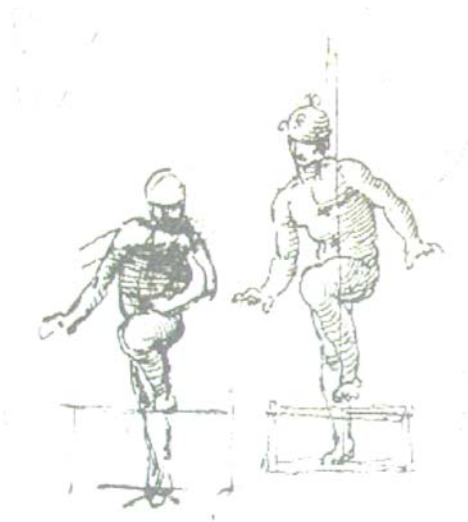


Reconnaissance d’objets et vision artificielle 2010

Motion and Human Actions I

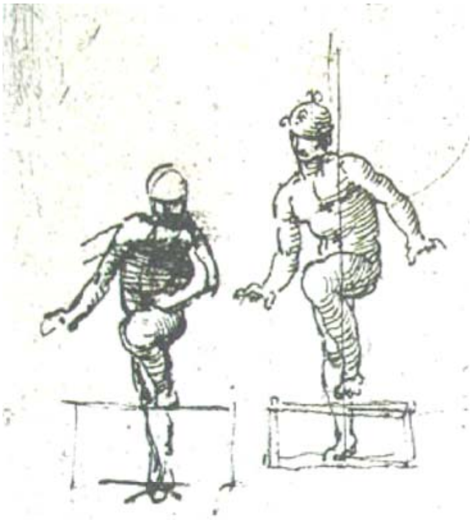
Ivan Laptev

ivan.laptev@inria.fr

INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d’Informatique, Ecole Normale Supérieure, Paris

Includes slides from: Ondra Chum, Alyosha Efros, Mark Everingham, Pedro Felzenszwalb, Rob Fergus, Kristen Grauman, Bastian Leibe, Ivan Laptev, Fei-Fei Li, Marcin Marszalek, Pietro Perona, Deva Ramanan, Bernt Schiele, Jamie Shotton, Andrea Vedaldi and Andrew Zisserman

Class overview



Motivation

Historic review
Modern applications

Human Pose Estimation

Pictorial structures
Learning models from image data
Recent advances
Datasets and challenges

Appearance-based methods

Motion history images
Active shape models
Tracking and motion priors

Motion-based methods

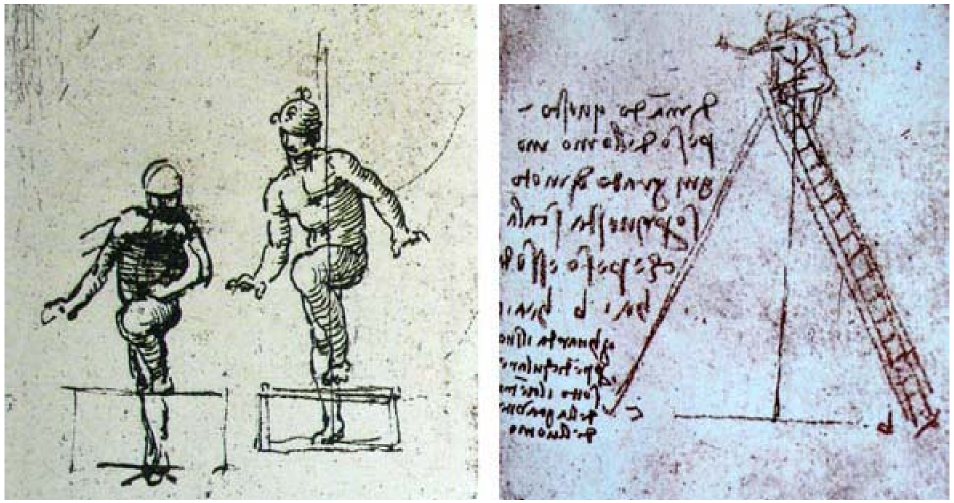
Generic and parametric Optical Flow
Motion templates

Motivation I: Artistic Representation

Early studies were motivated by human representations in Arts

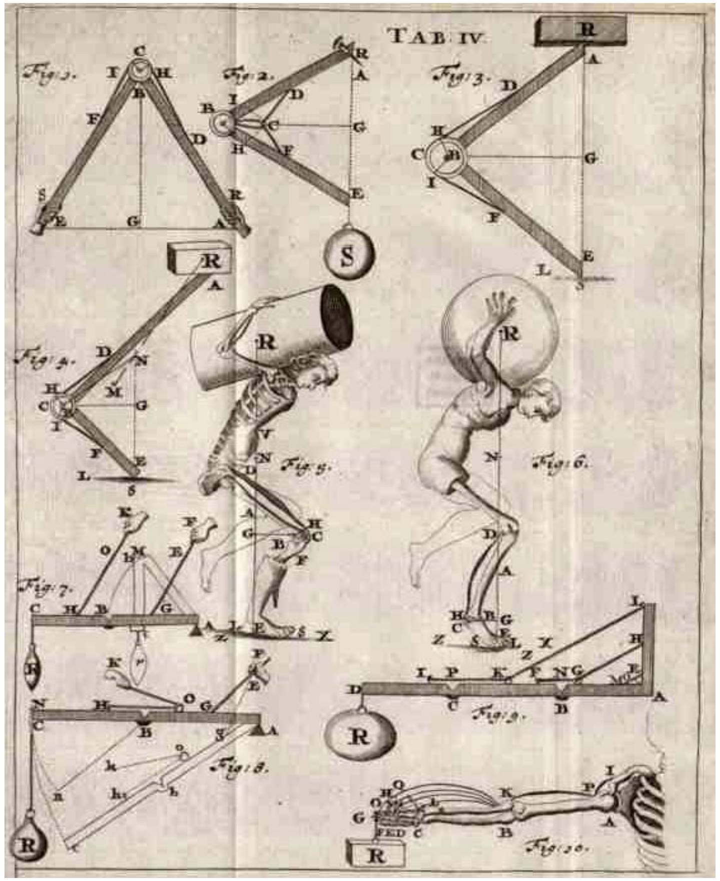
Da Vinci: “it is indispensable for a painter, to become totally familiar with the anatomy of nerves, bones, muscles, and sinews, such that he understands for their various motions and stresses, which sinews or which muscle causes a particular motion”

“I ask for the weight [pressure] of this man for every segment of motion when climbing those stairs, and for the weight he places on *b* and on *c*. Note the vertical line below the center of mass of this man.”



Leonardo da Vinci (1452–1519): A man going upstairs, or up a ladder.

Motivation II: Biomechanics

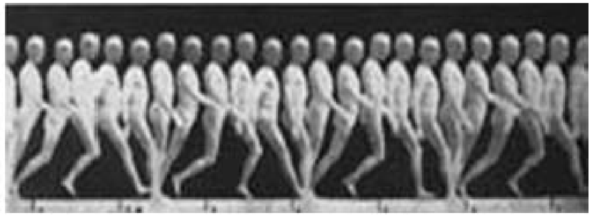


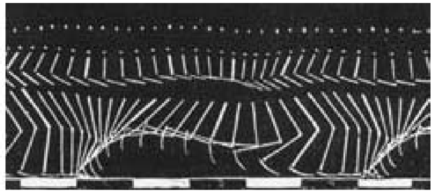
Giovanni Alfonso Borelli (1608–1679)

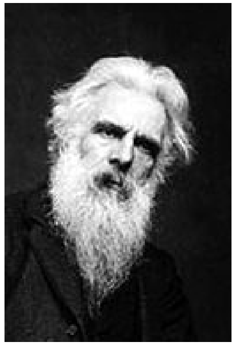
- The emergence of *biomechanics*
- Borelli applied to biology the analytical and geometrical methods, developed by Galileo Galilei
- He was the first to understand that bones serve as levers and muscles function according to mathematical principles
- His physiological studies included muscle analysis and a mathematical discussion of movements, such as running or jumping

Motivation III: Motion perception

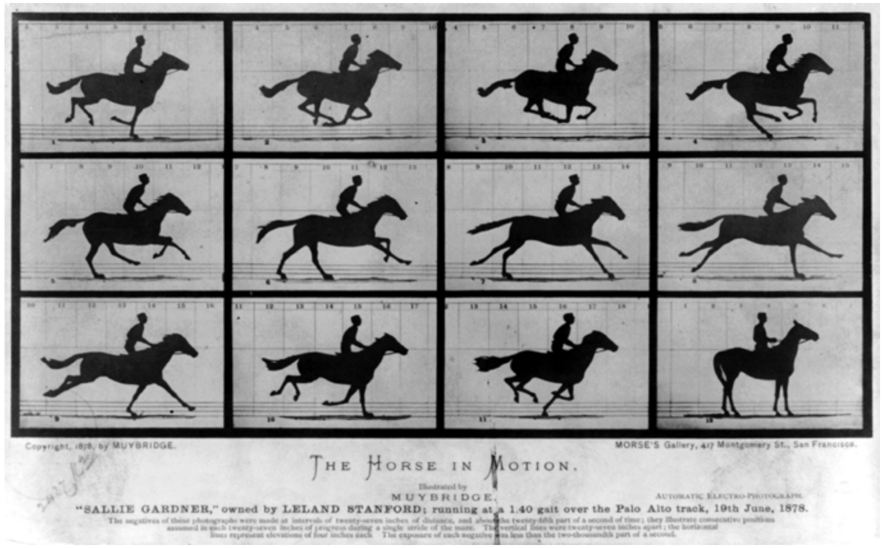
Etienne-Jules Marey: (1830–1904) made Chronophotographic experiments influential for the emerging field of *cinematography*





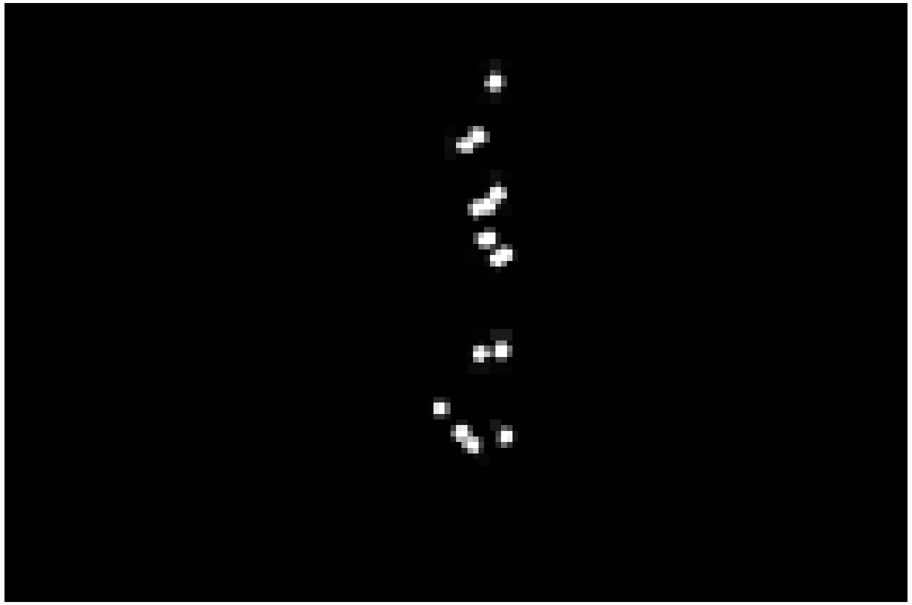


Eadweard Muybridge (1830–1904) invented a machine for displaying the recorded series of images. He pioneered motion pictures and applied his technique to movement studies



Motivation III: Motion perception

- Gunnar Johansson [1973] pioneered studies on the use of image sequences for a programmed human motion analysis
- “Moving Light Displays” (LED) enable identification of familiar people and the gender and inspired many works in computer vision.

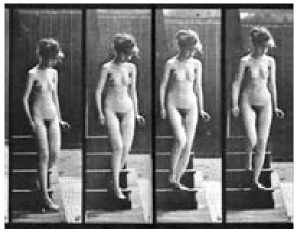


Gunnar Johansson, **Perception and Psychophysics,** 1973

Human actions: Historic overview

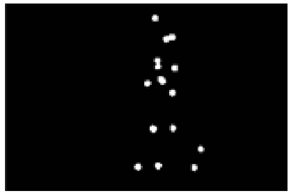
15th century
studies of
anatomy

17th century
emergence of
biomechanics



19th century
emergence of
cinematography

1973
studies of human
motion perception

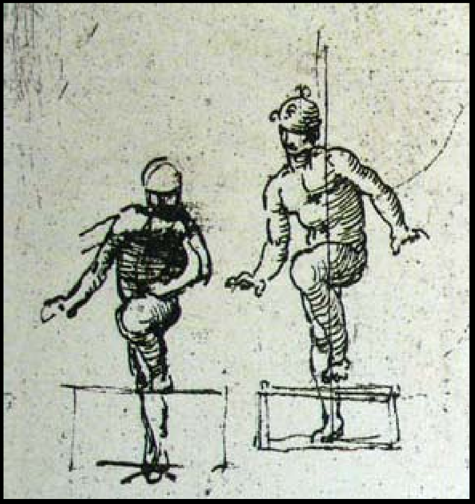


Modern computer vision

Modern applications: Motion capture and animation

Avatar (2009)

Modern applications: Motion capture and animation



Leonardo da Vinci (1452–1519)

Avatar (2009)

Modern applications: Video editing

Space-Time Video Completion
Y. Wexler, E. Shechtman and M. Irani, **CVPR** 2004

Modern applications: Video editing

Space-Time Video Completion
Y. Wexler, E. Shechtman and M. Irani, **CVPR** 2004

Modern applications: Video editing

Recognizing Action at a Distance
Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, **ICCV** 2003

Modern applications: Video editing

Recognizing Action at a Distance
Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, **ICCV** 2003

Applications: Unusual Activity Detection

e.g. for surveillance

Detecting Irregularities in Images and in Video
Boiman & Irani, **ICCV** 2005

Why automatic video understanding?

- Huge amount of video is available and growing

BBC Motion Gallery

TV-channels recorded
since 60's

>34K hours of video
upload every day

~30M surveillance cameras in US
=> ~700K video hours/day

Why automatic video understanding?

- Video indexing and search is useful in TV production, entertainment, education, social studies, security,…

TV & Web: e.g. *“Fight in a parlament”*

Home videos: e.g. *“My daughter climbing”*

Sociology research:

Manually analyzed smoking actions in 900 movies

Surveillance: e.g. *“Woman throws cat into wheelie bin”*
260K views in 7 days

- … how much is it about people?

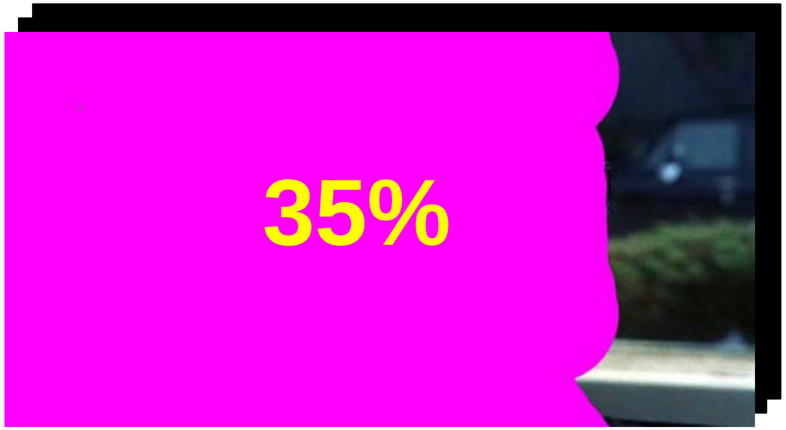
How many person-pixels are there?

Movies

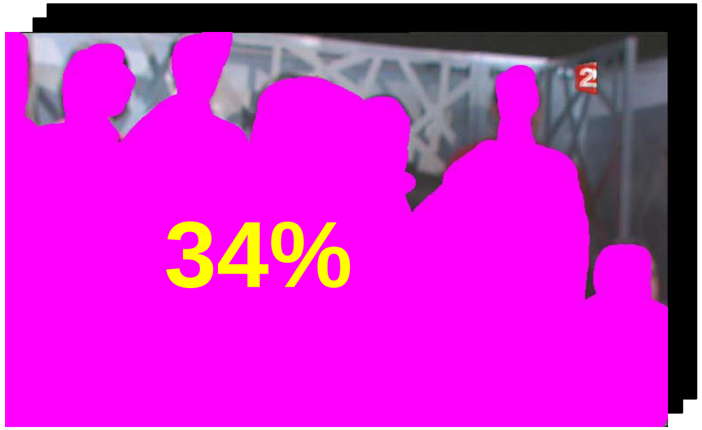
TV

YouTube

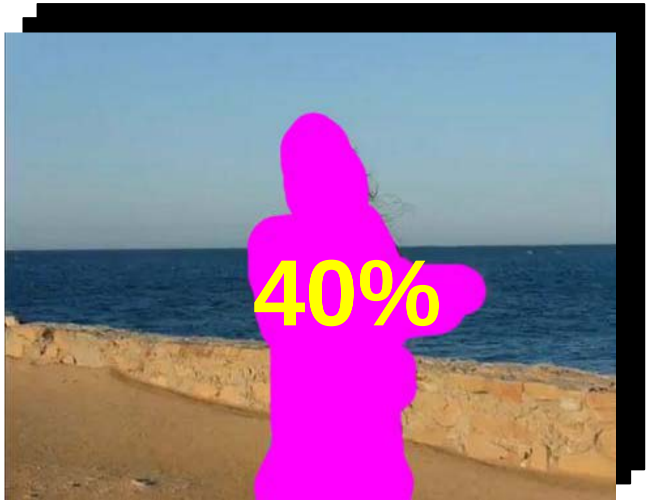
How many person-pixels are there?



Movies

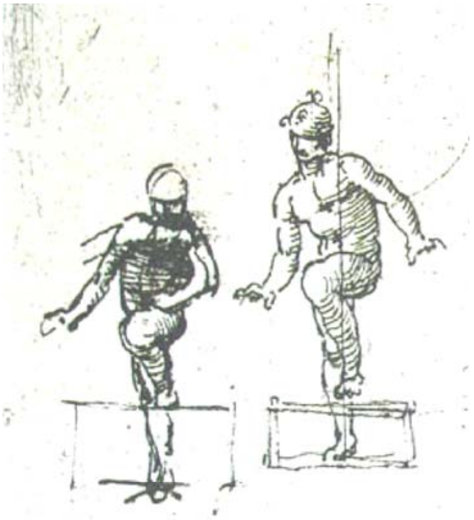


TV



YouTube

Class overview



Motivation
- Historic review
- Modern applications

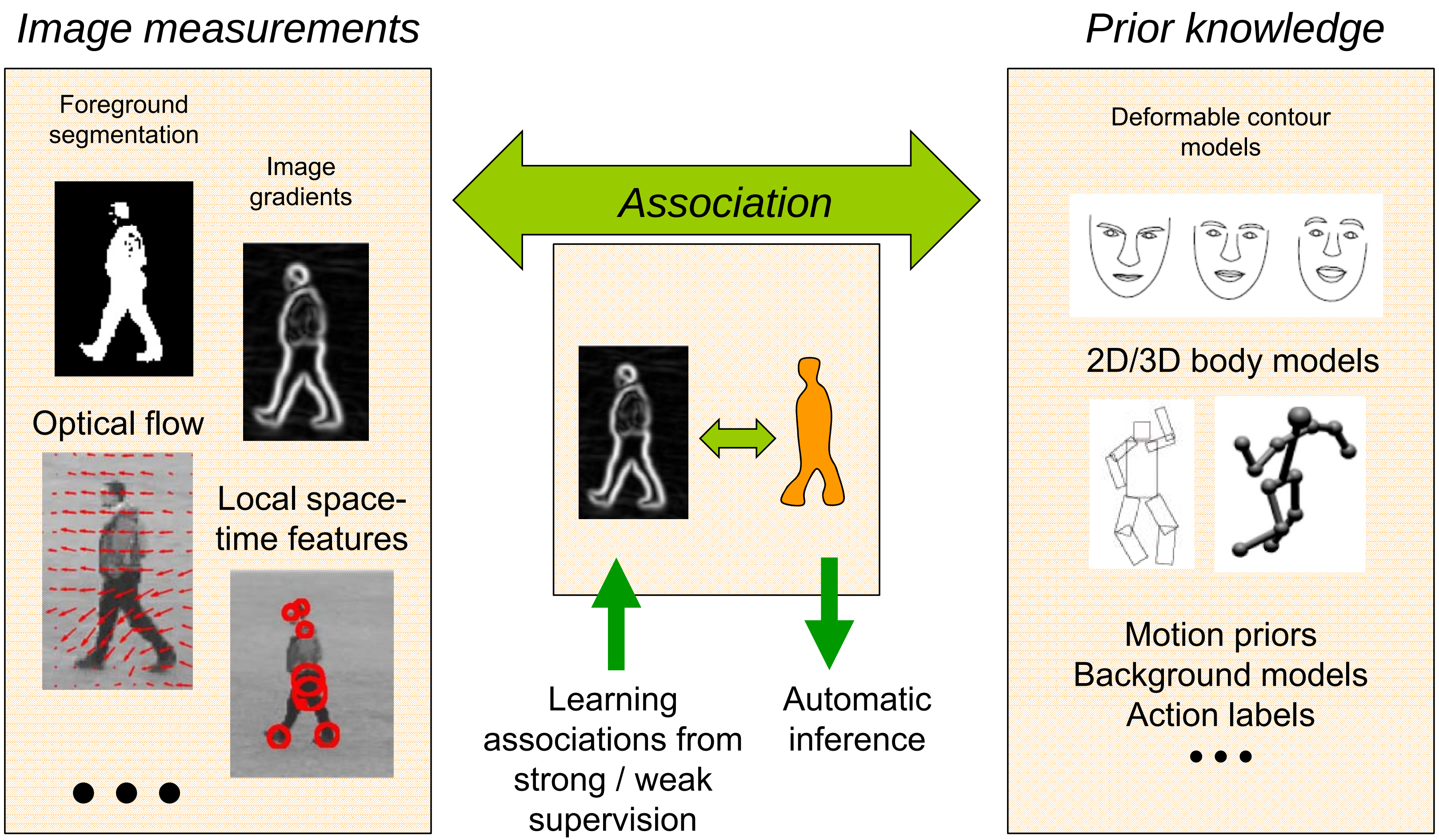
Human Pose Estimation
- Pictorial structures
- Learning models from image data
- Recent advances
- Datasets and challenges

Appearance-based methods
- Motion history images
- Active shape models
- Motion priors

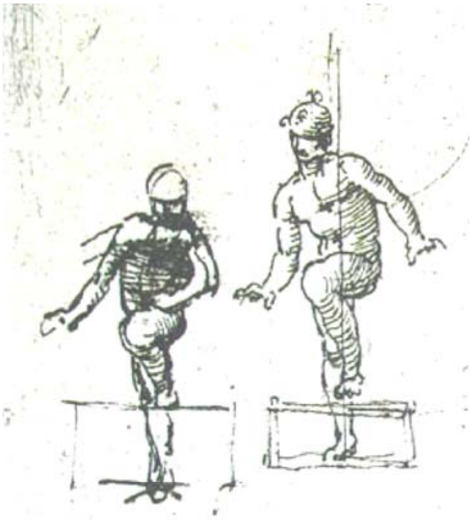
Motion-based methods
- Generic and parametric Optical Flow
- Motion templates

How to recognize actions?

Action understanding: Key components



Class overview



Motivation

Historic review
Modern applications

Human Pose Estimation

Pictorial structures
Learning models from image data
Recent advances
Datasets and challenges

Appearance-based methods

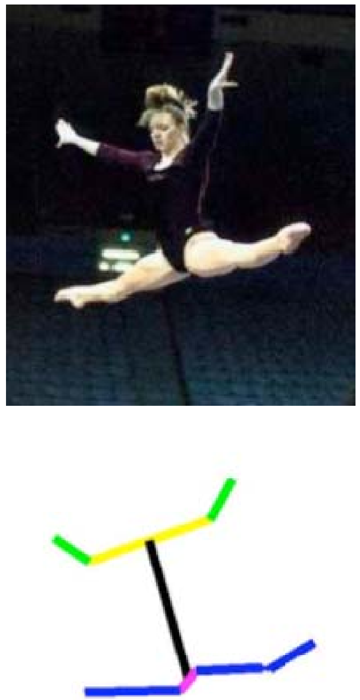
Motion history images
Active shape models
Motion priors

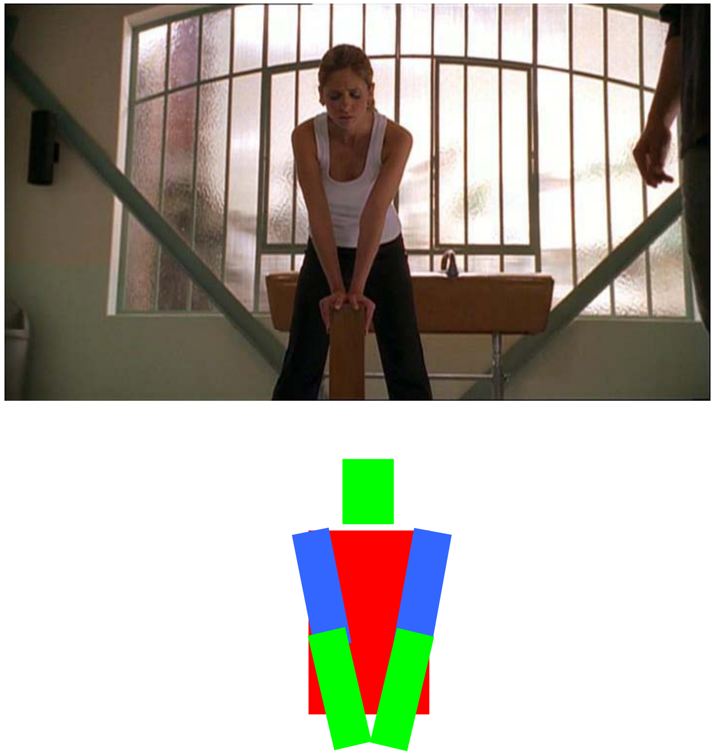
Motion-based methods

Generic and parametric Optical Flow
Motion templates

Objective and motivation

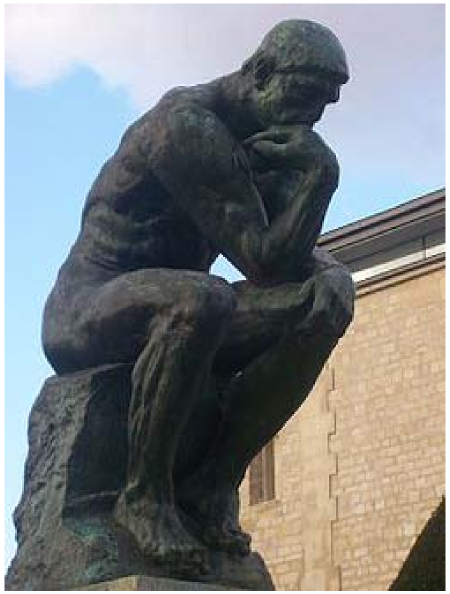
Determine human body pose (layout)





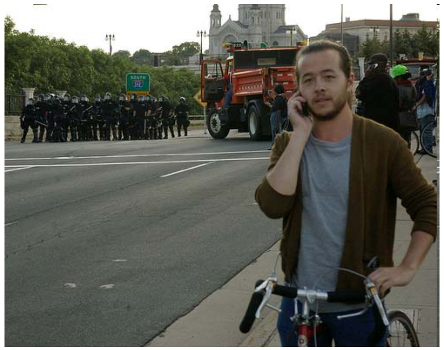
Why? To recognize poses, gestures, actions

Activities characterized by a pose



Activities characterized by a pose

Activities characterized by a pose

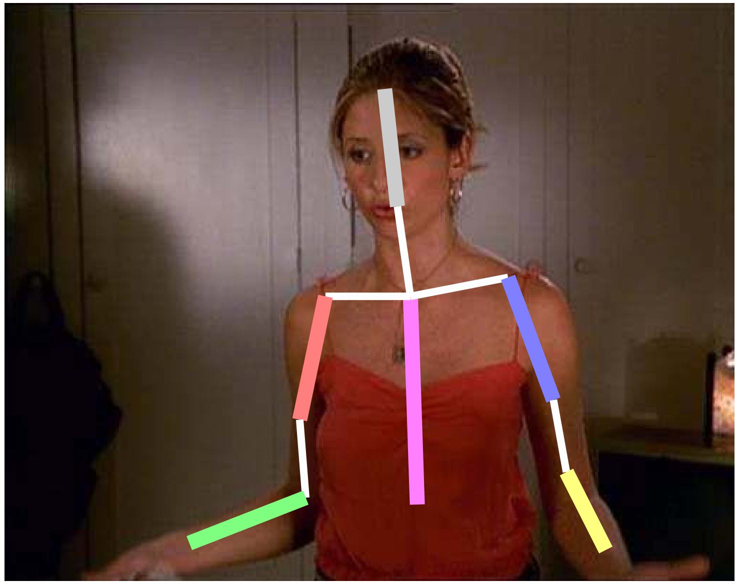


Challenges: articulations and deformations

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”

