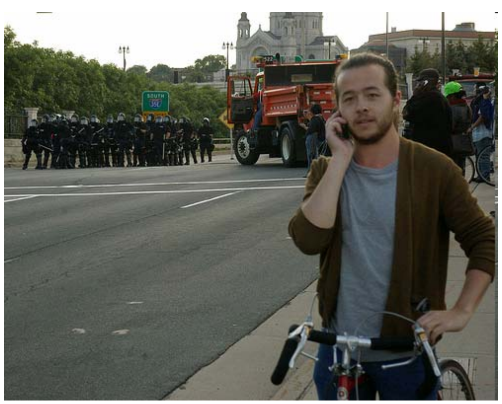
Activities characterized by a pose



Activities characterized by a pose

Activities characterized by a pose



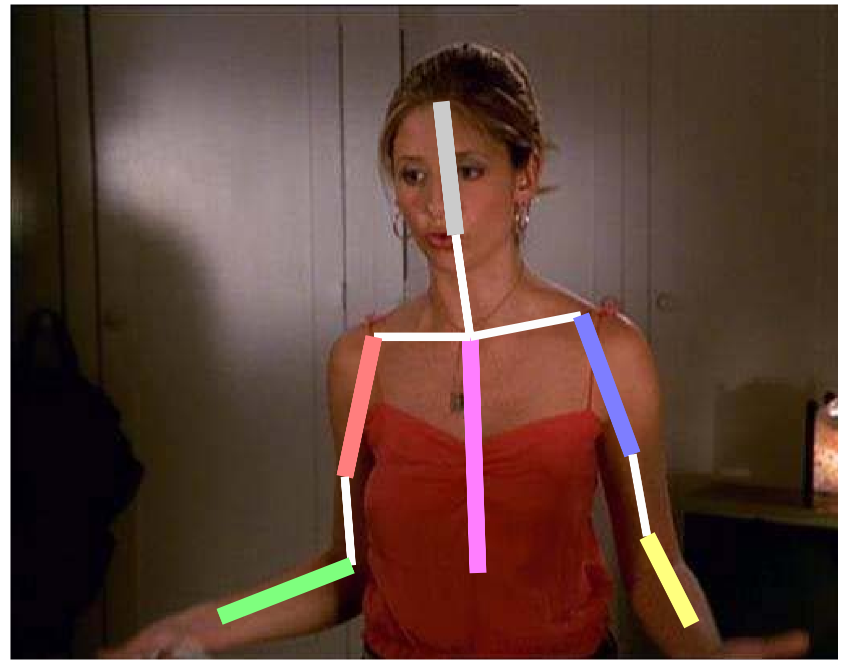


Challenges: articulations and deformations

Challenges: of (almost) unconstrained images




varying illumination and low contrast; moving camera and background; multiple people; scale changes; extensive clutter; any clothing

Outline

- Review of pictorial structures for articulated models

- Inference given the model: Strong supervision, full generative model – “Gold-standard model”

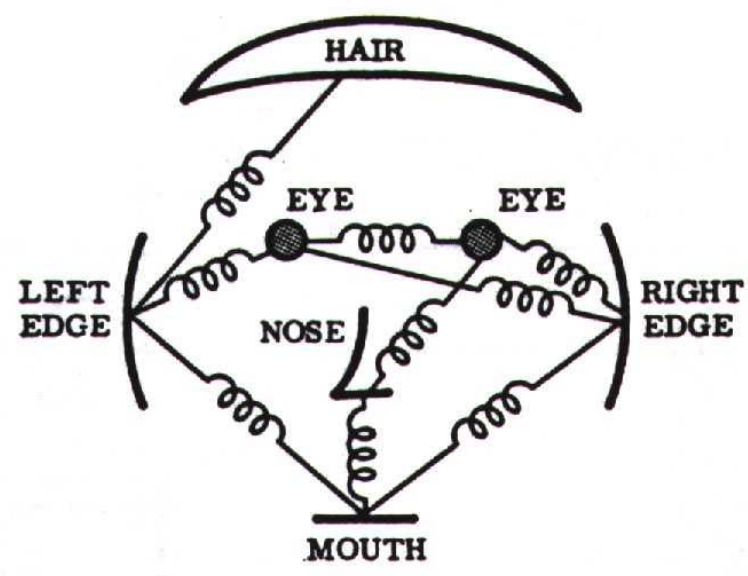
- Image parsing: learning the model for a specific image

- Recent advances

- Datasets and challenges

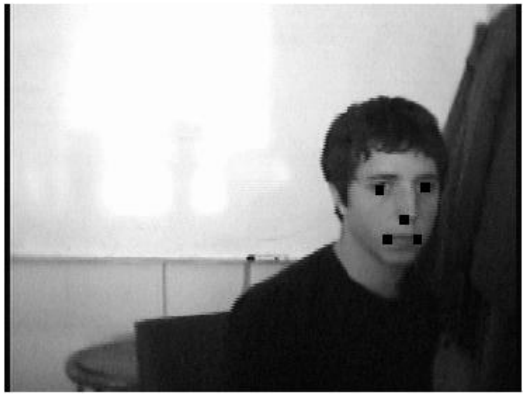
Pictorial Structures

- Intuitive model of an object
- Model has two components
  1. parts (2D image fragments)
  2. structure (configuration of parts)
- Dates back to Fischler & Elschlager 1973



Localize multi-part objects at arbitrary locations in an image

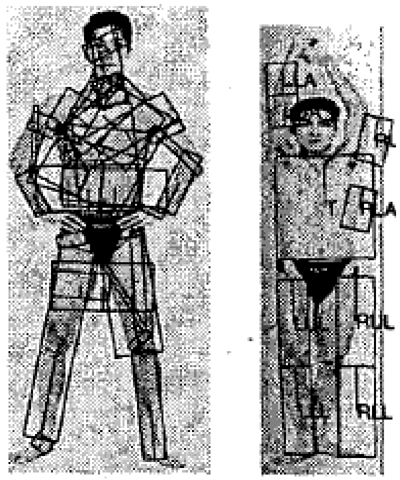
- Generic object models such as person or car
- Allow for articulated objects
- Simultaneous use of appearance and spatial information
- Provide efficient and practical algorithms



To fit model to image: minimize an energy (or cost) function that reflects both

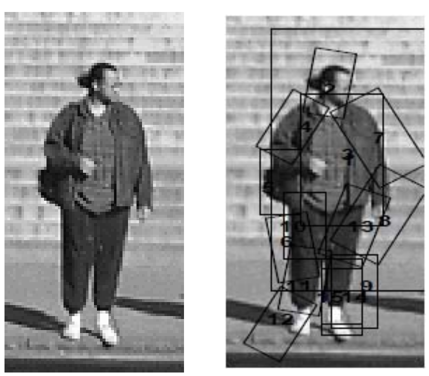
- Appearance: how well each part matches at given location
- Configuration: degree to which parts match 2D spatial layout

Long tradition of using pictorial structures for humans



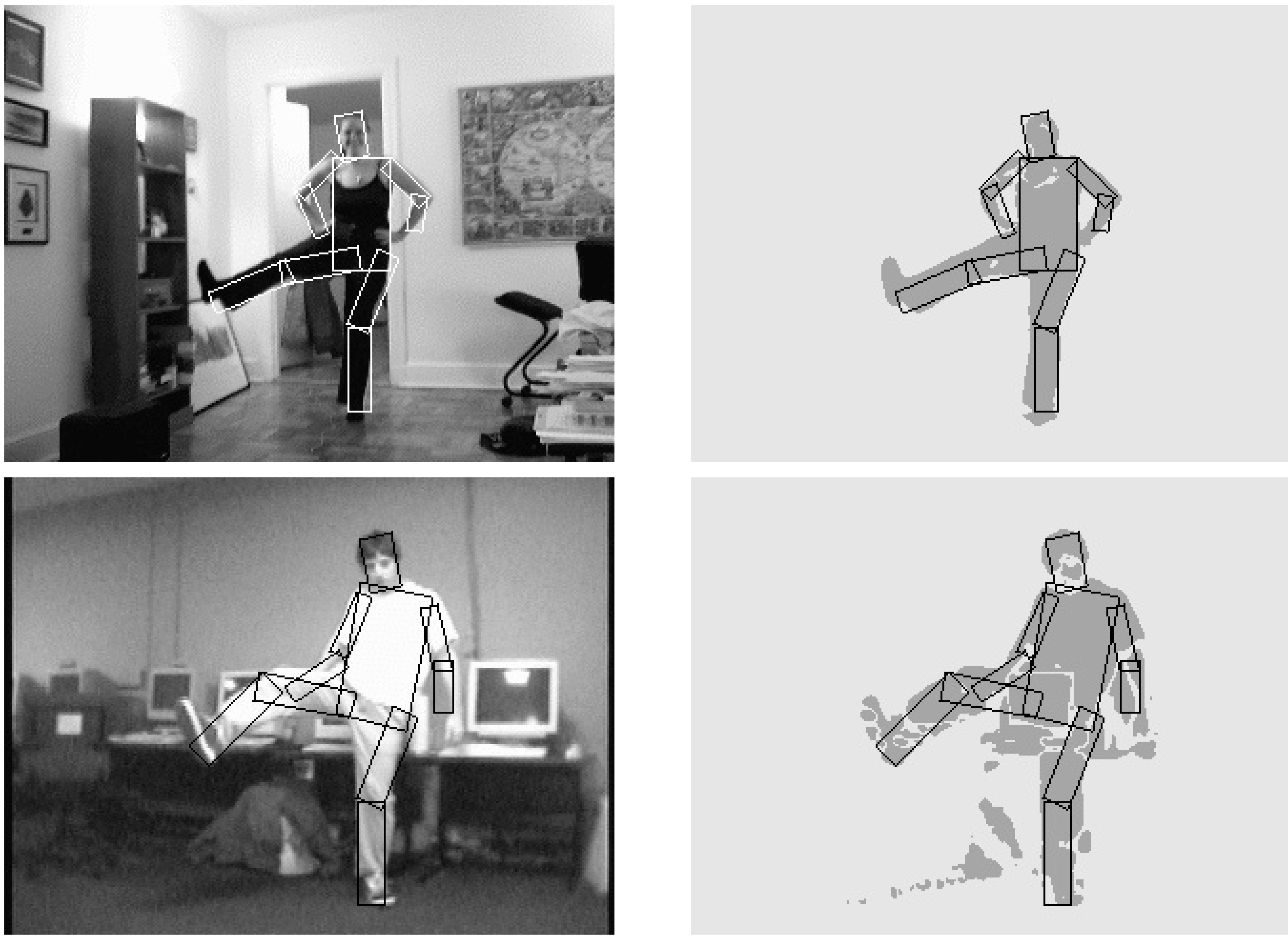
Finding People by Sampling
Ioffe & Forsyth, ICCV 1999

Pictorial Structure Models for Object Recognition
Felzenszwalb & Huttenlocher, 2000



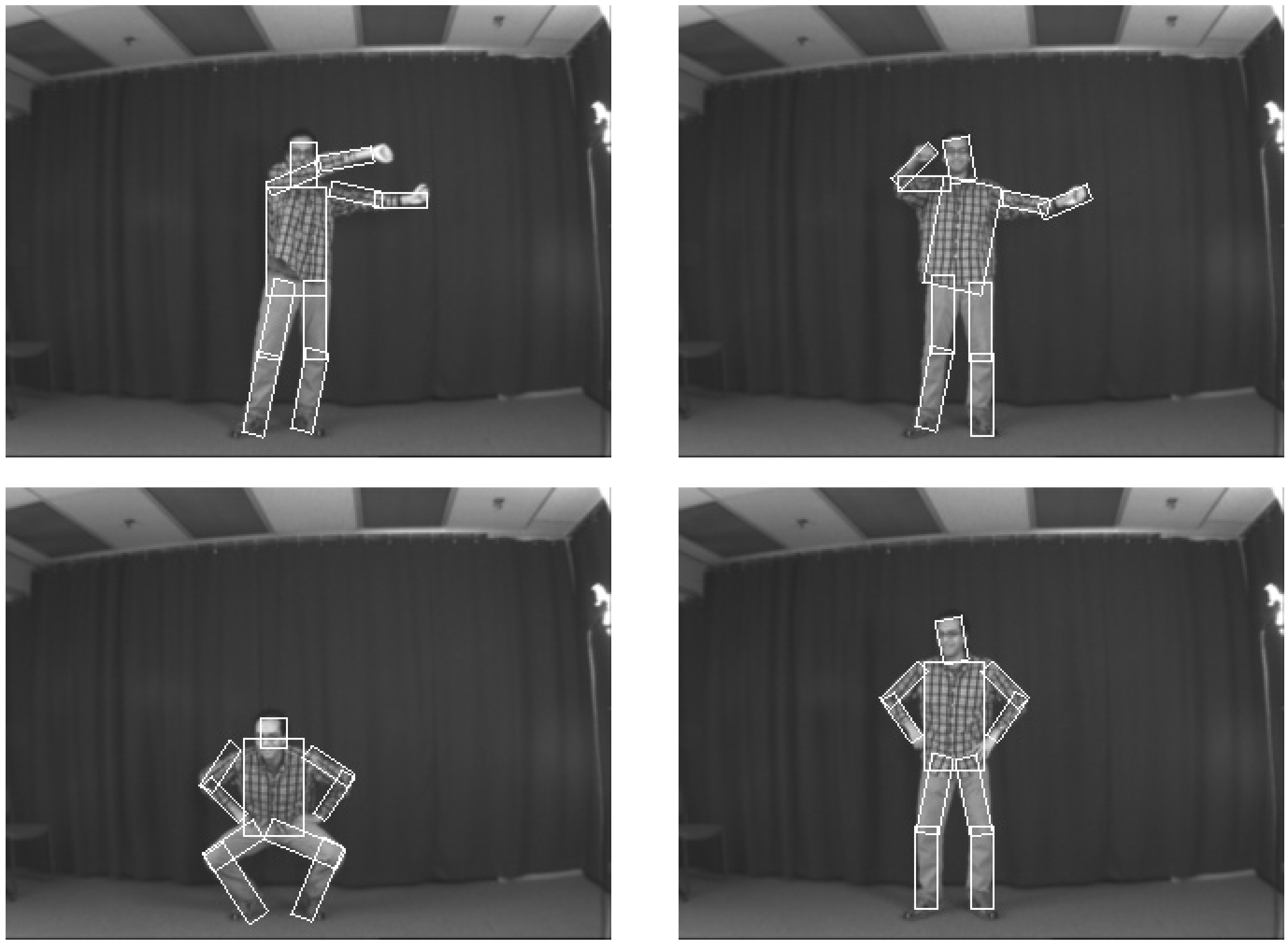
Learning to Parse Pictures of People
Ronfard, Schmid & Triggs, ECCV 2002

Felzenszwalb & Huttenlocher

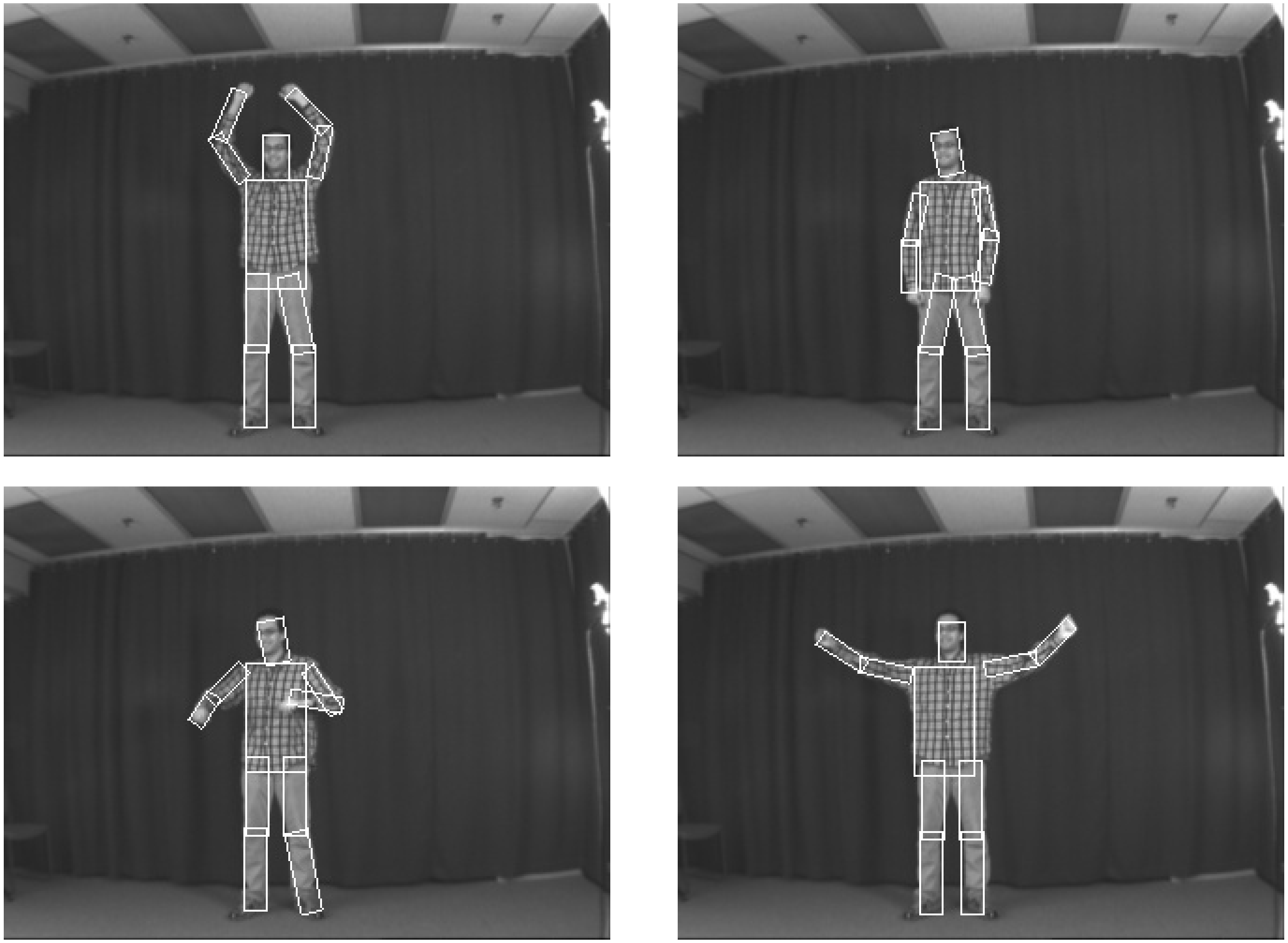


NB: requires background subtraction

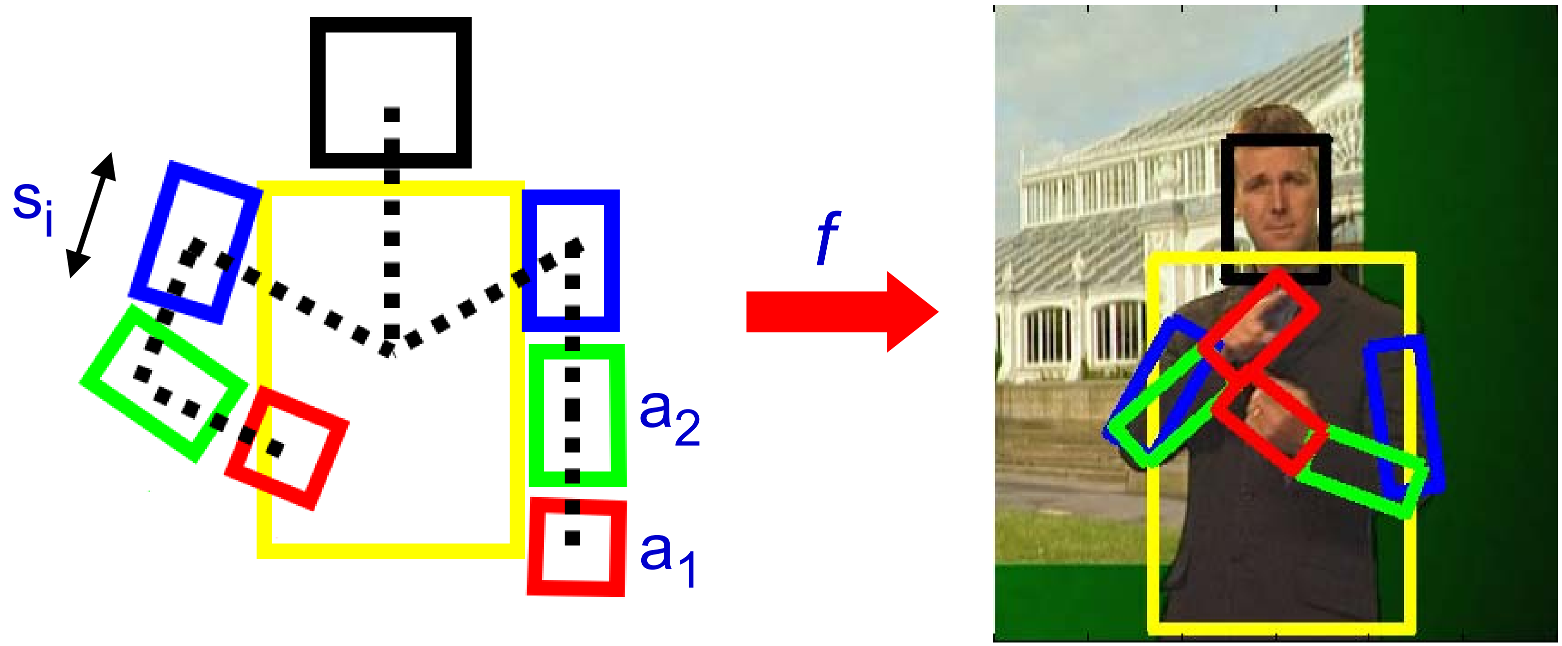
Variety of Poses



Variety of Poses



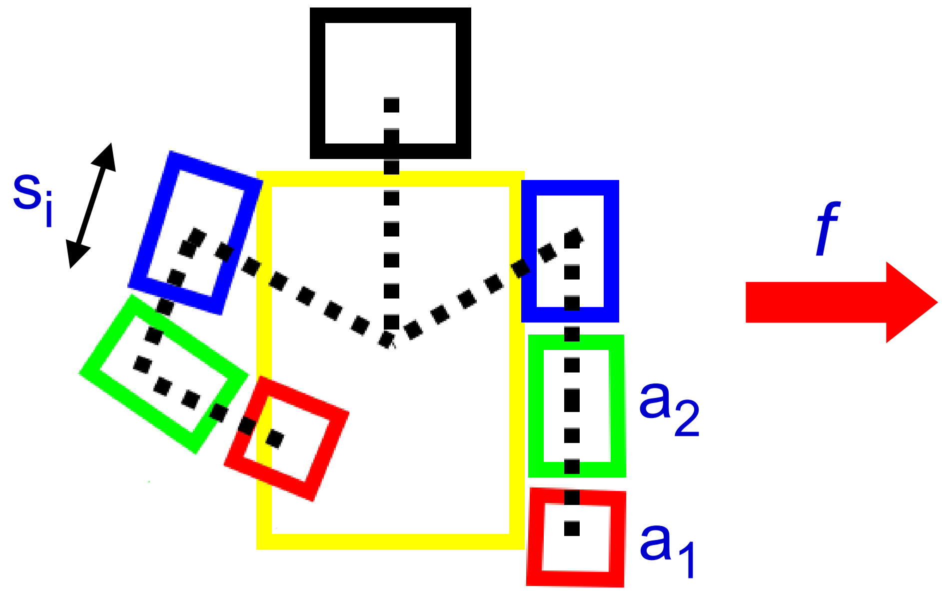
Objective: detect human and determine upper body pose (layout)