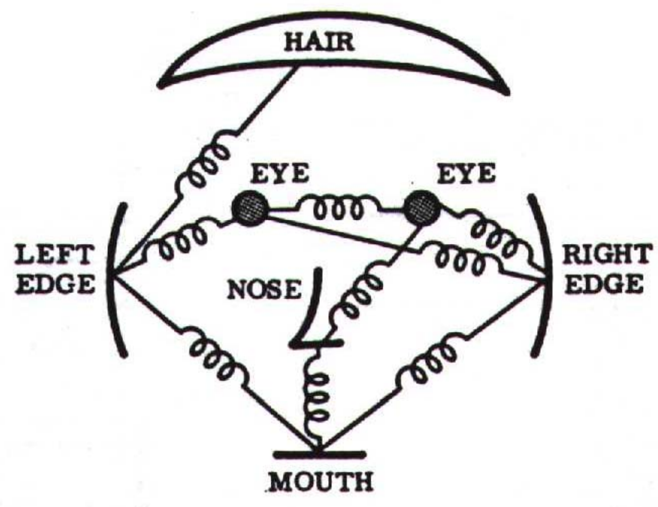
Image parsing: learning the model for a specific image

Recent advances

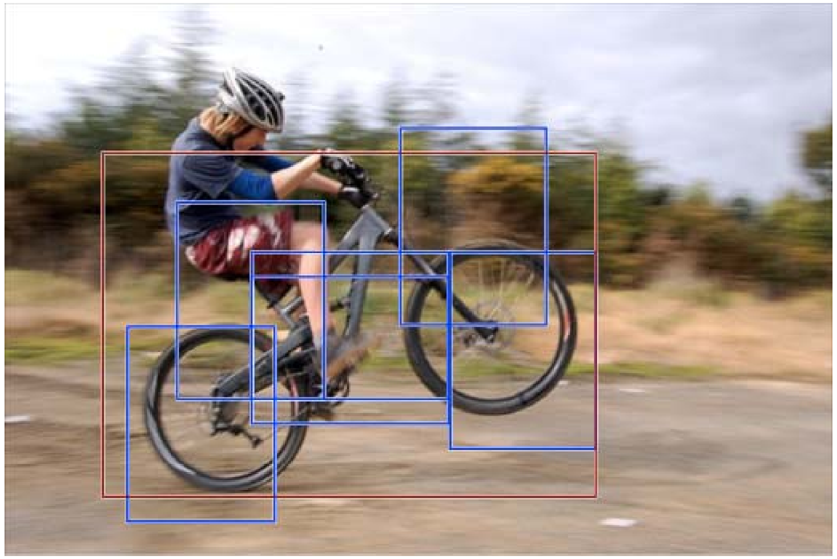
Datasets and challenges

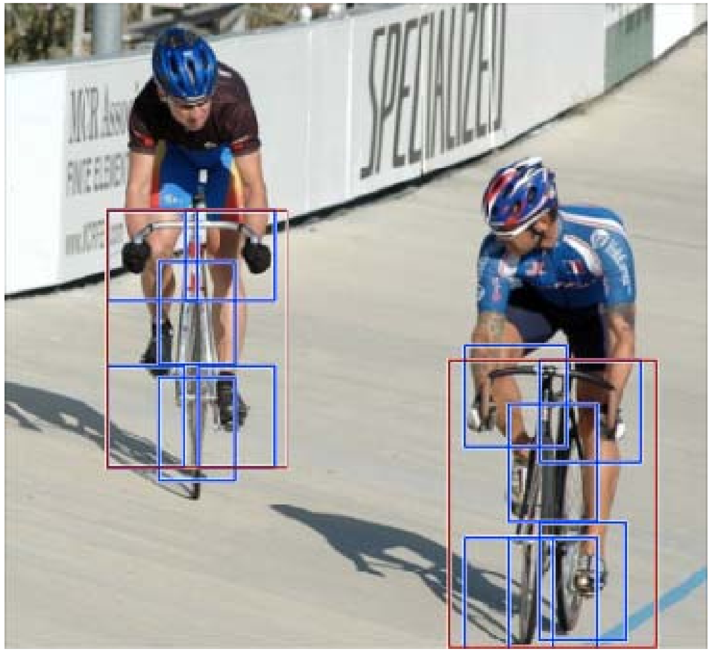
Pictorial Structures

- 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



From last lecture: objects





Mixture of deformable part-based models

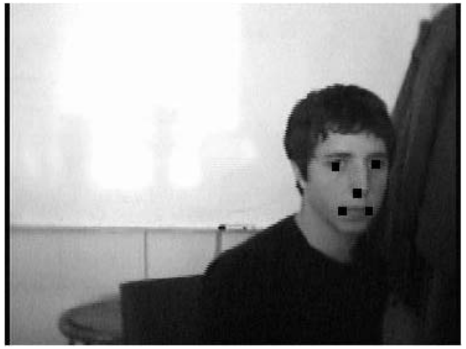
- One component per “aspect” e.g. front/side view

Each component has global template + deformable parts

Discriminative training from bounding boxes alone

Localize multi-part objects at arbitrary locations in an image

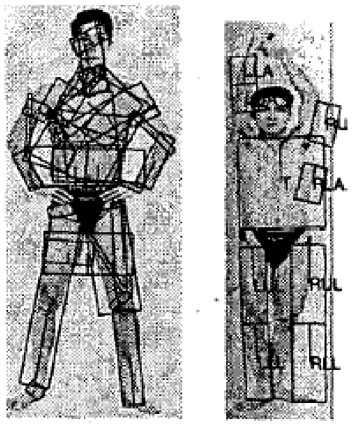
- Generic object models such as person or car
- Allow for articulated objects
- Simultaneous use of appearance and spatial information
- Provide efficient and practical algorithms



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

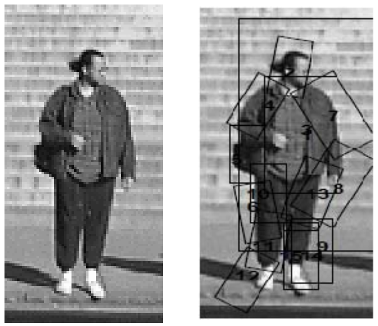
- Appearance: how well each part matches at given location
- Configuration: degree to which parts match 2D spatial layout

Long tradition of using pictorial structures for humans



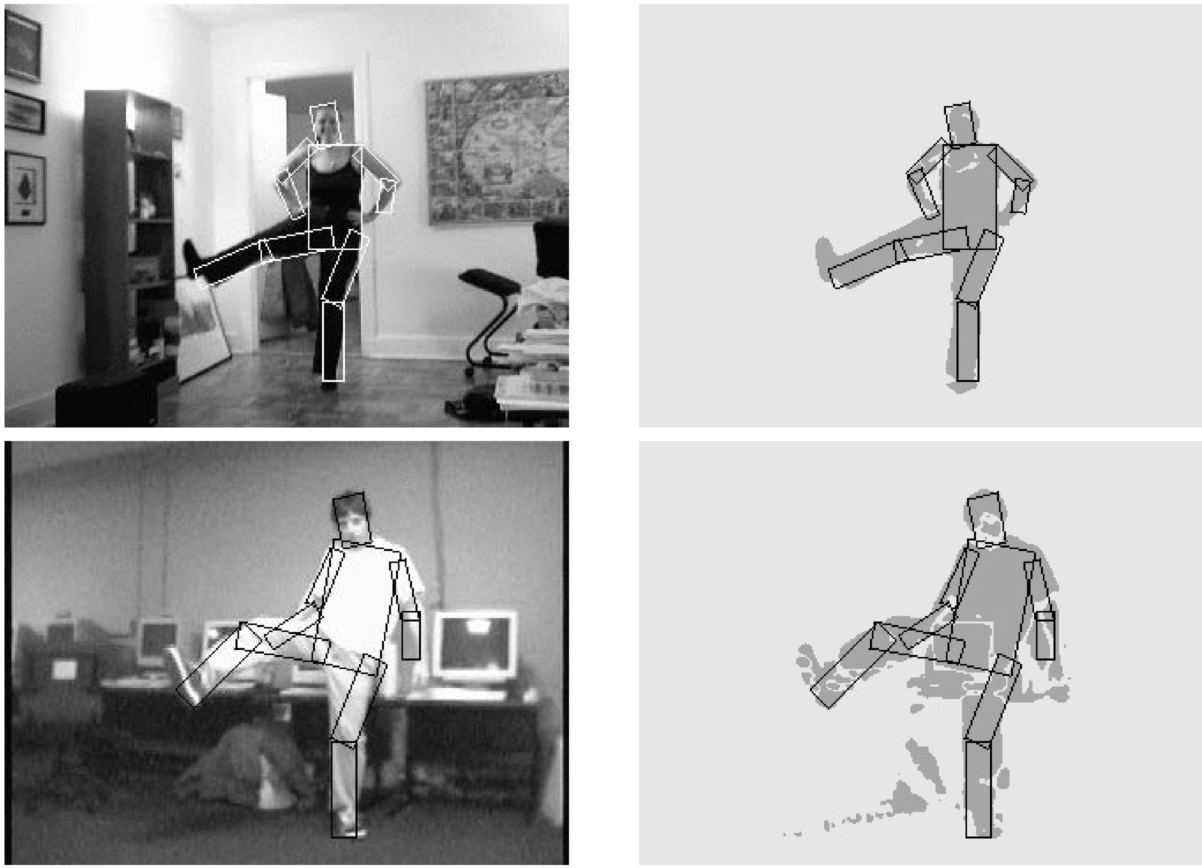
Finding People by Sampling
Ioffe & Forsyth, ICCV 1999

Pictorial Structure Models for Object Recognition
Felzenszwalb & Huttenlocher, 2000



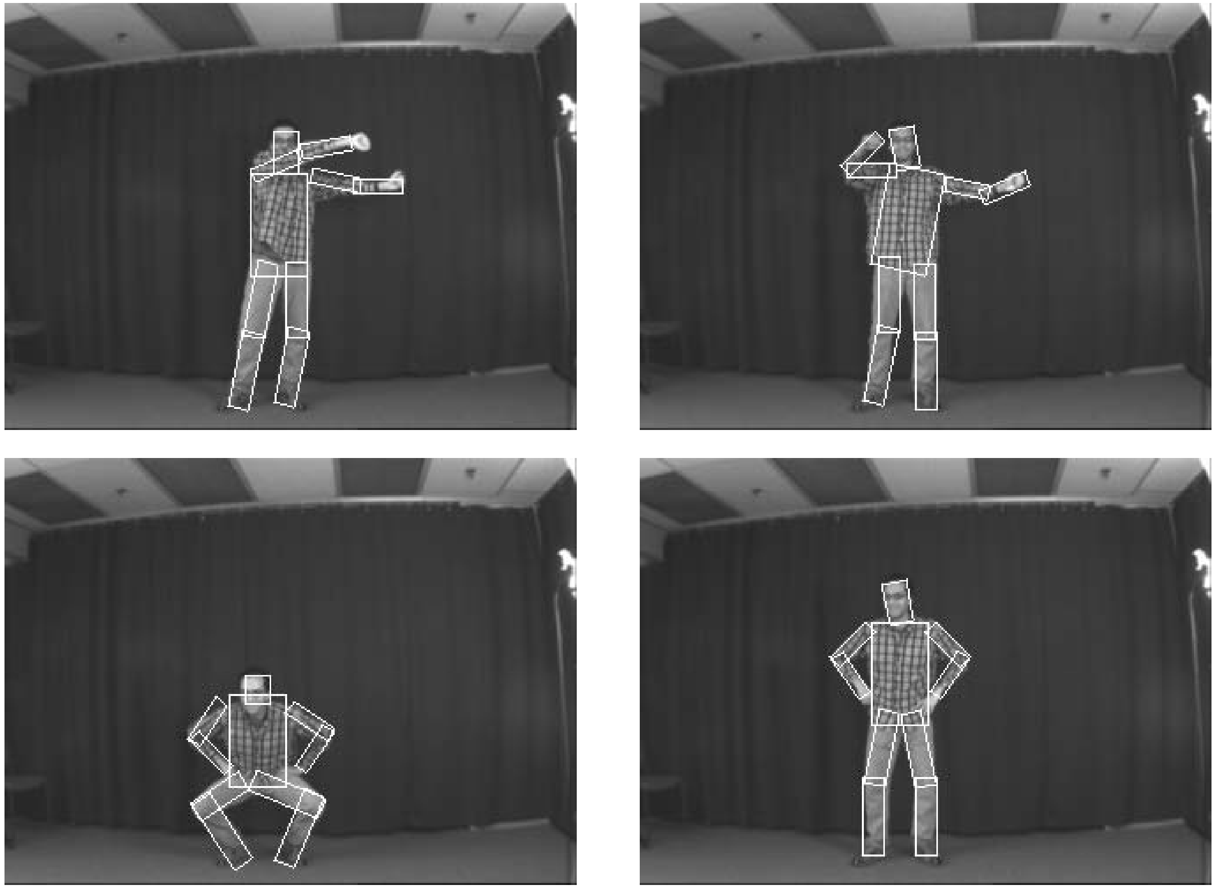
Learning to Parse Pictures of People
Ronfard, Schmid & Triggs, ECCV 2002

Felzenszwalb & Huttenlocher

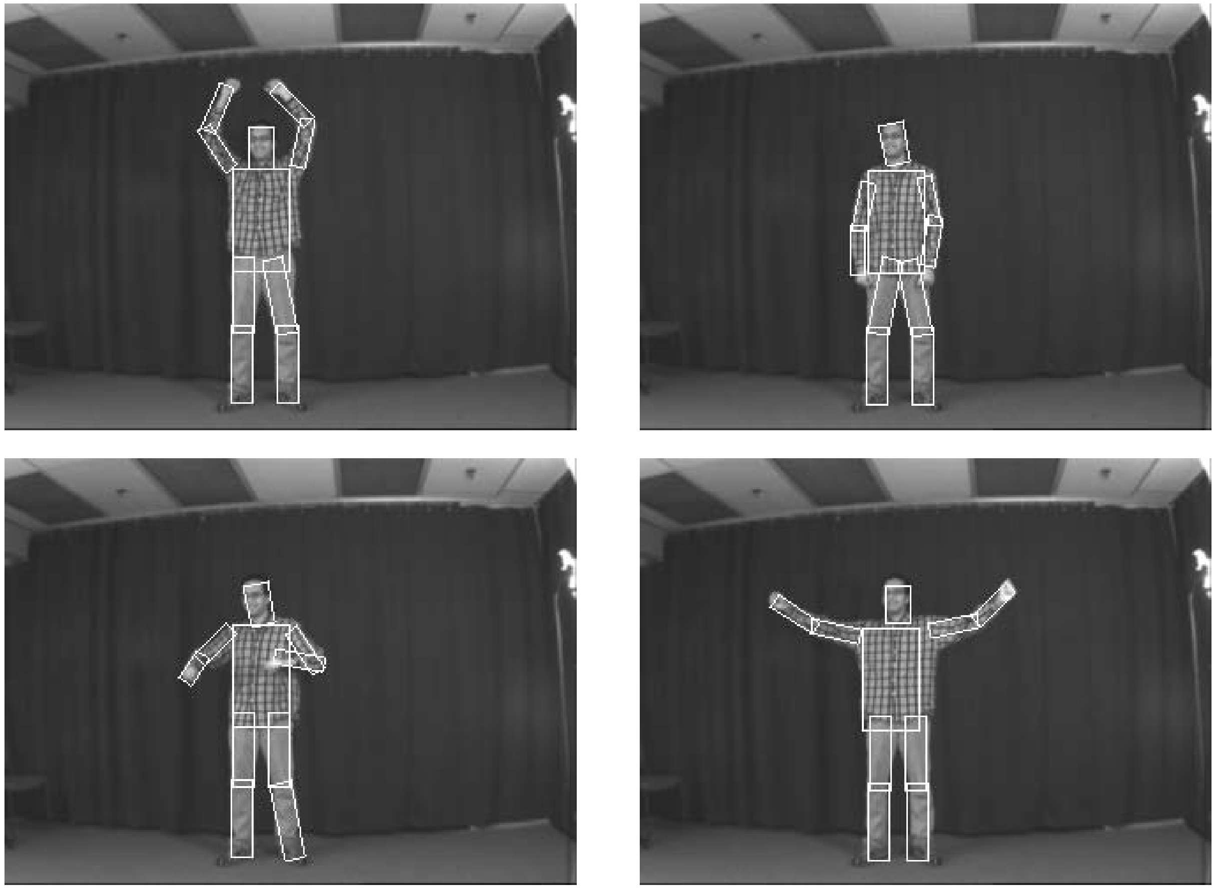


NB: requires background subtraction

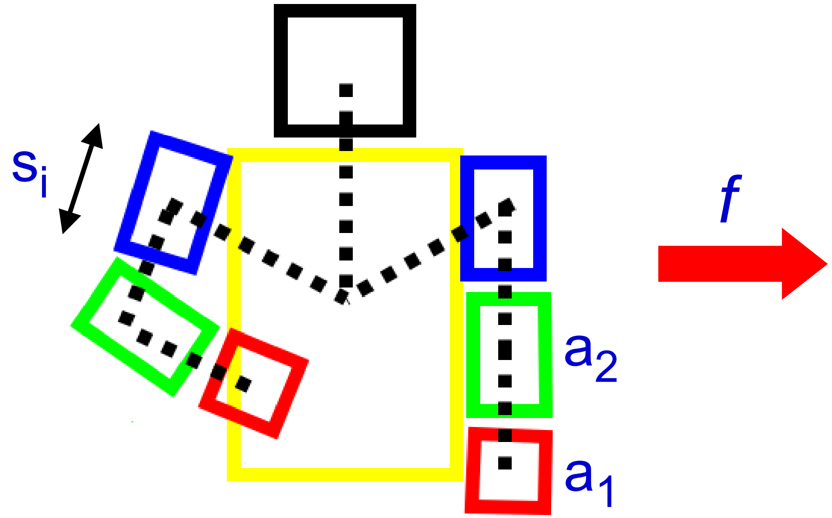
Variety of Poses

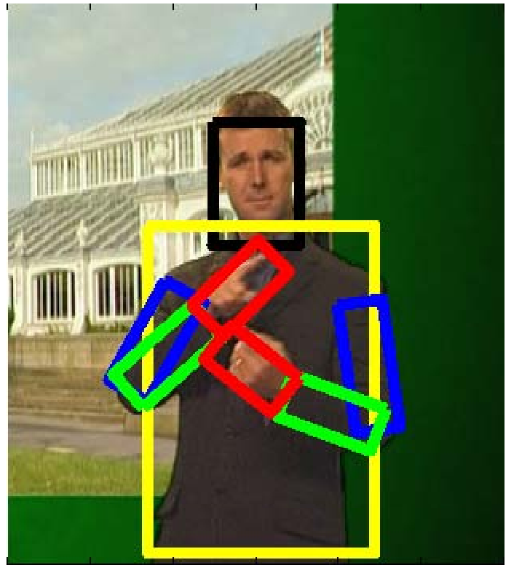


Variety of Poses



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

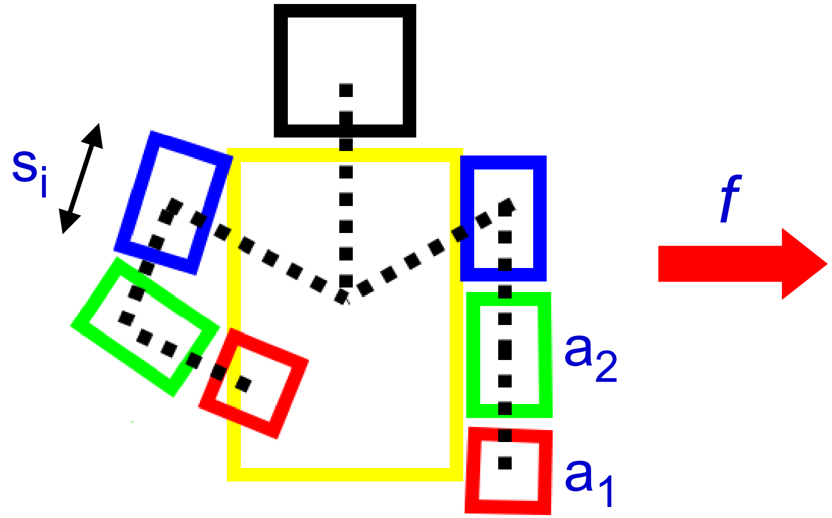


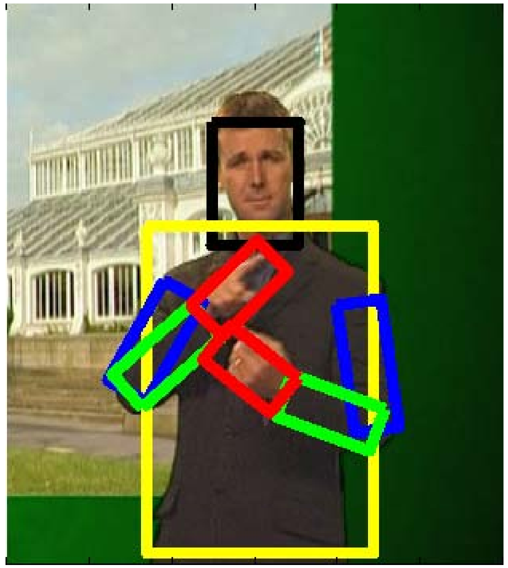


Model as a graph labelling problem

- 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:

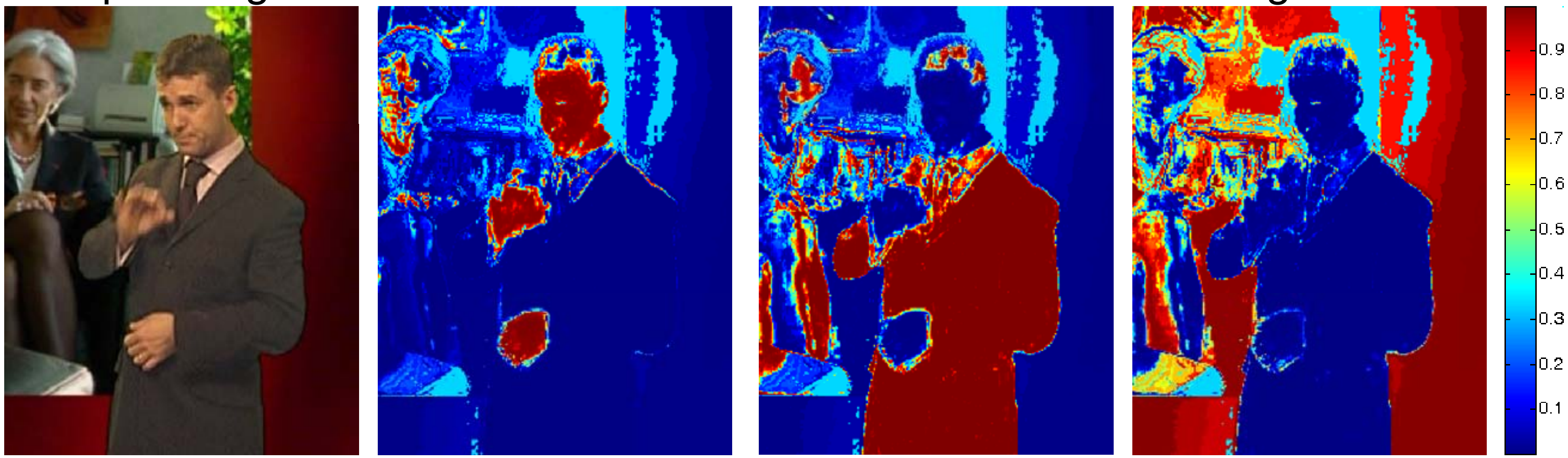
- colour
- HOG

for limbs/torso

- Fit model (inference) as labelling with lowest energy

Unary term: appearance feature I - colour

input image | skin | torso | background

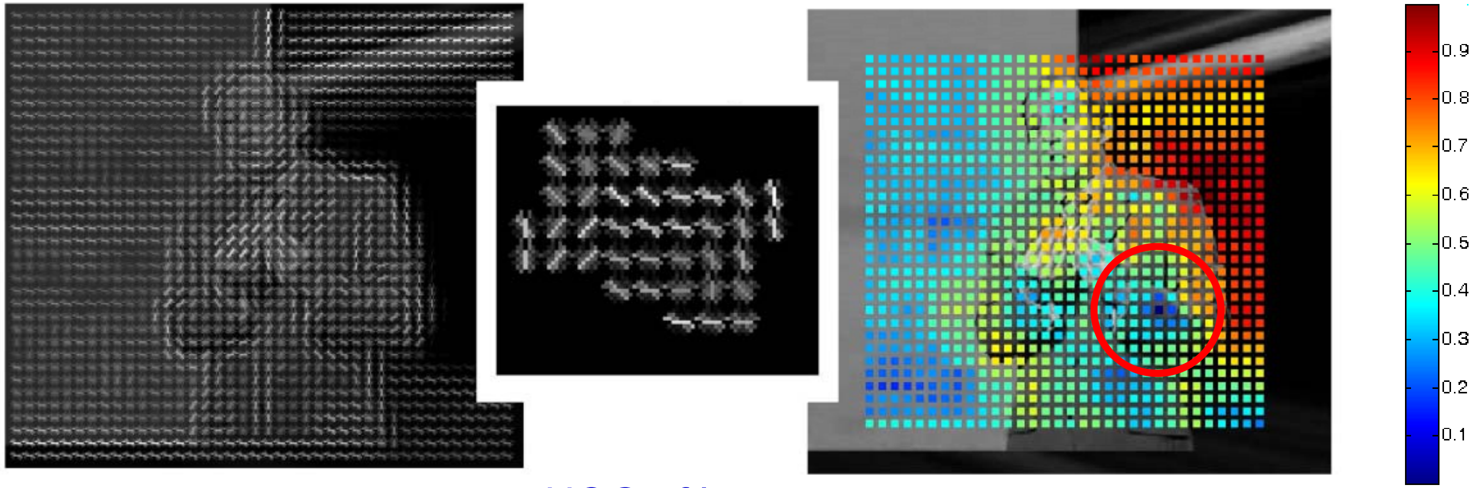


colour posteriors

Unary term: appearance feature II - HOG

Dalal & Triggs, CVPR 2005

Histogram of oriented gradients (HOG)



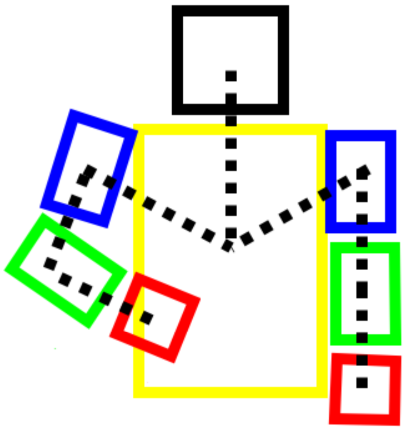
HOG of image

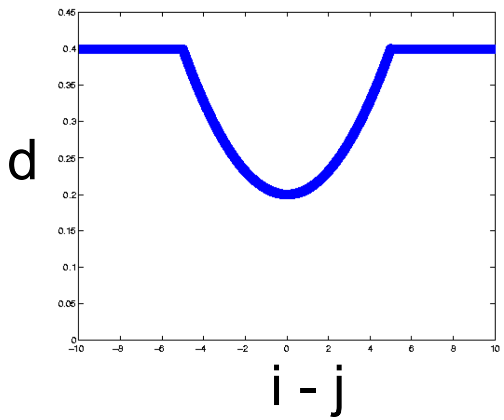
HOG of lower arm template (learned)

L2 Distance

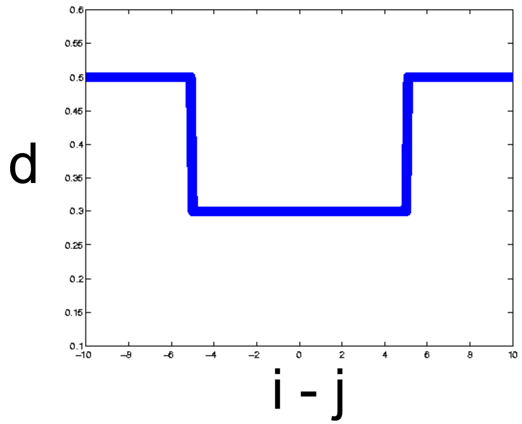
Pairwise terms: kinematic layout

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



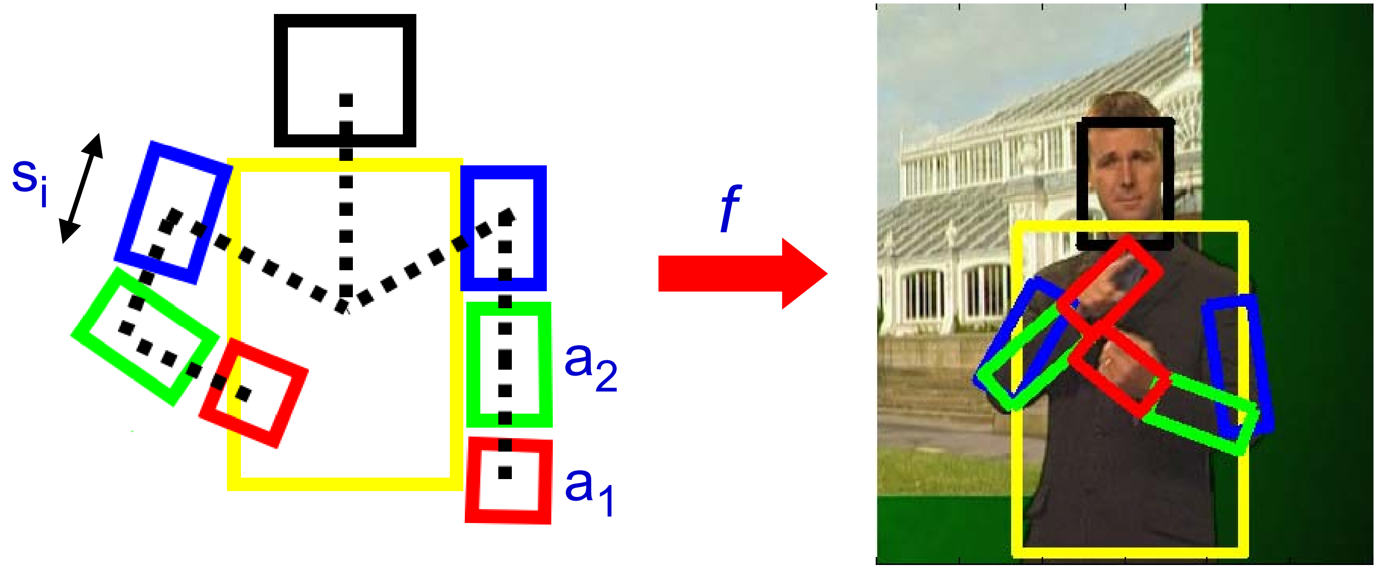


Truncated Quadratic



Potts

Pictorial structure model – CRF



- Each labelling has an energy (cost):

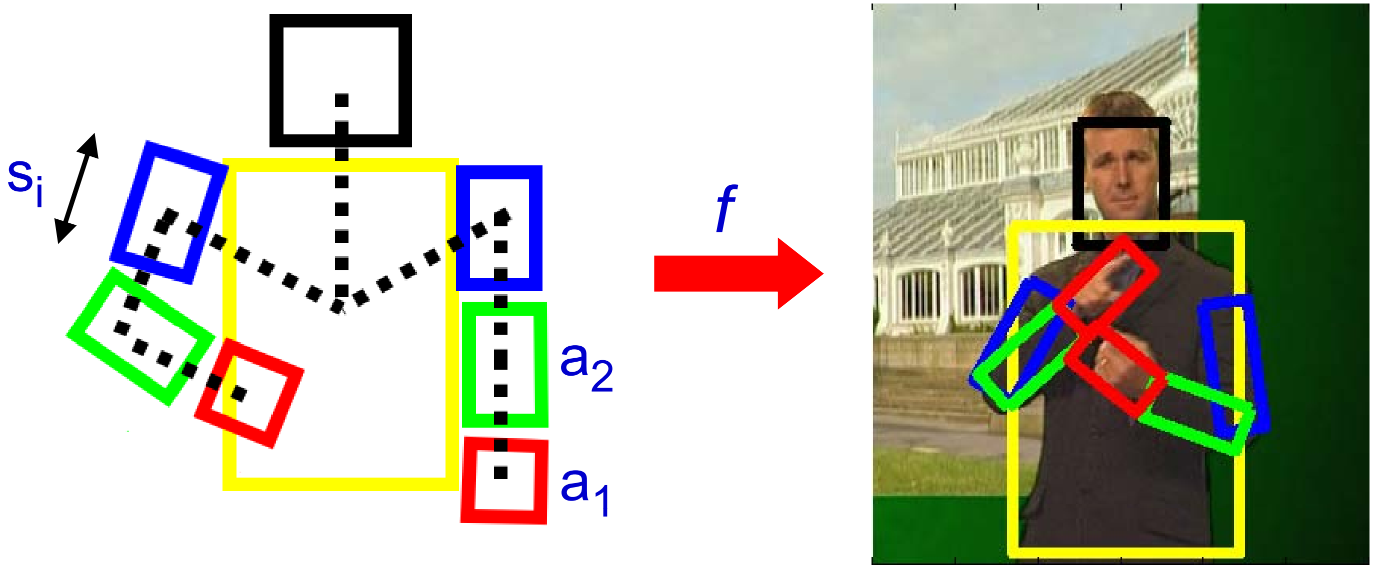
$$E(f) = \underbrace{\sum_{a \in \mathcal{V}} \theta_{a;f(a)}}_{\substack{\text{unary terms} \\ \text{(appearance)}}} + \underbrace{\sum_{(a,b) \in \mathcal{E}} \theta_{ab;f(a)f(b)}}_{\substack{\text{pairwise terms} \\ \text{(configuration)}}}$$

- Fit model (inference) as labelling with lowest energy

Features for unary:
- colour
- HOG

for limbs/torso

Complexity

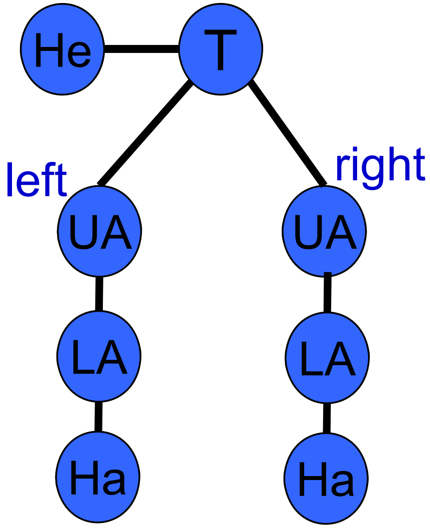


- n parts
- For each part there are h possible poses $\mathbf{p}_j = (x_j, y_j, \phi_j, s_j)$
- There are h^n possible labellings

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

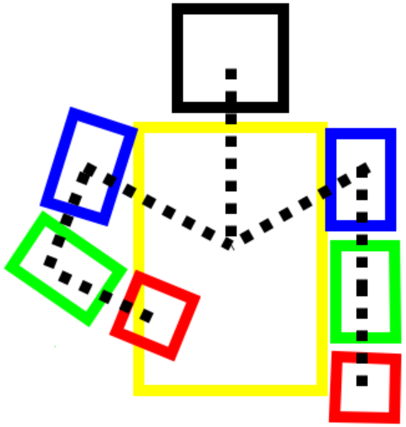
- Layout model defines the foreground,
i.e. it chooses the pixels to "explain"

- ignores skin and strong edge in background

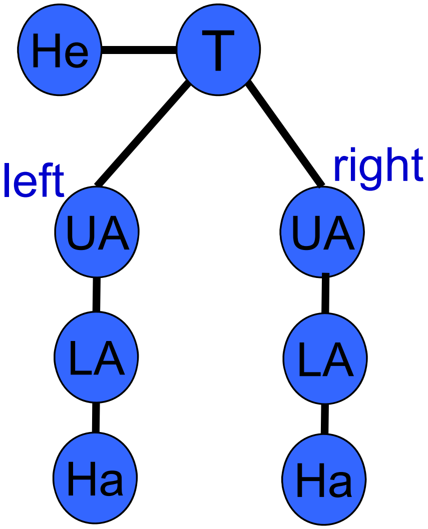
- "double counting"

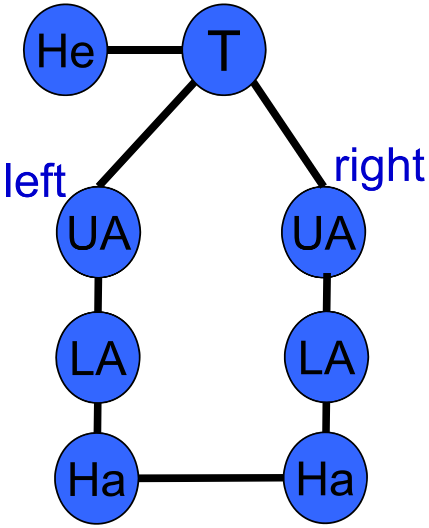
Generative model of foreground only

Kinematic structure vs graphical (independence) structure



Graph G = (V,E)





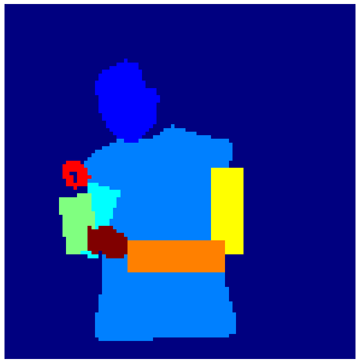
Requires more connections than a tree

And for the background problem

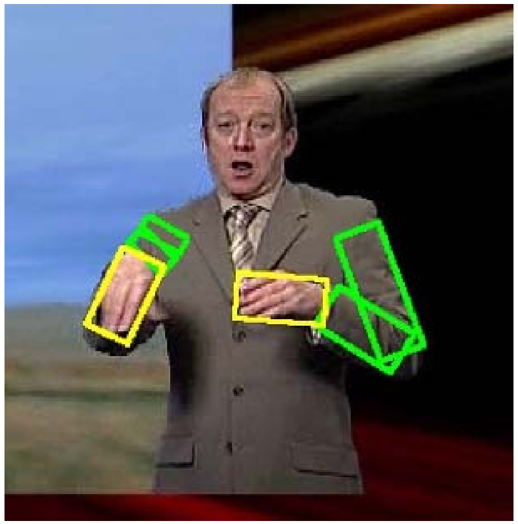
1. Add background model so that every pixel in region explained

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

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

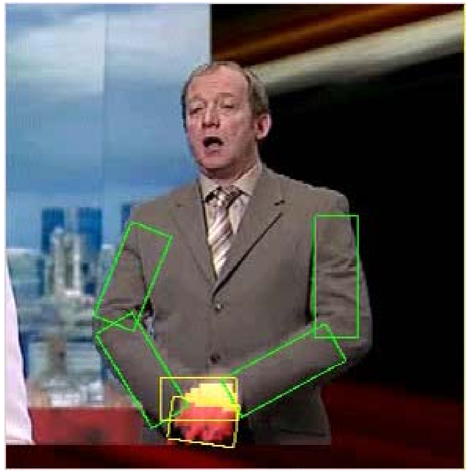
Patrick Buehler, Mark Everingham,

Daniel Huttenlocher, Andrew Zisserman

British Machine Vision Conference 2008

Objective

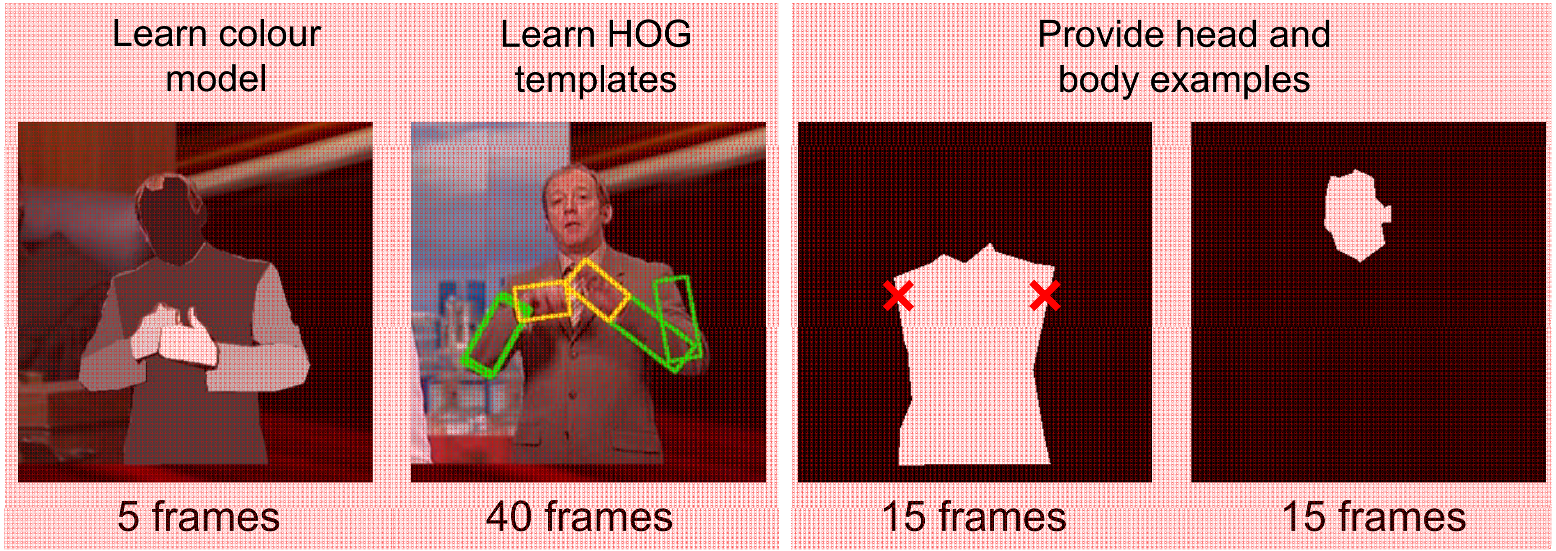
- Detect hands and arms of person signing British Sign Language
- Hour long sequences

- Strong but minimal supervision

Learning the model

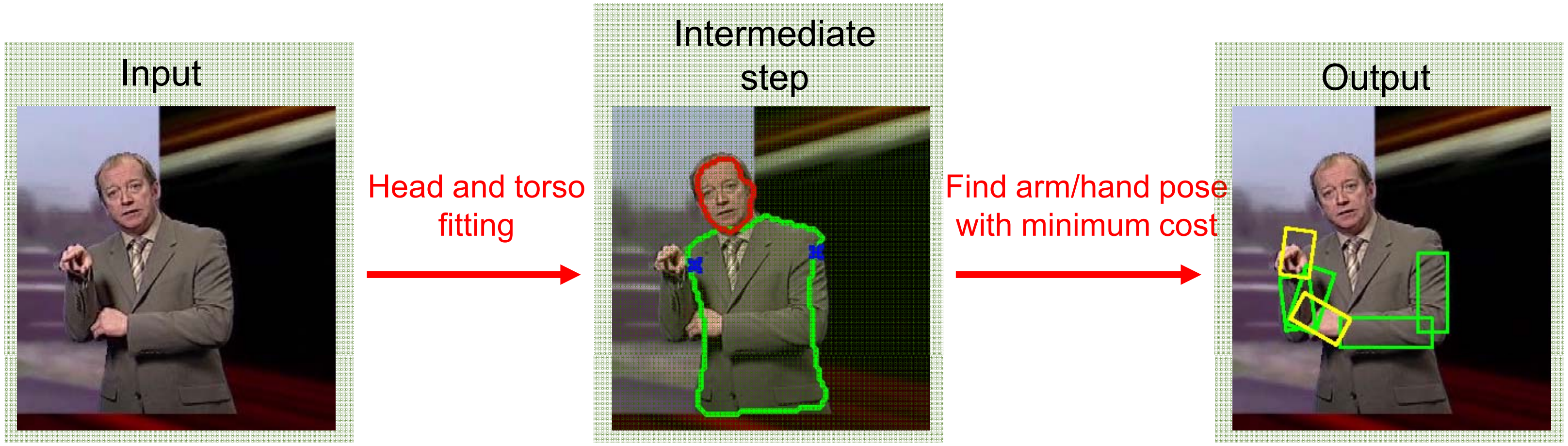
Strong supervision: manual input



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

samples

Input

Output

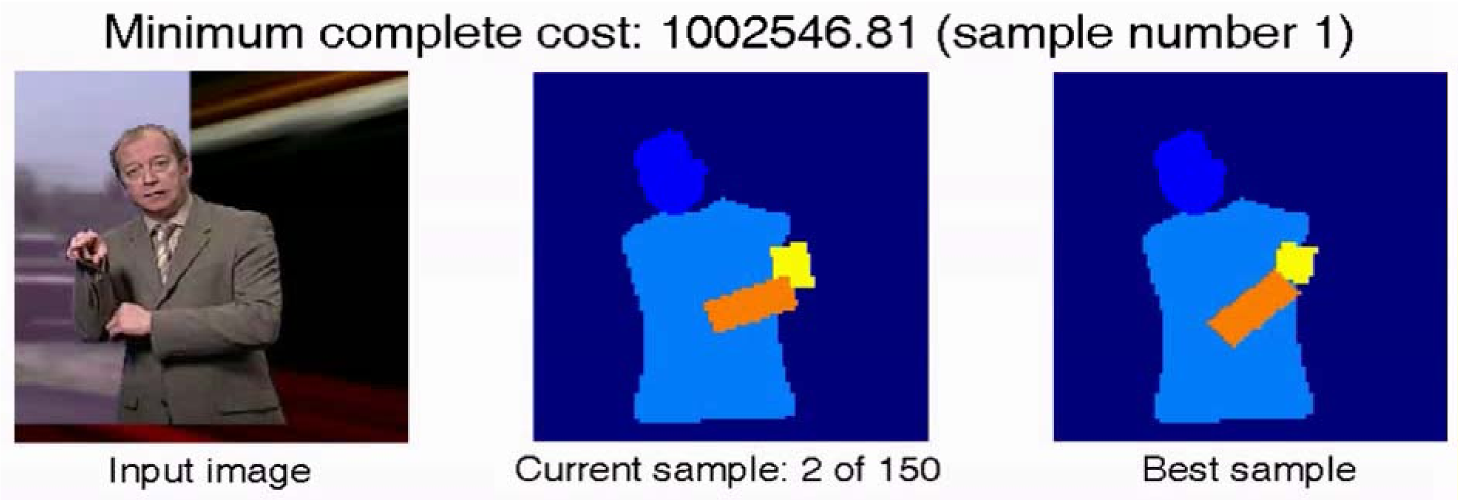
best arm candidate

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

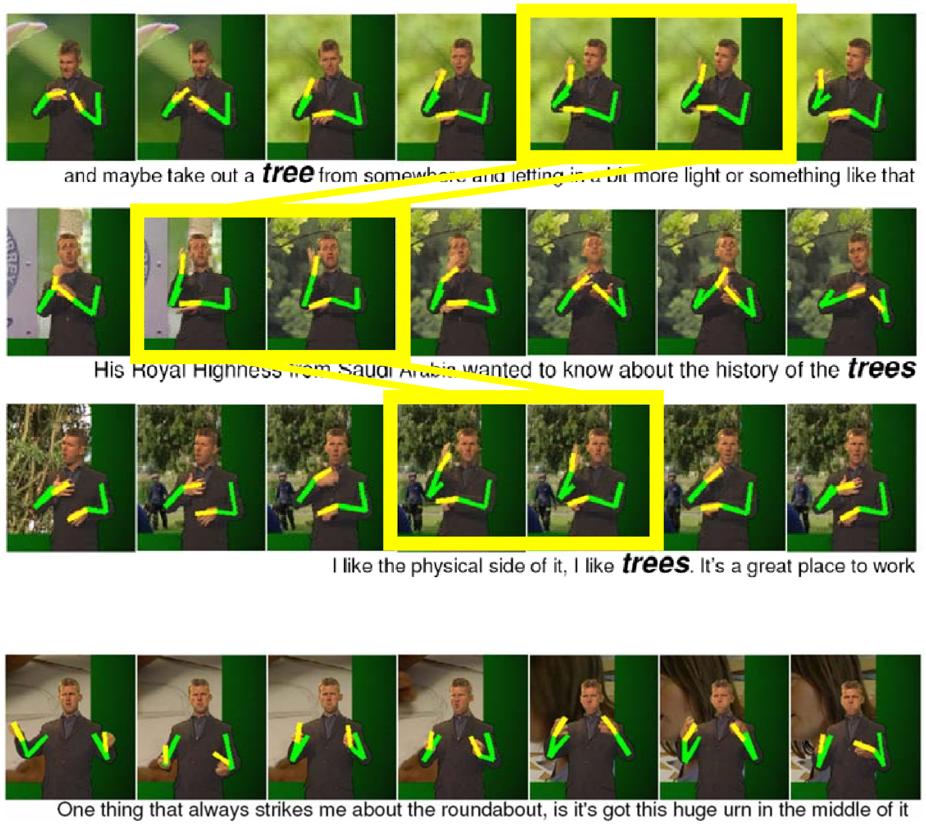
Use subtitles to find video sequences containing word. These are the positive training sequences. Use other sequences as negative training sequences.

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

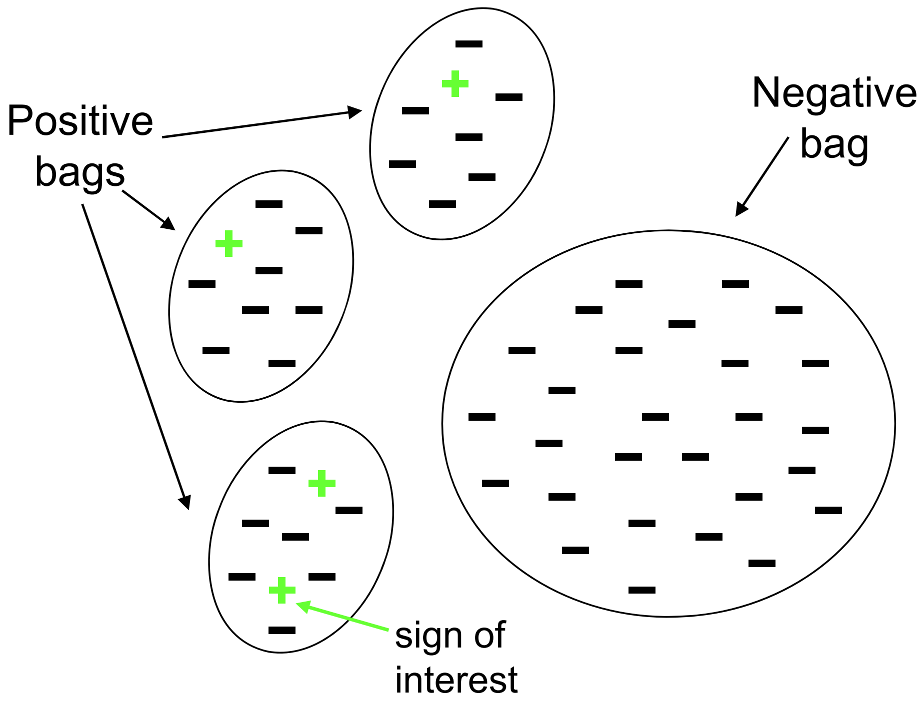
does not occur in the negative set

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

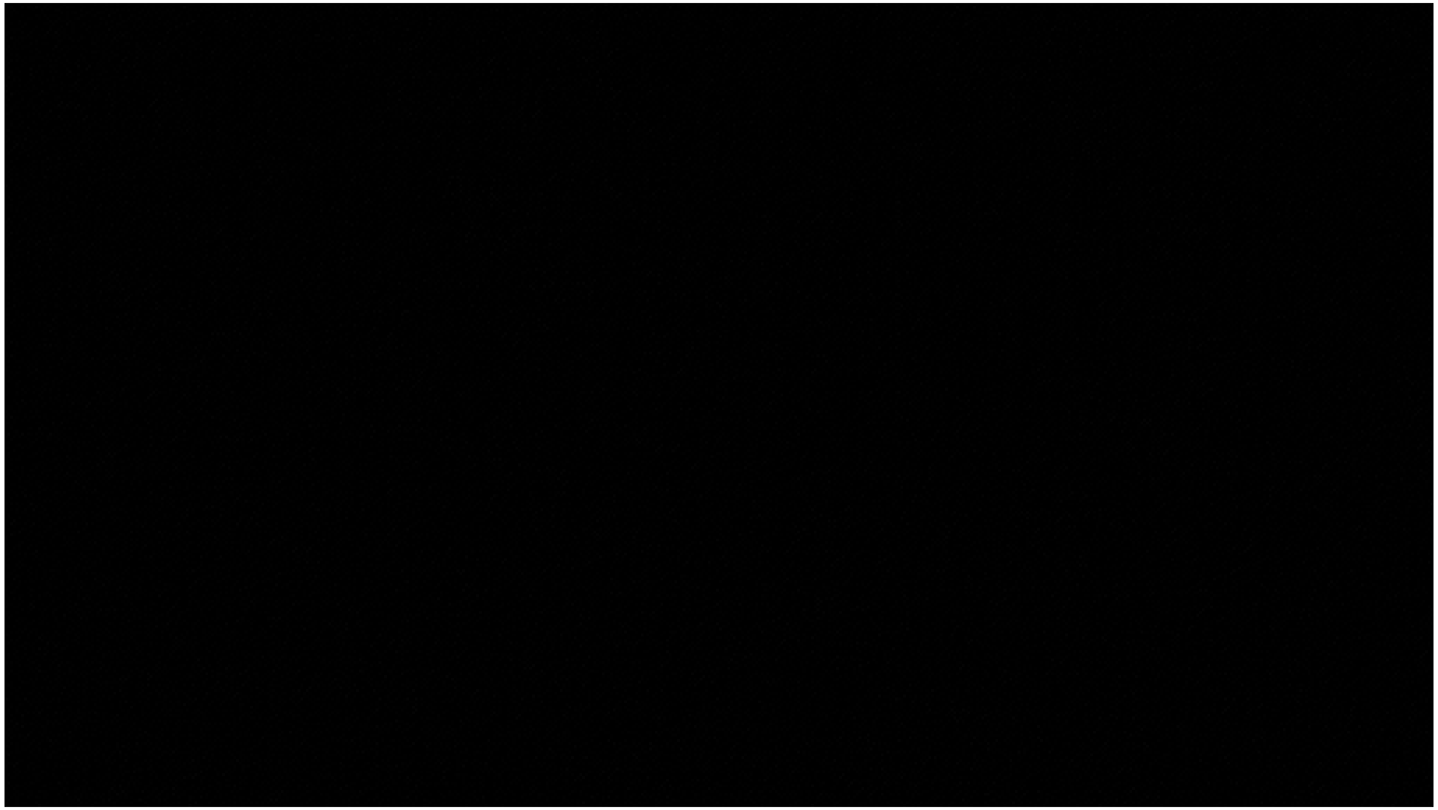
does not occur in the negative set

Multiple instance learning



Example

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



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”

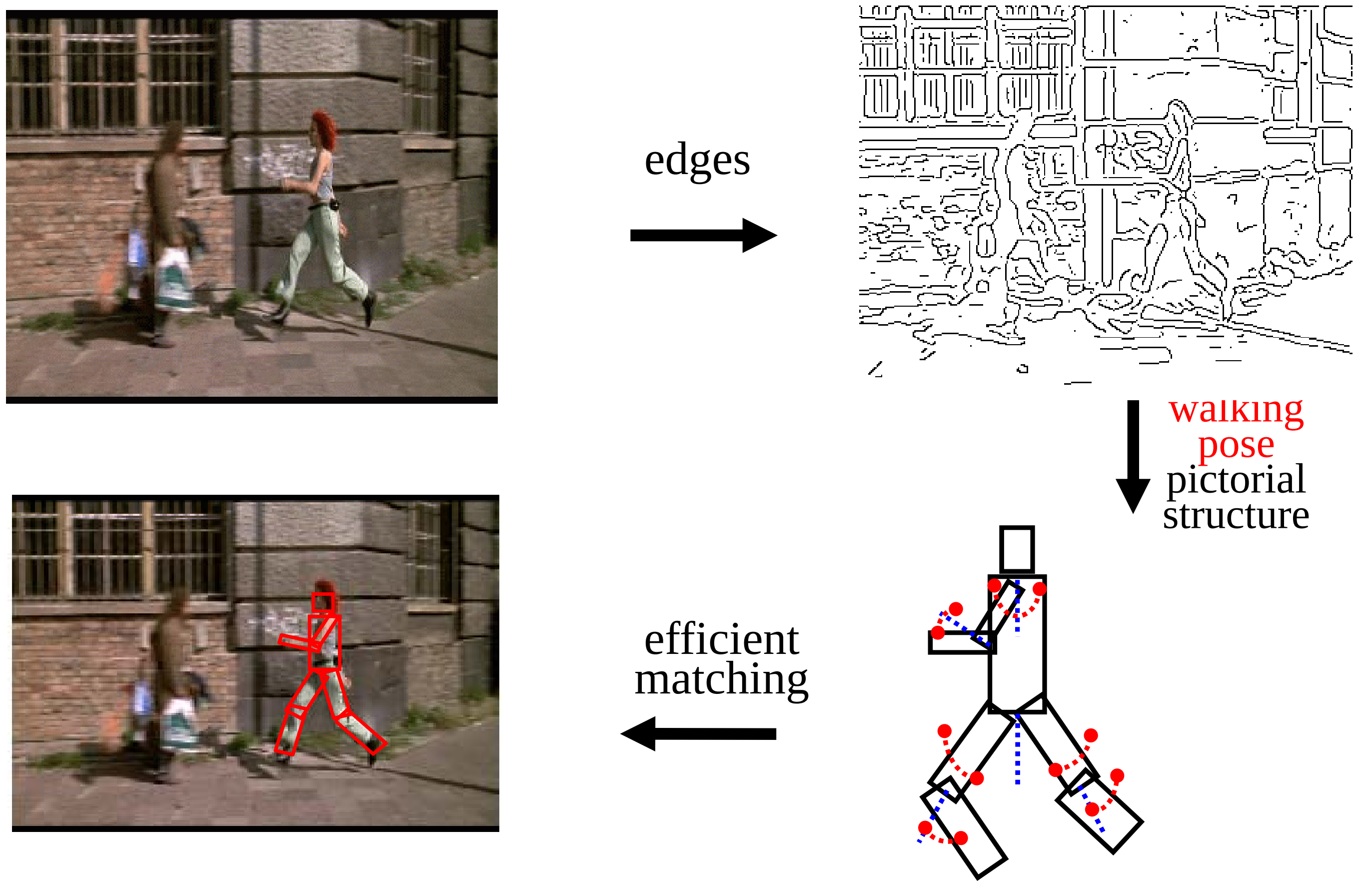
Image parsing: learning the model for a specific image