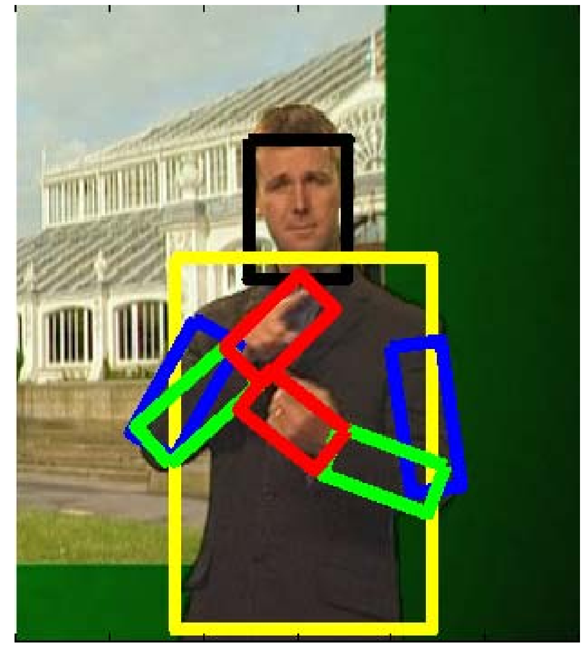


Model as a graph labelling problem

- Vertices $\mathcal{V}$ are parts, $a_i, i = 1, \cdots, n$
- Edges $\mathcal{E}$ are pairwise linkages between parts
- For each part there are h possible poses $\mathbf{p}_j = (x_j, y_j, \phi_j, s_j)$
- Label each part by its pose: $f : \mathcal{V} \longrightarrow \{1, \cdots, h\}$, i.e. part a takes pose $\mathbf{p}_{f(a)}$.

Pictorial structure model – CRF





- Each labelling has an energy (cost):

$$E(f) = \underbrace{\sum_{a \in \mathcal{V}} \theta_{a;f(a)}}_{\substack{\text{unary terms} \\ \text{(appearance)}}} + \underbrace{\sum_{(a,b) \in \mathcal{E}} \theta_{ab;f(a)f(b)}}_{\substack{\text{pairwise terms} \\ \text{(configuration)}}}$$

Features for unary:

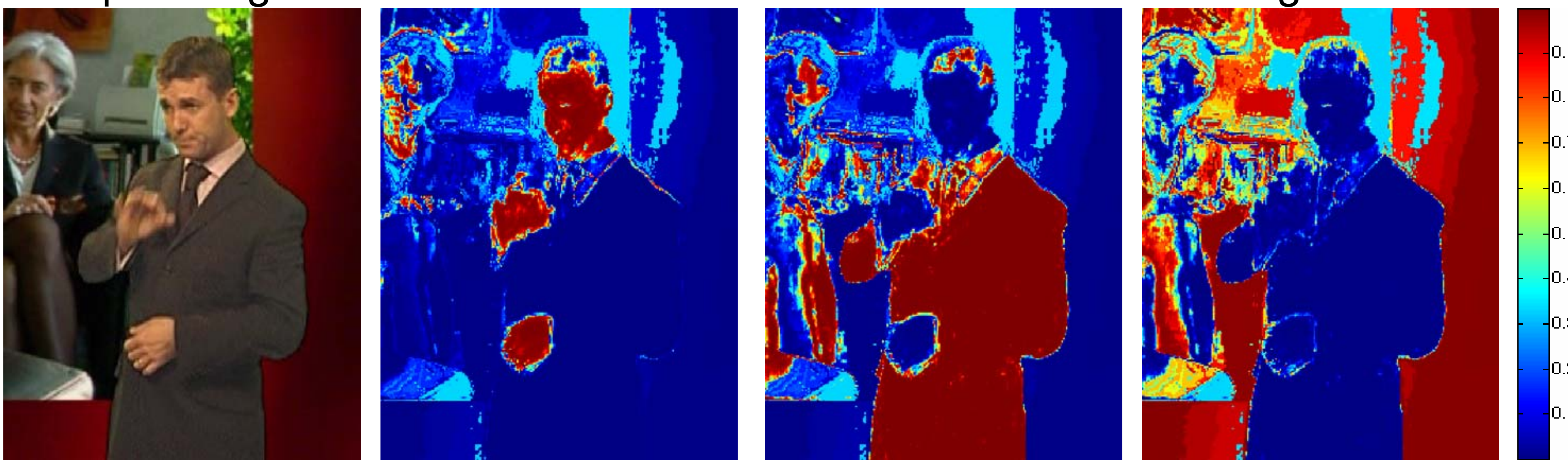
- colour
- HOG

for limbs/torso

- Fit model (inference) as labelling with lowest energy

Unary term: appearance feature I - colour

input image skin torso background

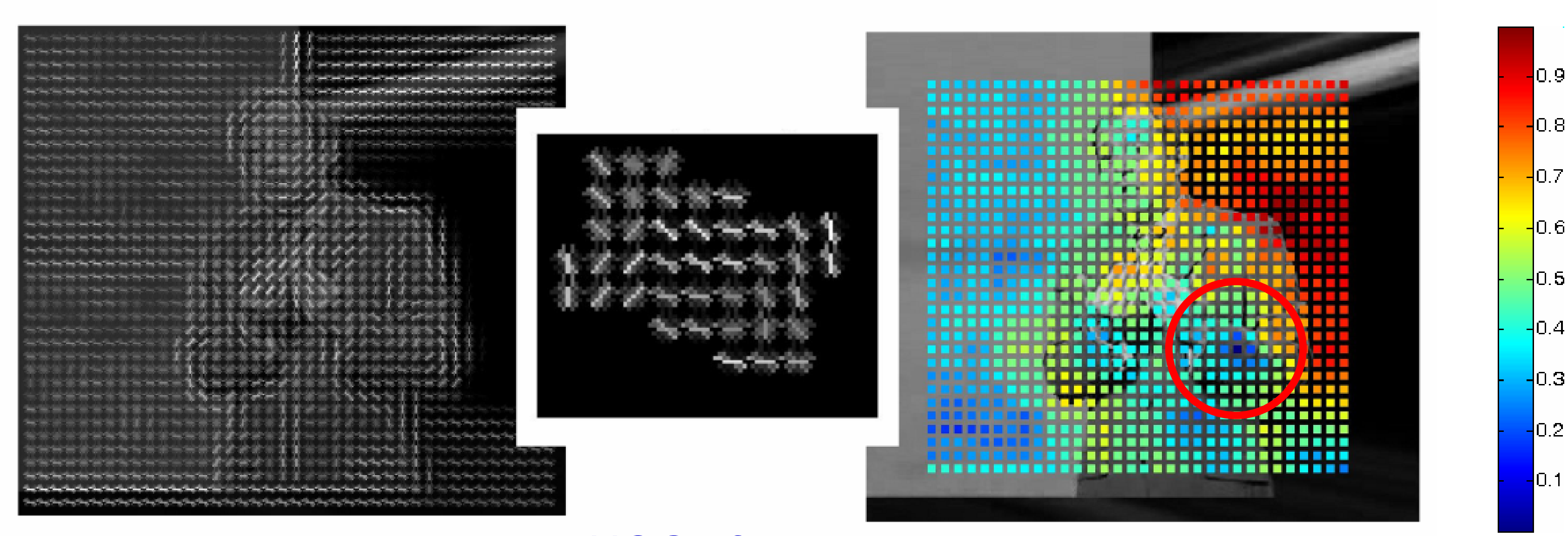


colour posteriors

Unary term: appearance feature II - HOG

Dalal & Triggs, CVPR 2005

Histogram of oriented gradients (HOG)



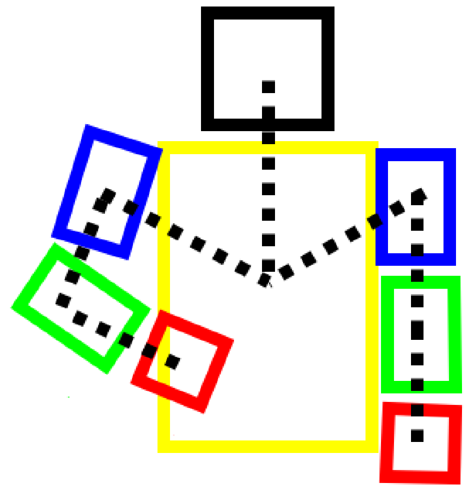
HOG of image

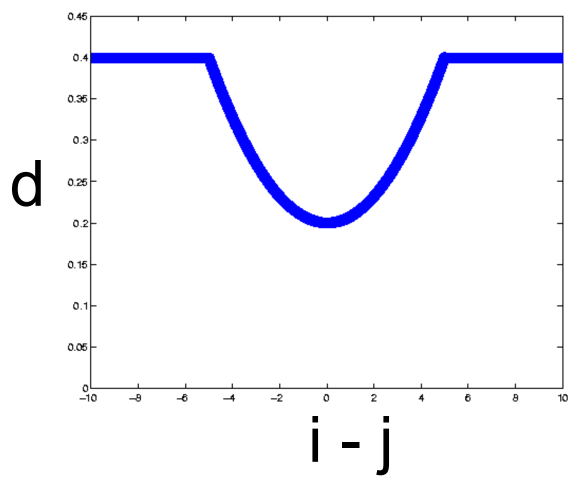
HOG of lower arm template (learned)

L2 Distance

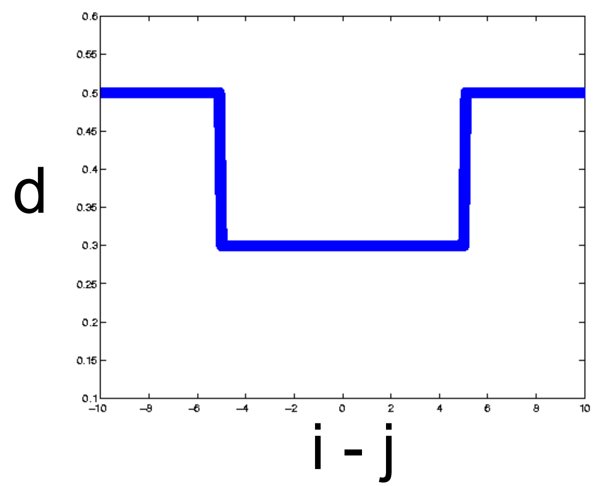
Pairwise terms: kinematic layout

$$\theta_{ab;ij} = w_{ab} d(|i-j|)$$



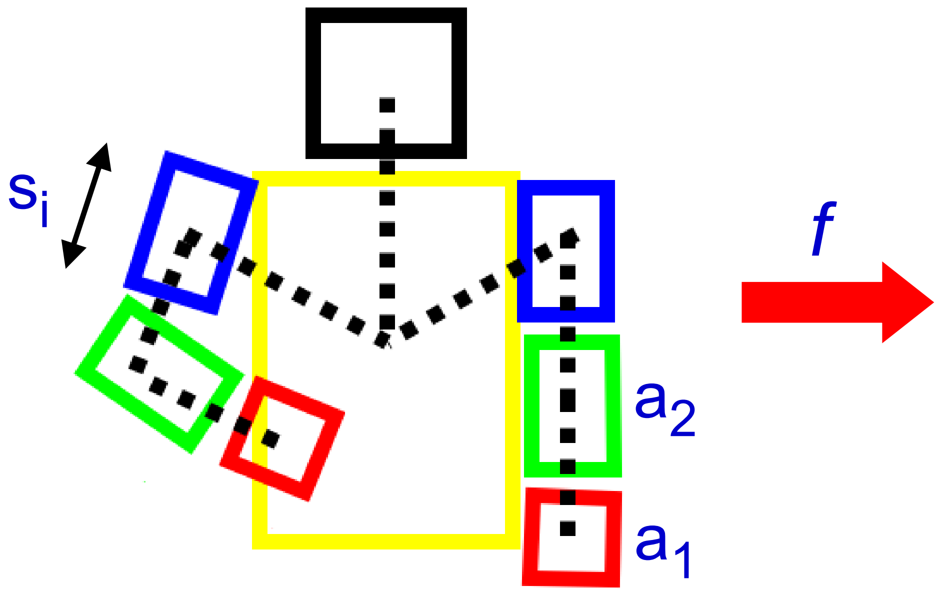


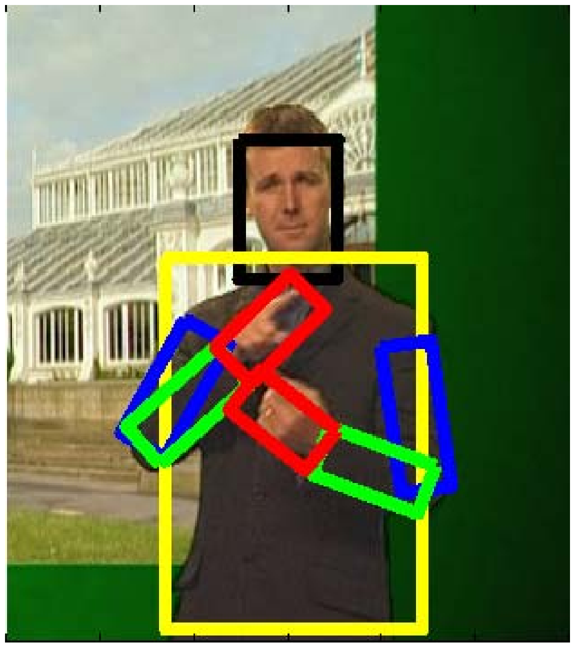
Truncated Quadratic



Potts

Pictorial structure model – CRF





Features for unary:

- colour
- HOG

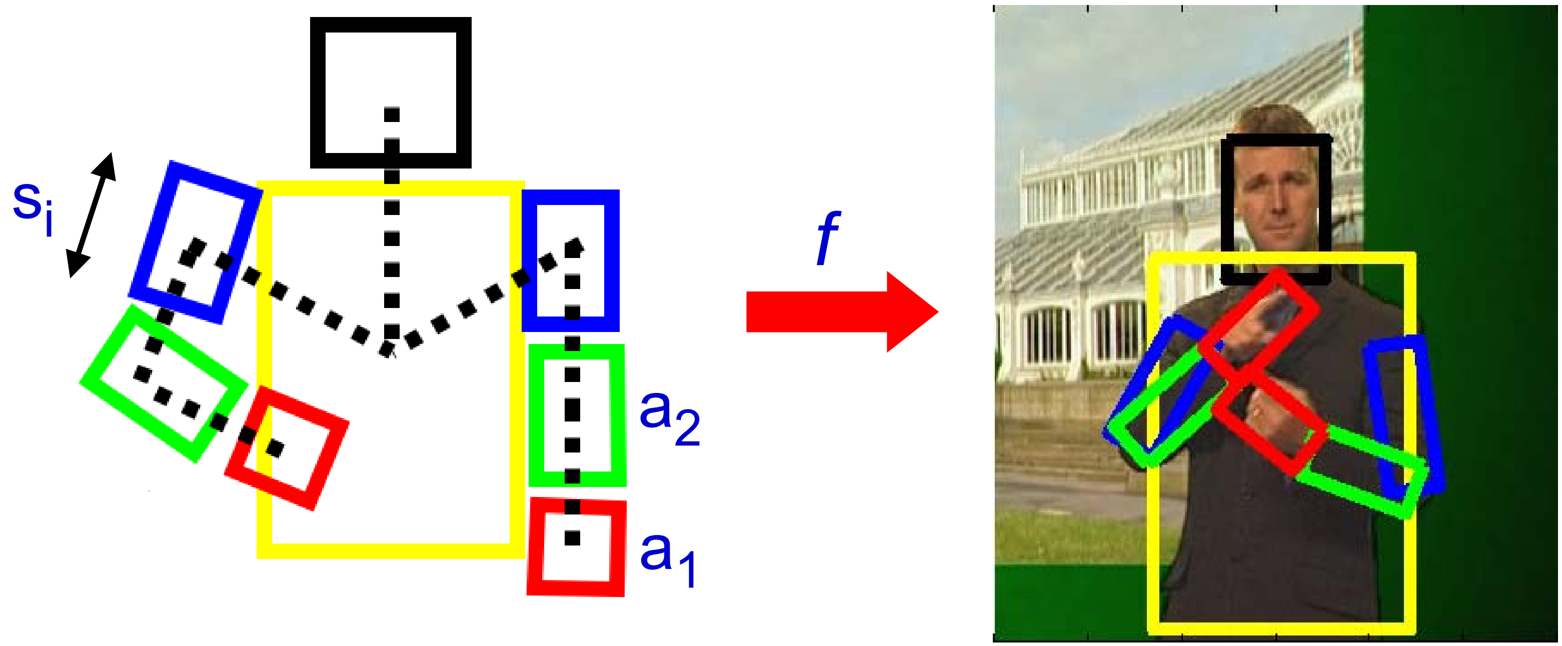
for limbs/torso

- Each labelling has an energy (cost):

$$E(f) = \underbrace{\sum_{a \in \mathcal{V}} \theta_{a;f(a)}}_{\substack{\text{unary terms} \\ \text{(appearance)}}} + \underbrace{\sum_{(a,b) \in \mathcal{E}} \theta_{ab;f(a)f(b)}}_{\substack{\text{pairwise terms} \\ \text{(configuration)}}}$$

- Fit model (inference) as labelling with lowest energy

Complexity

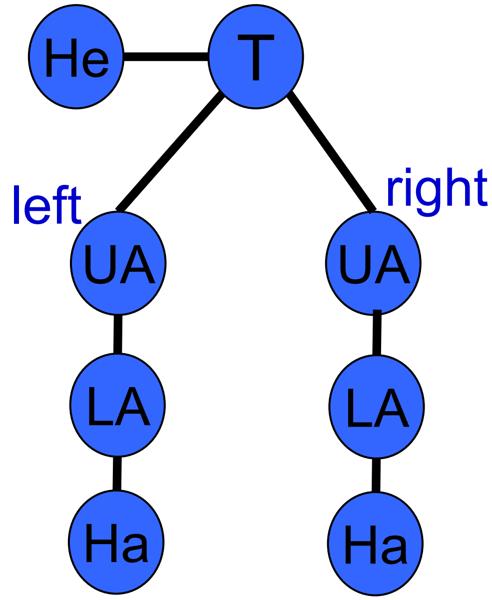


- n parts
- For each part there are h possible poses $\mathbf{p}_j = (x_j, y_j, \phi_j, s_j)$
- There are h^n possible labellings

Problem: any reasonable discretization (e.g. 12 scales and 36 angles for upper and lower arm, etc) gives a number of configurations 10^12 – 10^14

→ Brute force search not feasible

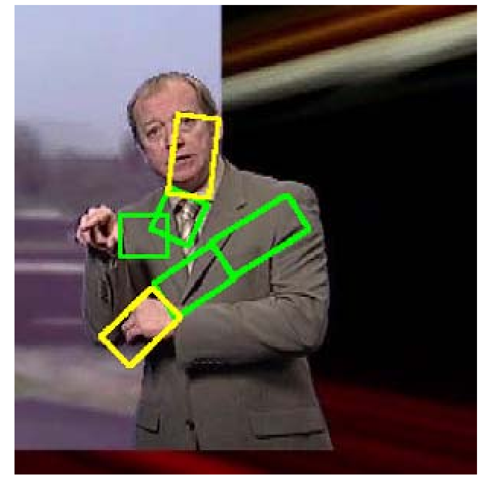
Are trees the answer?