Recent advances

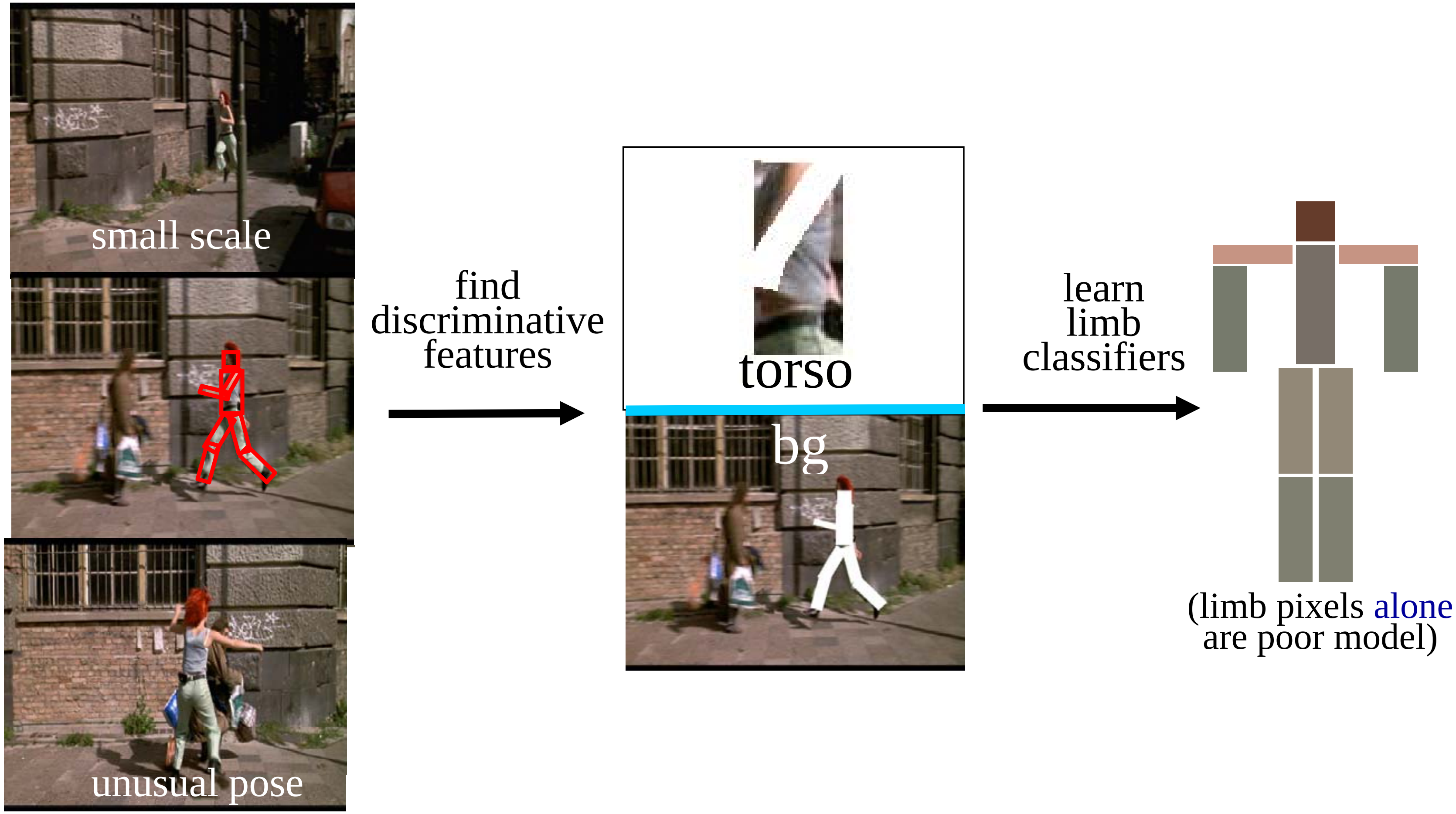
Datasets and challenges

Learning appearance models in videos

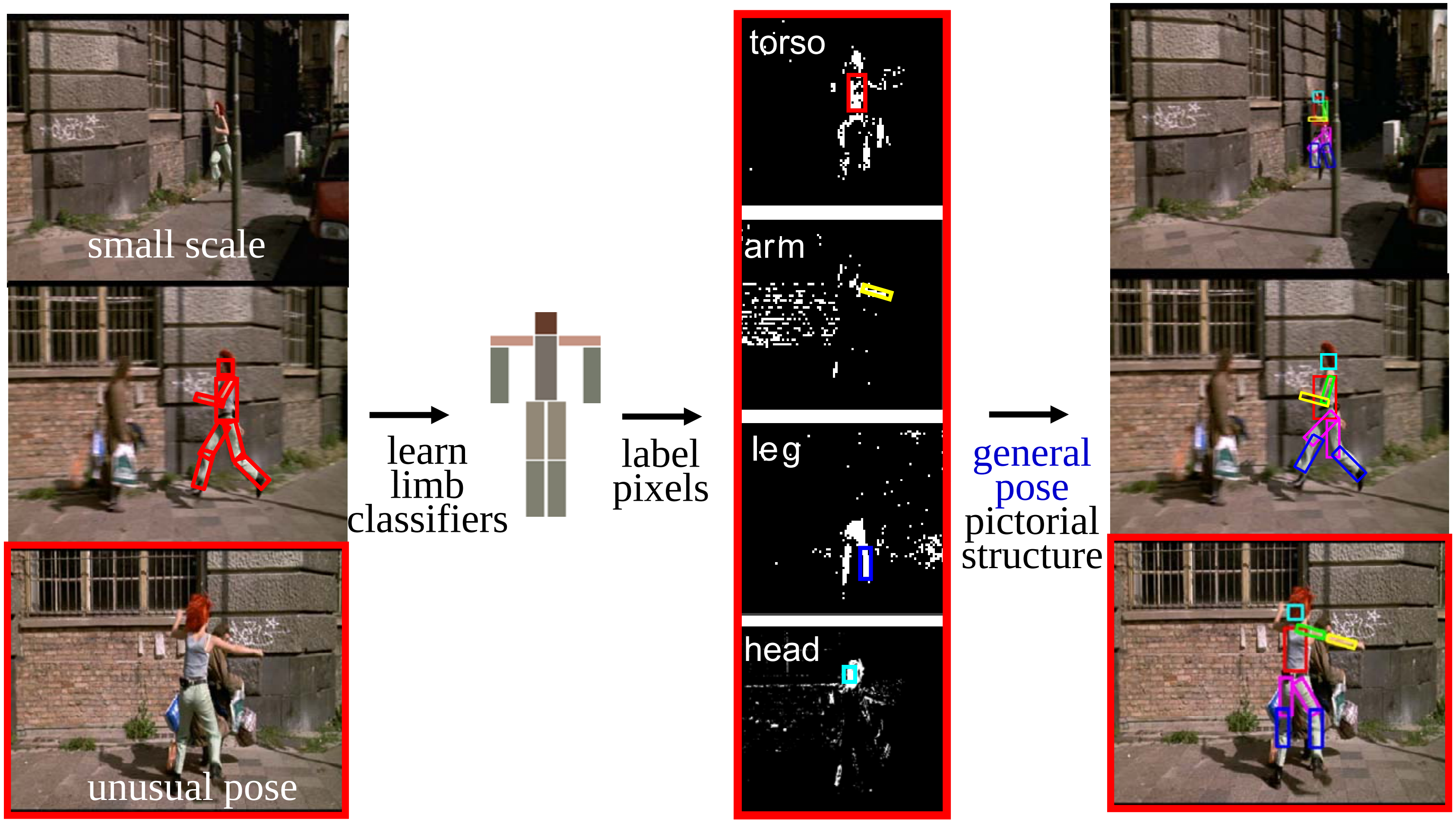
Strike a Pose: Tracking People by Finding Stylized Poses
Deva Ramanan, David Forsyth and Andrew Zisserman, CVPR 2005


edges
walking
pose
pictorial
structure
efficient
matching

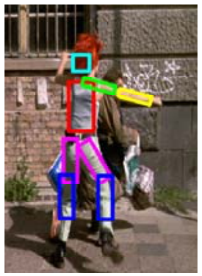
Build Model

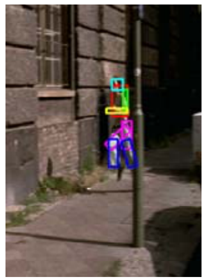

small scale
unusual pose
find discriminative features
torso
bg
learn limb classifiers
(limb pixels alone are poor model)

Build Model & Detect



Running Example





How well do classifiers generalize?

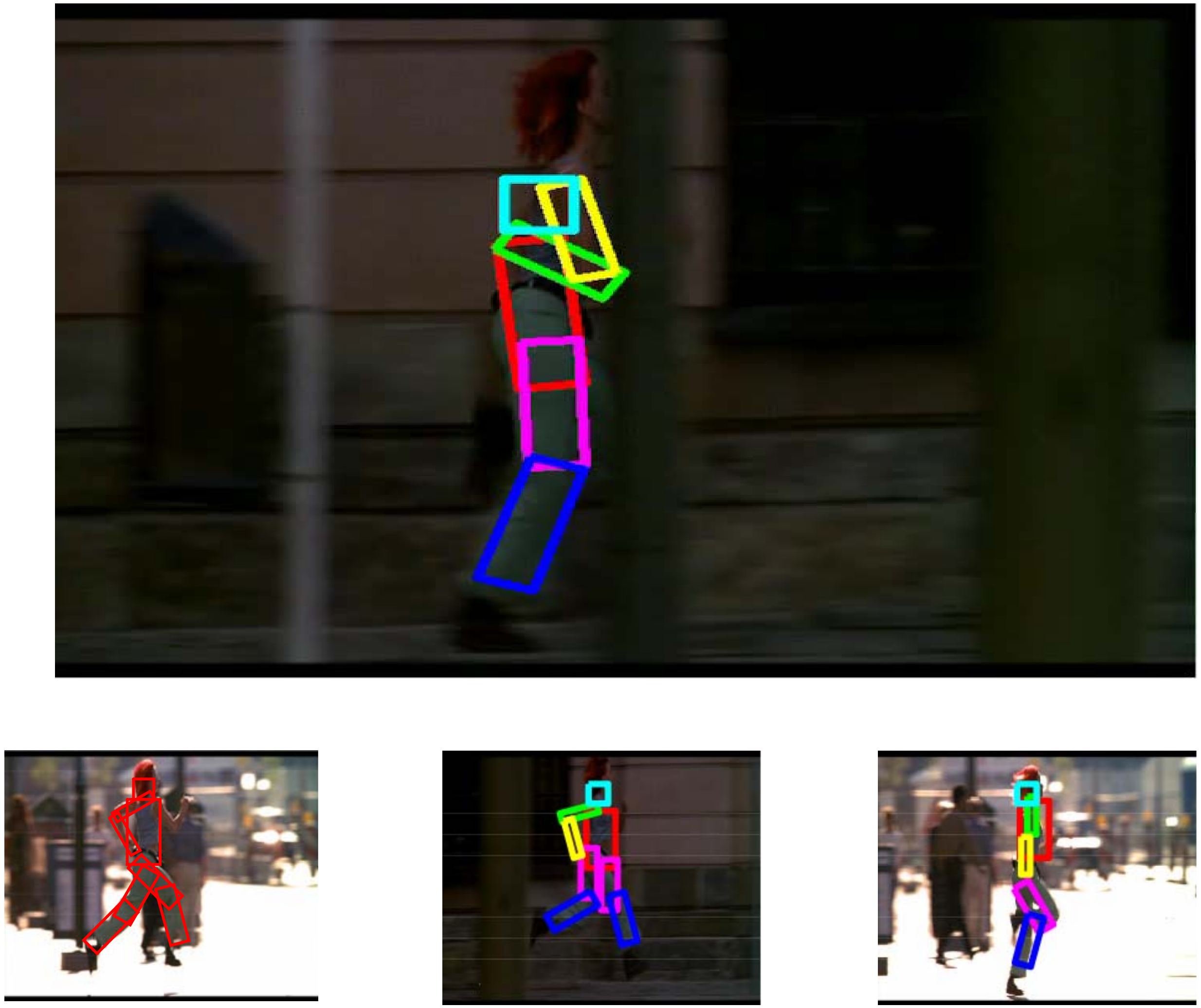
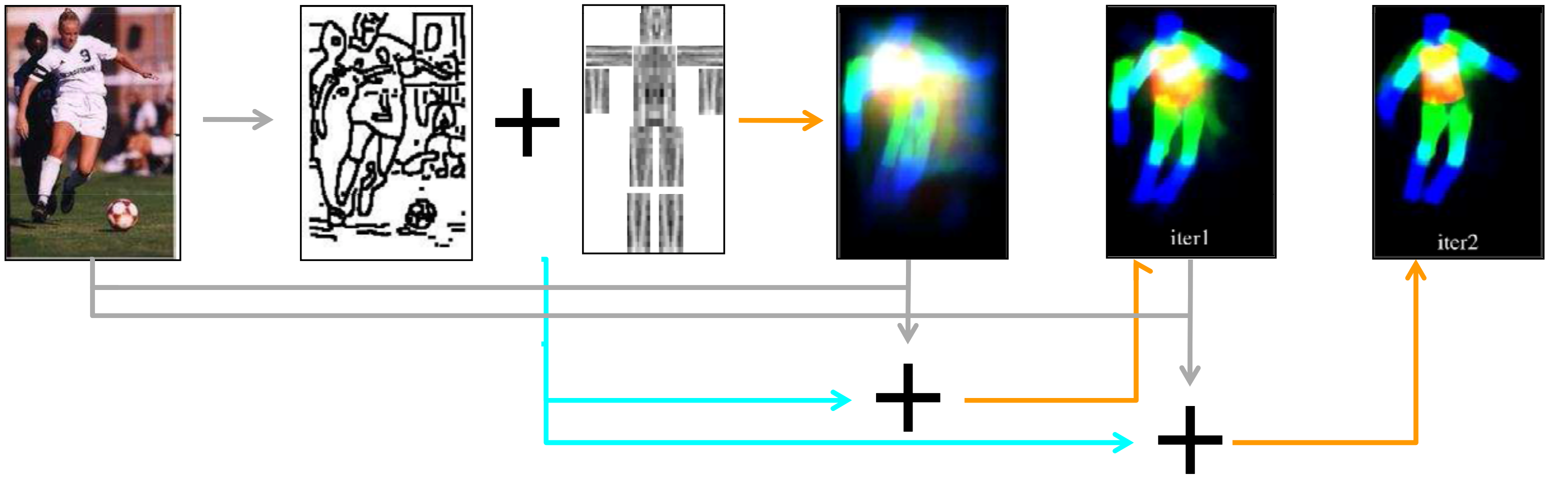


Image Parsing – Ramanan NIPS 06

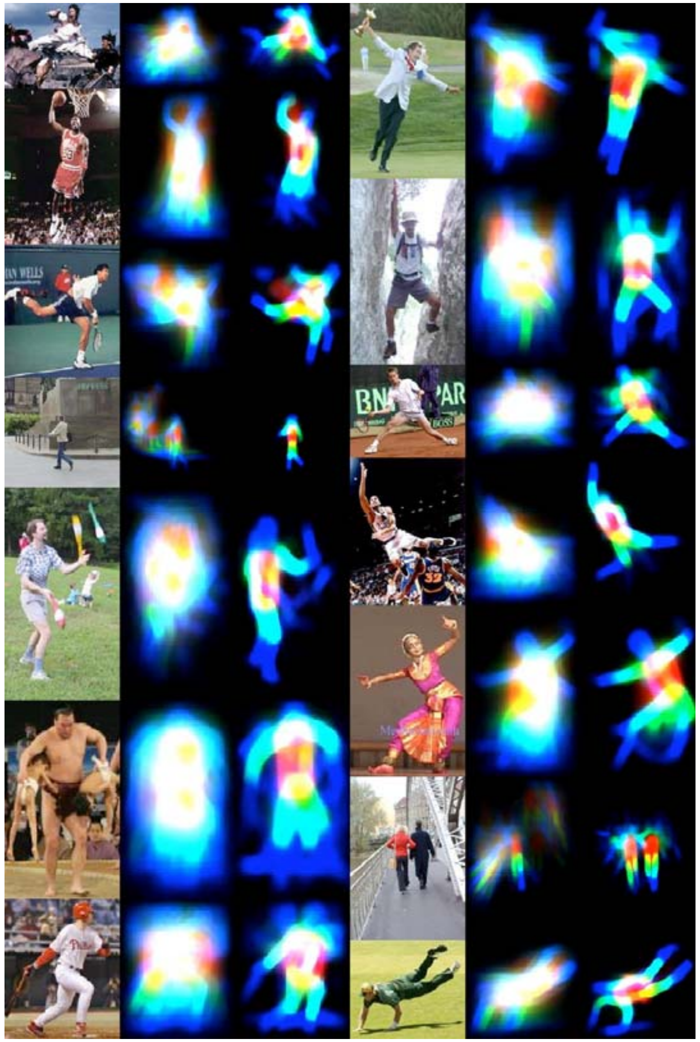
$$E(f) = \underbrace{\textstyle\sum_{a\in\mathcal{V}} \theta_{a;f(a)}}_{\substack{\text{unary terms}\\ \text{(edges/colour)}}} + \underbrace{\textstyle\sum_{(a,b)\in\mathcal{E}} \theta_{ab;f(a)f(b)}}_{\substack{\text{pairwise terms}\\ \text{(configuration)}}}$$

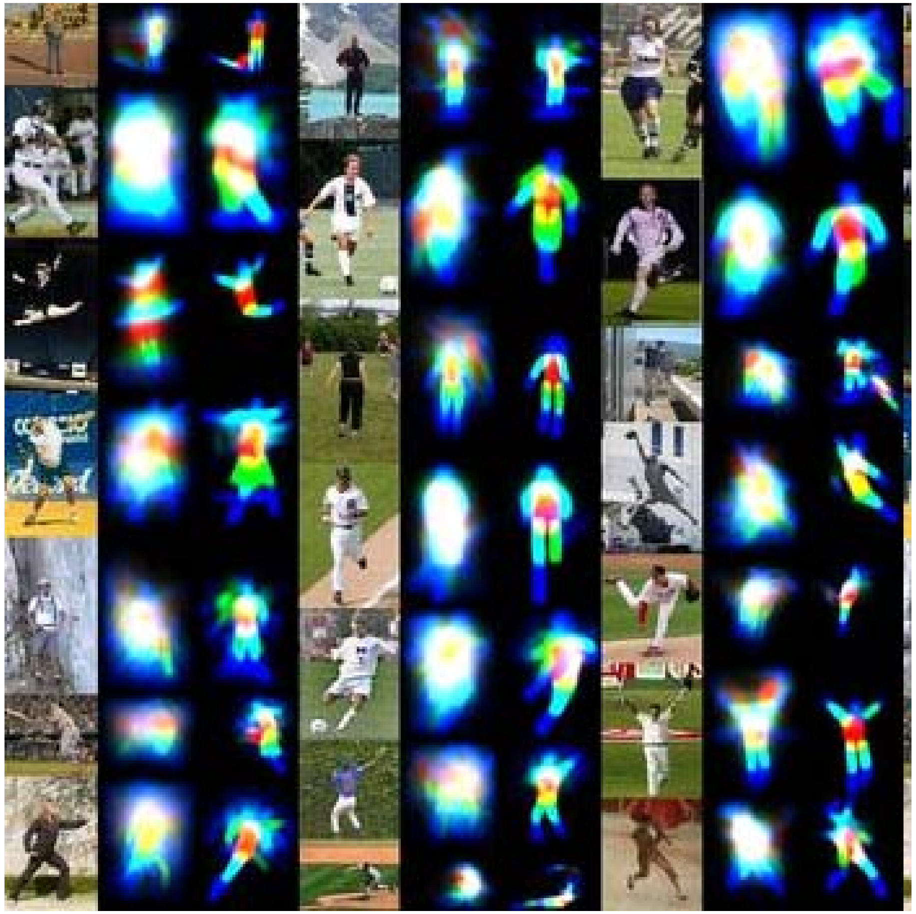
Learn image and person specific unary terms

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



77



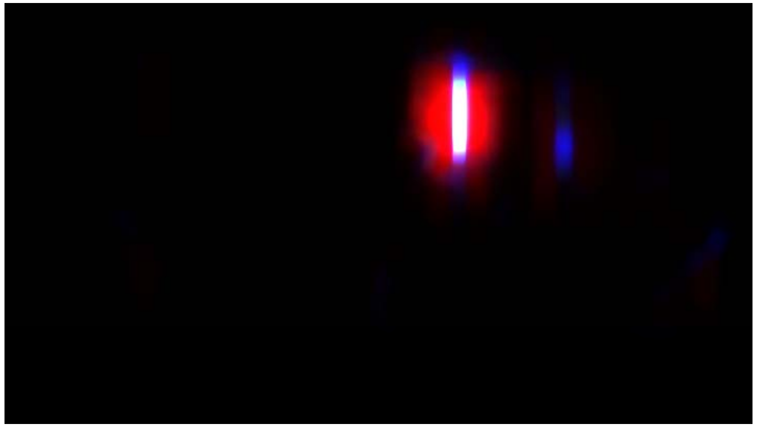


(Almost) unconstrained images

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

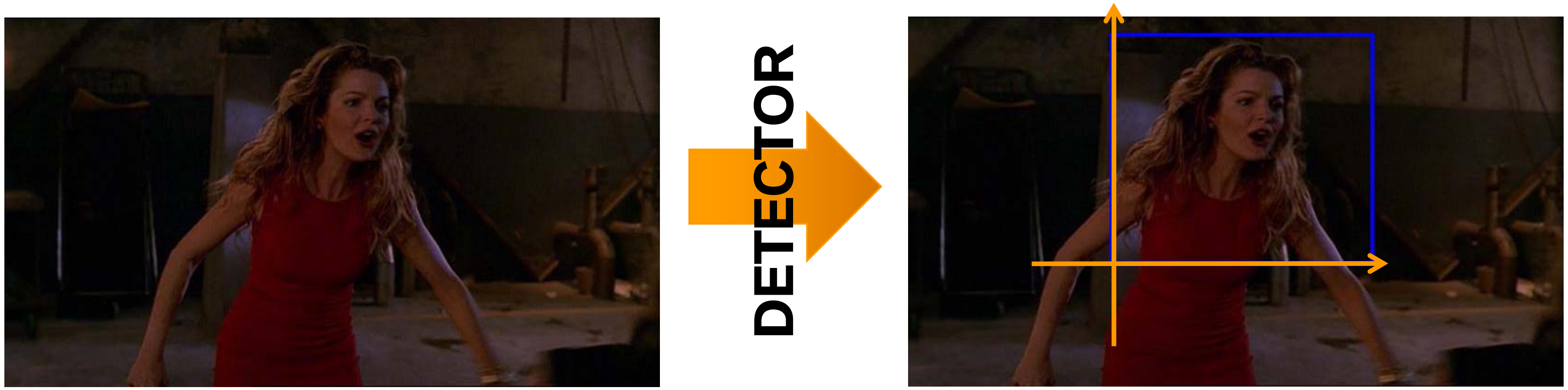
Vitto Ferrari, Manuel Marin-Jimenez, Andrew Zisserman

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

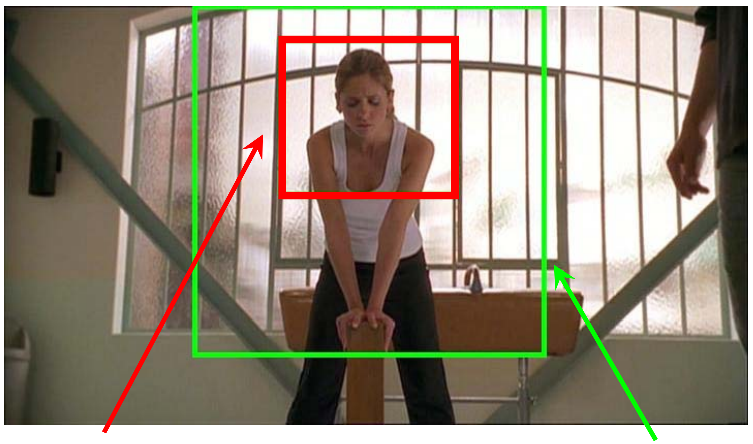
- **Detect person** – use upper body detector
- Use upper body region to restrict search
- Use colour segmentation to restrict search further
- Parsing pictorial structure by Ramanan NIPS 06

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

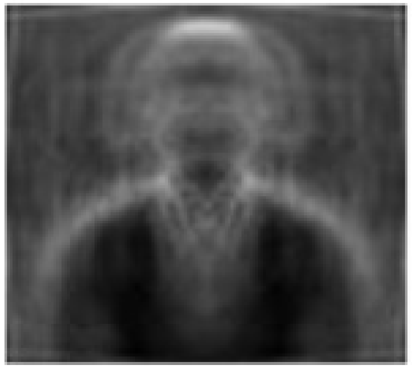
+ detects also back views

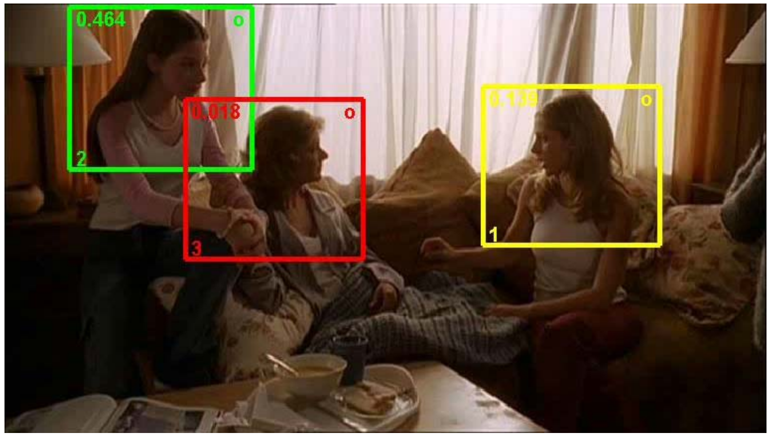
+ fast

- little info about pose (arms)

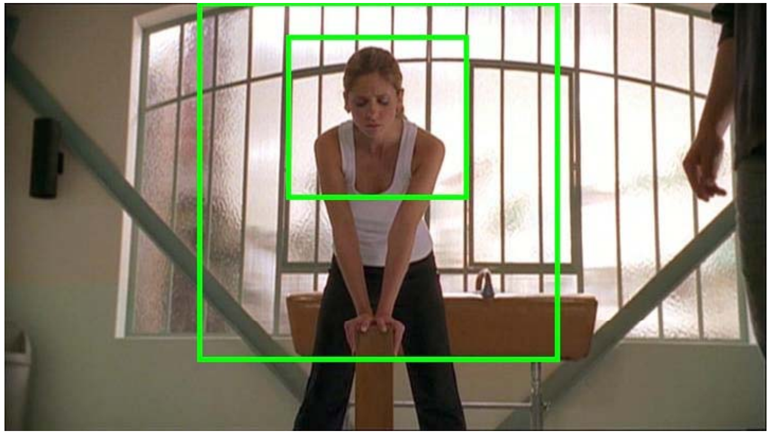
Upper body detector – using HOGs

average training data





Search space reduction by foreground highlighting



initialization

output

Idea

exploit knowledge about structure of search area to initialize Grabcut

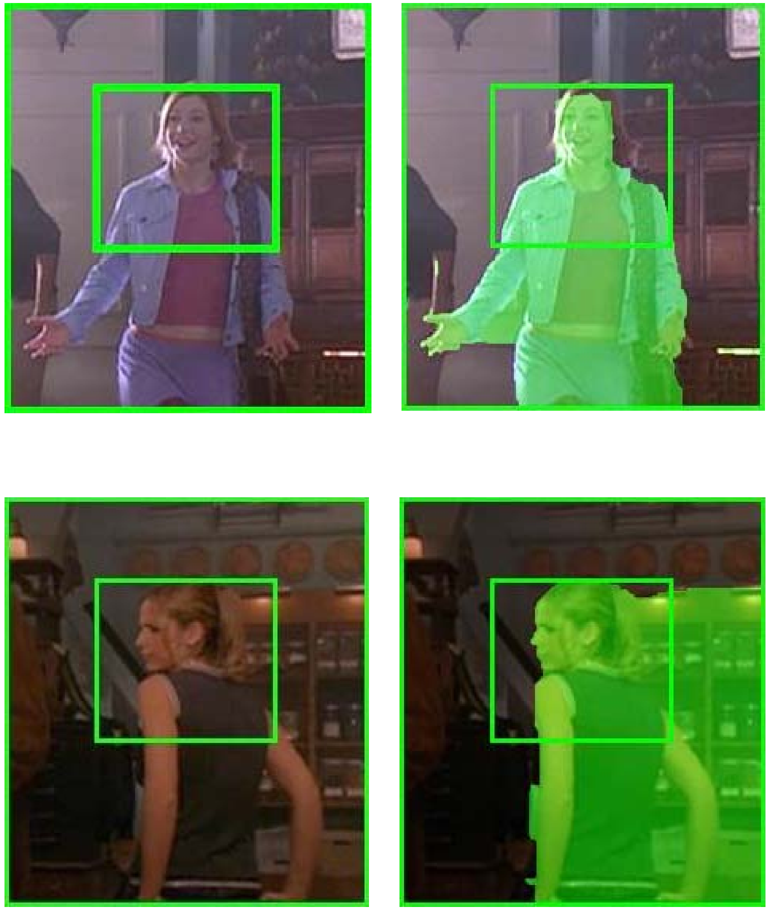
Initialization

- 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

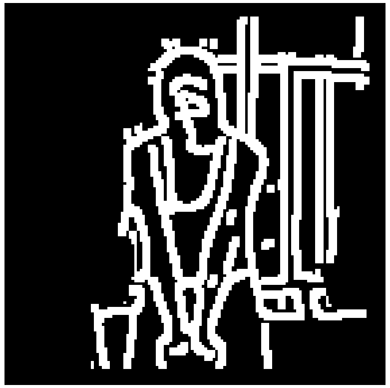
Initialization

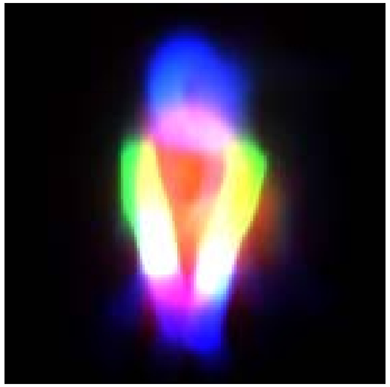
- learn fg/bg models from regions where person likely present/absent
- clamp central strip to fg
- don’t clamp bg (arms can be anywhere)

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

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

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)

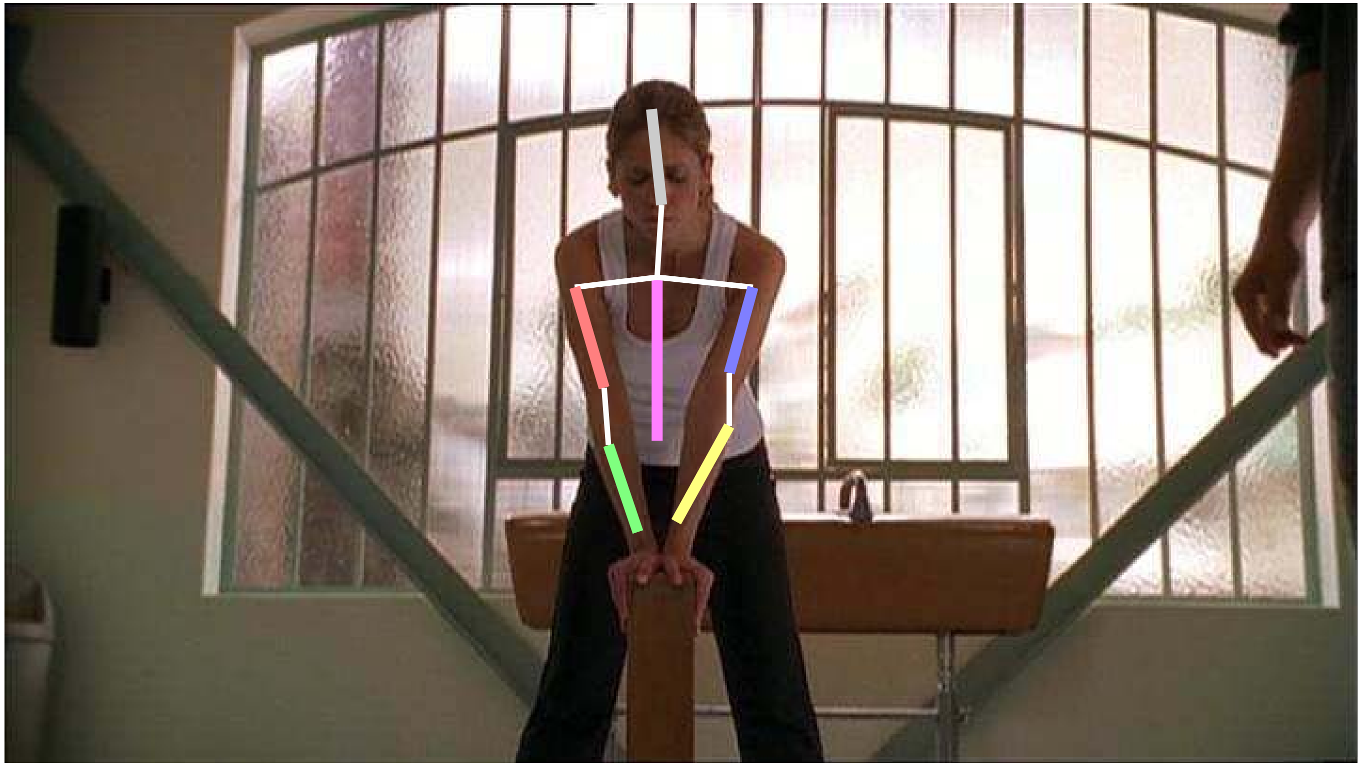
Failure of direct pose estimation

Ramanan NIPS 2006 unaided

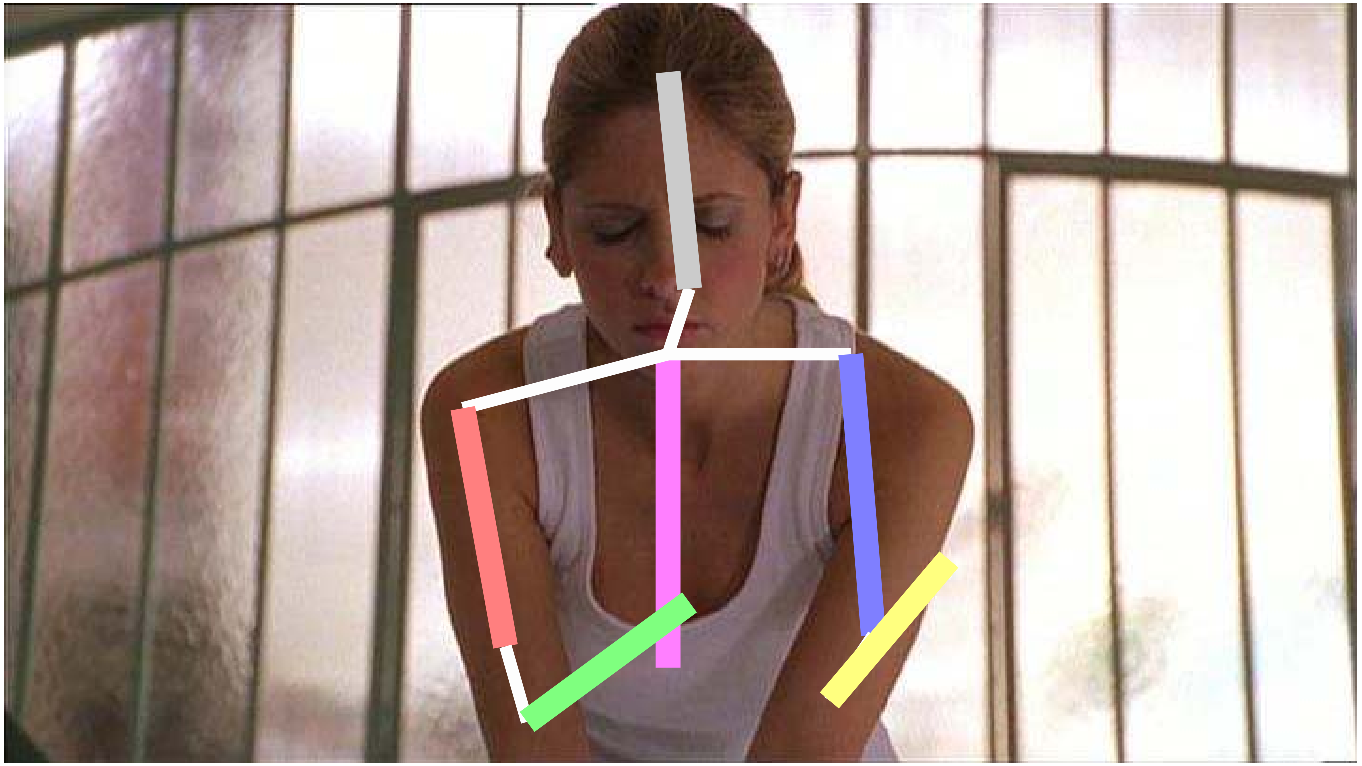
Results on Buffy frames

Results on PASCAL flickr images

What is missed?

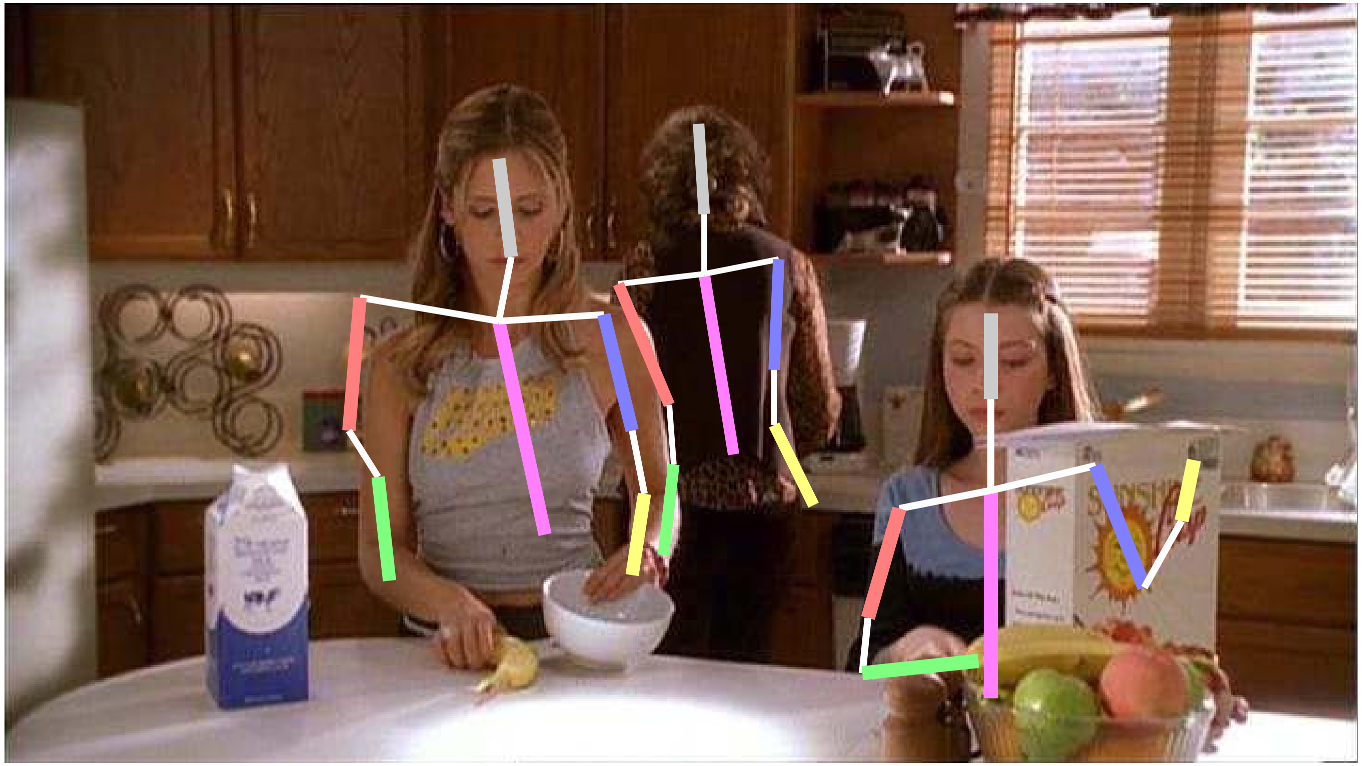


What is missed?



truncation is not modelled

What is missed?



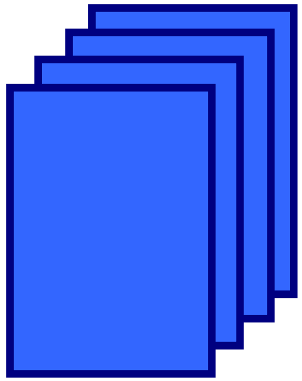
occlusion is not modelled

Application: Pose Search

Given user-selected
query frame+person …

query

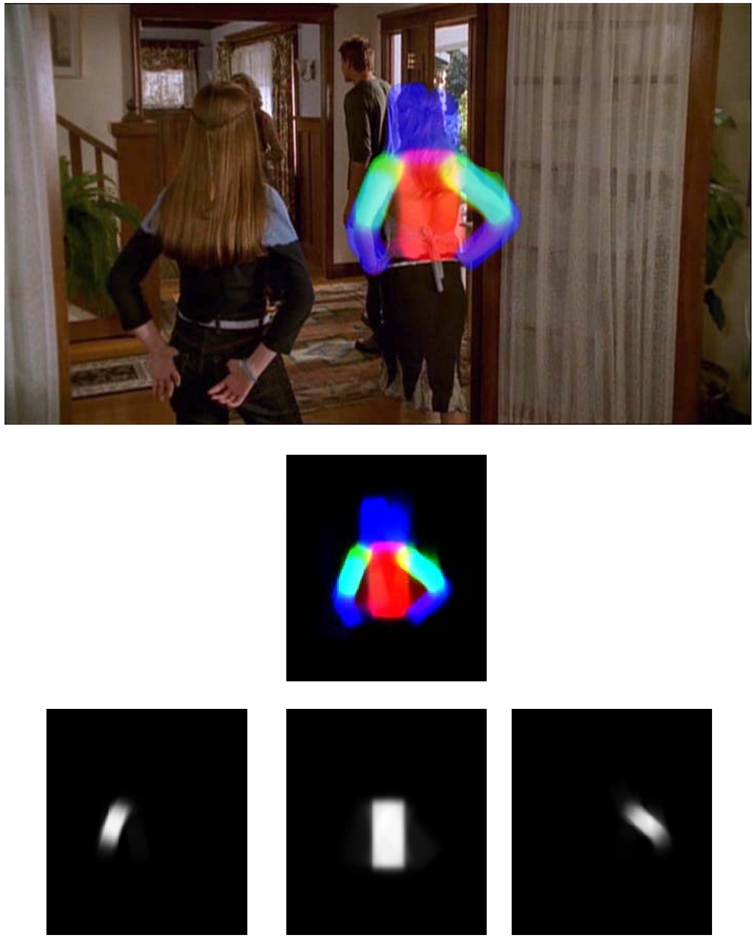
… retrieve shots with persons
in the same pose from video database



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

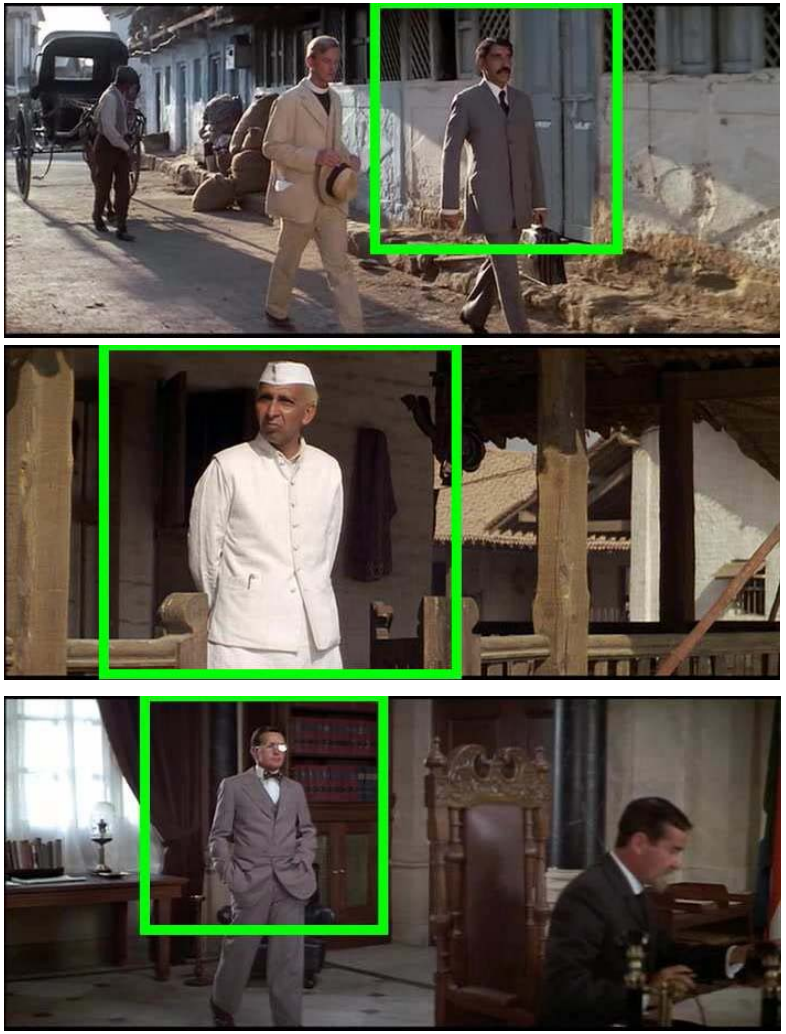
Pose Search

“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

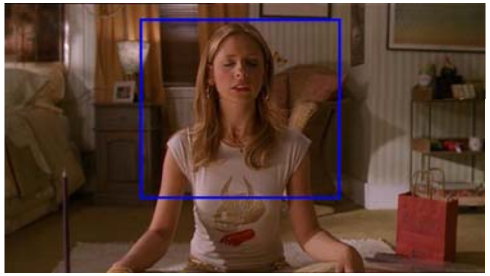
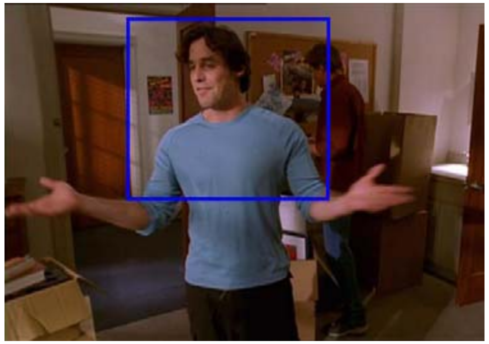
Better appearance models for pictorial structures

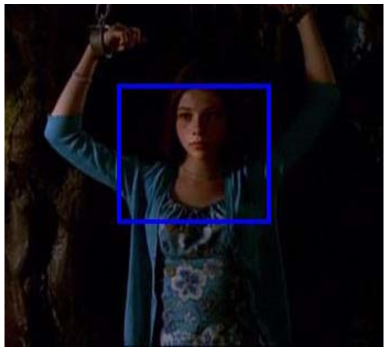
Marcin Eichner, Vittorio Ferrari
BMVC 2009

Better Appearance Models
Intuition 1

relative location (wrt detection window):

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



112

Better Appearance Models
Intuition 2

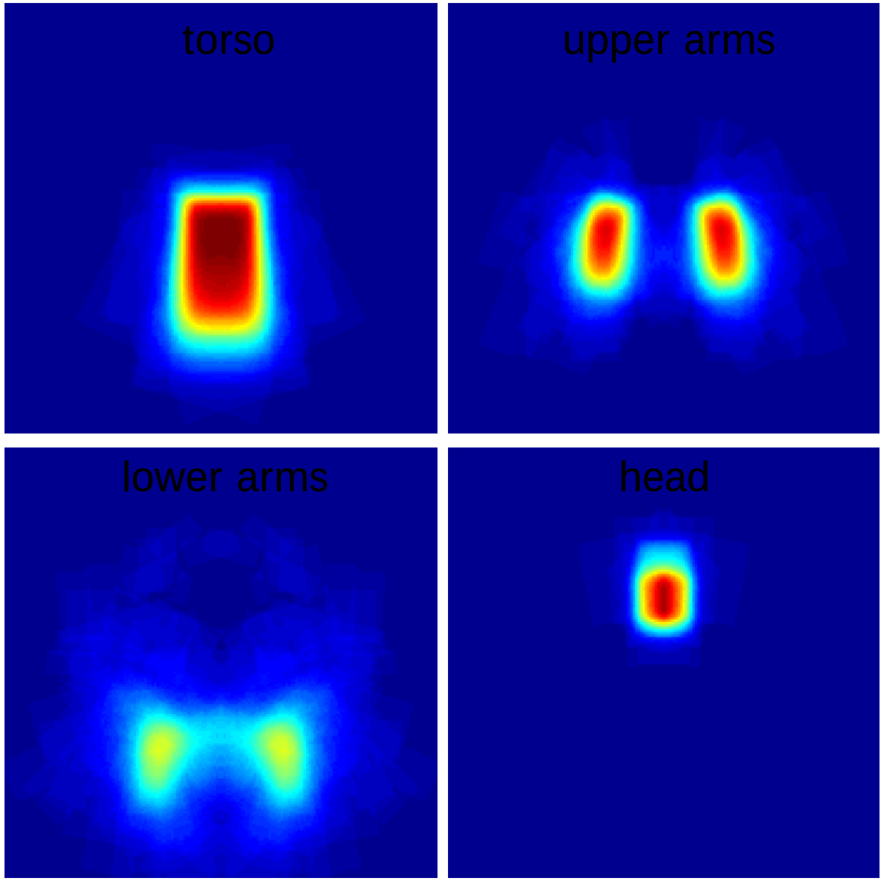
Appearance of different body parts is related

Use stable parts to improve the prediction of the unstable ones

Better Appearance Models – TRAINING Location Prior (LP)

LP encodes:

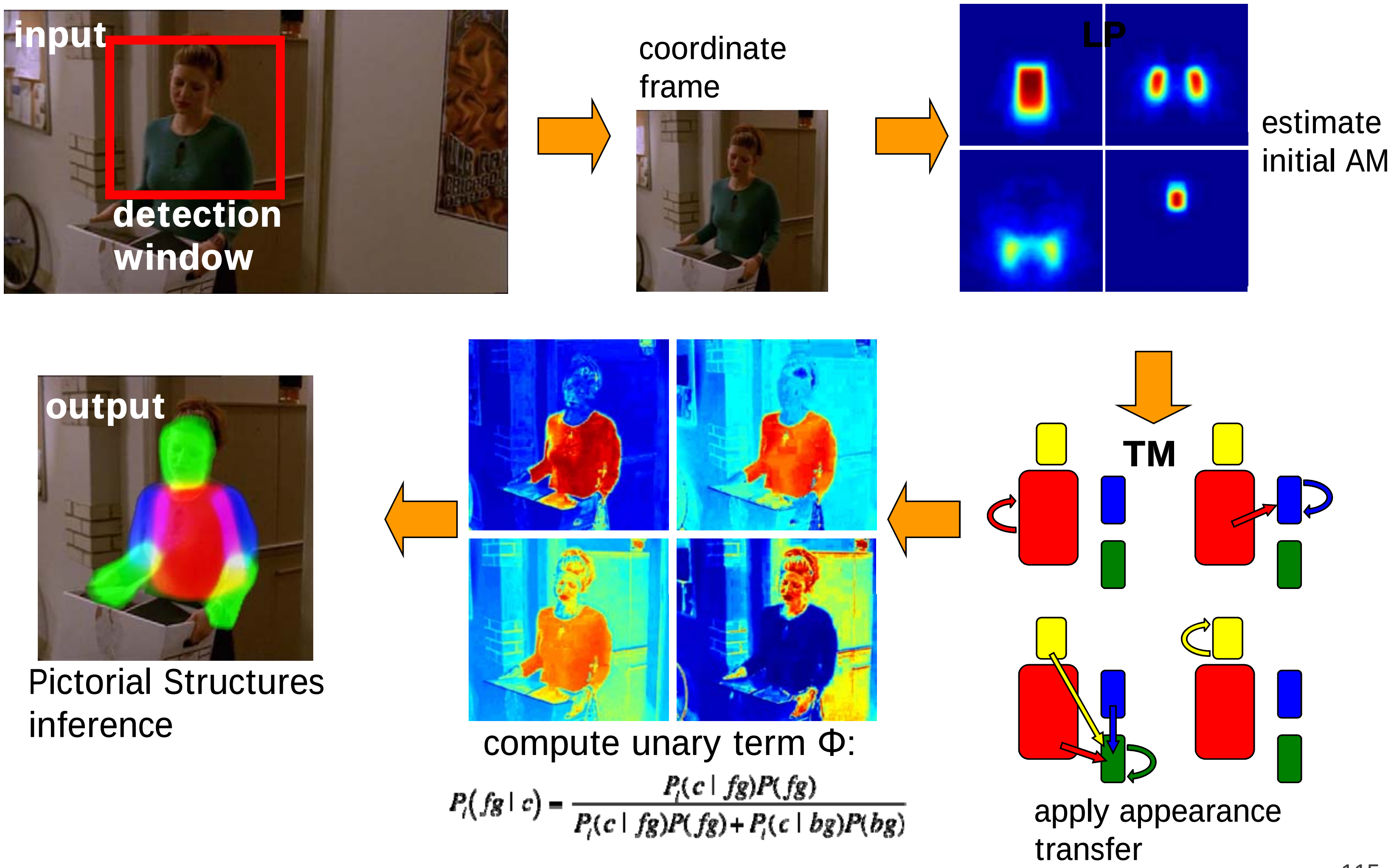
- variability of poses
- detection window inaccuracy



learnt location priors (PASCAL & Buffy 3,4)

114

Better Appearance Models – TEST



$$P_i(fg \mid c) = \frac{P_i(c \mid fg)P(fg)}{P_i(c \mid fg)P(fg) + P_i(c \mid bg)P(bg)}$$

115

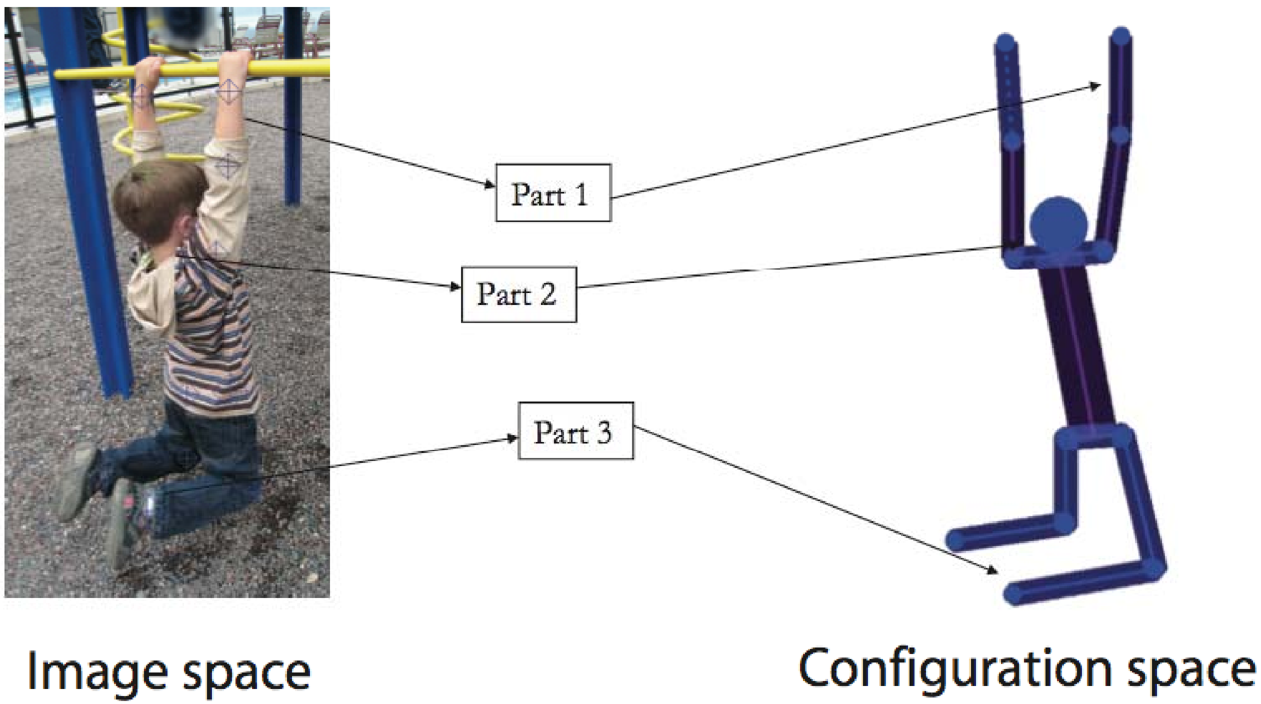
H3D: Humans in 3D

Lubomir Bourdev & Jitendra Malik

ICCV 2009

Robust detection is challenging and requires using parts

But how do we choose good parts?