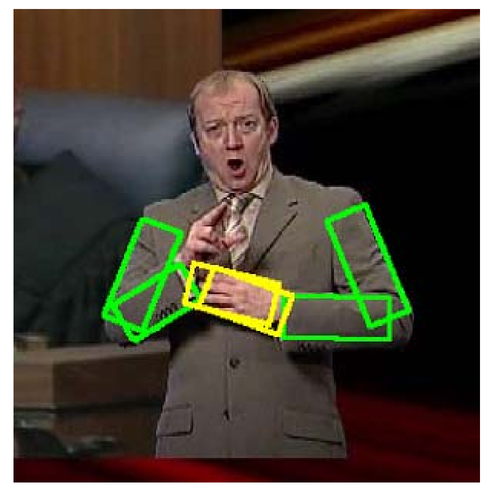
- With n parts and h possible discrete locations per part, $O(h^n)$
- For a tree, using dynamic programming this reduces to $O(nh^2)$
- If model is a tree and has certain edge costs, then complexity reduces to O(nh) using a distance transform [Felzenszwalb & Huttenlocher, 2000, 2005]

Problems with tree structured pictorial structures

- Layout model defines the foreground,
i.e. it chooses the pixels to "explain"



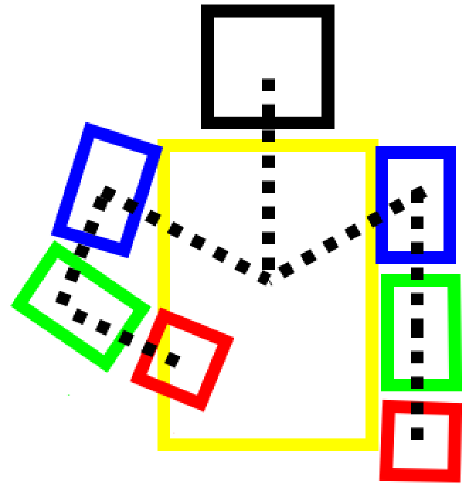
- ignores skin and strong edge in background



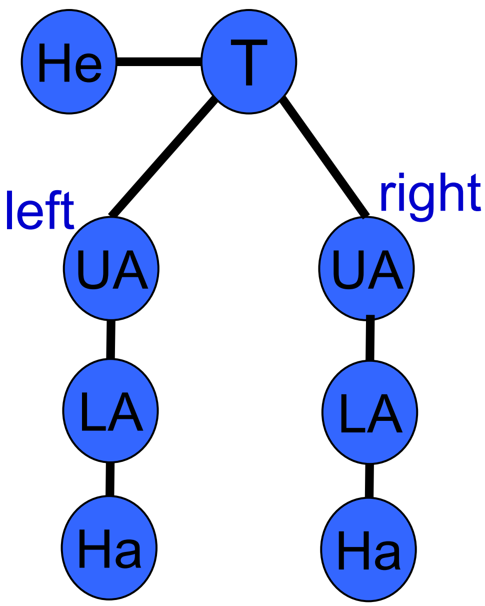
- "double counting"

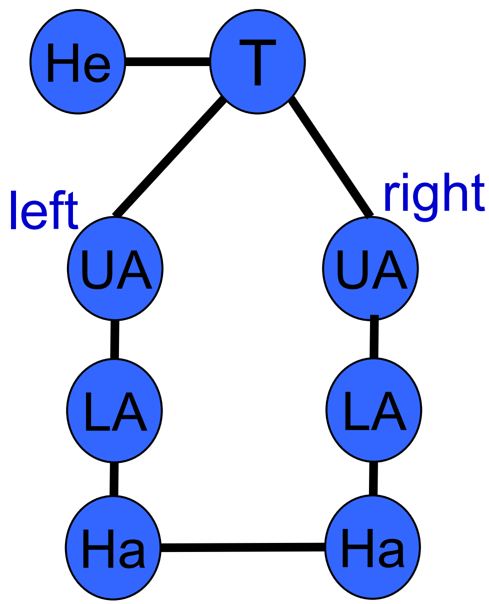
Generative model of foreground only

Kinematic structure vs graphical (independence) structure



Graph G = (V,E)





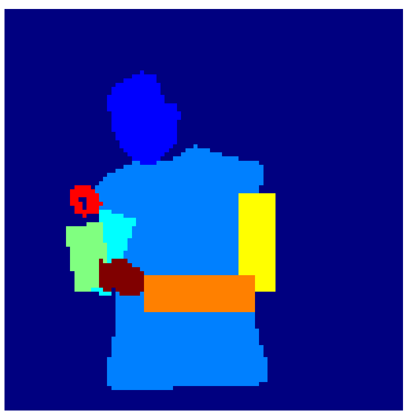
Requires more connections than a tree

And for the background problem

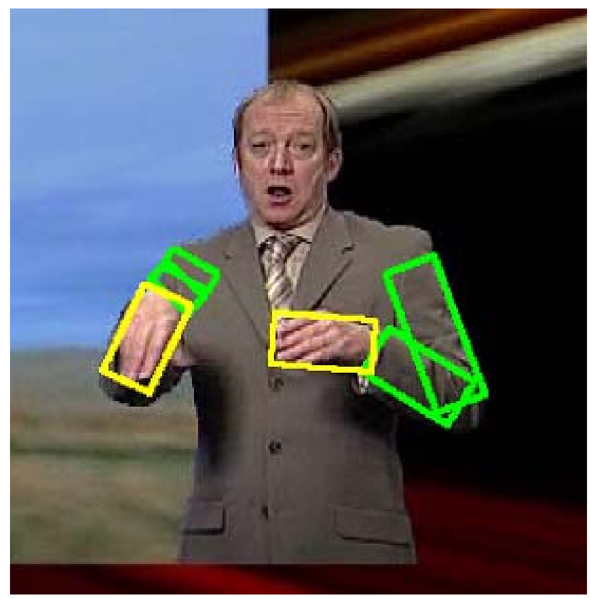
1. Add background model so that every pixel in region explained

$$E_{\text{full}} = E(f) + \sum_{\text{pixels } \mathbf{x}_i \text{ not in } f} E(\mathbf{x}_i | \text{bgcol})$$

2. *f* lays out parts in back-to-front depth order (painter's algorithm)



Colour is pixel-wise labelling
by parts (back-to-front)



Generative model of entire region

Outline

- Review of pictorial structures for articulated models
- Inference given the model: Strong supervision, full generative model – "Gold-standard model"
- Image parsing: learning the model for a specific image
- Recent advances
- Datasets and challenges

Long Term Arm and Hand Tracking for Continuous Sign Language TV Broadcasts

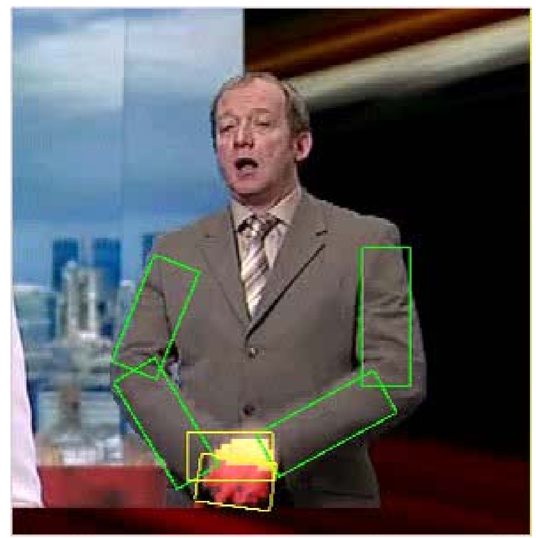
Patrick Buehler, Mark Everingham,

Daniel Huttenlocher, Andrew Zisserman

British Machine Vision Conference 2008

Objective

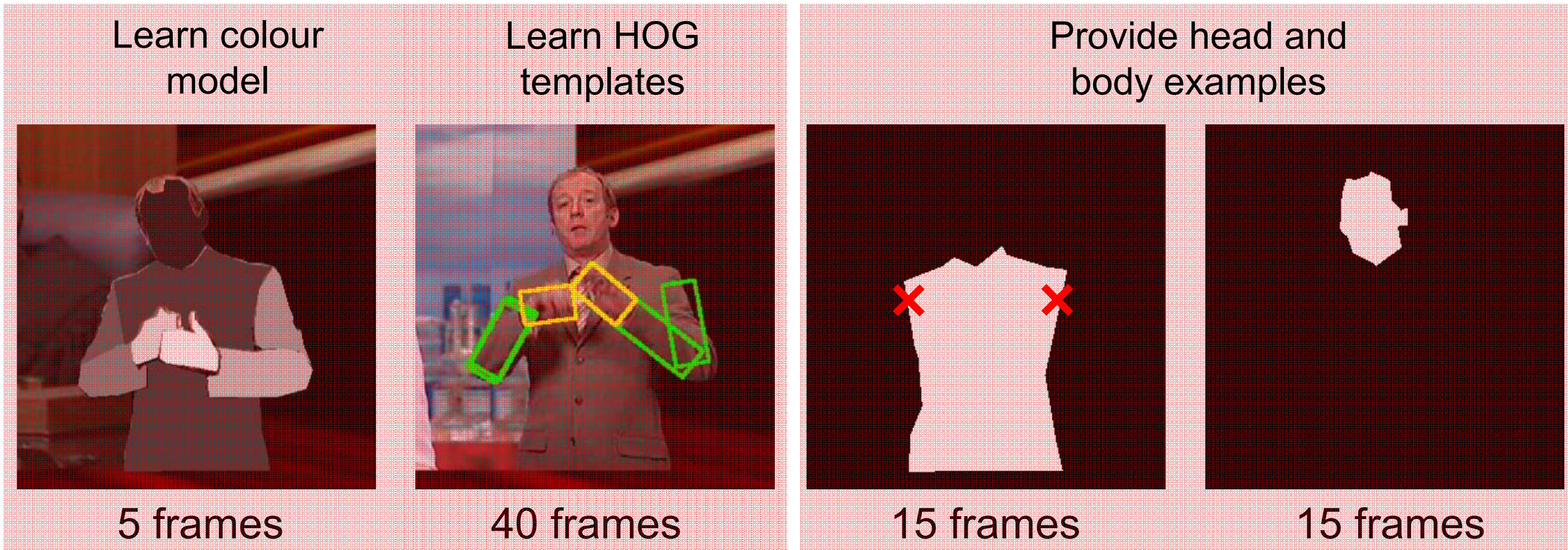
- Detect hands and arms of person signing British Sign Language
- Hour long sequences

- Strong but minimal supervision

Learning the model

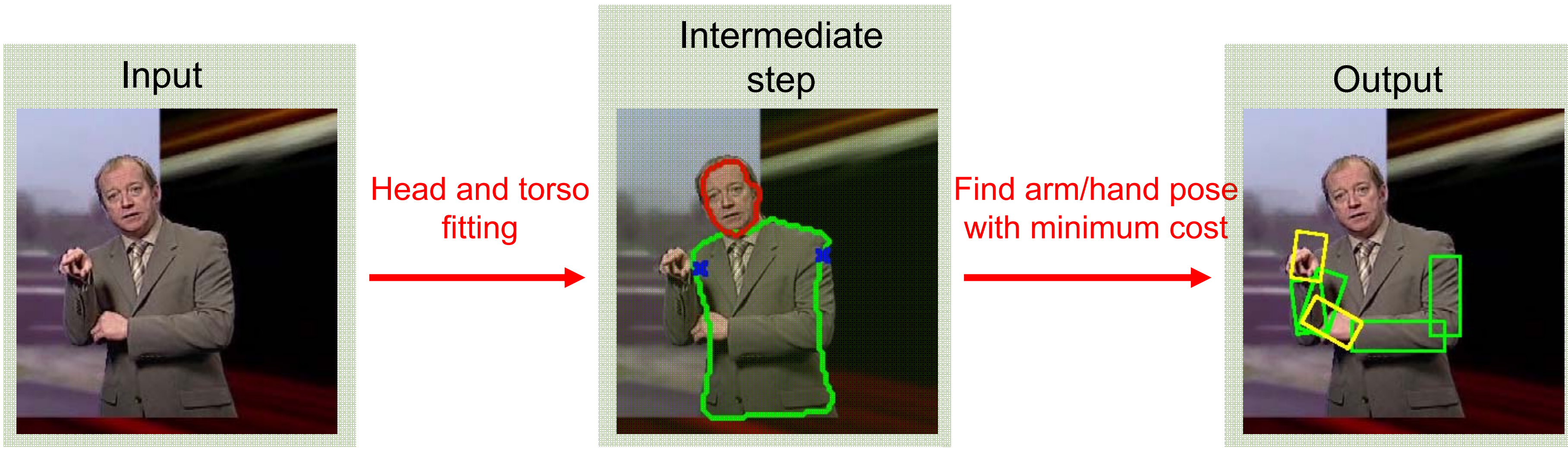
Strong supervision: manual input



40 annotated frames per video, used for pose estimation in > 50,000 frames

Inference (model fitting)

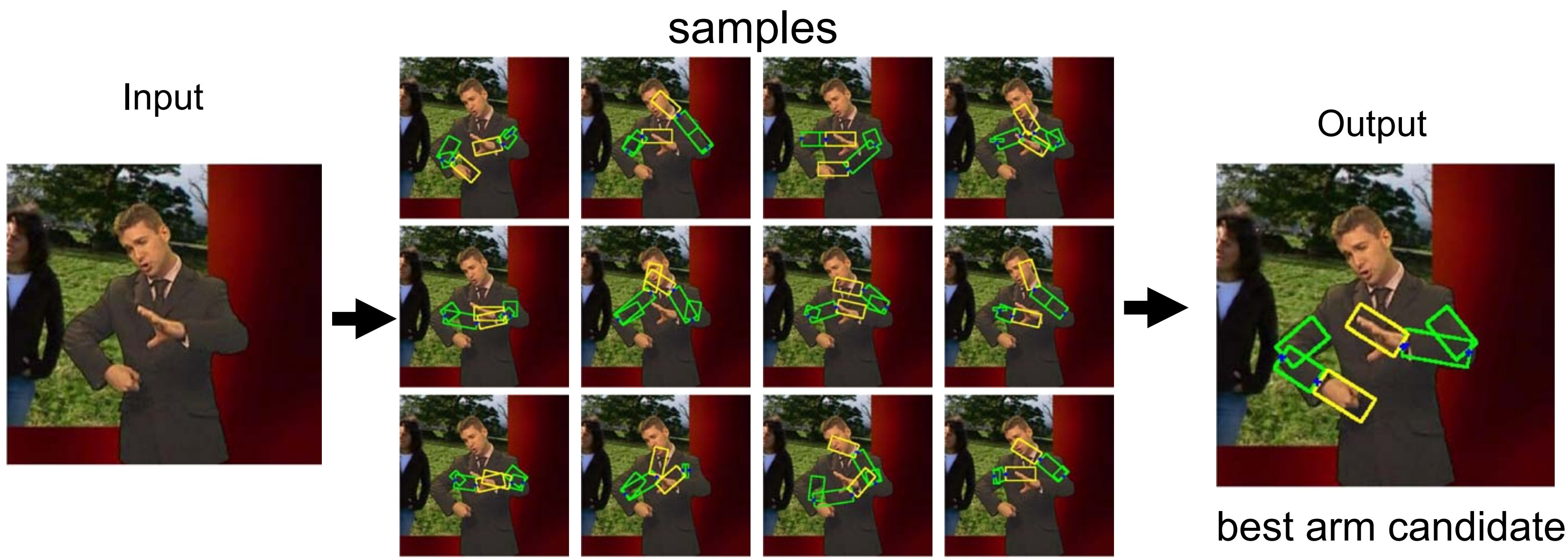
- Fit head and torso *[Navaratnam et al. 2005]*
- Then**:** arms and hands



Problem: Brute force search is still not feasible

Model fitting by sampling

- Sample configurations from inexpensive model
- Evaluate configuration using full model

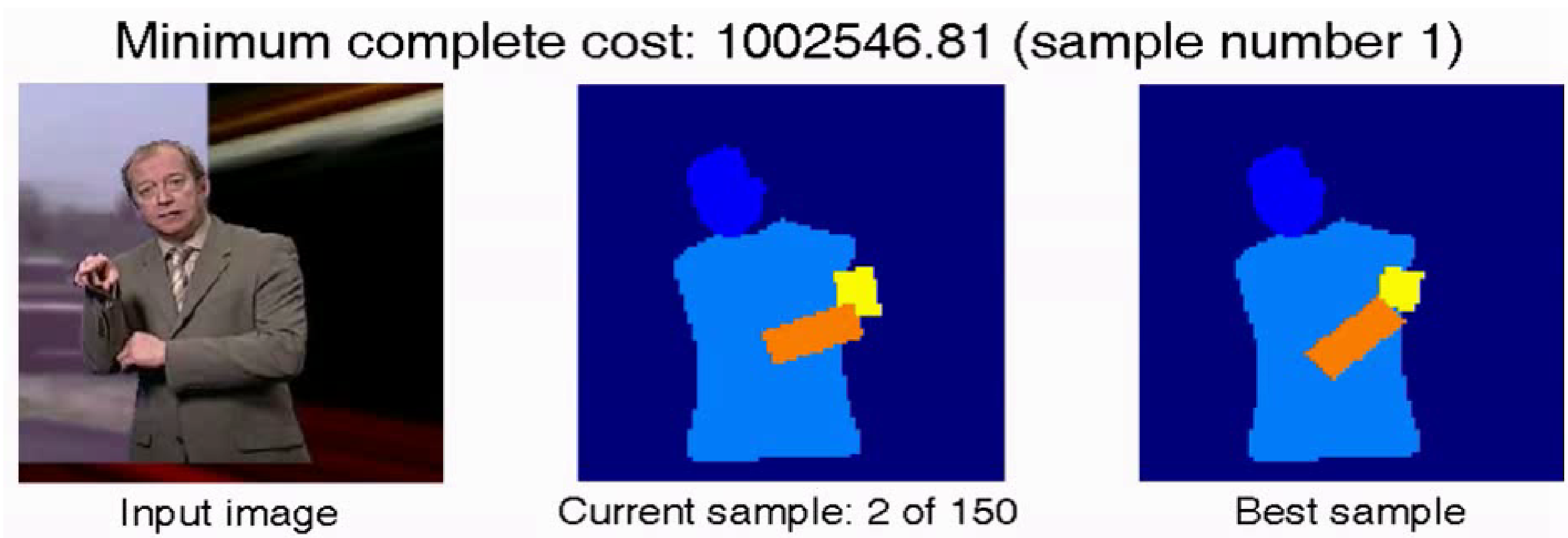


For sampling use tree structured pictorial Structures:

- [Felzenszwalb & Huttenlocher 2000, 2005]
- Complexity linear in the number of parts → O(nh)
- Pr(f | data): Sample from max-marginal with heuristics 1000 times
- cf Felzenszwalb & Huttenlocher 2005 sampled from marginal

Model fitting by sampling

- Sample configurations from inexpensive tree structured model
- Evaluate configuration using full model



Example results

Pose estimation results

Application

Learning sign language by watching TV (using weakly aligned subtitles)

Patrick Buehler

Mark Everingham

Andrew Zisserman

CVPR 2009

Objective

Learn signs in British Sign Language (BSL) corresponding to text words:

- Training data from TV broadcasts with simultaneous signing
- Supervision solely from sub-titles

Input: video + subtitle

Output: automatically learned signs (4x slow motion)

Office

Government

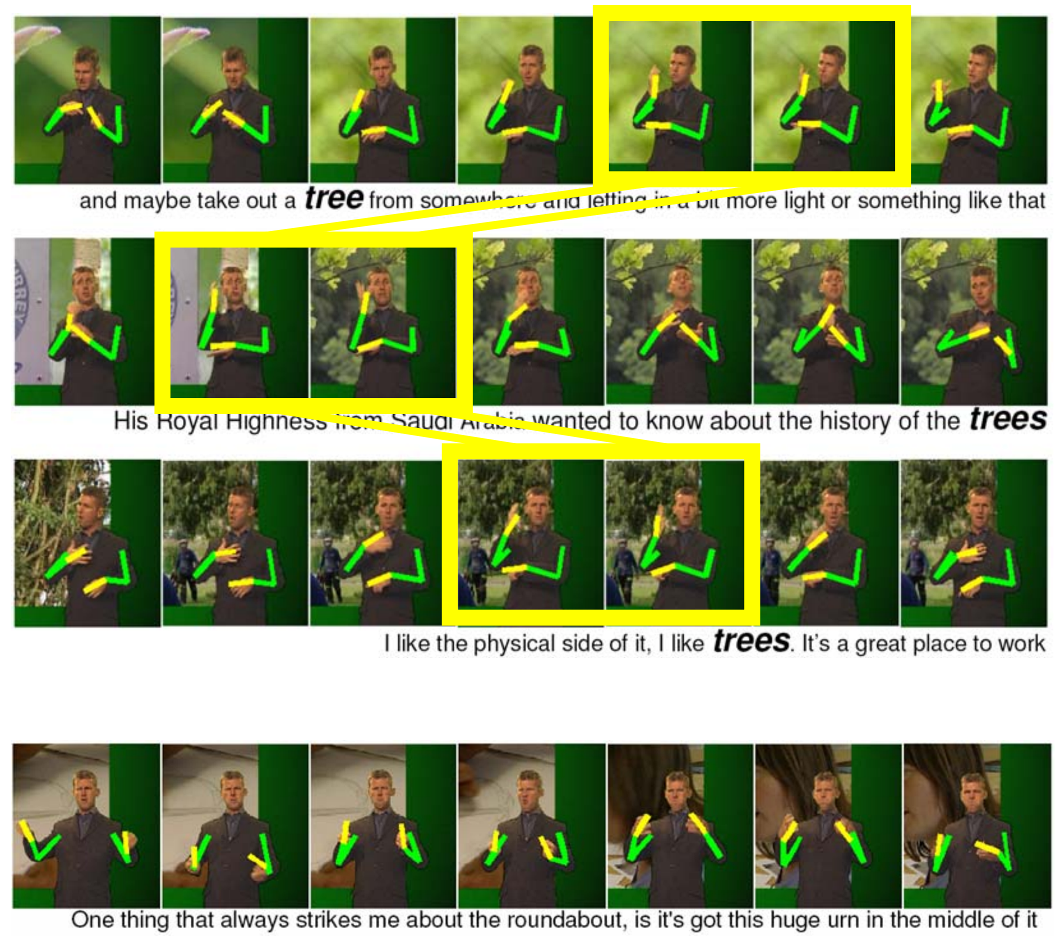
Use subtitles to find video sequences containing word. These are the positive training sequences. Use other sequences as negative training sequences.