Parts clustered in config space

Generalized Cylinders
[Nevatia, Binford AI77]

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

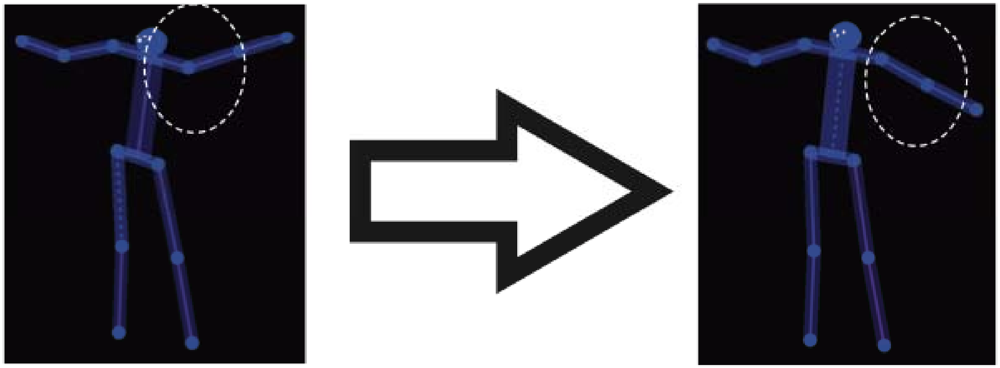
Parts clustered in image space

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:**



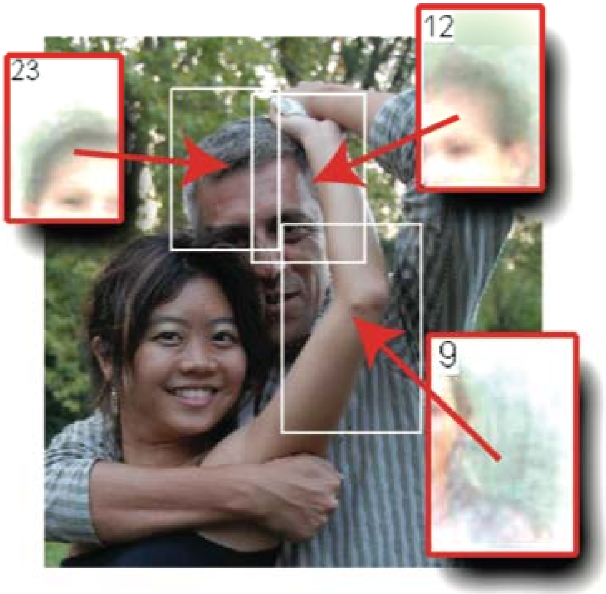
2. **Use it to generate similar examples given a query:**

query Match 1 Match 2 Weaker Match

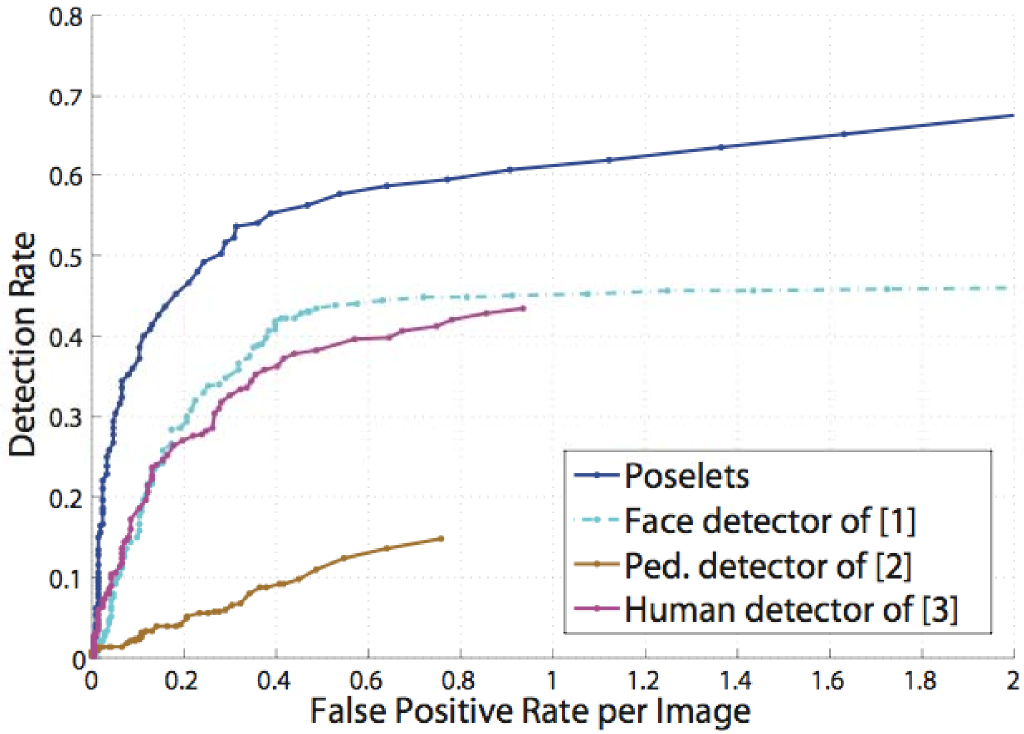
Average image for 100 poselets

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**



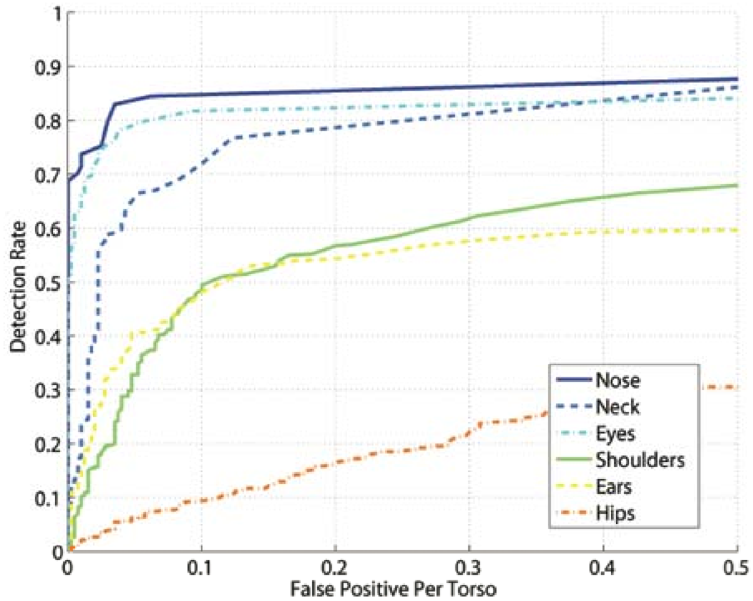
[1] L.Bourdev and J.Brandt, *Robust Object Detection using a Soft Cascade*, CVPR05
[2] N.Dalal and B.Triggs, *Histograms of Oriented Gradients for Human Detection*, CVPR05
[3] P.Felzenszwalb, D.Mcallester and D.Ramanan, *A Discriminatively Trained, Multiscale, Deformable Part Model*, CVPR08

- **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”

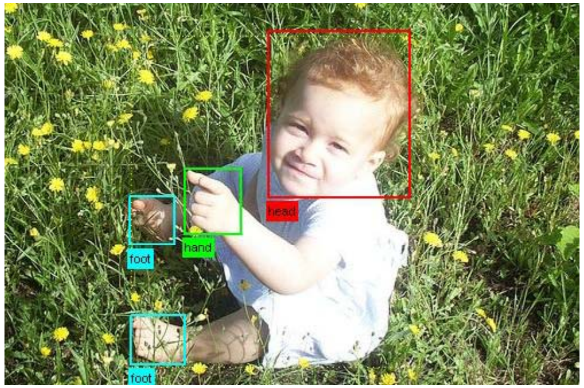
Image parsing: learning the model for a specific image

Recent advances

Datasets and challenges

Datasets & Evaluation

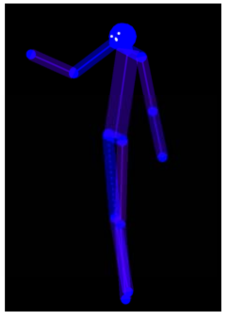
Some efforts evaluating person image parsing



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)

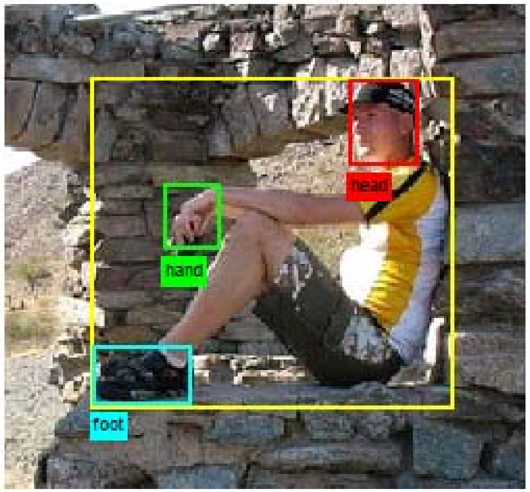
Mark Everingham, Luc Van Gool
Chris Williams, John Winn
Andrew Zisserman

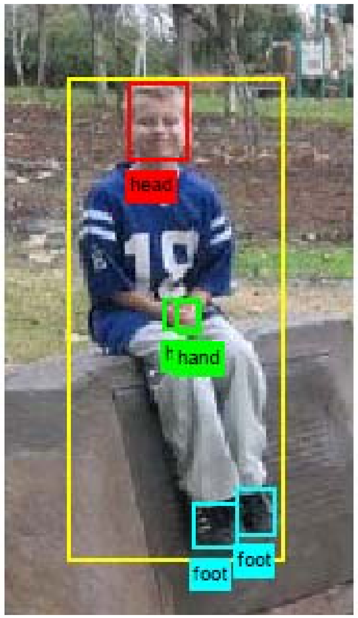
PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

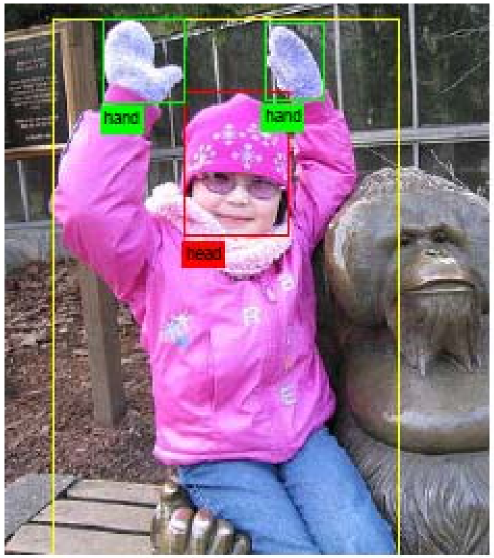
Person Layout Taster

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)





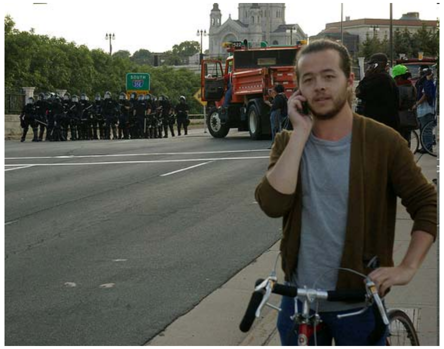


Human Action Classes Taster

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

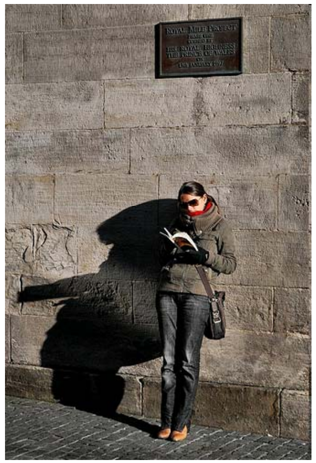
- choice of classes governed by availability from flickr
- evaluation is by AP on each class
- 50-90 training images for each class

phoning

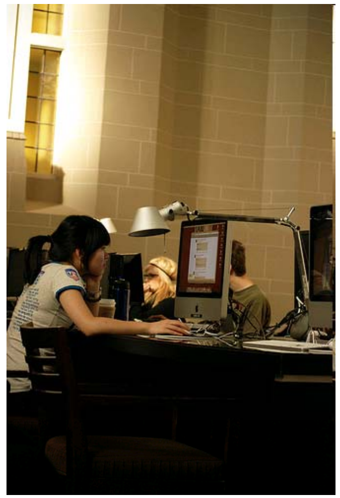


playing instrument

reading



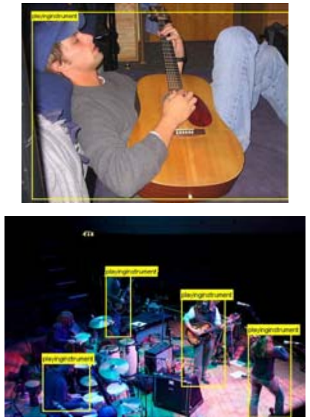
working on computer



Nine Action Classes

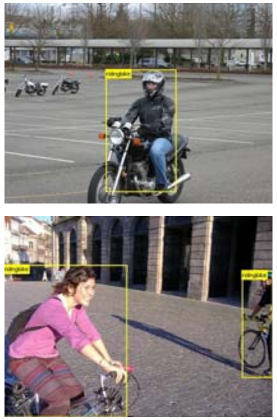
Phoning

Playing Instrument

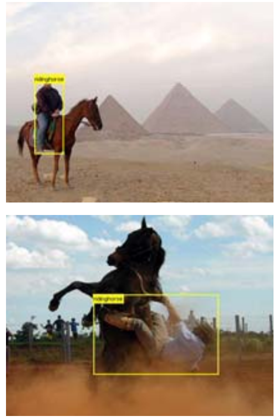


Reading

Riding Bike



Riding Horse

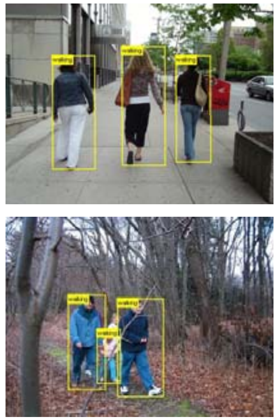


Running

Taking Photo

Using Computer

Walking



Dataset Statistics

Images collected from flickr using action queries

- Disjoint to main challenge dataset

	Training	Testing
Images	454	454
Objects	608	613

- 50-100 training objects per class
- Only subset of people are annotated (bounding box + action)
- All people in dataset are labelled with exactly one action class
  - In future actions will not be mutually exclusive (or complete?)

Methods

Comp9 (Train on VOC data): 11 Methods, 8 Groups

- Image classification within bounding box
  - > SVM, bag of words/spatial pyramid
  - > Multiple features: SIFT, PHOG, color SIFT, etc.
- Context (image, bounding box, neighbouring region)
- Classification of multiple figure-ground segmentations
- Combined image classification and part-based detection

Comp10 (Train on own data): 1 Method

- Poselets, object context

AP by Class/Method

Comp9 results

	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking
BONN_ACTION	47.5	51.1	31.9	64.5	69.1	78.5	32.4	53.9	61.1
CVC_BASE	56.2	56.5	34.7	75.1	83.6	86.5	25.4	60.0	69.2
CVC_SEL	49.8	52.8	34.3	74.2	85.5	85.1	24.9	64.1	72.5
INRIA_SPM_HT	53.2	53.6	30.2	78.2	88.4	84.6	30.4	60.9	61.8
NUDT_SVM_WHGO_SIFT_CENTRIST_LLM	47.2	47.9	24.5	74.2	81.0	79.5	24.9	58.6	71.5
SURREY_MK_KDA	52.6	53.5	35.9	81.0	89.3	86.5	32.8	59.2	68.6
UCLEAR_SVM_DOSP_MULTFEATS	47.0	57.8	26.9	78.8	89.7	87.3	32.5	60.0	70.1
UMCO_DHOG_KSVM	53.5	43.0	32.0	67.9	68.8	83.0	34.1	45.9	60.4
WILLOW_A_SVMSIFT_1-A_LSVM	49.2	37.7	22.2	73.2	77.1	81.7	24.3	53.7	56.9
WILLOW_LSVM	40.4	29.9	32.2	53.5	62.2	73.6	17.6	45.8	41.5
WILLOW_SVMSIFT	47.9	29.1	21.7	53.5	76.7	78.3	26.0	42.9	56.4

(1st, 2nd, 3rd place)

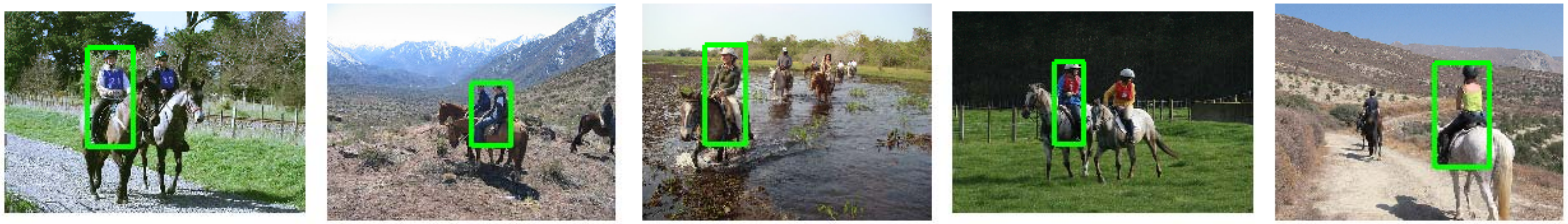
Comp10 results

	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking
BERKELEY_POSELETS_ACTION	45.9	45.8	23.7	79.9	87.6	83.1	26.2	44.9	66.6

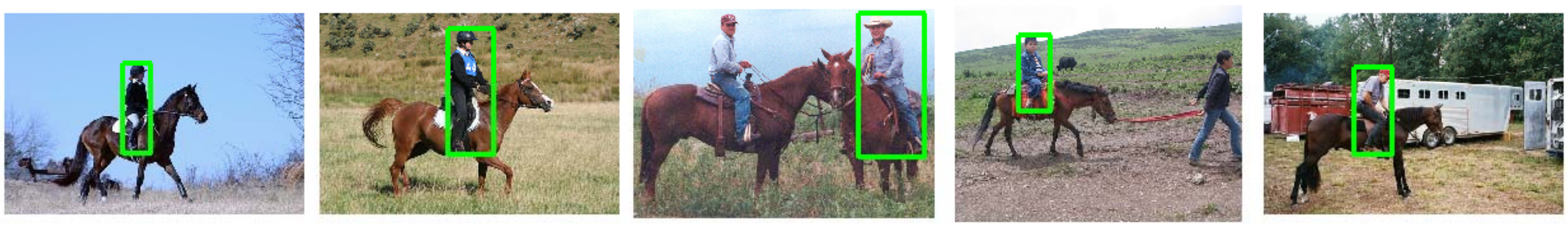
“True Positives”: Riding Horse

UCLEAR_SVM_DOSP_MULTFEATS

SURREY_MK_KDA



INRIA_SPM_HT



“False Negatives”: Riding Horse

UCLEAR_SVM_DOSP_MULTFEATS

SURREY_MK_KDA

INRIA_SPM_HT

“False Positives”: Riding Horse

UCLEAR_SVM_DOSP_MULTFEATS

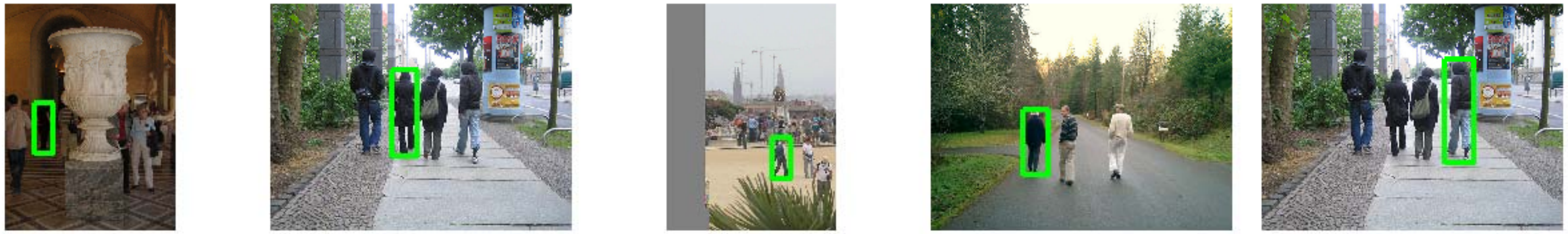
SURREY_MK_KDA

INRIA_SPM_HT

“True Positives”: Walking

CVC_SEL

NUDT_SVM_WHGO_SIFT_CENTRIST_LLM



UCLEAR_SVM_DOSP_MULTFEATS

“False Negatives”: Walking

CVC_SEL

NUDT_SVM_WHGO_SIFT_CENTRIST_LLM

UCLEAR_SVM_DOSP_MULTFEATS

“False Positives”: Walking

CVC_SEL

NUDT_SVM_WHGO_SIFT_CENTRIST_LLM

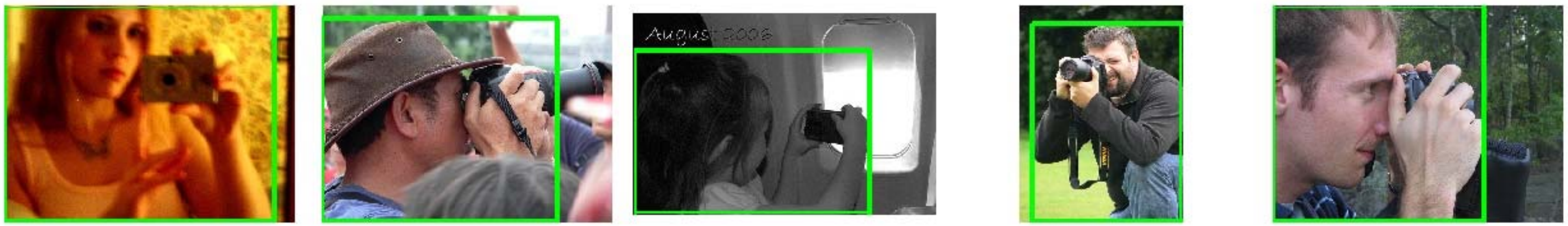
UCLEAR_SVM_DOSP_MULTFEATS

“True Positives”: Taking Photo

UMCO_DHOG_KSVM

SURREY_MK_KDA

UCLEAR_SVM_DOSP_MULTFEATS



“False Negatives”: Taking Photo

UMCO_DHOG_KSVM

SURREY_MK_KDA

UCLEAR_SVM_DOSP_MULTFEATS

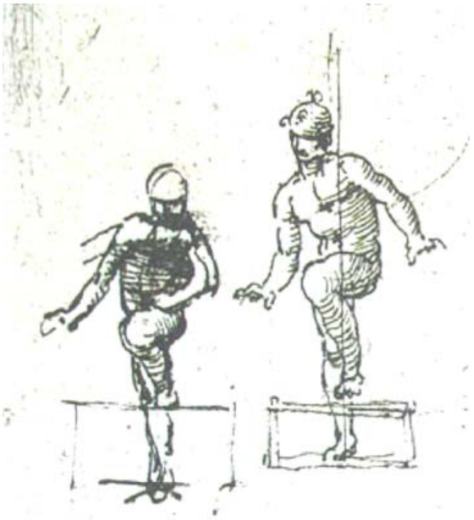
“False Positives”: Taking Photo

UMCO_DHOG_KSVM

SURREY_MK_KDA

UCLEAR_SVM_DOSP_MULTFEATS

Class overview



Motivation

Historic review
Modern applications

Human Pose Estimation

Pictorial structures
Learning models from image data
Recent advances
Datasets and challenges

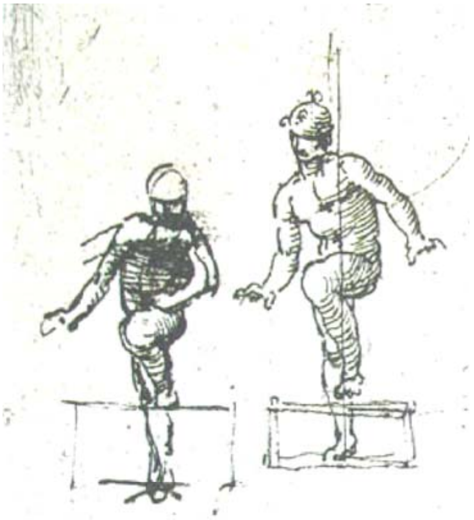
Appearance-based methods

Motion history images
Active shape models
Motion priors

Motion-based methods

Generic and parametric Optical Flow
Motion templates

Class overview



Motivation

Historic review
Modern applications

Human Pose Estimation

Pictorial structures
Learning models from image data
Recent advances
Datasets and challenges

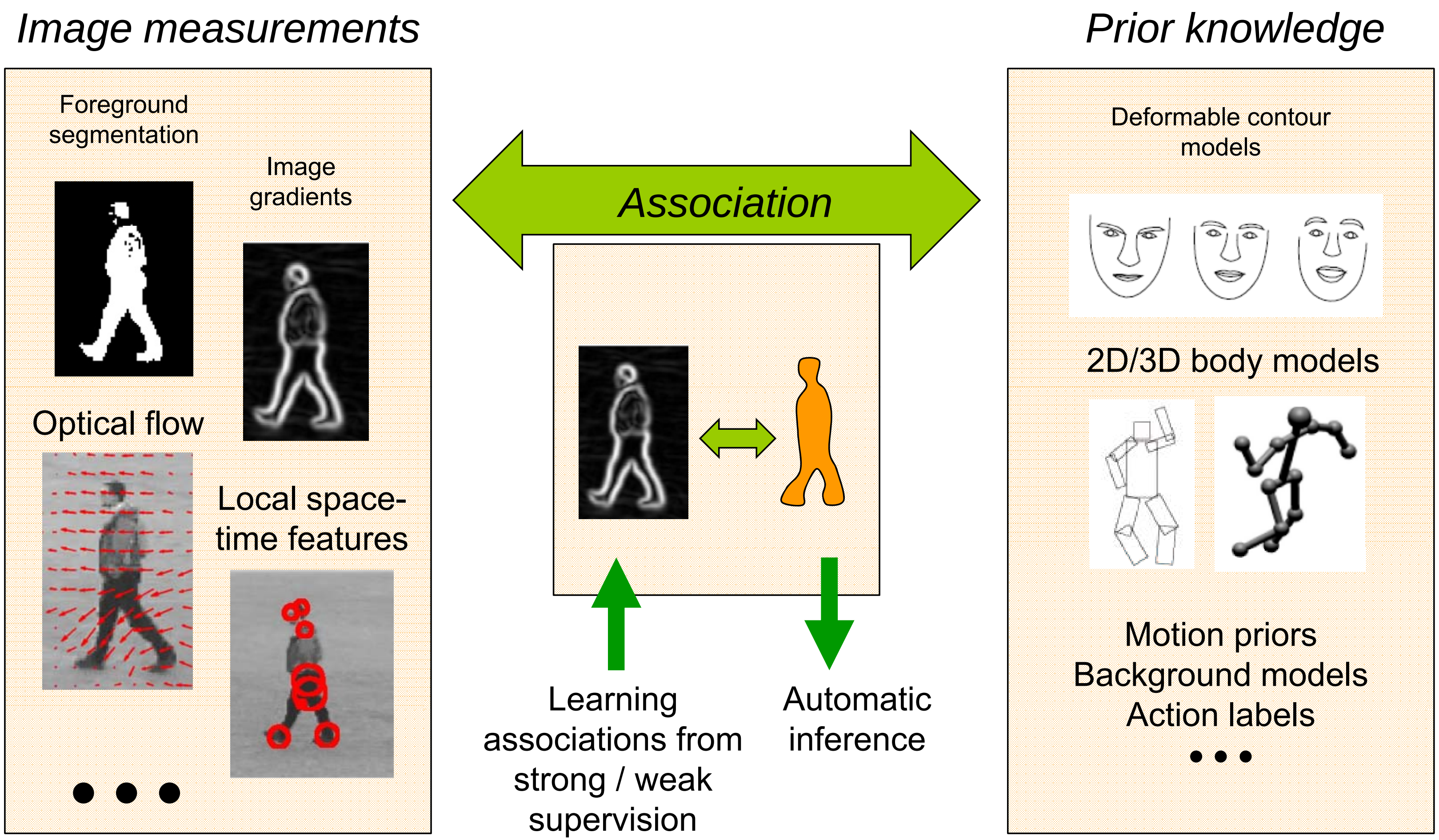
Appearance-based methods

Motion history images
Active shape models
Motion priors

Motion-based methods

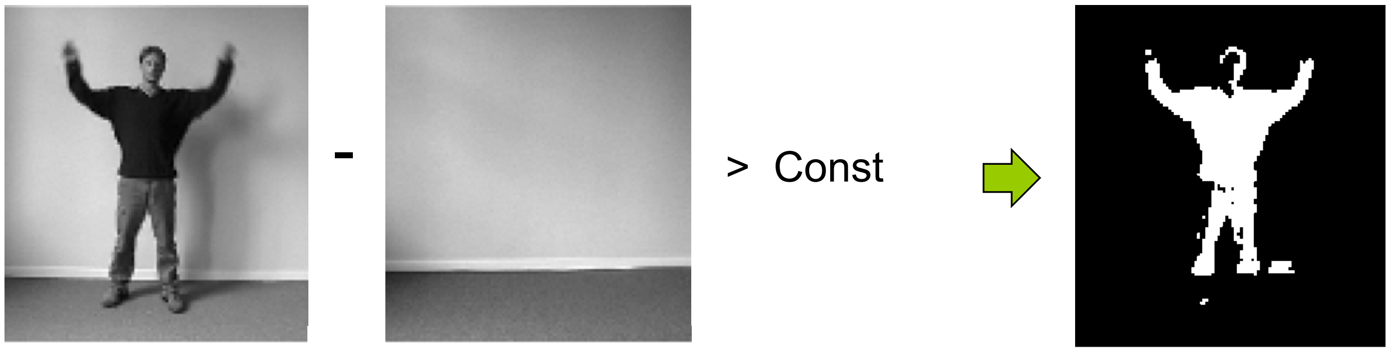
Generic and parametric Optical Flow
Motion templates

Action understanding: Key components



Foreground segmentation

Image differencing: a simple way to measure motion/change

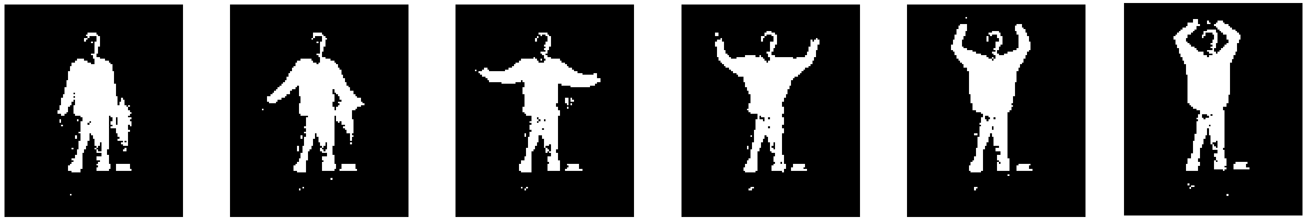


Better Background / Foreground separation methods exist:

- Modeling of color variation at each pixel with Gaussian Mixture
- Dominant motion compensation for sequences with moving camera
- Motion layer separation for scenes with non-static backgrounds

Temporal Templates

$D(x, y, t) \quad t = 1, \ldots, T$



Idea: summarize motion in video in a
Motion History Image (MHI):

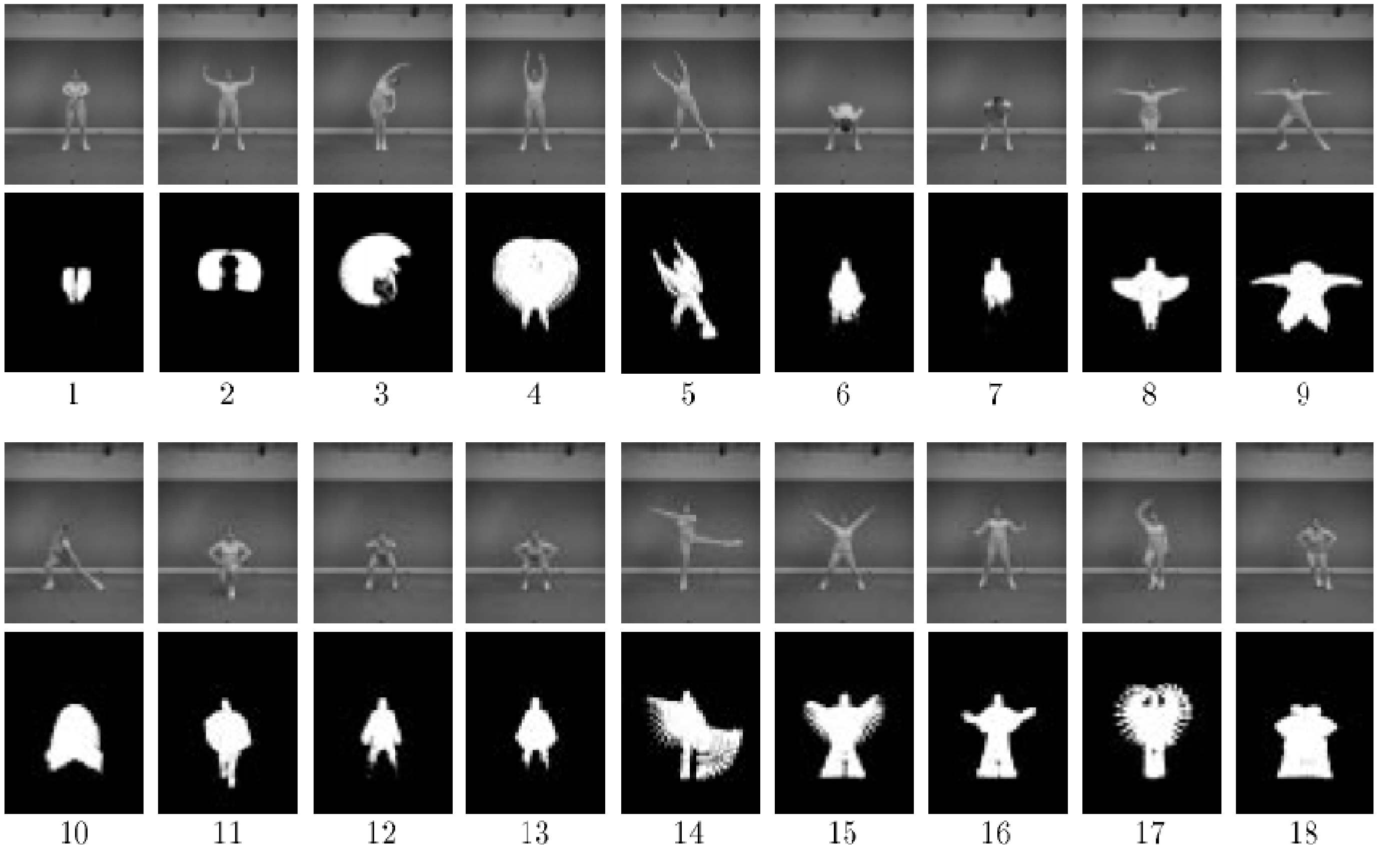
$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max\ (0, H_\tau(x, y, t-1) - 1) & \\ \quad \text{otherwise} & \end{cases}$$

Descriptor: Hu moments of different orders

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q \rho(x, y)\, dx\, dy$$

[A.F. Bobick and J.W. Davis, PAMI 2001]

Aerobics dataset



Nearest Neighbor classifier: 66% accuracy

Temporal Templates: Summary

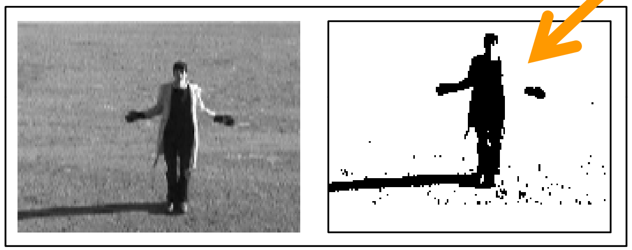
Pros:

+ Simple and fast
+ Works in controlled settings

Not all shapes are valid
➡ Restrict the space
of admissible silhouettes

Cons:

- Prone to errors of background subtraction



Variations in light, shadows, clothing…

What is the background here?

- Does not capture *interior* motion and shape

Silhouette
tells little
about actions

Active Shape Models of Cootes et al.

Point Distribution Model

- Represent the shape of samples by a set of corresponding points or *landmarks*

$$\mathbf{x} = (x_1, \dots, x_n, y_1, \dots, y_n)^T$$

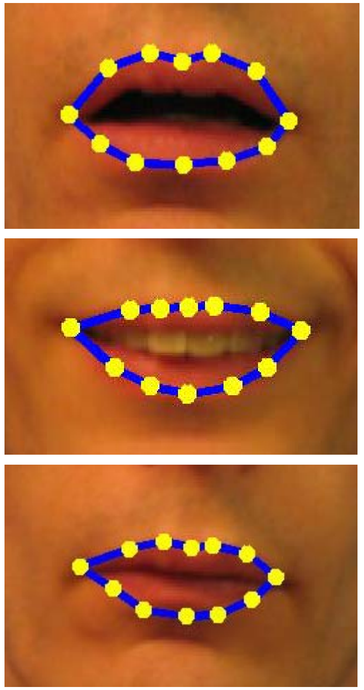
- Assume each shape can be represented by the linear combination of basis shapes

$$\mathbf{\Phi} = (\phi_1 | \phi_2 | \dots | \phi_t)$$

such that $\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b}$

for mean shape $\bar{\mathbf{x}} = \frac{1}{s} \sum_{i=1}^{s} \mathbf{x}_i$

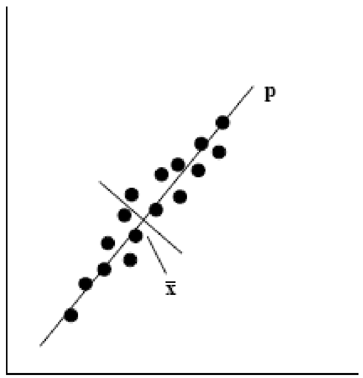
and some parameters $\mathbf{b}$

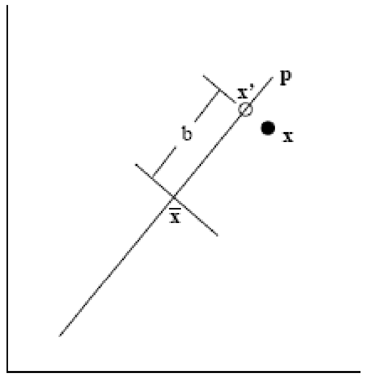


Active Shape Models of Cootes et al.

- Basis shapes can be found as the main modes of variation of in the training data.

2D Example:
(each point can be thought as a shape in N-Dim space)





Principle Component Analysis (PCA):

Covariance matrix $\mathbf{S} = \frac{1}{s-1}\sum_{i=1}^{s}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$

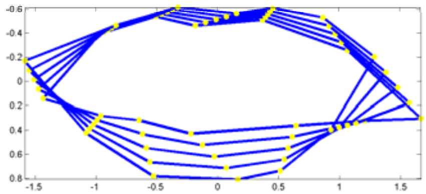
Eigenvectors $\mathbf{\Phi} = (\phi_1|\phi_2|\ldots|\phi_t)$ eigenvalues $\lambda_1, \ldots, \lambda_t$

Active Shape Models of Cootes et al.

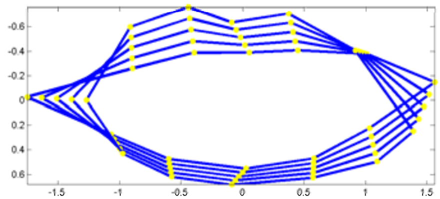
- Back-project from shape-space $\mathbf{b}$ to image space $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b}$

  ➡ Three main modes of lips-shape variation:

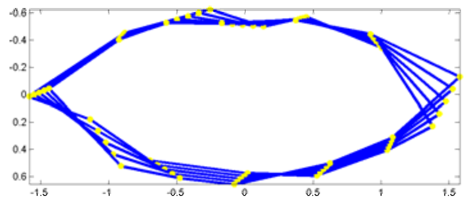
$\mathbf{b} = (\mu\lambda_1, 0, 0, ...)^\top$



$\mathbf{b} = (0, \mu\lambda_2, 0, 0, ...)^\top$

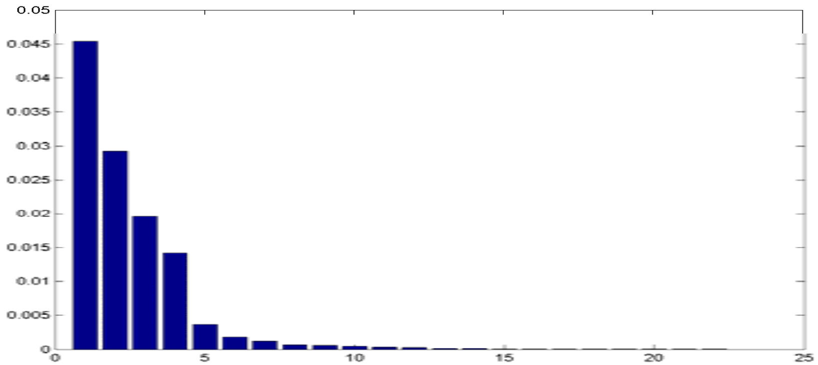


$\mathbf{b} = (0, 0, \mu\lambda_3, 0, 0, ...)^\top$



$\mu = -3, 1.5, 0, 1.5, 3$

Distribution of eigenvalues: $\lambda_1, \lambda_2, \lambda_3, ...$



A small fraction of basis shapes (eigenvecors) accounts for the most of shape variation (=> landmarks are redundant)

Active Shape Models of Cootes et al.

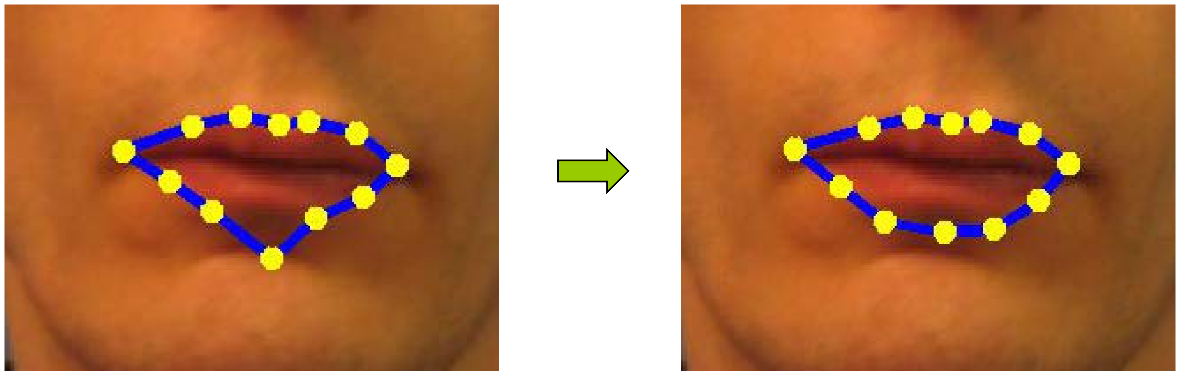
- $\mathbf{\Phi}$ is orthonormal basis, therefore $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^{\top}$

$\Rightarrow$ Given estimate of $\mathbf{x}$ we can recover shape parameters $\mathbf{b}$

$$\mathbf{b} = \mathbf{\Phi}^{\top}(\mathbf{x} - \bar{\mathbf{x}})$$

- Projection onto the shape-space serves as a *regularization*

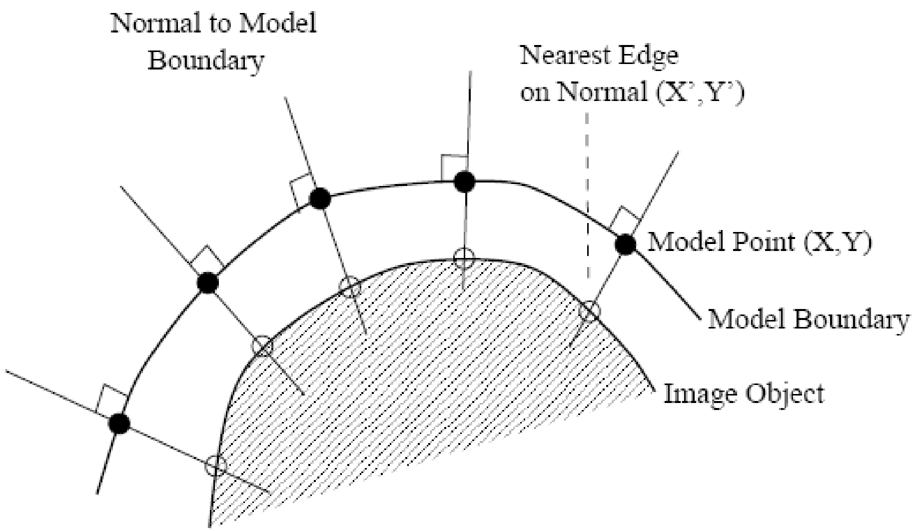
$$\mathbf{x} \quad \Rightarrow \quad \mathbf{b} = \mathbf{\Phi}^{\top}(\mathbf{x} - \bar{\mathbf{x}}) \quad \Rightarrow \quad \mathbf{x}_{\mathbf{reg}} = \bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b}$$



Active Shape Models of Cootes et al.

How to use Active Shape Models for shape estimation?

- Given initial guess of model points $\mathbf{x}$ estimate new positions $\mathbf{x}'$ using local image search, e.g. locate the closest edge point



- Re-estimate shape parameters

$$\mathbf{b}' = \mathbf{\Phi}^\top (\mathbf{x}' - \bar{\mathbf{x}})$$

Active Shape Models of Cootes et al.

- To handle translation, scale and rotation, it is useful to normalize $\mathbf{x}$ prior to shape estimation:

  $$\mathbf{x} = \mathbf{T}(\bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b})$$

  using similarity transformation

  $$\mathbf{T}(\mathbf{x}_{\text{norm}}) = \begin{pmatrix} a & c \\ -c & a \end{pmatrix} \mathbf{x} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

  A simple way to estimate $\mathbf{T}$ is to assign (t_x, t_y) and a to the mean position and the standard deviation of points in $\mathbf{x}$ respectively and set $c = 0$. For more sophisticated normalization techniques see:

  http://www.isbe.man.ac.uk/~bim/Models/app_model.ps.gz

  Note: model parameters $\bar{\mathbf{x}}$, $\mathbf{\Phi}$ have to be computed using *normalized* image point coordinates $\mathbf{x}_{\text{norm}} = T^{-1}(\mathbf{x})$

Active Shape Models of Cootes et al.

- Iterative ASM alignment algorithm
    1. Initialize with the reasonable guess of $\mathbf{T}$ and $\mathbf{b} = \mathbf{0}^\top$
    2. Estimate $\mathbf{x}'$ from image measurements
    3. Re-estimate $\mathbf{T}, \mathbf{b}$
    4. Unless $\mathbf{T}, \mathbf{b}$ converged, repeat from step 2

Example: face alignment

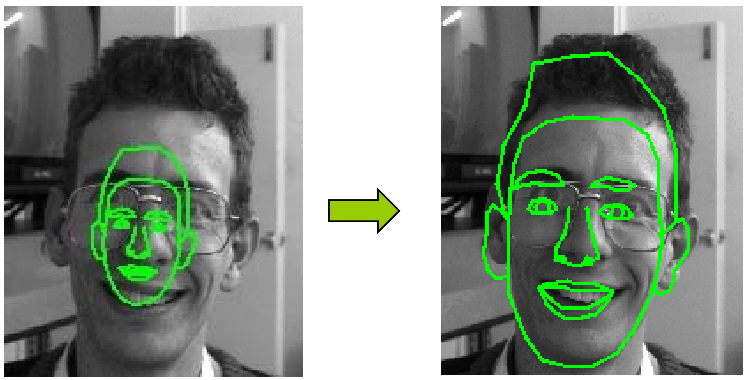
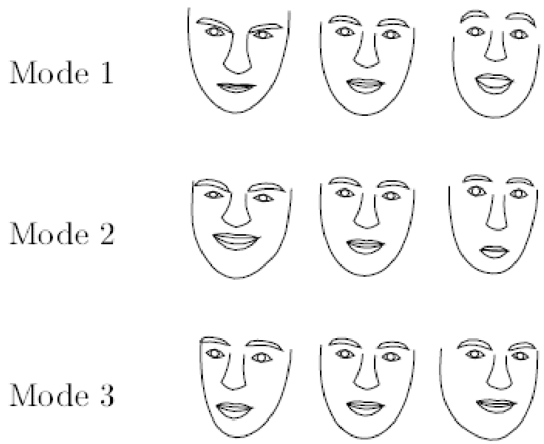


Illustration of face shape space



Active Shape Models: Their Training and Application
T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, **CVIU** 1995

Active Shape Model tracking

Aim: to track ASM of time-varying shapes, e.g. human silhouettes

- Impose time-continuity constraint on model parameters.
  For example, for shape parameters $\mathbf{b}$:

$$b_i^{(k)} = b_i(k-1) + w_i^{k-1}$$

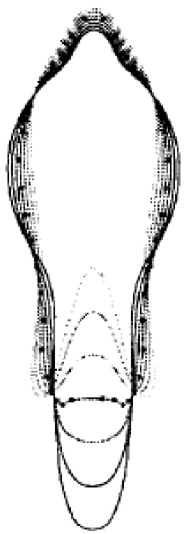
$$w_i \sim \mathcal{N}(0, \mu\lambda_i) \quad \text{Gaussian noise}$$

  For similarity transformation $\mathbf{T}$

$$a^{(k)} = a^{(k-1)} + w_a^{k-1}, \quad w_a = \mathcal{N}(0, \sigma_a)$$

$$t_{x|y}^{(k)} = t_{x|y}^{(k-1)} + v_{x|y}^{(k-1)} + w_{x|y}^{k-1}, \quad w_{x|y} = \mathcal{N}(0, \sigma_{x|y})$$

  More complex dynamical models possible



- Update model parameters at each time frame using e.g. Kalman filter

Person Tracking

Learning flexible models from image sequences
A. Baumberg and D. Hogg, **ECCV** 1994

Person Tracking

Learning flexible models from image sequences
A. Baumberg and D. Hogg, **ECCV** 1994

Active Shape Models: Summary

Pros:

+ Shape prior helps overcoming segmentation errors
+ Fast optimization
+ Can handle interior/exterior dynamics

Cons:

- Optimization gets trapped in local minima
- Re-initialization is problematic

Possible improvements:

- Learn and use motion priors, possibly specific to different actions

Motion priors

- Accurate motion models can be used both to:
  - ❖ Help accurate tracking
  - ❖ Recognize actions

- Goal: formulate motion models for different types of actions and use such models for action recognition

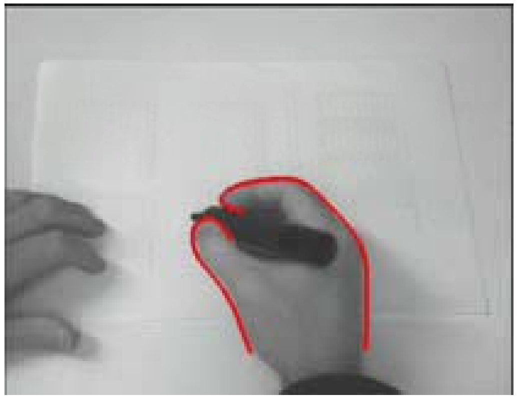
Example:

Drawing with 3 action modes

line drawing

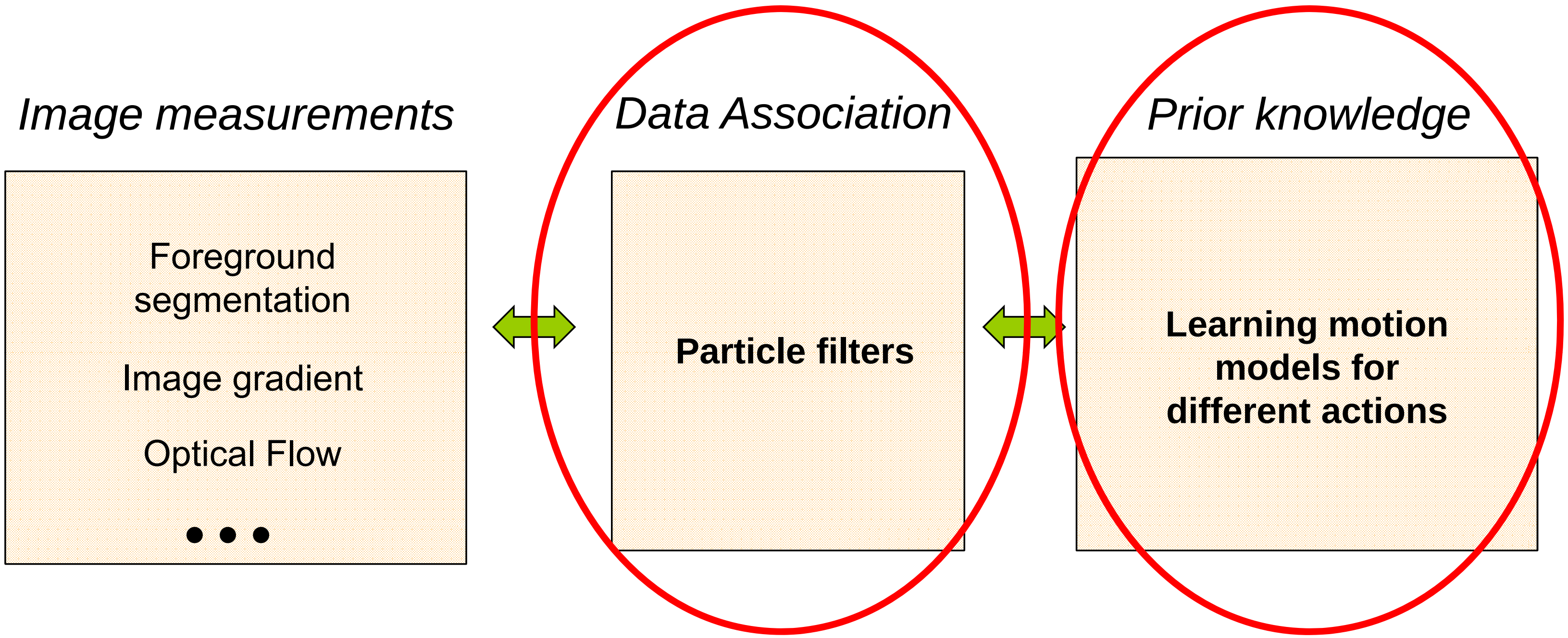
scribbling

idle



[M. Isard and A. Blake, ICCV 1998]

Incorporating motion priors



Bayesian Tracking

General framework: recognition by synthesis;
generative models;
finding best explanation of the data

Notation:

$\mathbf{Z}_i$ image data at time i

$\mathbf{X}_i$ model parameters at time i (e.g. shape and its dynamics)

$p(\mathbf{X}_i)$ prior density for $\mathbf{X}_i$

$p(\mathbf{Z}_i|\mathbf{X}_i)$ likelihood of data for the given model configuration

We search posterior defined by the Bayes' rule

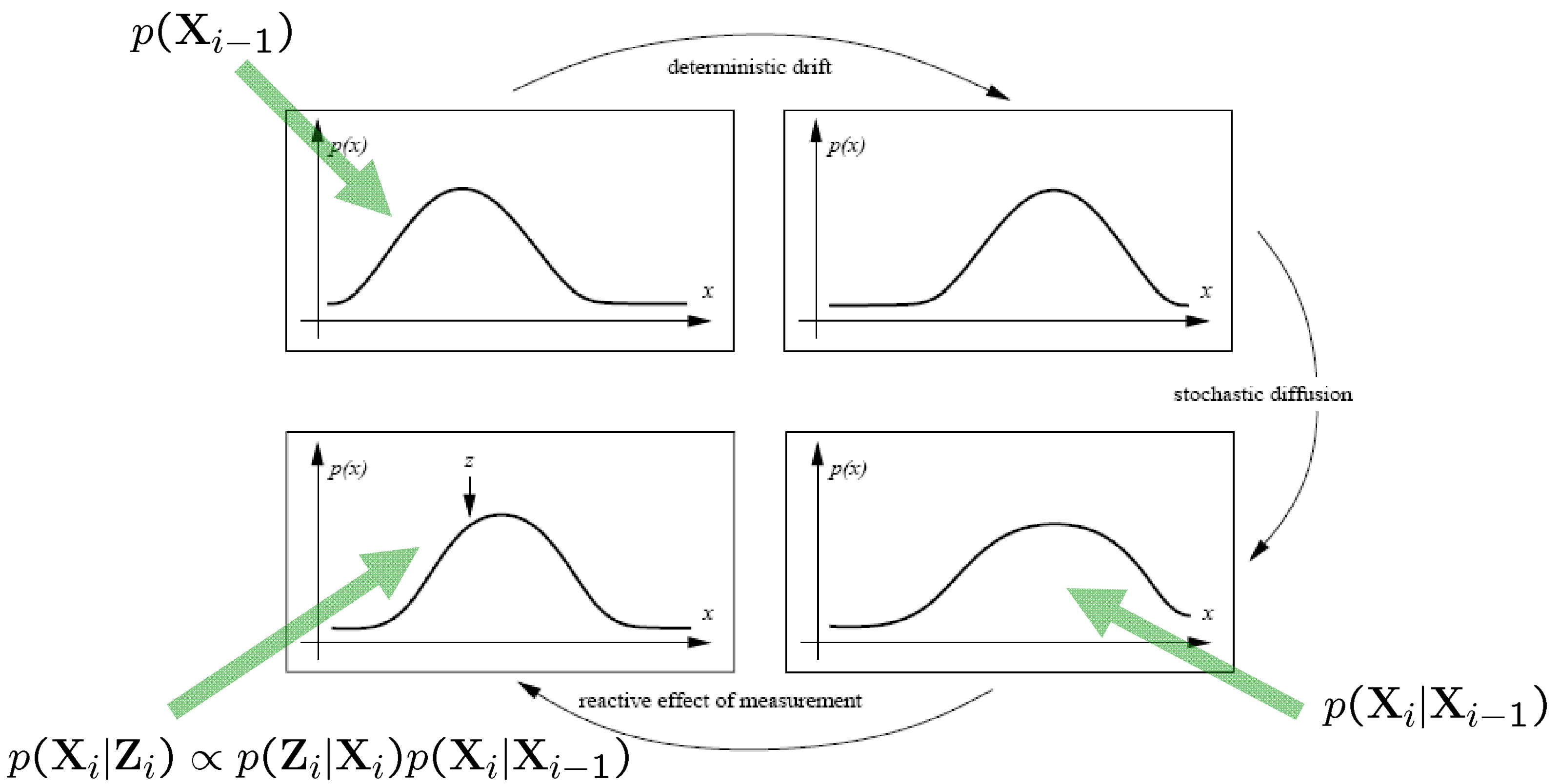
$$p(\mathbf{X}|\mathbf{Z}) \propto \mathbf{p}(\mathbf{Z}|\mathbf{X})\mathbf{p}(\mathbf{X})$$