Overview

Given an English word e.g. “tree” what is the corresponding British Sign Language sign?

positive sequences

negative set



Use sliding window to choose sub-sequence of poses in one positive sequence and determine if

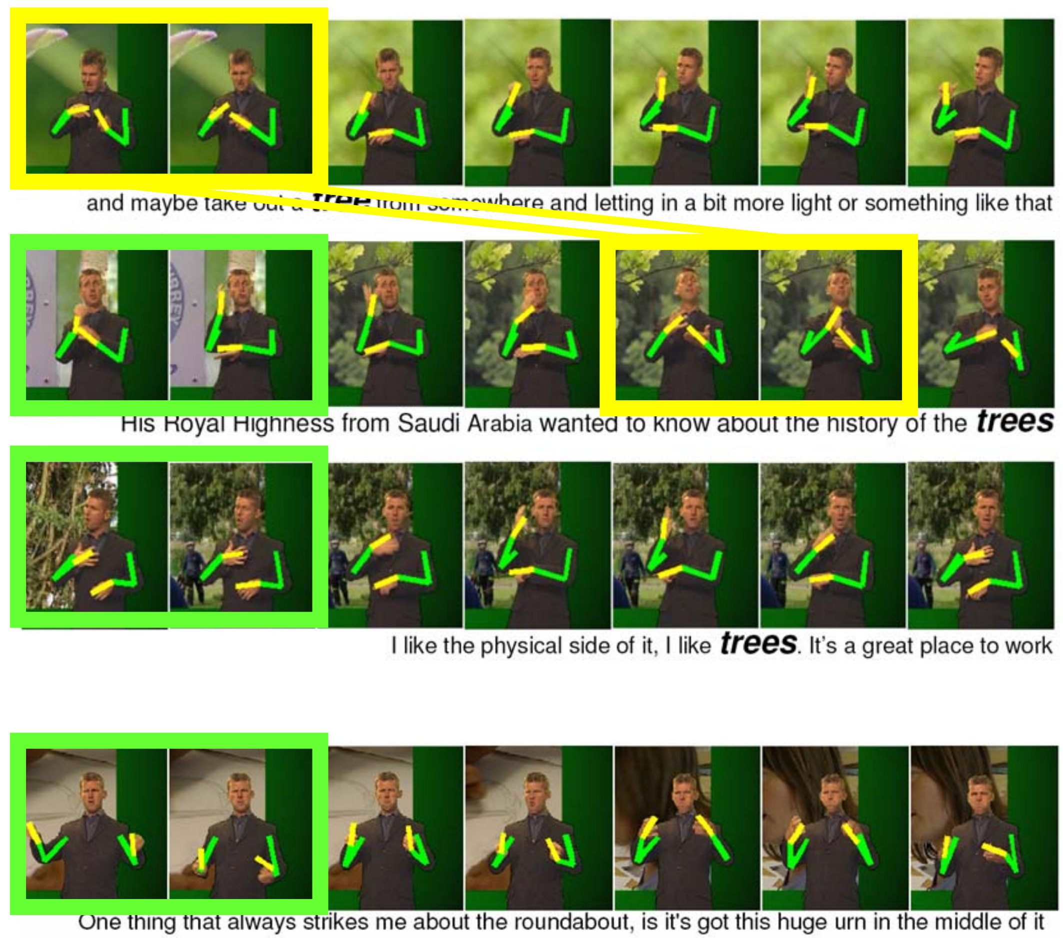
same sub-sequence of poses occurs in other positive sequences somewhere, but

does not occur in the negative set

1st sliding window

positive sequences

negative set



Use sliding window to choose sub-sequence of poses in one positive sequence and determine if

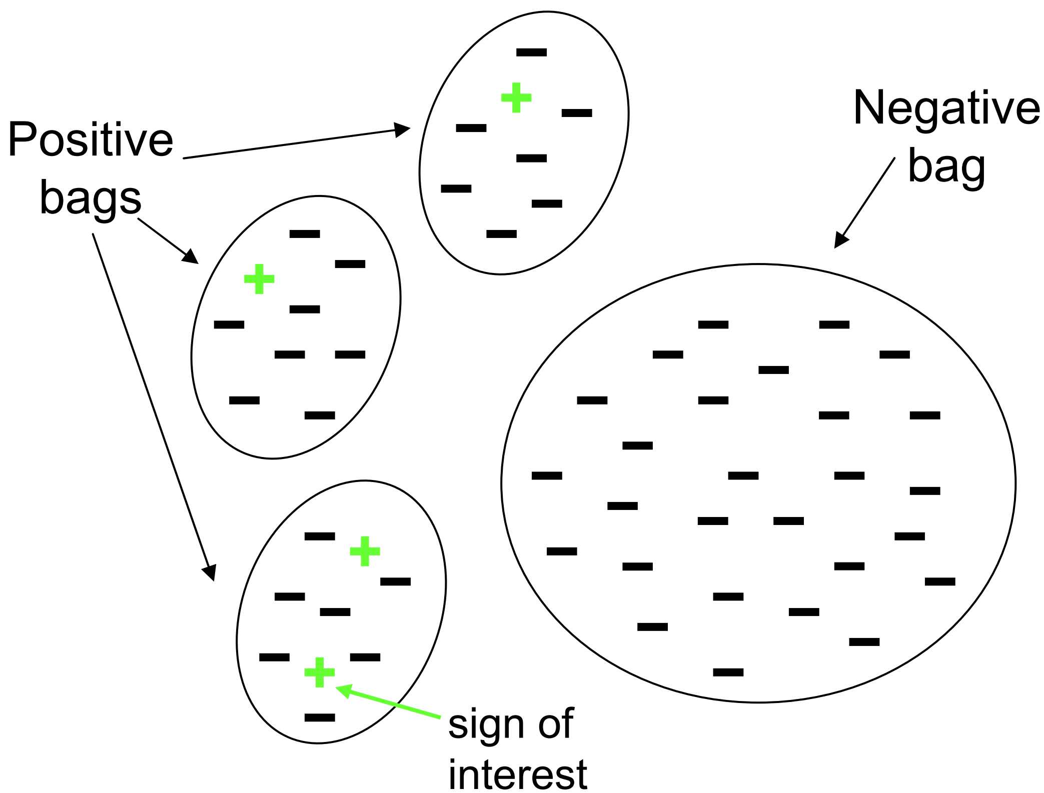
same sub-sequence of poses occurs in other positive sequences somewhere, but

does not occur in the negative set

positive sequences

negative set

Multiple instance learning



Evaluation

Good results for a variety of signs:

Signs where hand movement is important

↓

Navy

Prince

Signs where hand shape is important

↓

Lung

Garden

Signs where both hands are together

↓

Fungi

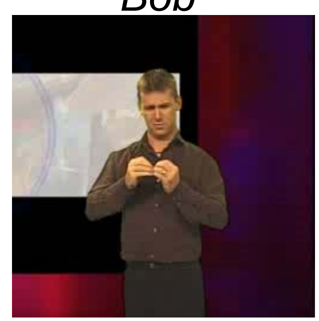
Golf

Signs which are finger-- spelled

↓

Kew

Bob



Signs which are perfomed in front of the face

↓

Whale

Rose

Summary

Given a good appearance model and proper account of foreground and background, then problems such as occlusion and ordering can be resolved. The cost of inference still remains though.

Next:

- How to obtain models automatically in videos and images
- If the appearance features are discriminative, how far can one go with foreground only pictorial structures and tree based inference?

Outline

- Review of pictorial structures for articulated models

- Inference given the model: Strong supervision, full generative model – “Gold-standard model”

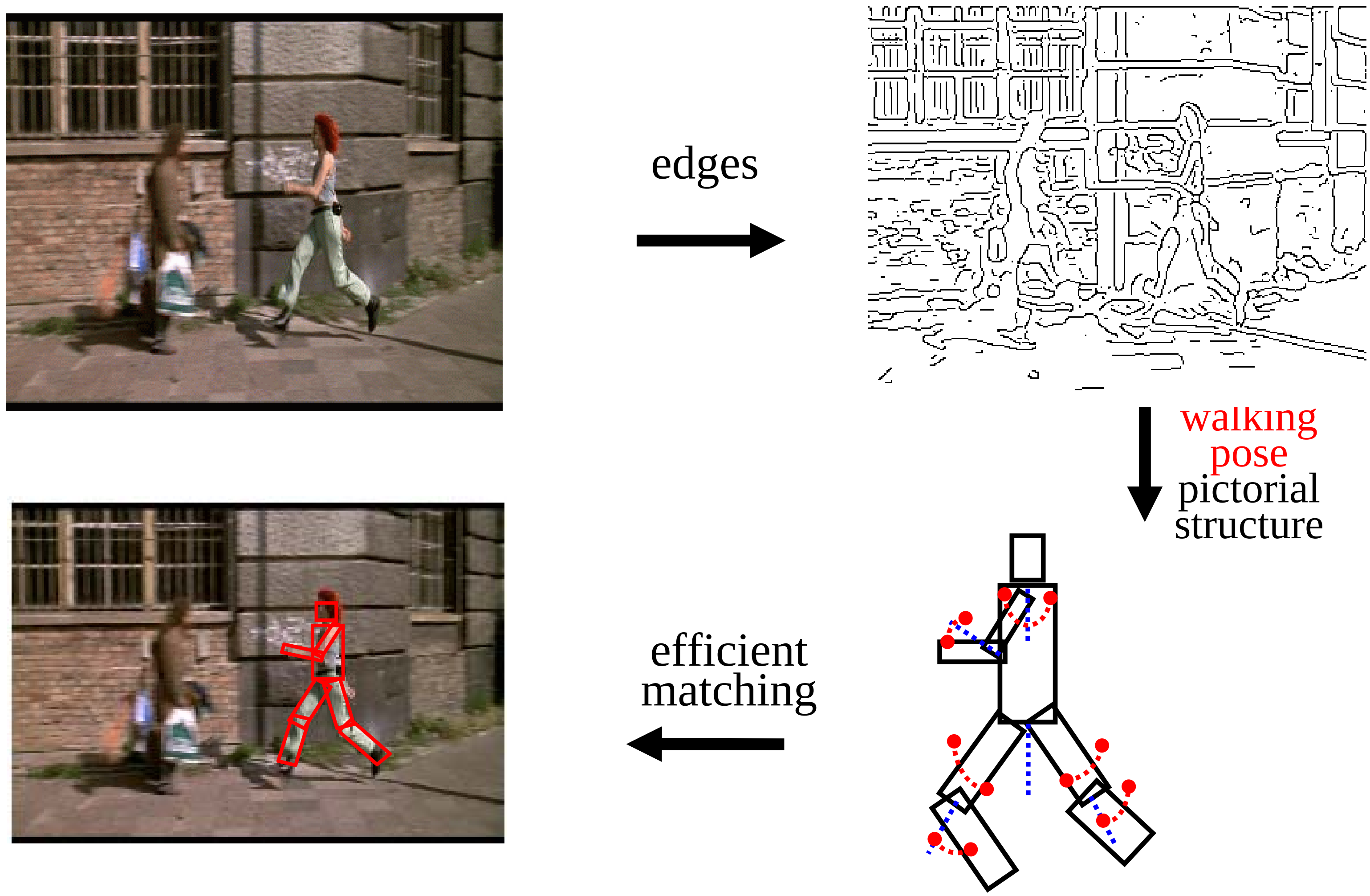
- Image parsing: learning the model for a specific image

- Recent advances

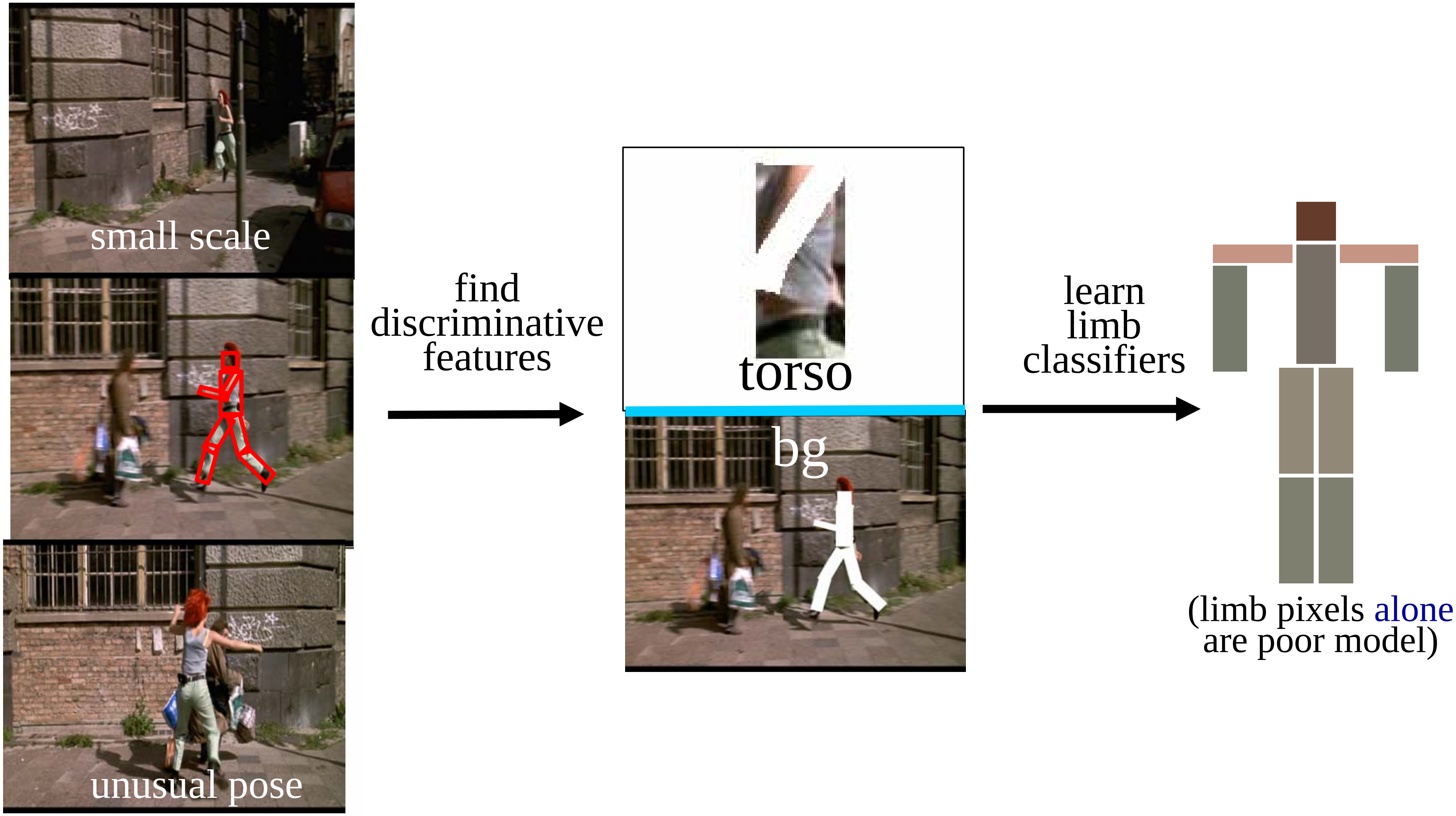
- Datasets and challenges

Learning appearance models in videos

Strike a Pose: Tracking People by Finding Stylized Poses
Deva Ramanan, David Forsyth and Andrew Zisserman, CVPR 2005

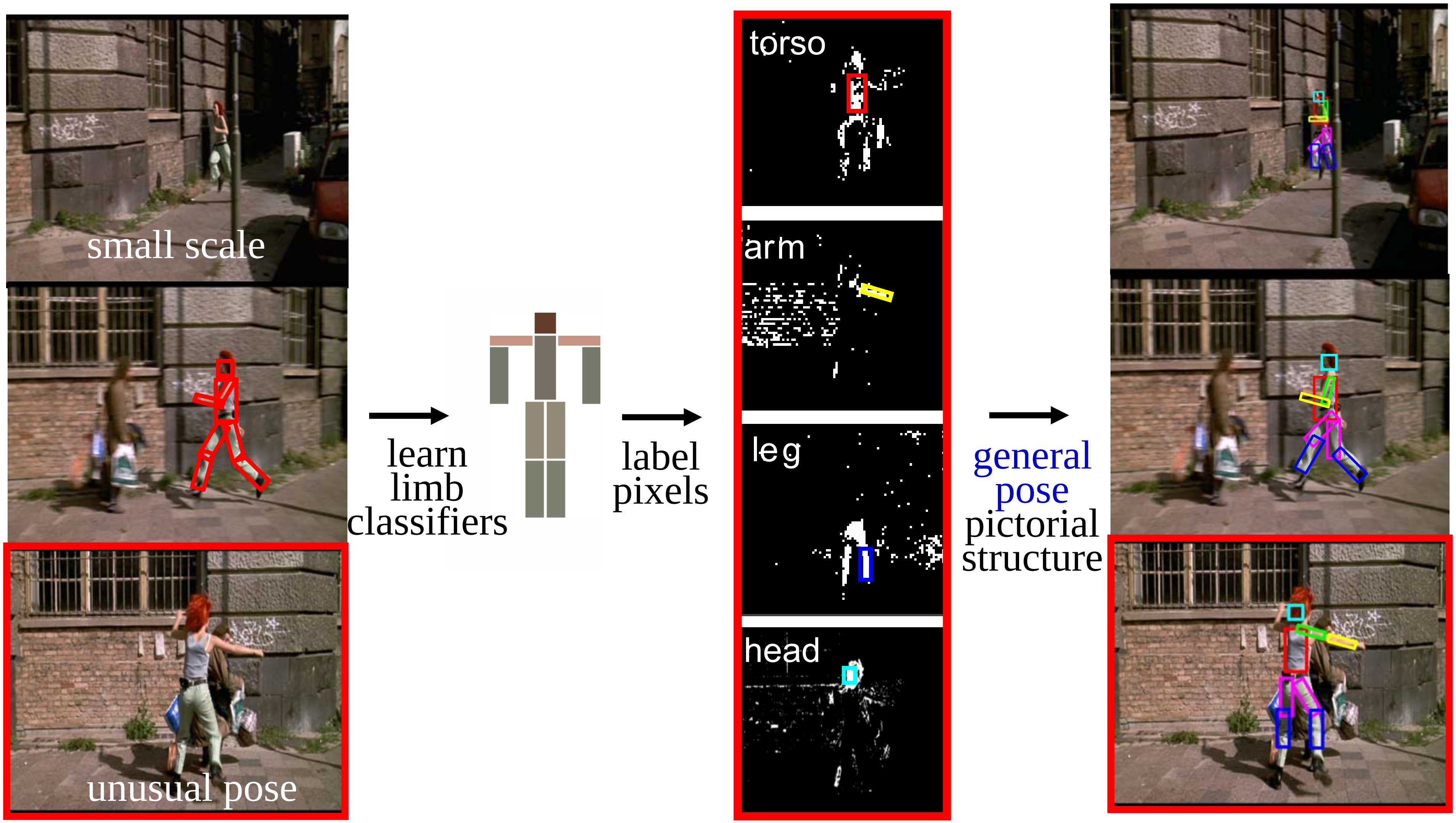

edges
walking
pose
pictorial
structure
efficient
matching

Build Model

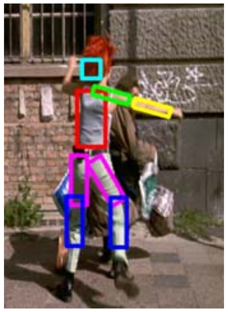


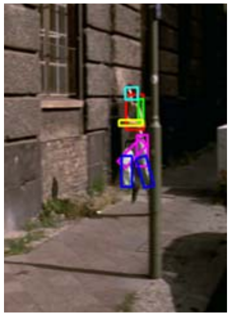
small scale
unusual pose
find discriminative features
torso
bg
learn limb classifiers
(limb pixels alone are poor model)

Build Model & Detect

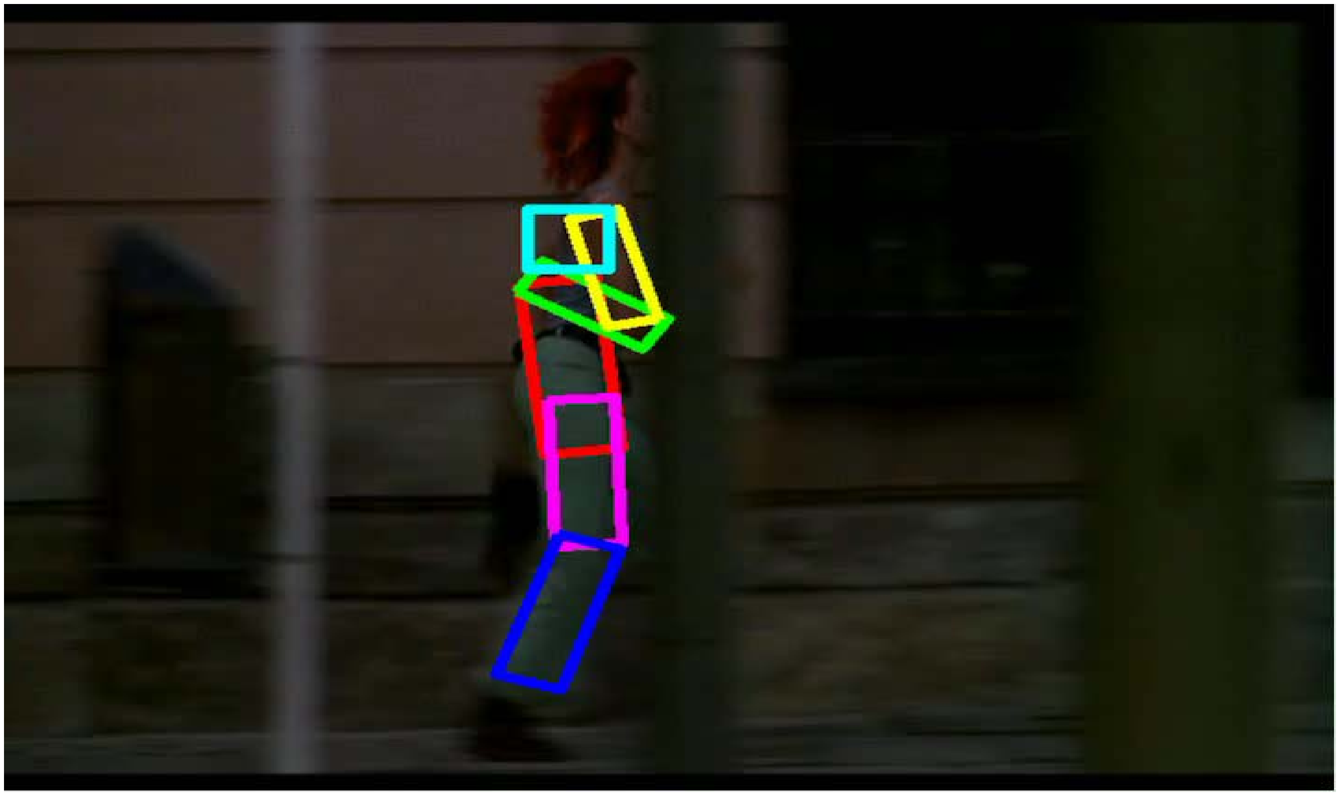


Running Example





How well do classifiers generalize?



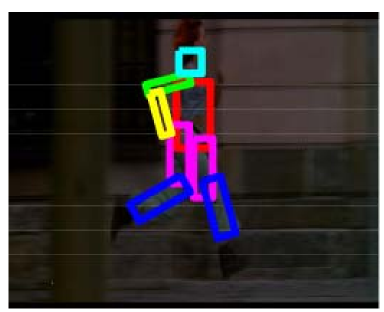
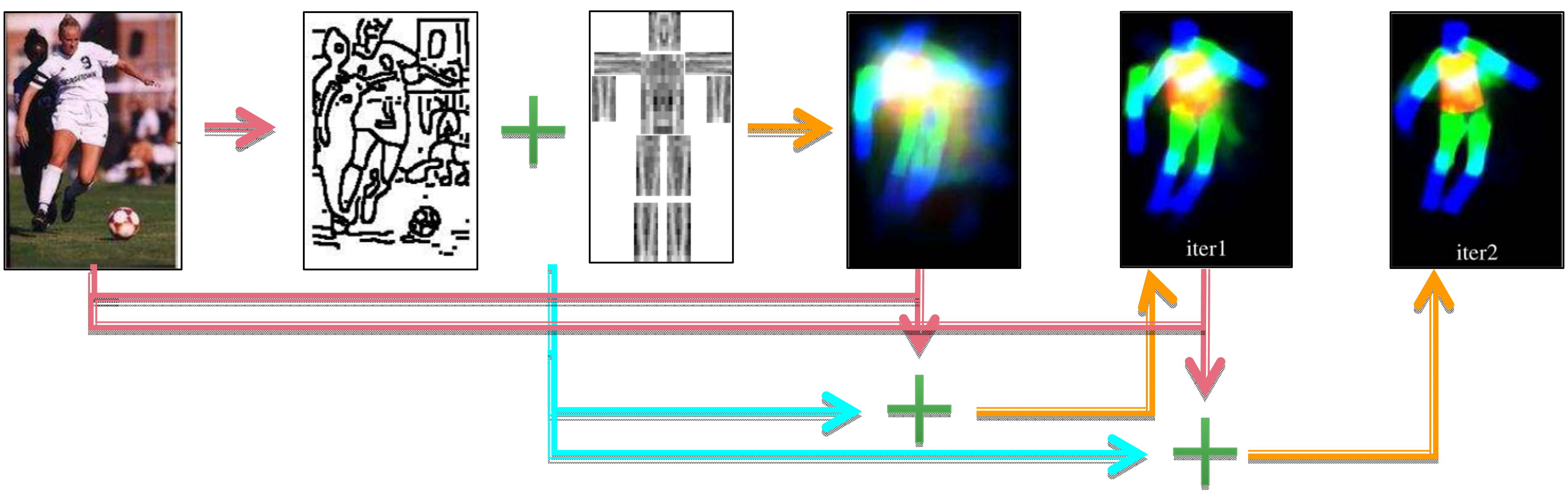


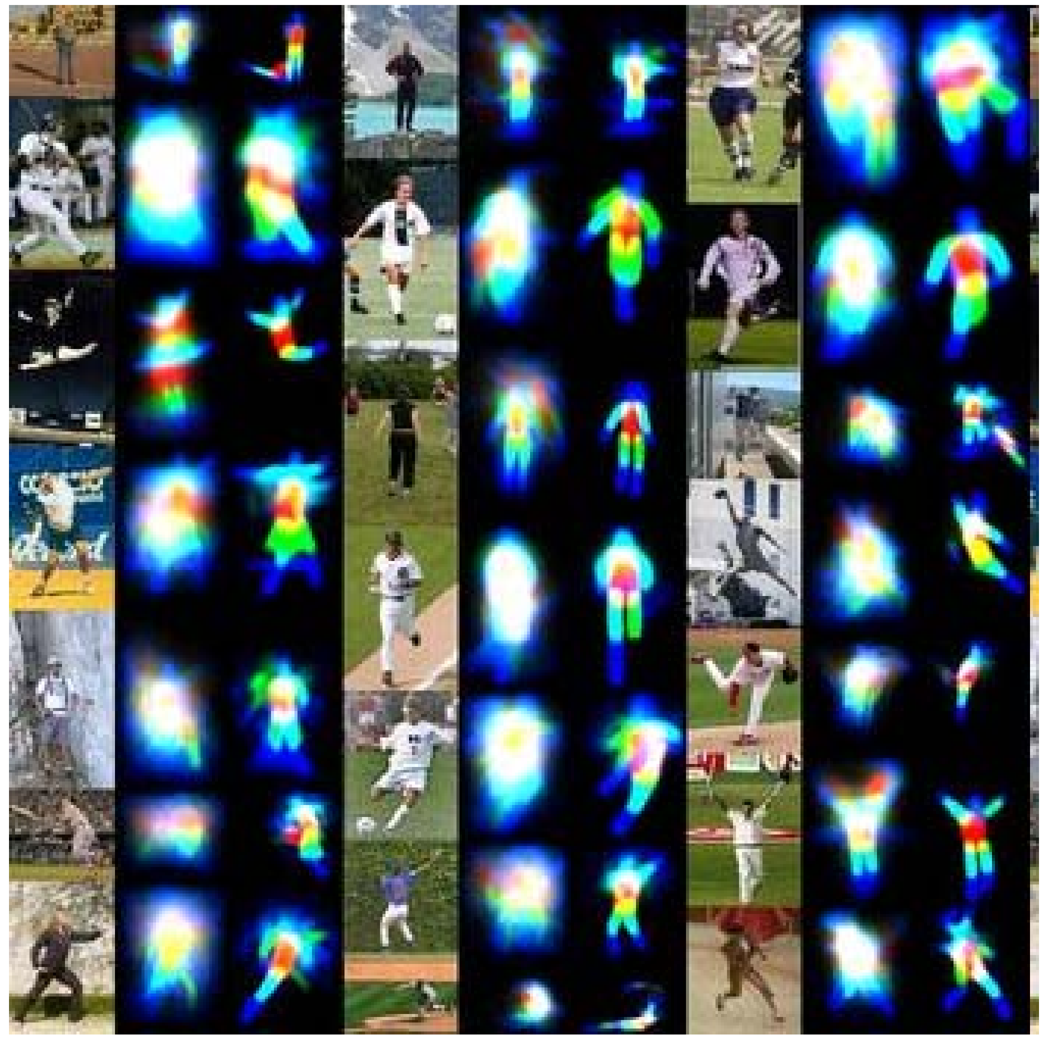
Image Parsing – Ramanan NIPS 06

$$E(f) = \underbrace{\textstyle\sum_{a \in \mathcal{V}} \theta_{a;f(a)}}_{\text{unary terms (edges/colour)}} + \underbrace{\textstyle\sum_{(a,b) \in \mathcal{E}} \theta_{ab;f(a)f(b)}}_{\text{pairwise terms (configuration)}}$$

Learn image and person specific unary terms

- initial iteration → edges
- following iterations → edges & colour





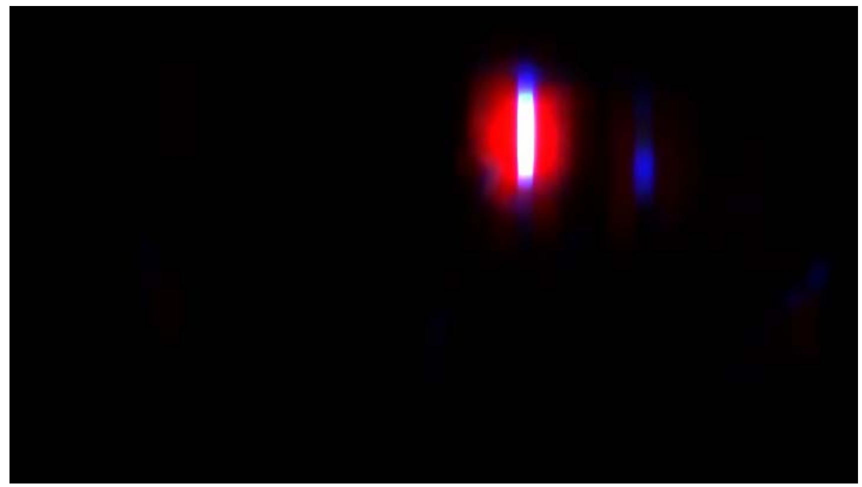
(Almost) unconstrained images




Extremely difficult when knowing nothing about appearance/pose/location

Failure of direct pose estimation

Ramanan NIPS 2006 unaided



Not powerful enough for a cluttered image where size is not given

Progressive search space reduction for human pose estimation

Vitto Ferrari, Manuel Marin-Jimenez, Andrew Zisserman

CVPR 2008/2009

Restrict search space using detector

Find (x,y,s) coordinate frame for a person

detection window (upper-body, face etc.)

Ferrari et al. 08, Andriluka et al. 09, Gammeter et al. 08

Learn an image and person specific model

Supervision

- None

Weaker model

- Tree structured graphical model
- Overlap not modelled
- Single scale parameter
- No background model

Inference

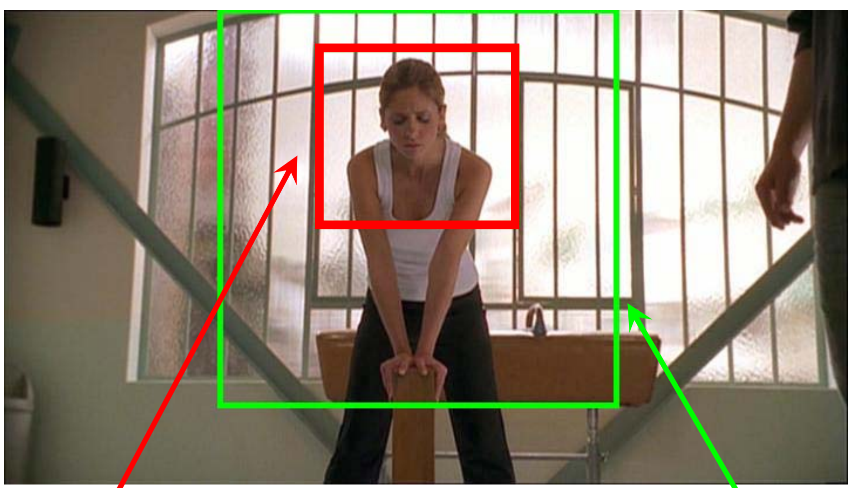
- **Detect person** – use upper body detector
- Use upper body region to restrict search
- Use colour segmentation to restrict search further
- Parsing pictorial structure by Ramanan NIPS 06

Search space reduction by upper body human detection

(1) detect human; (2) reduce search from h^n

Train

Test



Idea

get approximate location and scale with a detector generic over pose and appearance

Building an upper-body detector

- based on Dalal and Triggs CVPR 2005

- train = 96 frames X 12 perturbations

Benefits for pose estimation

+ fixes scale of body parts

+ sets bounds on x,y locations

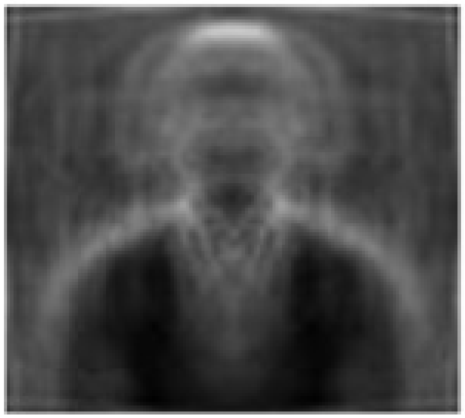
+ detects also back views

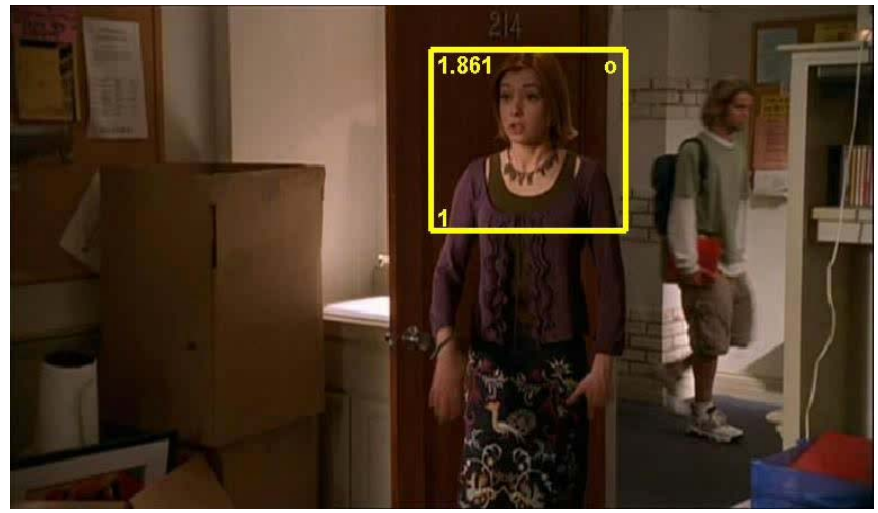
+ fast

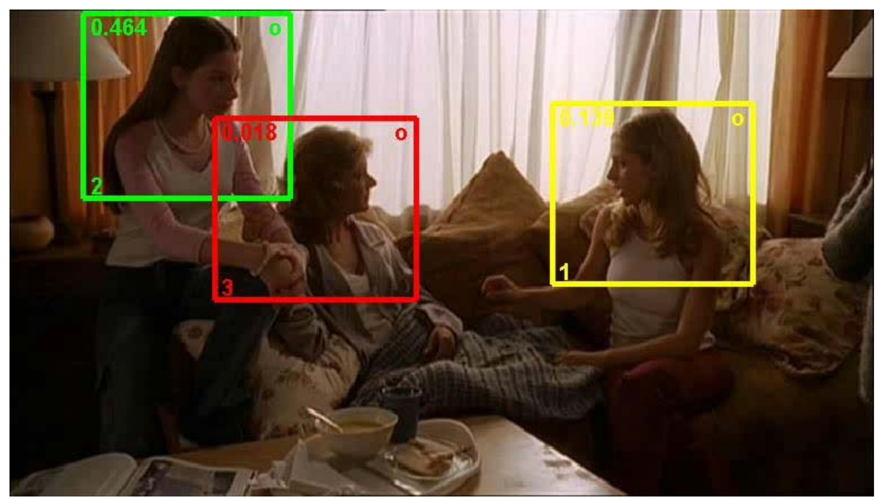
- little info about pose (arms)

Upper body detector – using HOGs

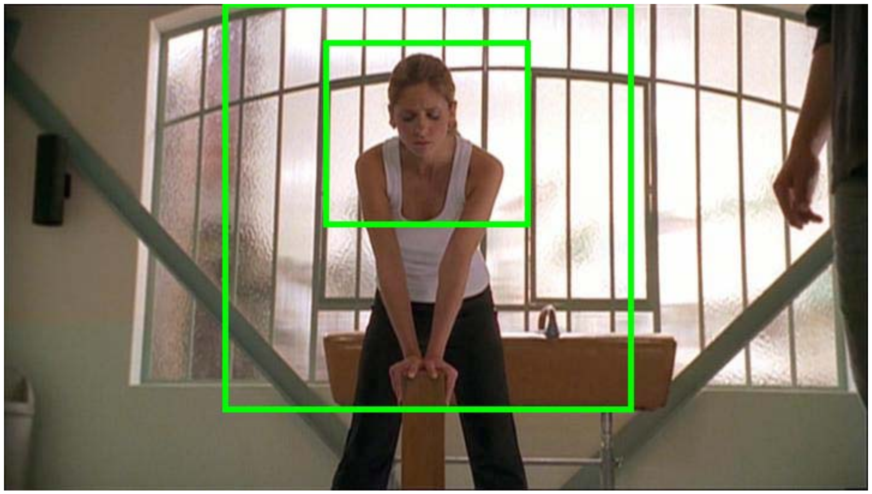
average training data







Search space reduction by foreground highlighting



initialization

output

Idea

exploit knowledge about structure of search area to initialize Grabcut

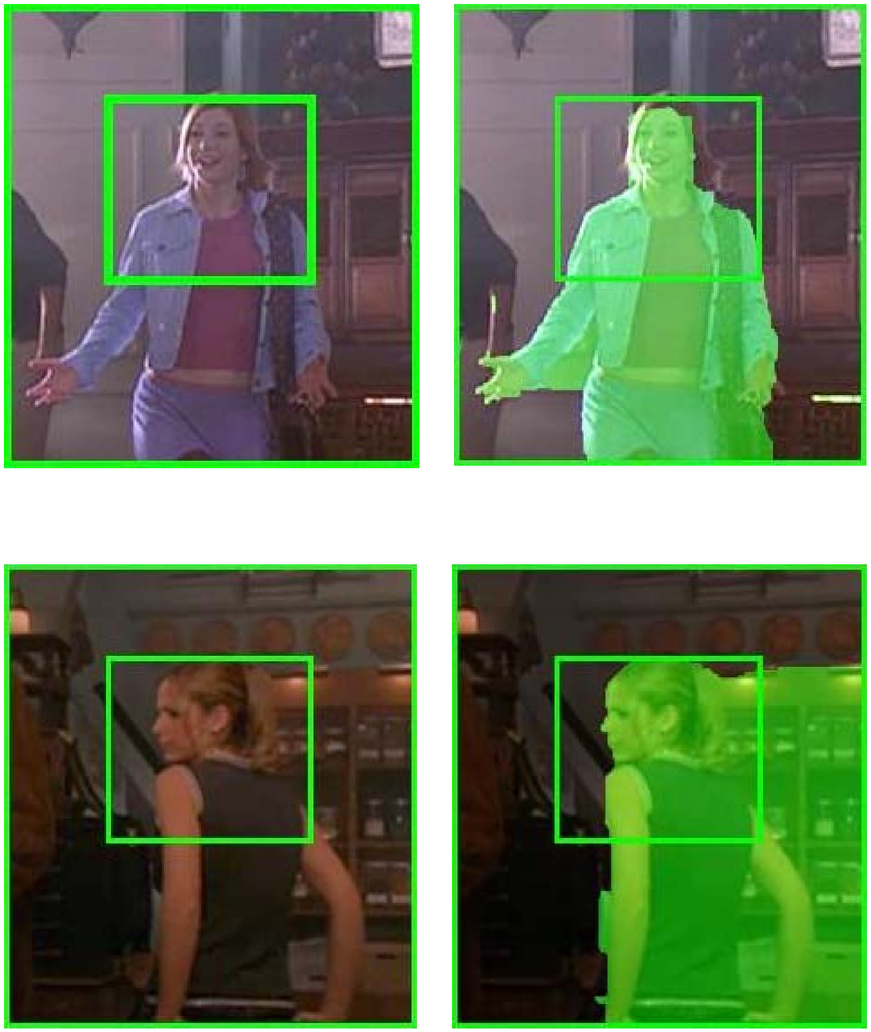
Initialization

- learn fg/bg models from regions where person likely present/absent
- clamp central strip to fg
- don't clamp bg (arms can be anywhere)

Benefits for pose estimation

+ further reduce clutter
+ conservative (no loss 95.5% times)
+ needs no knowledge of background
+ allows for moving background

Search space reduction by foreground highlighting



Idea

exploit knowledge about structure of search area to initialize Grabcut

Initialization

- learn fg/bg models from regions where person likely present/absent
- clamp central strip to fg
- don’t clamp bg (arms can be anywhere)

Benefits for pose estimation

+ further reduce clutter
+ conservative (no loss 95.5% times)
+ needs no knowledge of background
+ allows for moving background

Pose estimation by image parsing - Ramanan NIPS 06

edge parse

appearance

edge + col parse

Goal

estimate posterior of part configuration

$$E(f) = \underbrace{\textstyle\sum_{a \in \mathcal{V}} \theta_{a;f(a)}}_{\text{unary terms (edges/colour)}} + \underbrace{\textstyle\sum_{(a,b) \in \mathcal{E}} \theta_{ab;f(a)f(b)}}_{\text{pairwise terms (configuration)}}$$

Algorithm

1. inference with edges unary
2. learn appearance models of body parts and background
3. inference with edges + colour unary

Advantages of space reduction

+ much more robust

+ much faster (10x-100x)

Failure of direct pose estimation

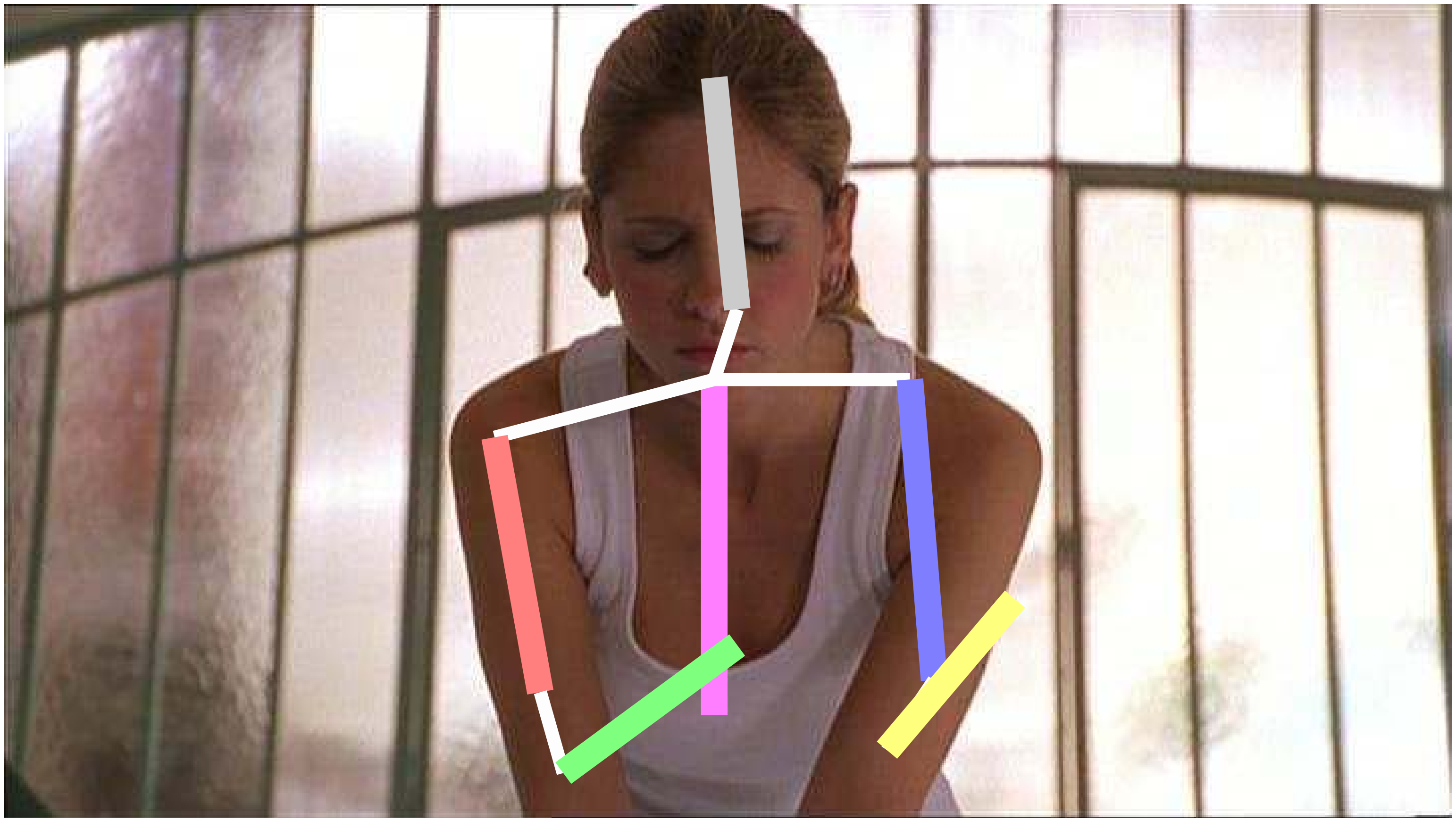
Ramanan NIPS 2006 unaided

Results on Buffy frames

Results on PASCAL flickr images

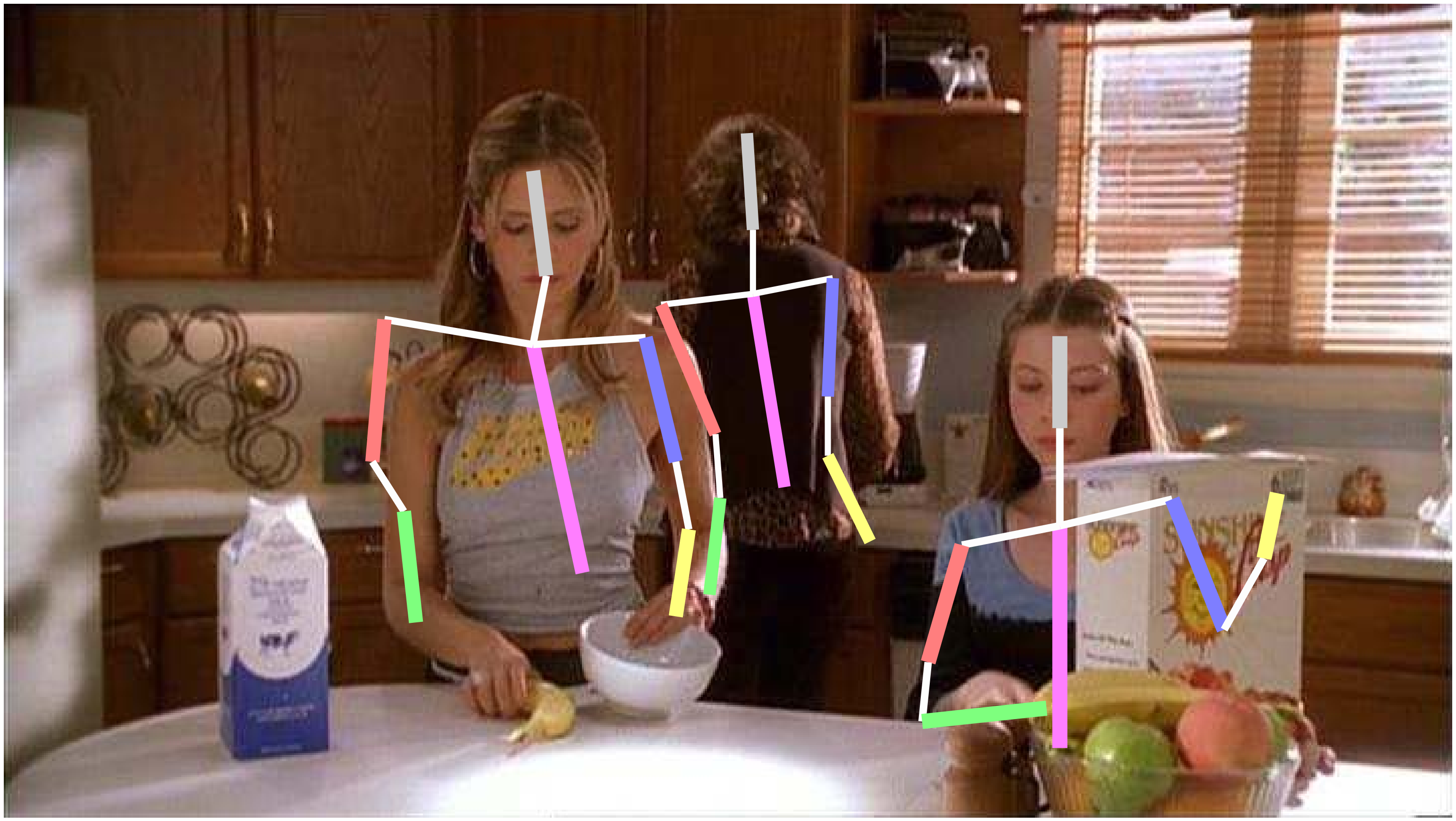
What is missed?

What is missed?



truncation is not modelled

What is missed?



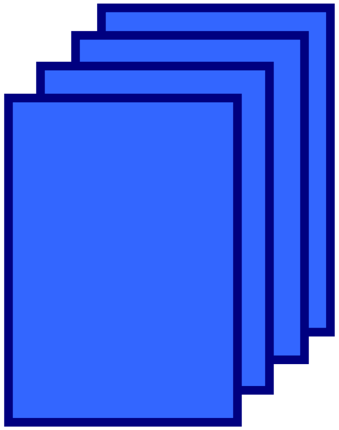
occlusion is not modelled

Application: Pose Search

Given user-selected
query frame+person …

query

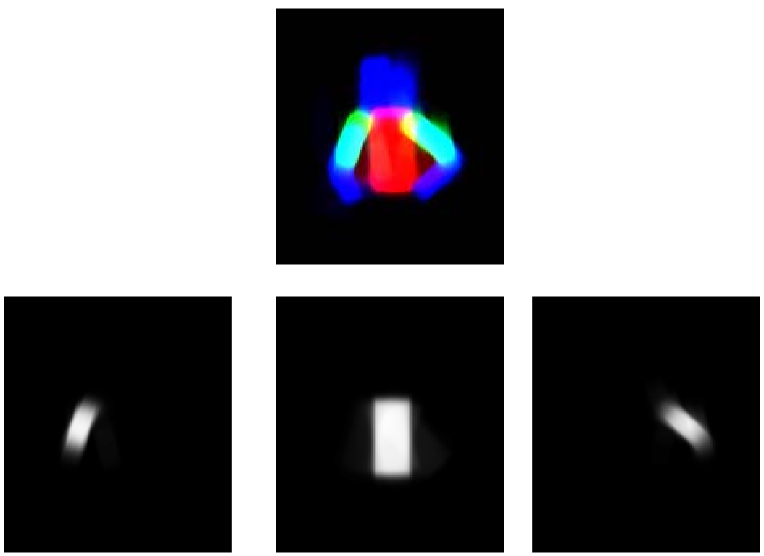
… retrieve shots with persons
in the same pose from video database



video database

CVPR 2009

Pose Search



Pose descriptors

- soft-segmentations of body parts
- distributions over orient+location for parts and pairs of parts

Similarity measures

- dot-product (= soft intersection)
- Batthacharrya / Chi-square

Processing

Off-line:

- Detect upper bodies in every frame
- Link (track) upper body detections
- Estimate upper body pose for each frame of track
- Compute descriptor (vector) for each upper body pose

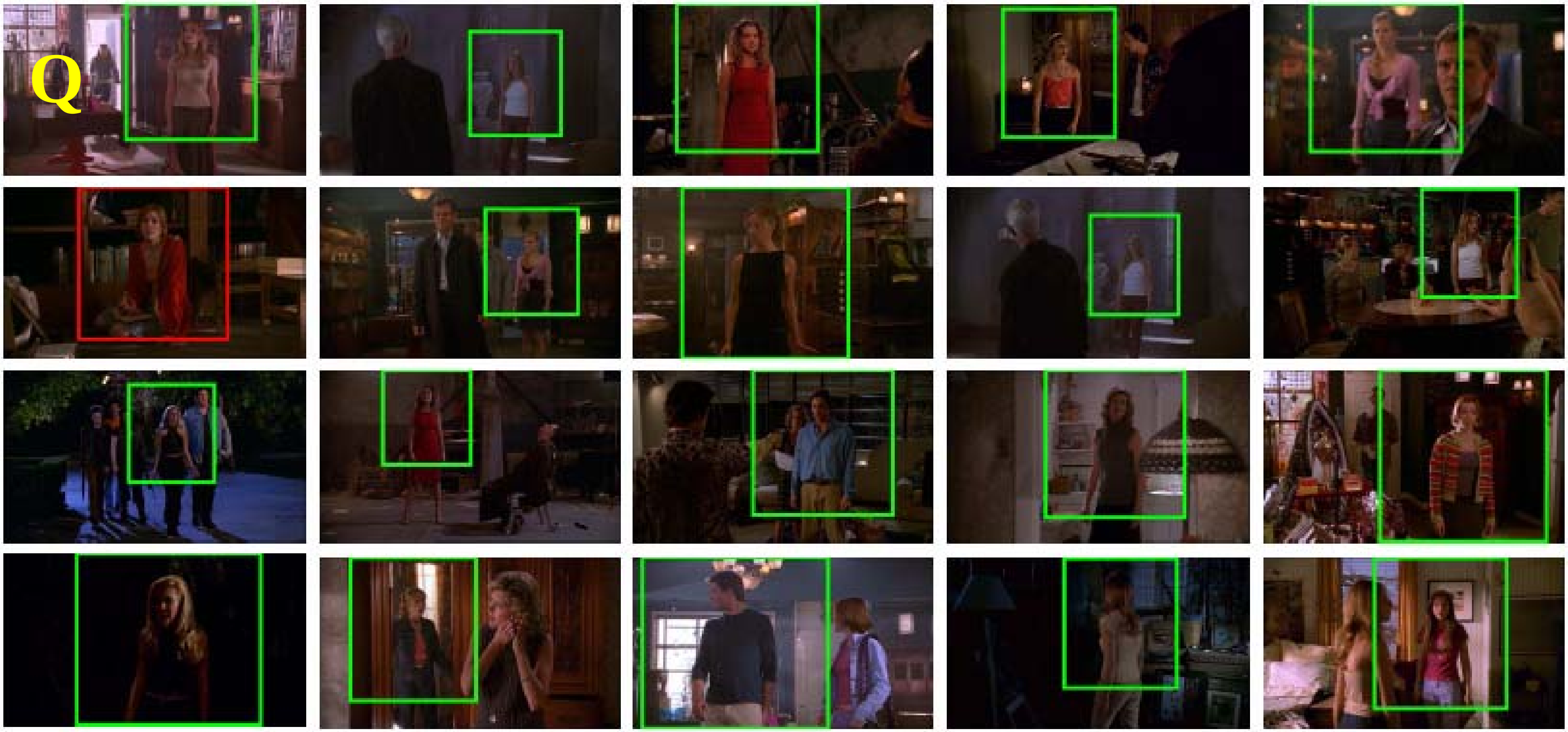
Run-time:

- Rank each track by its similarity to the query pose

Pose Search

“hips pose”

Pose Search



“rest pose”

Pose Search

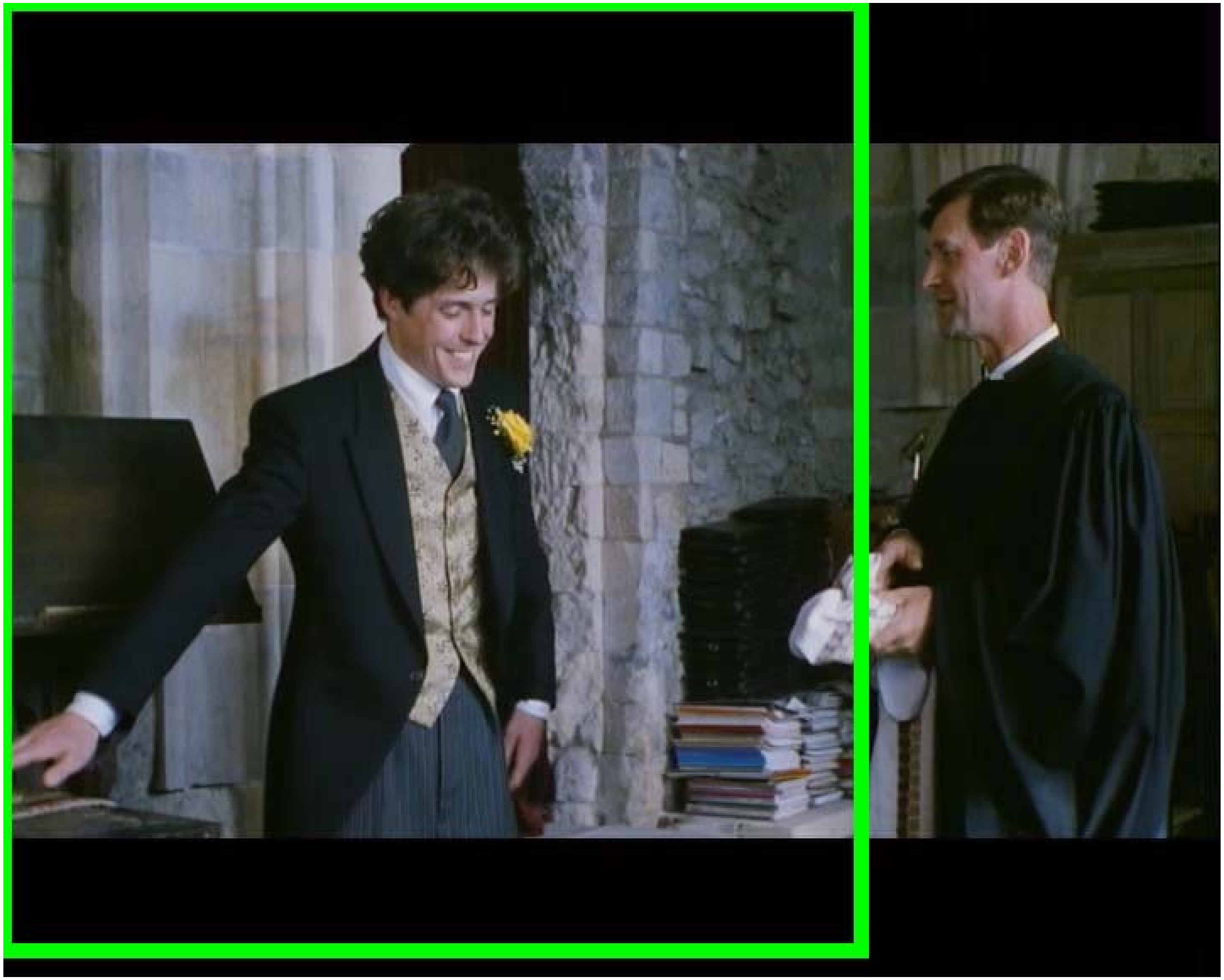
“rest pose”

Other poses – query interesting pose

Hollywood movies – Query on Gandhi, Search Hugh Grant opus

Other poses – query interesting pose

Hollywood movies – Query on Gandhi, Search Hugh Grant opus



Outline

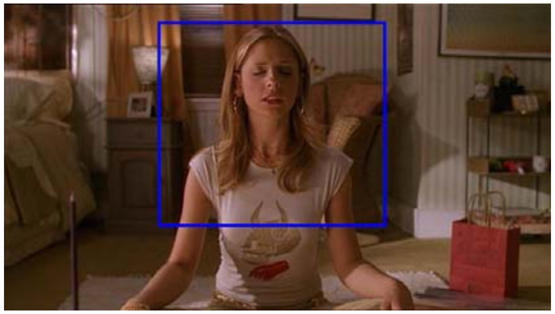
- Review of pictorial structures for articulated models
- Inference given the model: Strong supervision, full generative model – “Gold-standard model”
- Image parsing: learning the model for a specific image
- Recent advances
- Datasets and challenges

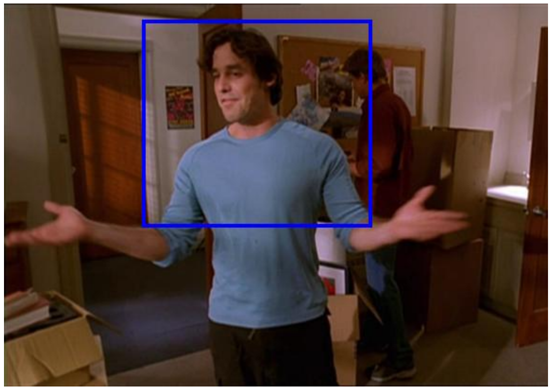
Better appearance models for pictorial structures

Marcin Eichner, Vittorio Ferrari

BMVC 2009

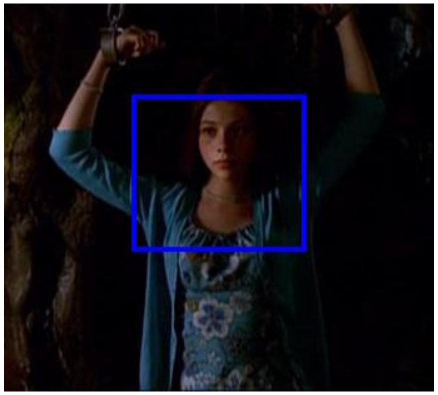
Better Appearance Models
Intuition 1

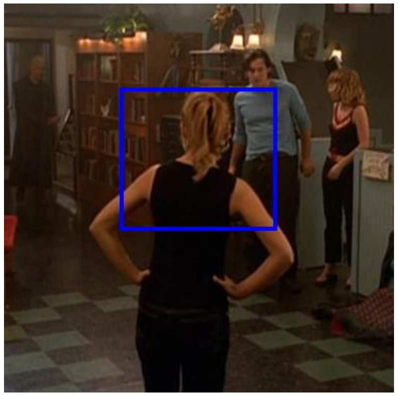




relative location (wrt detection window):

- stable, e.g. head, torso
- unstable, e.g. upper/lower arms





Better Appearance Models
Intuition 2

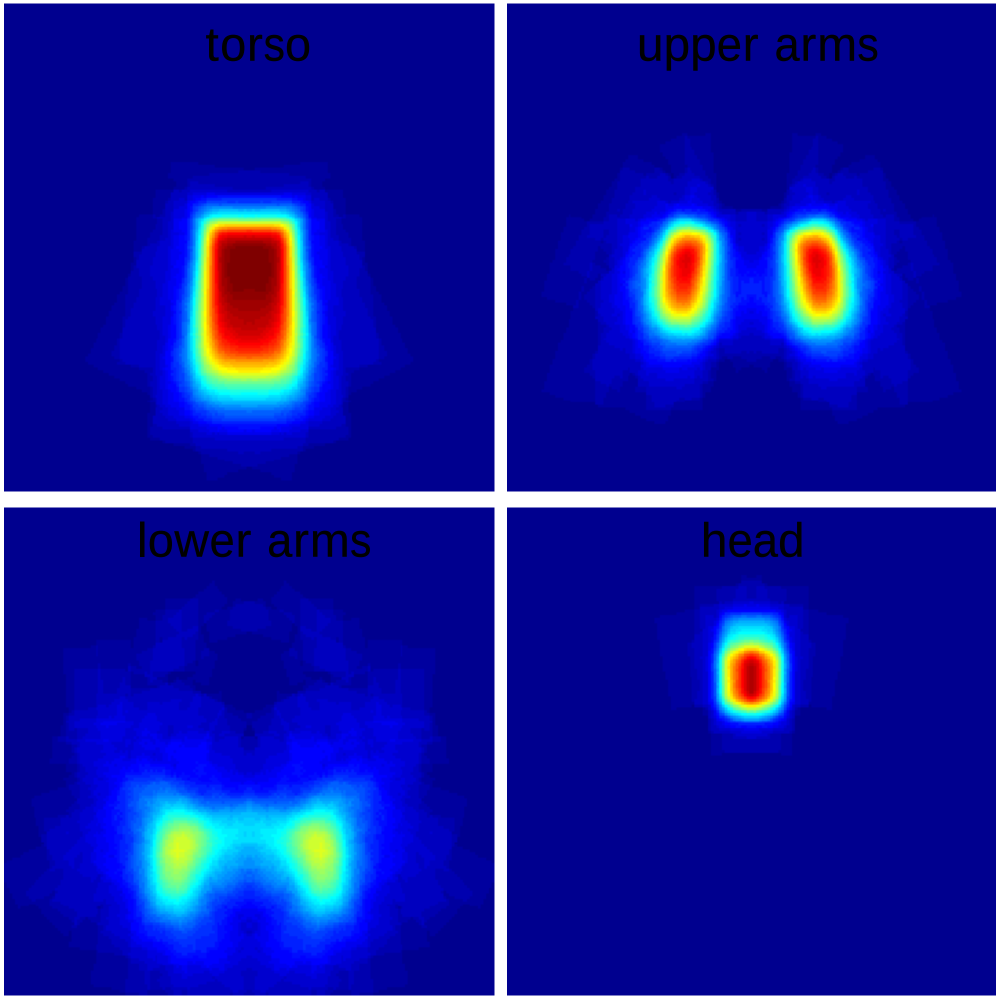
Appearance of different body parts is related

Use stable parts to improve the prediction of the unstable ones

Better Appearance Models – TRAINING Location Prior (LP)

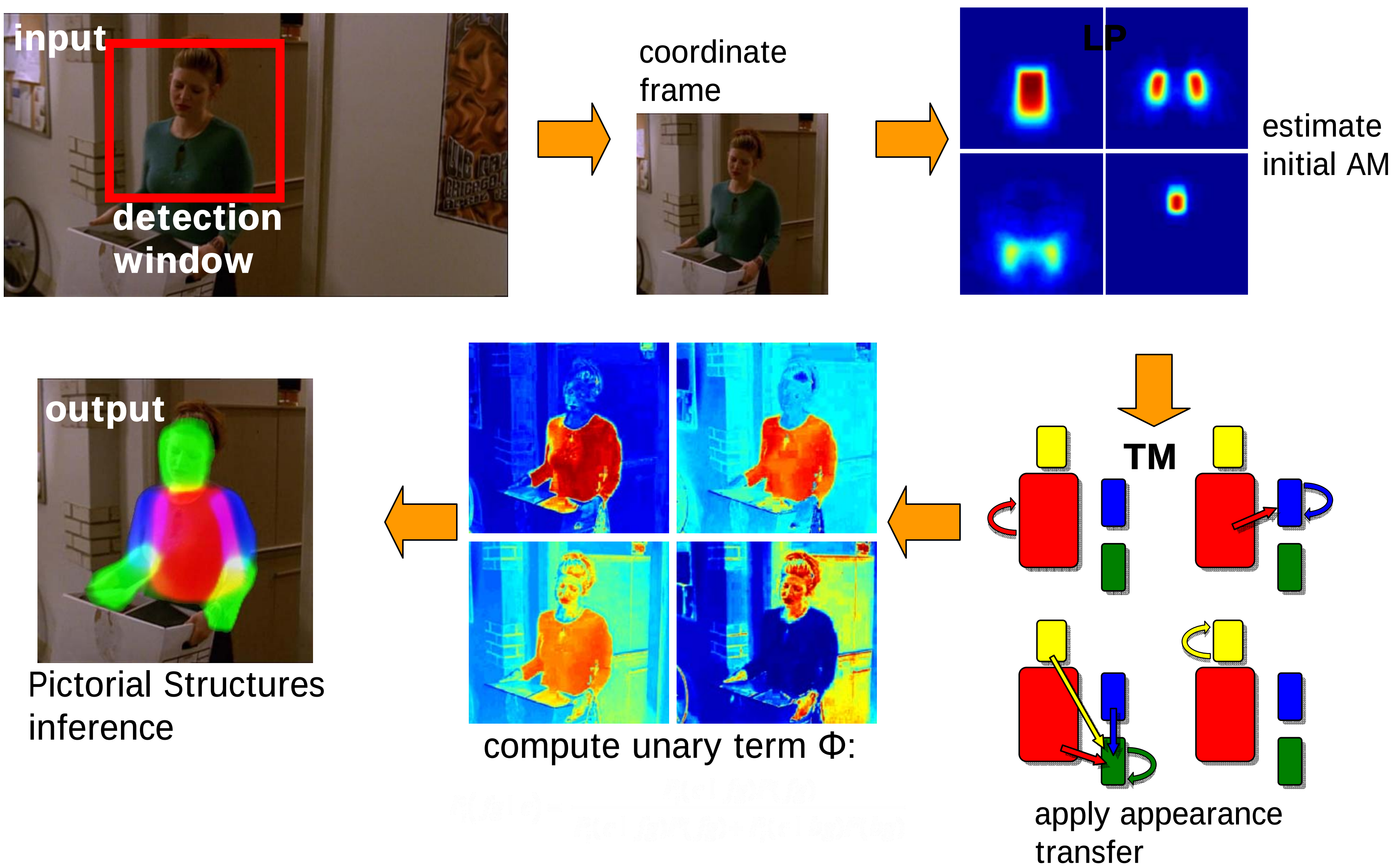
LP encodes:

- variability of poses
- detection window inaccuracy



learnt location priors (PASCAL & Buffy 3,4)

Better Appearance Models – TEST



Efficient Discriminative Learning of Parts-based Models

Pawan Kumar, Phil Torr, Andrew Zisserman

ICCV 2009

Learning a discriminative model

Supervision

- bounding rectangles for limbs for positive examples

Weak model

- Tree structured graphical model
- Parts labelled as occluded or not
- Scale of parts known

Discriminative learning

- Similar to Max-margin Markov network
- But much more efficient inference

Problem formulation

Energy of a labelling:

$$E(f) = \sum_{a \in \mathcal{V}} \overline{\theta}_{a;f(a)} + \sum_{(a,b) \in \mathcal{E}} \overline{\theta}_{ab;f(a)f(b)} + b = \mathbf{w}^\top \boldsymbol{\theta}_f + b$$

Assume the following form

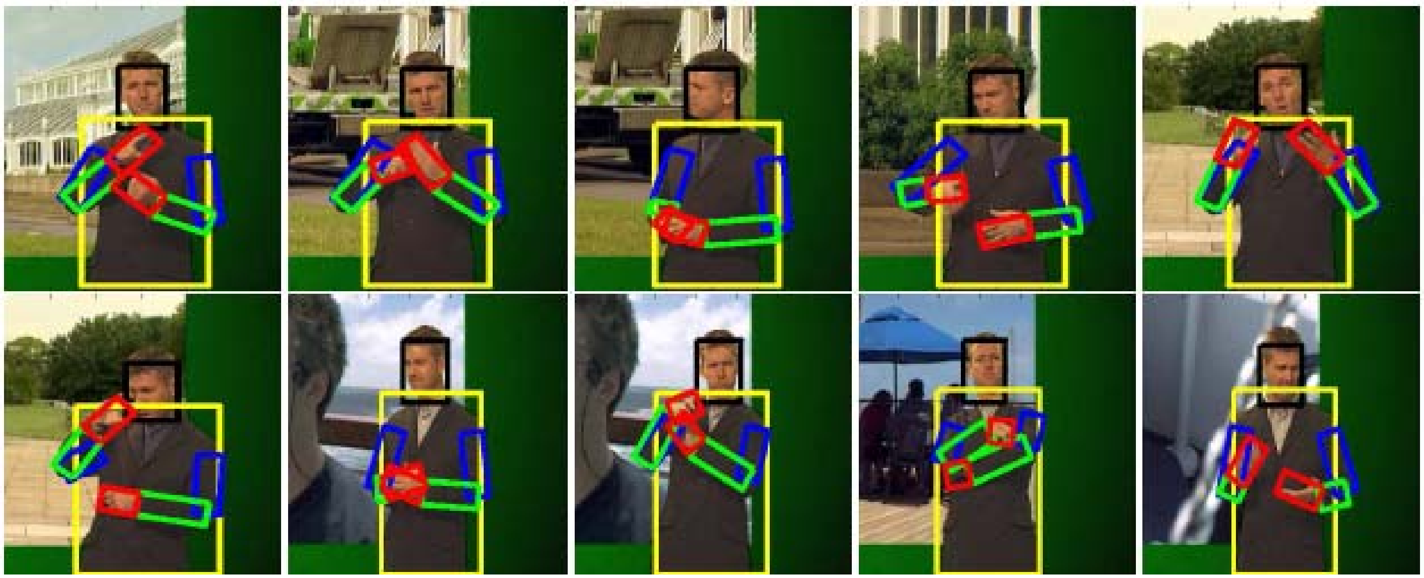
- Unary term: $\overline{\theta}_{a;f(a)} = \mathbf{w}_a^\top \boldsymbol{\theta}_{a;f(a)}$,
- Pairwise term: $\overline{\theta}_{ab;f(a)f(b)} = \mathbf{w}_{ab}^\top \boldsymbol{\theta}_{ab;f(a)f(b)}$,

where

- $\boldsymbol{\theta}_{a;f(a)}$ is the feature vector for part a (HOG + colour)
- $\boldsymbol{\theta}_{ab;f(a)f(b)}$ are the pairwise features (spatial configuration)
- We want to learn the parameters $\mathbf{w} = (\mathbf{w}_a; \mathbf{w}_{ab})$ and b

Training data

Positive examples



Negative examples

- **all** other configurations

Wide margin formulation for learning

$$(\mathbf{w}^*, b^*) = \quad \arg\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C(\textstyle\sum_k \xi^k + \sum_l \xi^l),$$

$$\text{s.t.} \quad \mathbf{w}^\top \boldsymbol{\theta}_+^k + b \geq 1 - \xi^k, \forall k \text{ (positive examples)},$$

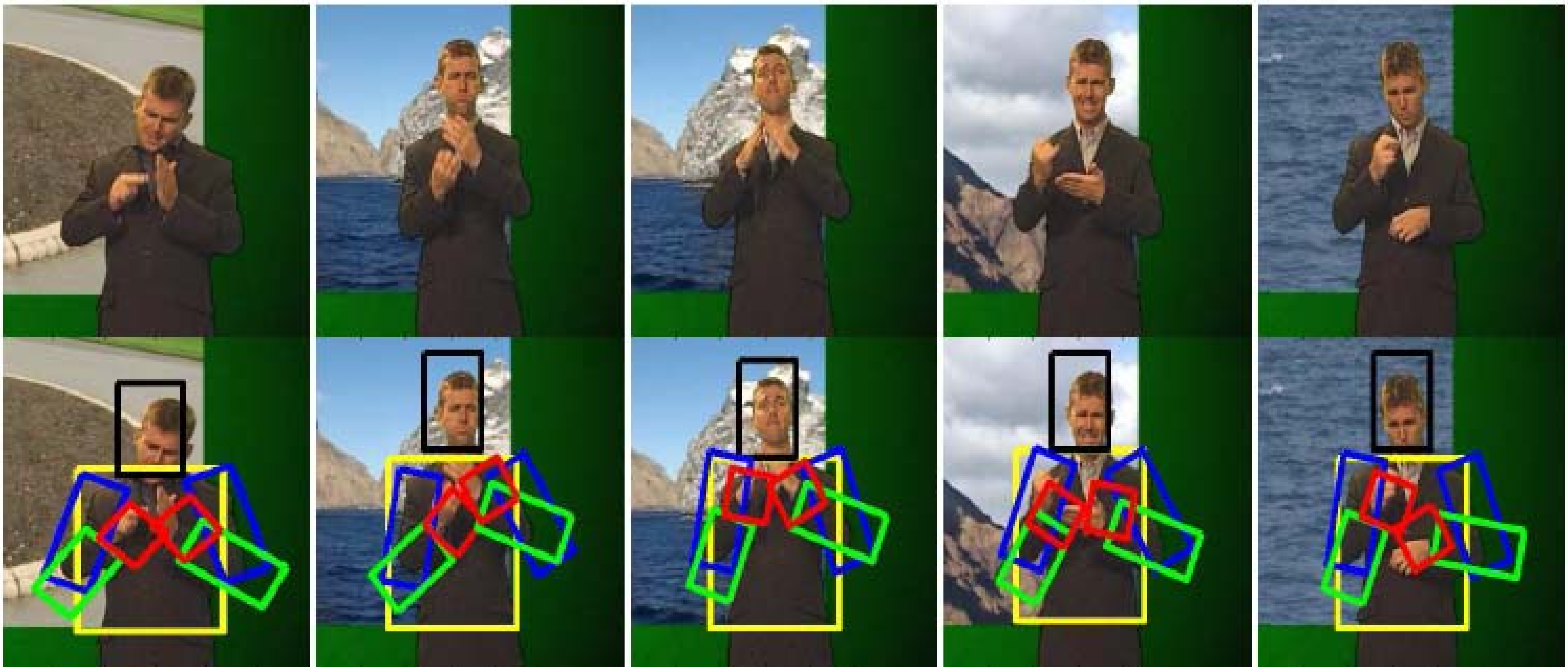
$$\mathbf{w}^\top \boldsymbol{\theta}_-^l + b \leq -1 + \xi^l, \forall l \text{ (negative examples)},$$

$$\xi^k \geq 0, \forall k, \xi^l \geq 0, \forall l.$$

Convex formulation. Similar to:

- Tsochantaridis, Hofmann, Joachims, & Altun. Support vector learning for interdependent and structured output spaces.ICML, 2004.
- (supervised version of) Felzenszwalb, McAllester, & Ramanan. A discriminatively trained, multiscale, deformable part model. CVPR, 2008

Results



H3D: Humans in 3D

Lubomir Bourdev & Jitendra Malik

ICCV 2009

Robust detection is challenging and requires using parts

But how do we choose good parts?

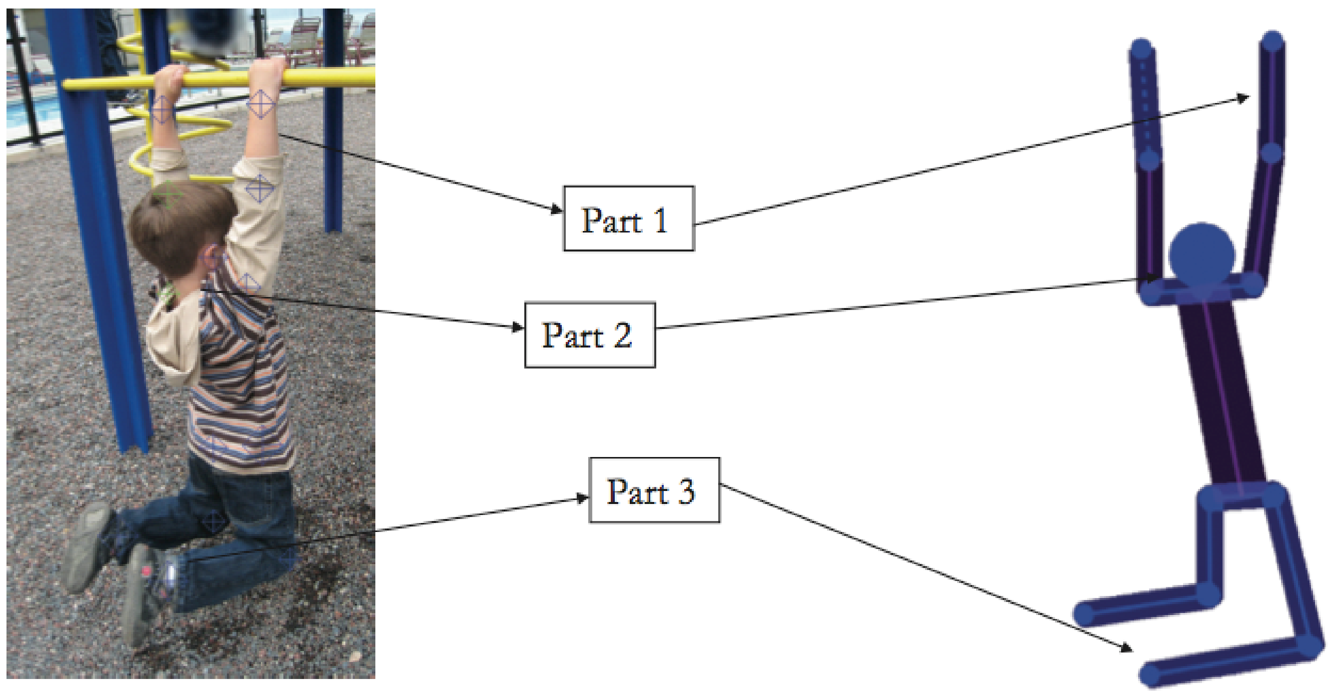


Image space

Configuration space

Parts clustered in config space

Generalized Cylinders
[Nevatia, Binford AI77]

Pictorial Structures
[Felzenszwalb, Huttenlocher IJCV05]
[Andriluka, Roth, Schiele CVPR09]
[Ramanan NIPS06]

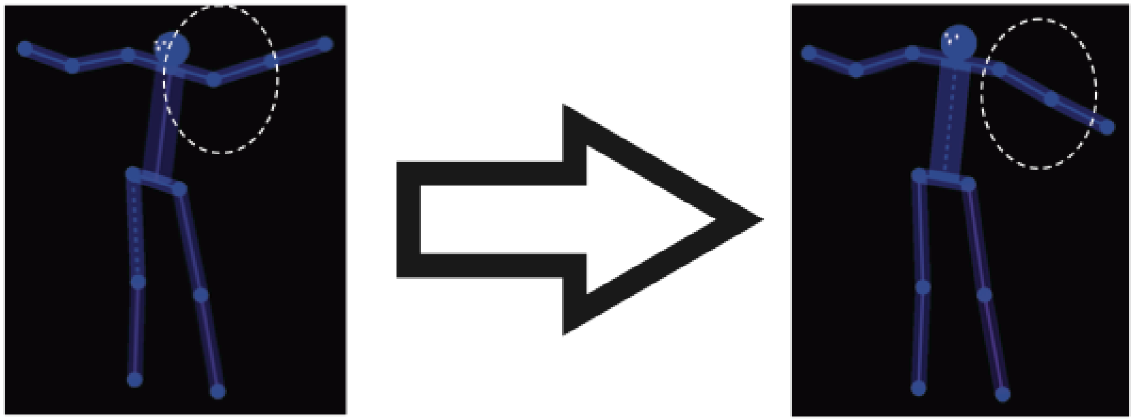
Parts clustered in image space

Holistic Methods (pedestrians)
[Dalal, Triggs CVPR05]
[Oren et al CVPR97]

Learning Parts from the Image
[Leibe et al ECCV04]
[Fergus et al, CVPR03]
[Mori, Malik, ECCV02]

Our approach combines the strengths of both prior research directions

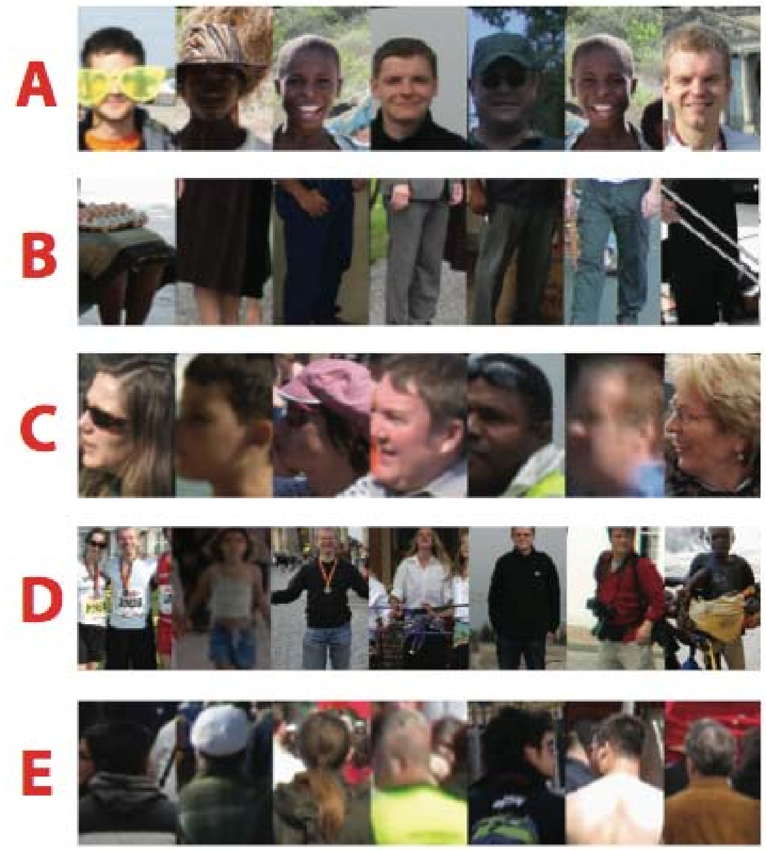
1. Define a configuration-space distance between two poses at a given region:



2. Use it to generate similar examples given a query:

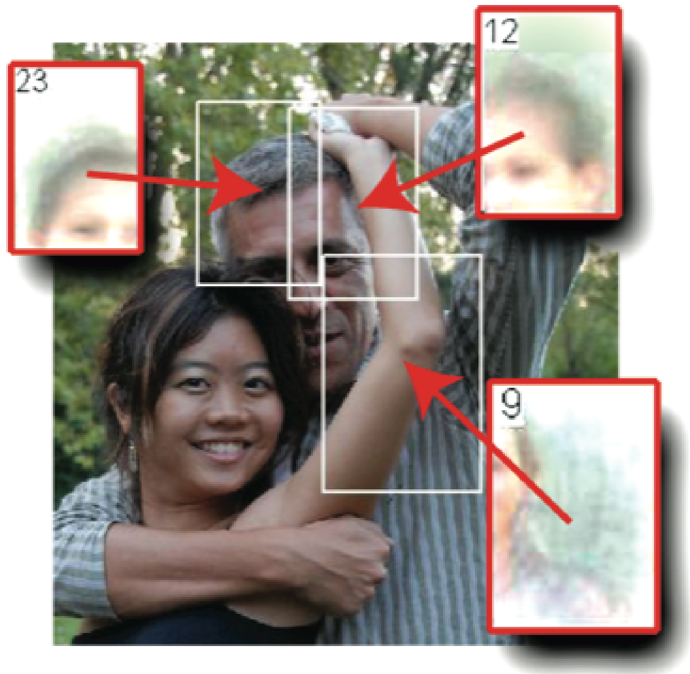
query Match 1 Match 2 Weaker Match

Average image for 100 poselets

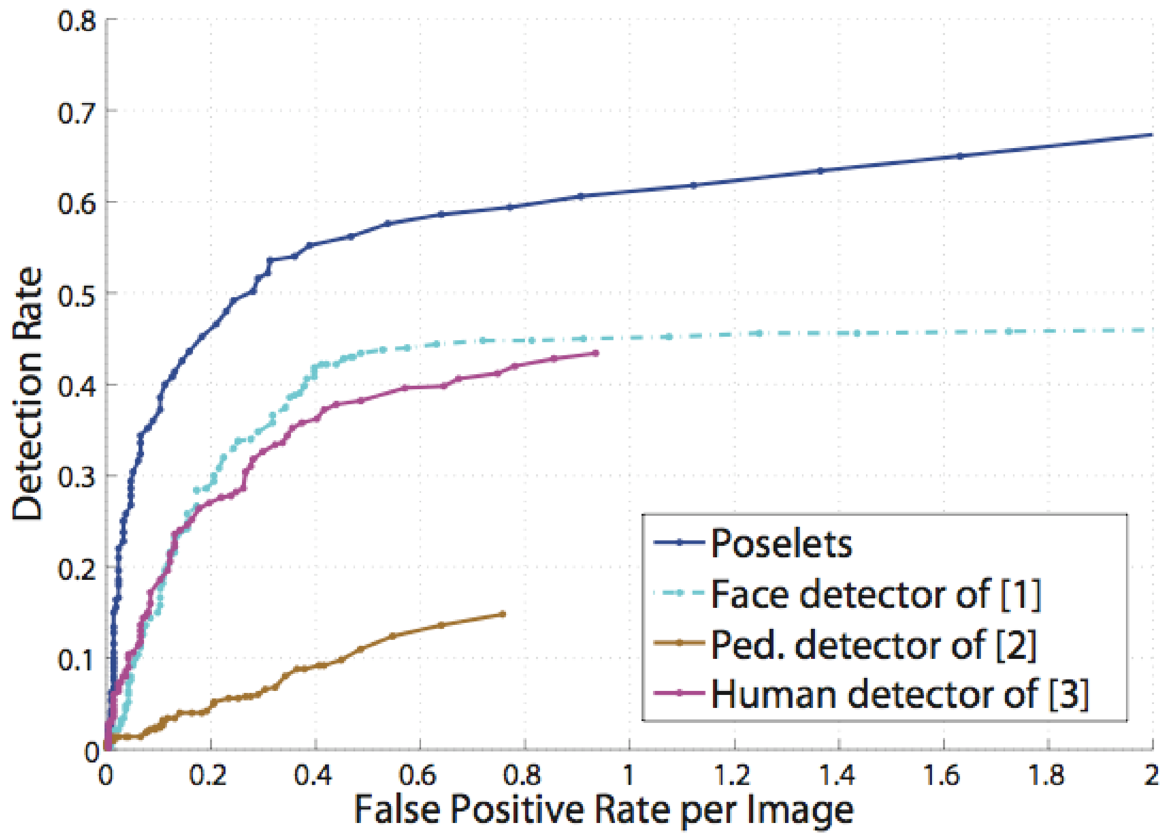


Examples from some of them

4. Combine them with Max-Margin Hough Transform (Maji/Malik CVPR09) to vote for torso, or bounds, or keypoint locations



- **Human torso detection on H3D test set**

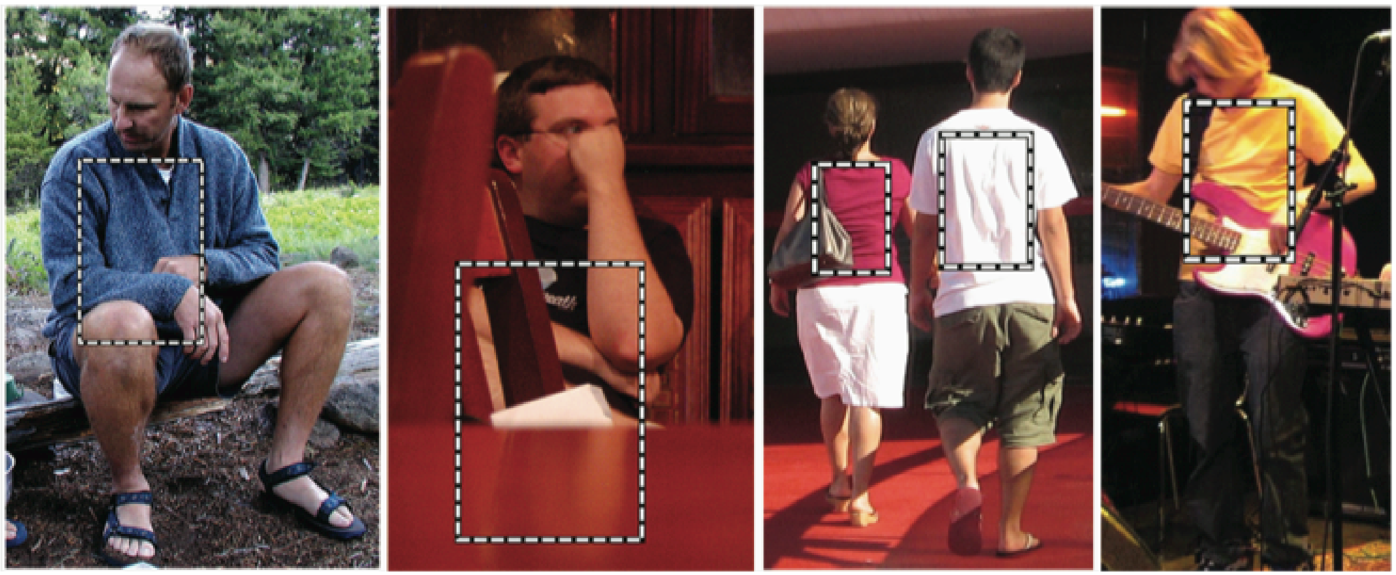


[1] L.Bourdev and J.Brandt, *Robust Object Detection using a Soft Cascade*, CVPR05

[2] N.Dalal and B.Triggs, *Histograms of Oriented Gradients for Human Detection*, CVPR05

[3] P.Felzenszwalb, D.Mcallester and D.Ramanan, *A Discriminatively Trained, Multiscale, Deformable Part Model*, CVPR08

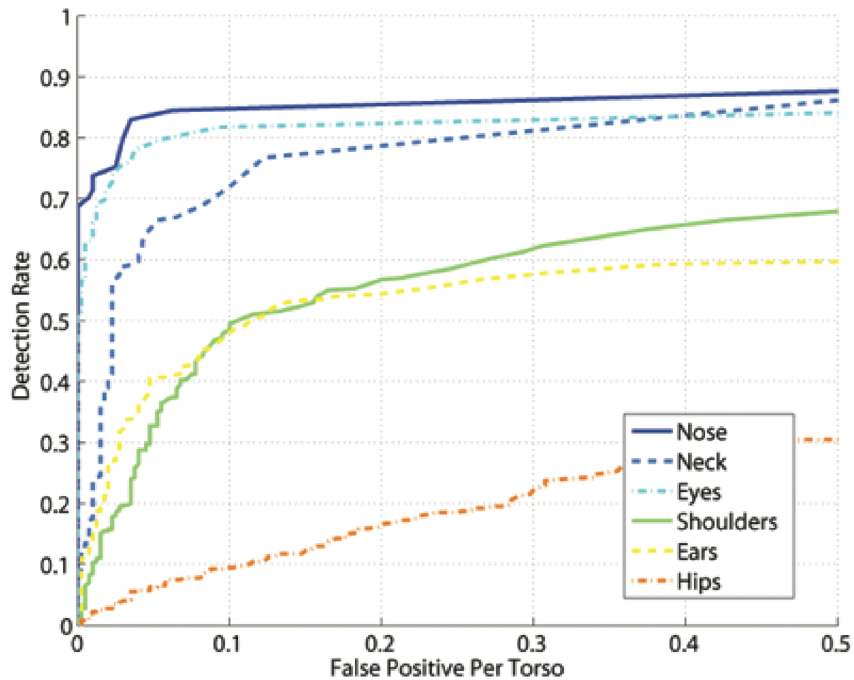
- **Examples of torso detections from H3D**



- **Detecting person bounds with PASCAL VOC 2007**

AP =0.394

Detecting keypoints



ROC for localizing keypoints, conditioned on torso detection

Further ideas:

• Human Pose Estimation Using Consistent Max-Covering, Hao Jiang, ICCV 09

• Max-margin hidden conditional random fields for human action recognition, Yang Wang and Greg Mori, CVPR 09

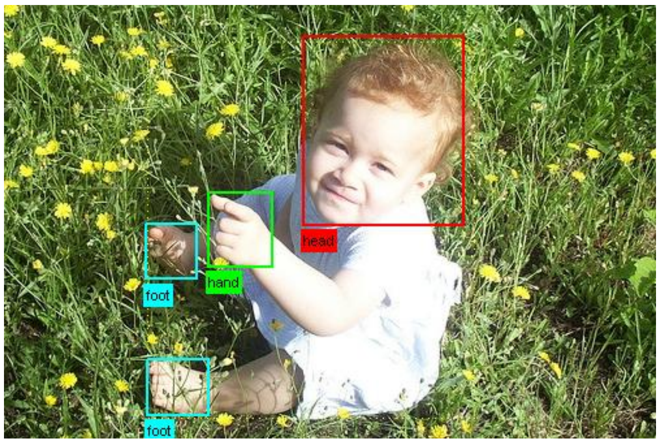
• Adaptive pose priors for pictorial structures, B. Sapp, C. Jordan, and B. Taskar, CVPR 10

Outline

- Review of pictorial structures for articulated models
- Inference given the model: Strong supervision, full generative model – “Gold-standard model”
- Image parsing: learning the model for a specific image
- Recent advances
- Datasets and challenges

Datasets & Evaluation

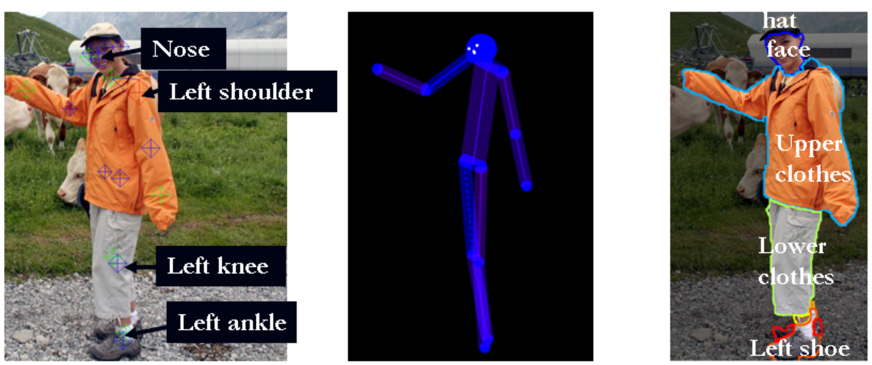
Some efforts evaluating person image parsing



PASCAL VOC "Person Layout"

Oxford Buffy Stickmen

276 frames x 6 = 1656 body parts (sticks)



Keypoint Annotations 3D Pose Region Labels

Berkeley H3D

ETHZ Pascal stickmen set

549 images x6 = 3294 body parts (sticks)

The PASCAL Visual Object Classes Challenge 2010 (VOC2010)

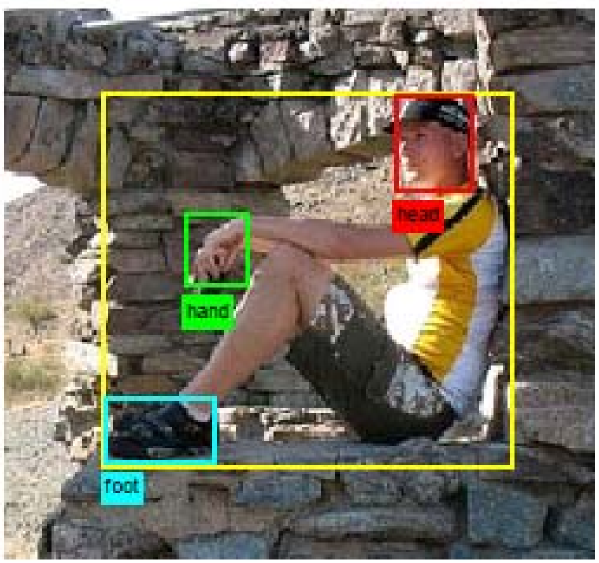
Mark Everingham, Luc Van Gool
Chris Williams, John Winn
Andrew Zisserman

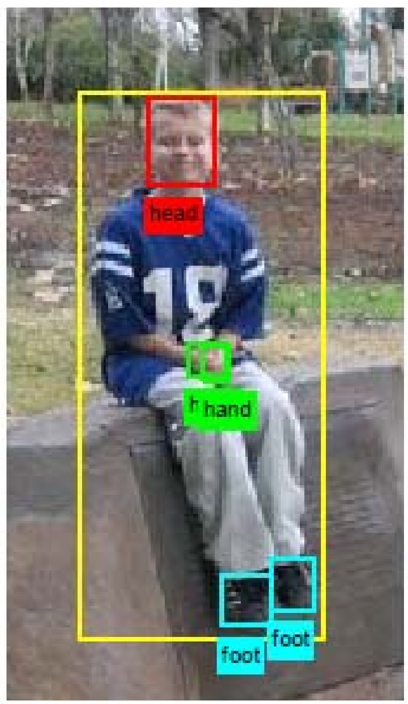
PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

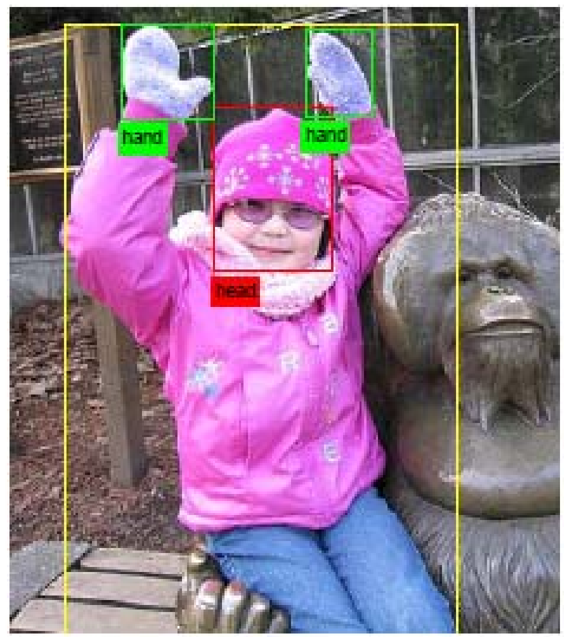
Person Layout Taster

Given the bounding box of a person, predict the visibility and positions of head, hands and feet.

- About 600 training examples
- But can also use any training data (not overlapping with test set)





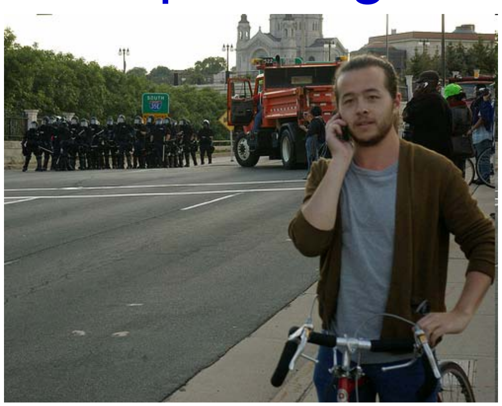


Human Action Classes Taster

Given the bounding box of a person, determine which, if any, of 9 action classes apply

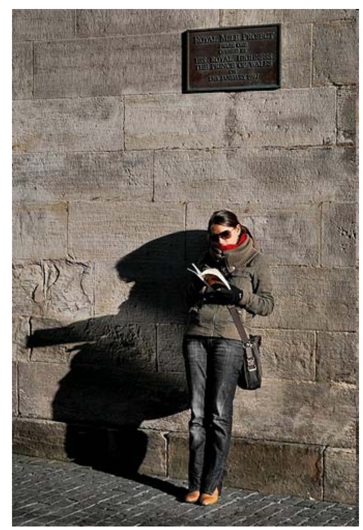
- suggested by Ivan Laptev
- choice of classes governed by availability from flickr
- evaluation is by AP on each class
- 50-90 training images for each class

phoning

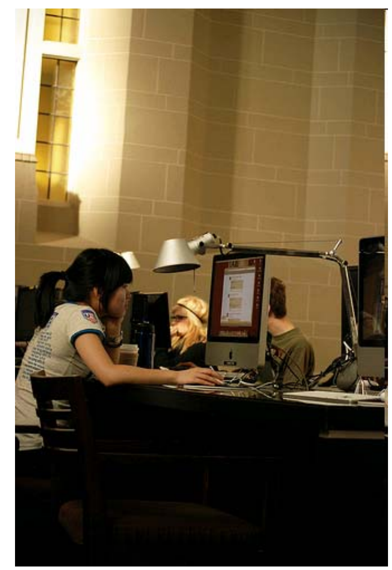


playing instrument

reading



working on computer



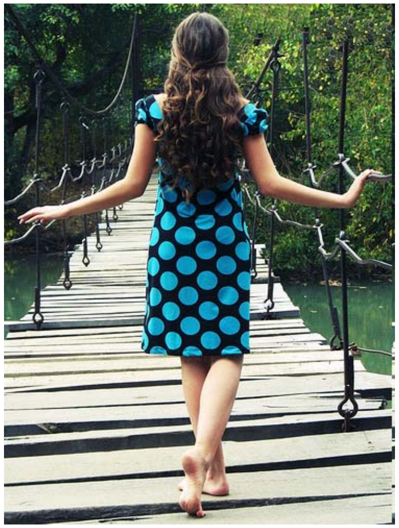
Human Action Classes Taster continued

riding horse

riding bike

running

walking



taking photo