For tracking the Markov assumption gives the prior $p(\mathbf{X}_i|\mathbf{X}_{i-1})$

Temporal update rule: $p(\mathbf{X}_i|\mathbf{Z}_i) \propto p(\mathbf{Z}_i|\mathbf{X}_i)p(\mathbf{X}_i|\mathbf{X}_{i-1})$

Kalman Filtering

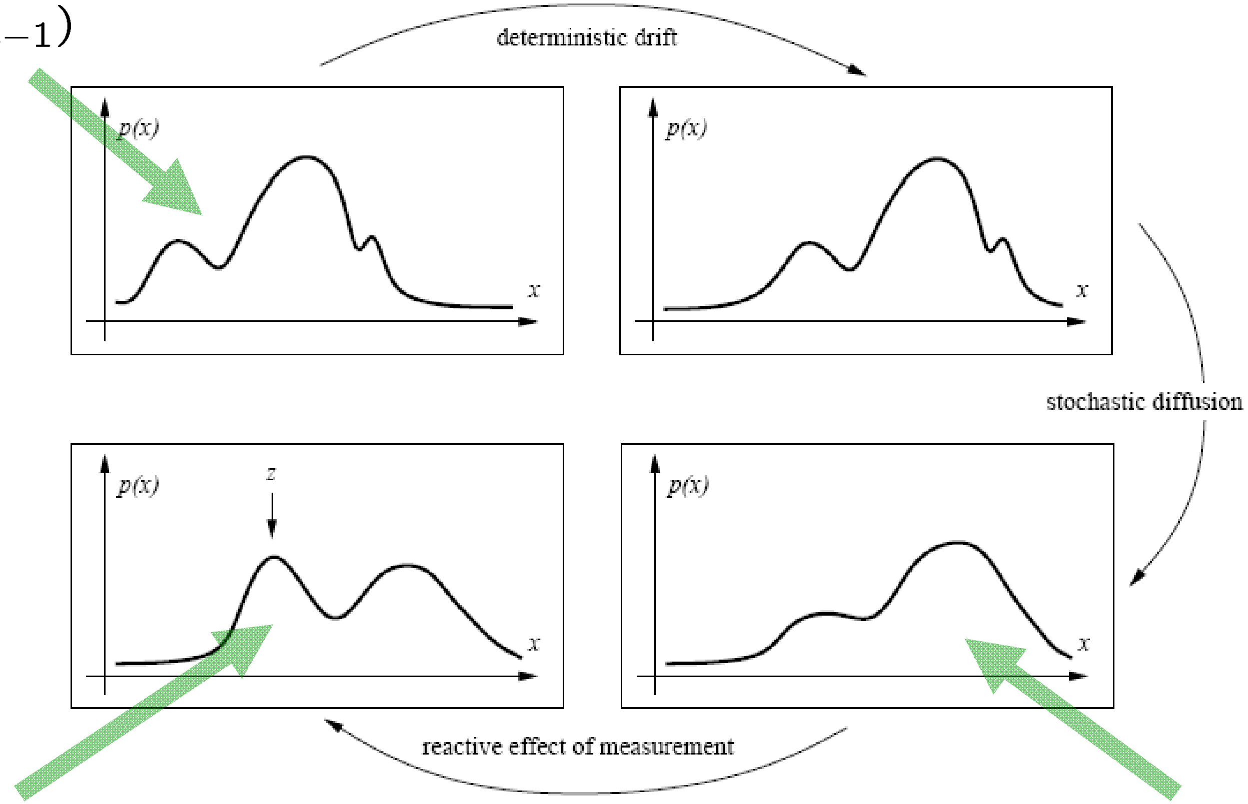
If all probability densities are uni-modal, specifically Gussians, the posterior can be evaluated in the closed form



Particle Filtering

In reality probability densities are almost always *multi-modal*

$p(\mathbf{X}_{i-1})$

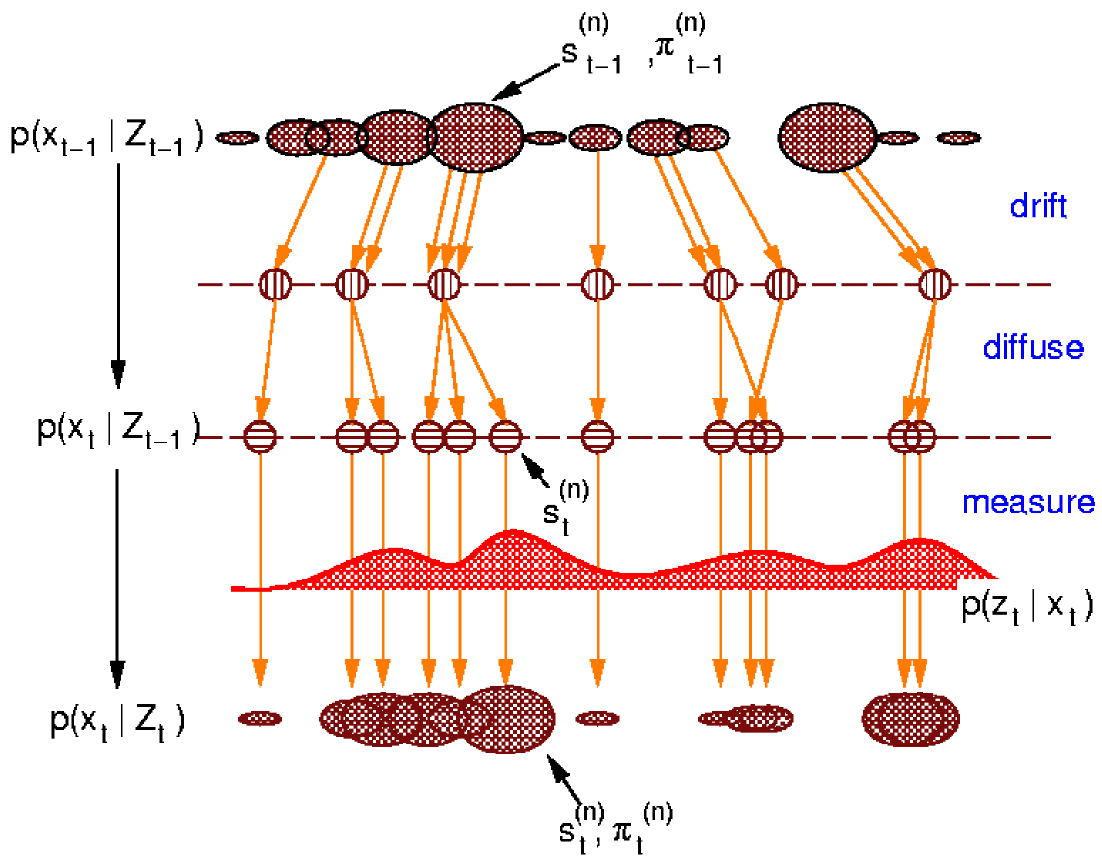


$p(\mathbf{X}_i|\mathbf{Z}_i) \propto p(\mathbf{Z}_i|\mathbf{X}_i)p(\mathbf{X}_i|\mathbf{X}_{i-1})$ $\qquad$ $p(\mathbf{X}_i|\mathbf{X}_{i-1})$

Particle Filtering

In reality probability densities are almost always *multi-modal*

➡ Approximate distributions with weighted particles



Particle Filtering

Tracking examples:

$\mathbf{x}$ describes leave shape

$\mathbf{x}$ describes head shape

CONDENSATION - conditional density propagation for visual tracking
A. Blake and M. Isard **IJCV** 1998

Learning dynamic prior

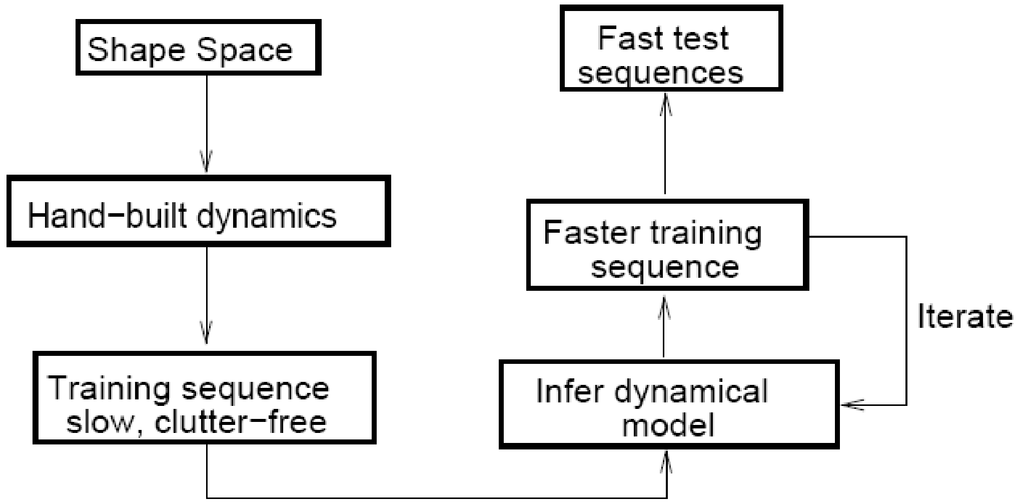
- Dynamic model: 2nd order Auto-Regressive Process

State $\mathcal{X}_k = \begin{pmatrix} \mathbf{X}_{k-1} \\ \mathbf{X}_k \end{pmatrix}$

Update rule: $\mathcal{X}_k - \overline{\mathcal{X}} = A(\mathcal{X}_{k-1} - \overline{\mathcal{X}}) + B\mathbf{w}_k$

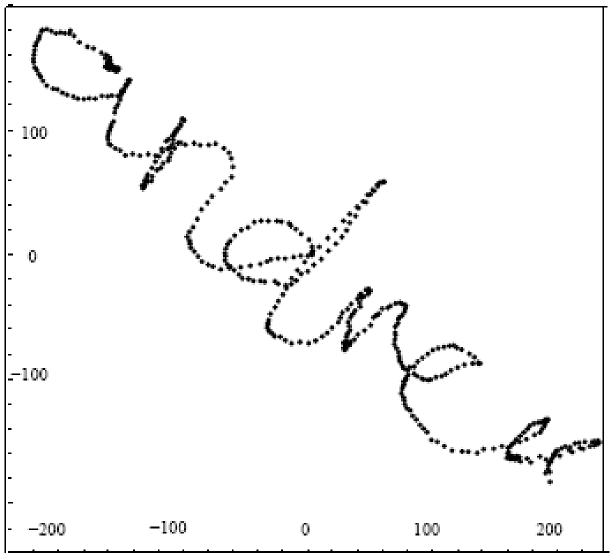
Model parameters: $A = \begin{pmatrix} 0 & I \\ A_2 & A_1 \end{pmatrix}, \quad \overline{\mathcal{X}} = \begin{pmatrix} \overline{\mathbf{X}} \\ \overline{\mathbf{X}} \end{pmatrix}$ and $B = \begin{pmatrix} 0 \\ B_0 \end{pmatrix}$

Learning scheme:

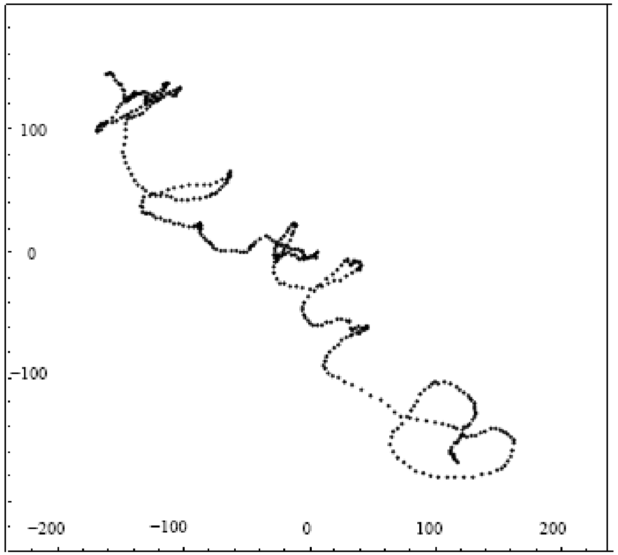


Learning dynamic prior

Learning point sequence



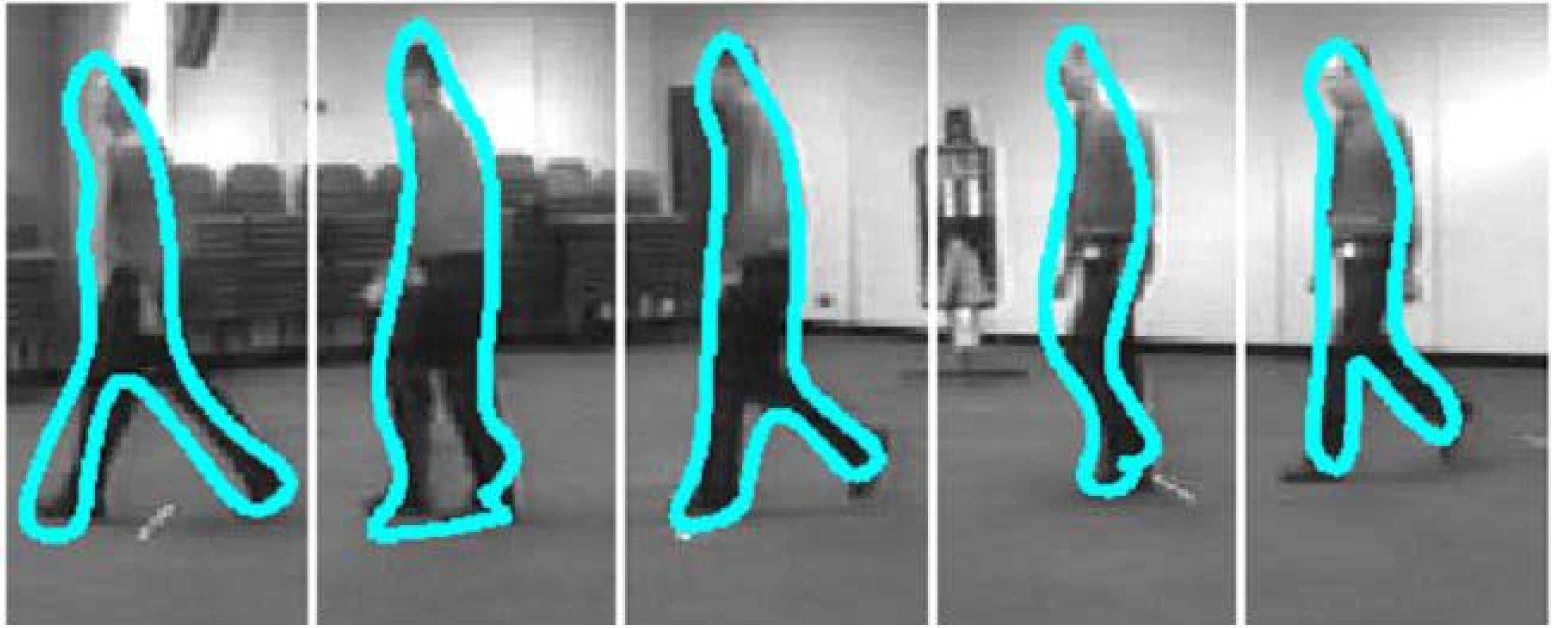
Random simulation of the learned dynamical model



Statistical models of visual shape and motion
A. Blake, B. Bascle, M. Isard and J. MacCormick, **Phil.Trans.R.Soc. 1998**

Learning dynamic prior

Random simulation of the learned gate dynamics



Dynamics with discrete states

Introduce "mixed" state $\mathcal{X}_k^+ = \begin{pmatrix} \mathcal{X}_k \\ y_k \end{pmatrix}$

Continuous state space (as before)

Discrete variable identifying dynamical model $y_k = 1, 2, ..., n$

Transition probability matrix

$$P(y_k = j | y_{k-1} = i) = T_{i,j},$$

or more generally $P(y_k = j | y_{k-1} = i, \mathcal{X}_{k-1}) = T_{i,j}(\mathcal{X}_{k-1})$

Incorporation of the mixed-state model into a particle filter is straightforward, simply use $\mathcal{X}_k^+$ instead of $\mathcal{X}_k$ and the corresponding update rules

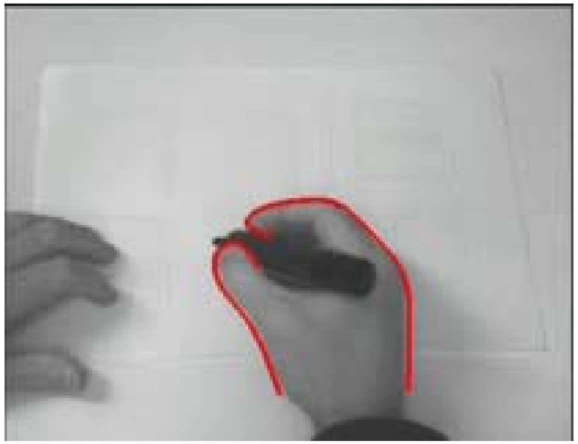
Dynamics with discrete states

Example: Drawing

Transition probability matrix

$$T = \begin{array}{cc} & \begin{array}{ccc} \text{line} & \text{idle} & \text{scribbling} \end{array} \\ & \begin{pmatrix} 0.9800 & 0.0015 & 0.0185 \\ 0.0850 & 0.9000 & 0.0150 \\ 0.0050 & 0.0150 & 0.9800 \end{pmatrix} \begin{array}{l} \text{line} \\ \text{idle} \\ \text{scribbling} \end{array} \end{array}$$

Result: simultaneously improved tracking and gesture recognition



line drawing

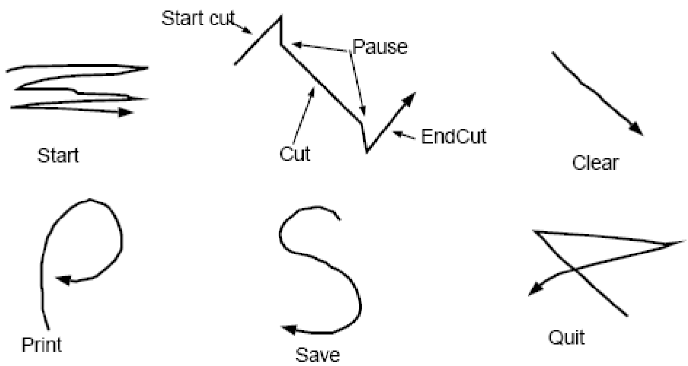
scribbling

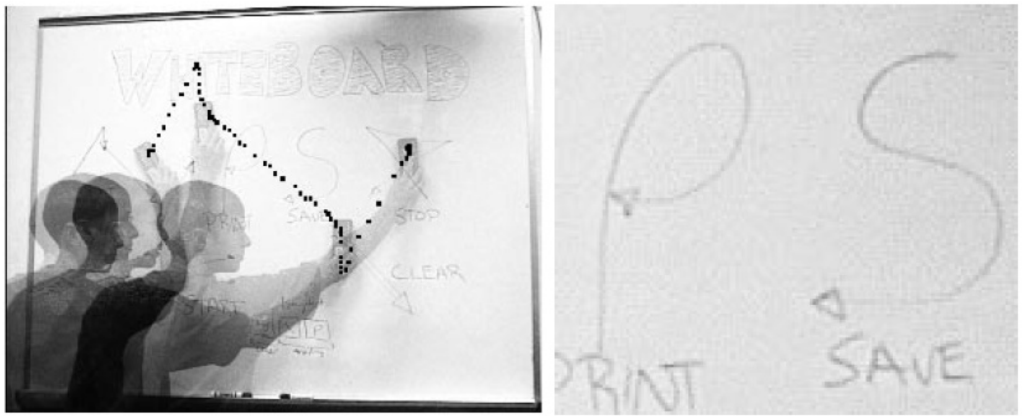
idle

A mixed-state Condensation tracker with automatic model-switching
M. Isard and A. Blake, **ICCV** 1998

Dynamics with discrete states

Similar illustrated on gesture recognition in the context of a visual black-board interface





[M.J. Black and A.D. Jepson, ECCV 1998]

Motion priors & Trackimg: Summary

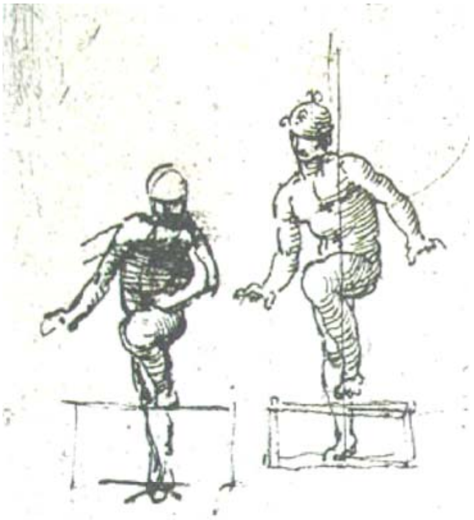
Pros:

+ more accurate tracking using specific motion models
+ Simultaneous tracking and motion recognition with discrete state dynamical models

Cons:

- Local minima is still an issue
- Re-initialization is still an issue

Class overview



Motivation

Historic review
Modern applications

Human Pose Estimation

Pictorial structures
Learning models from image data
Recent advances
Datasets and challenges

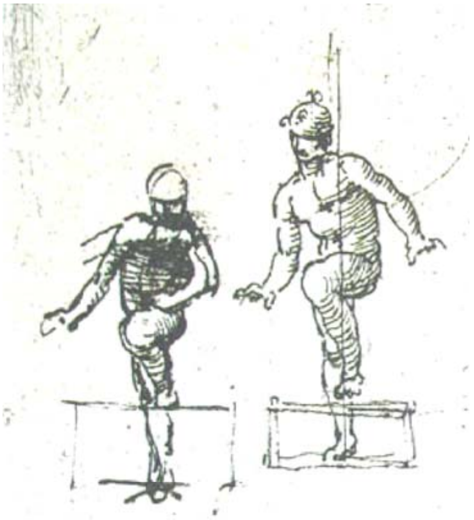
Appearance-based methods

Motion history images
Active shape models
Motion priors

Motion-based methods

Generic and parametric Optical Flow
Motion templates

Class overview



Motivation

Historic review
Modern applications

Human Pose Estimation

Pictorial structures
Learning models from image data
Recent advances
Datasets and challenges

Appearance-based methods

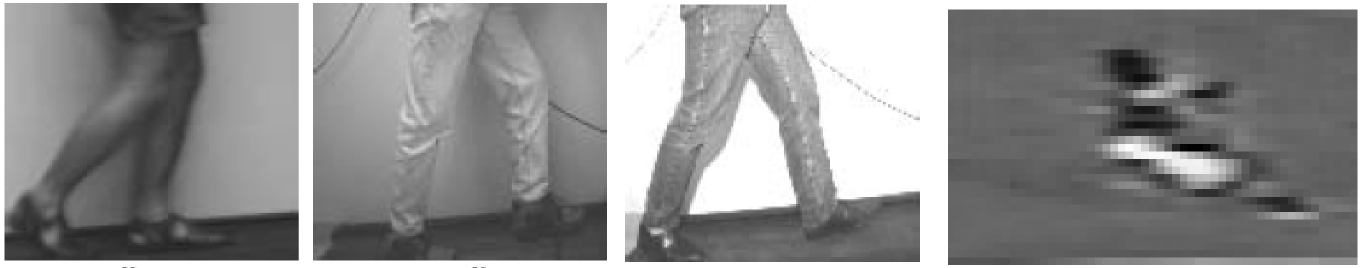
Motion history images
Active shape models
Motion priors

Motion-based methods

Generic and parametric Optical Flow
Motion templates

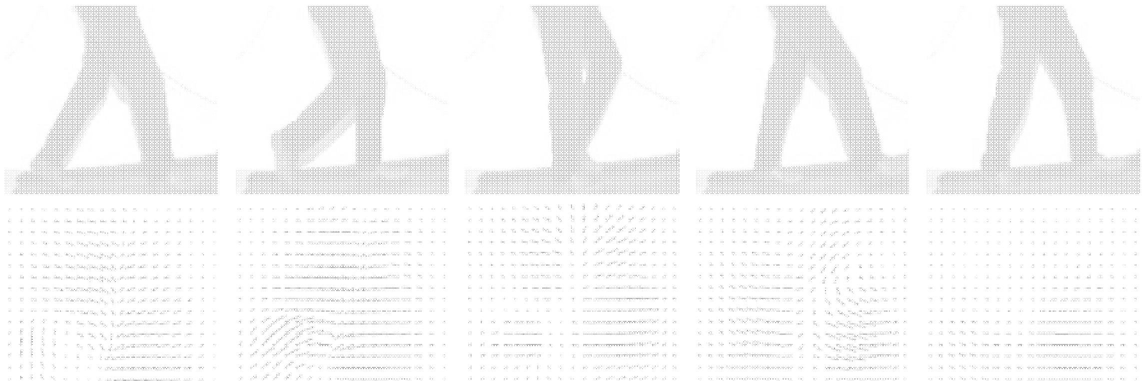
Shape and Appearance vs. Motion

- Shape and appearance in images depends on many factors: clothing, illumination contrast, image resolution, etc…



[Efros et al. 2003]

- Motion field (in theory) is invariant to shape and can be used directly to describe human actions

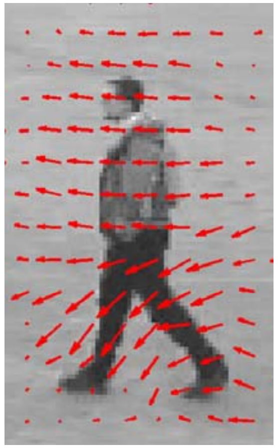


Motion estimation: Optical Flow

- Classic problem of computer vision [Gibson 1955]
- Goal: estimate motion field

  How? We only have access to image pixels

  ⇒ Estimate pixel-wise correspondence between frames = Optical Flow

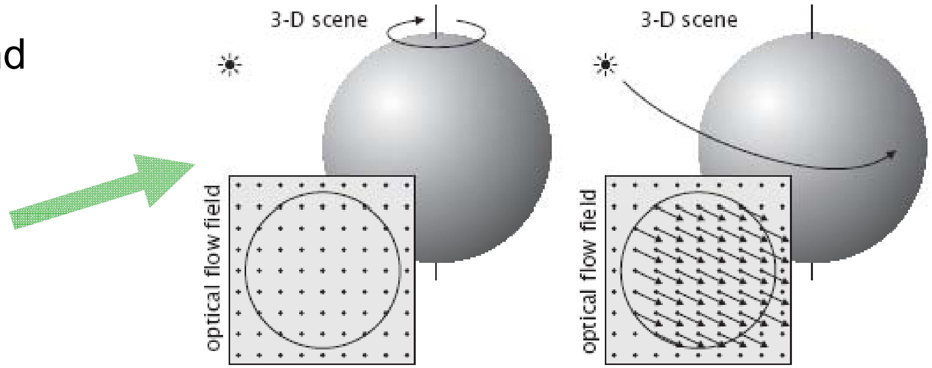


- Brightness Change assumption: corresponding pixels preserve their intensity (color)

  ❖ Useful assumption in many cases

  ❖ Breaks at occlusions and illumination changes

  ❖ Physical and visual motion may be different



Generic Optical Flow

- Brightness Change Constraint Equation (BCCE)

$$(\nabla I)^\top \mathbf{v} + I_t = 0$$

$\mathbf{v} = (v_x, v_y)^\top$ Optical flow

$\nabla I = (I_x, I_y)^\top$ Image gradient

One equation, two unknowns => cannot be solved directly

⇨ Integrate several measurements in the local neighborhood and obtain a *Least Squares Solution* [Lucas & Kanade 1981]

$$< \nabla I (\nabla I)^\top > \mathbf{v} = - < \nabla I I_t >$$

$$\begin{pmatrix} < I_x^2 > & < I_x I_y > \\ < I_x I_y > & < I_y^2 > \end{pmatrix} \mathbf{v} = - \begin{pmatrix} < I_x I_t > \\ < I_y I_t > \end{pmatrix}$$

Second-moment matrix, the same one used to compute Harris interest points!

$< \cdot >$ Denotes integration over a spatial (or spatio-temporal) neighborhood of a point

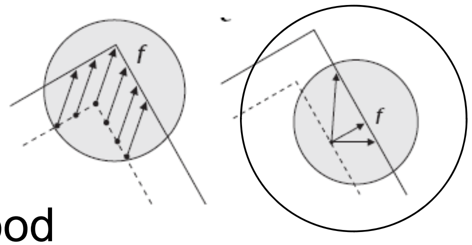
Generic Optical Flow

- The solution of $< \nabla I (\nabla I)^\top > \mathbf{v} = - < \nabla I I_t >$ assumes
  1. Brightness change constraint holds in $< \cdot >$
  2. Sufficient variation of image gradient in $< \cdot >$
  3. Approximately constant motion in $< \cdot >$

  Motion estimation becomes *inaccurate* if any of assumptions 1-3 is violated.

- Solutions:

  (2) Insufficient gradient variation known as *aperture problem*

  ➡ Increase integration neighborhood

  (3) Non-constant motion in $< \cdot >$

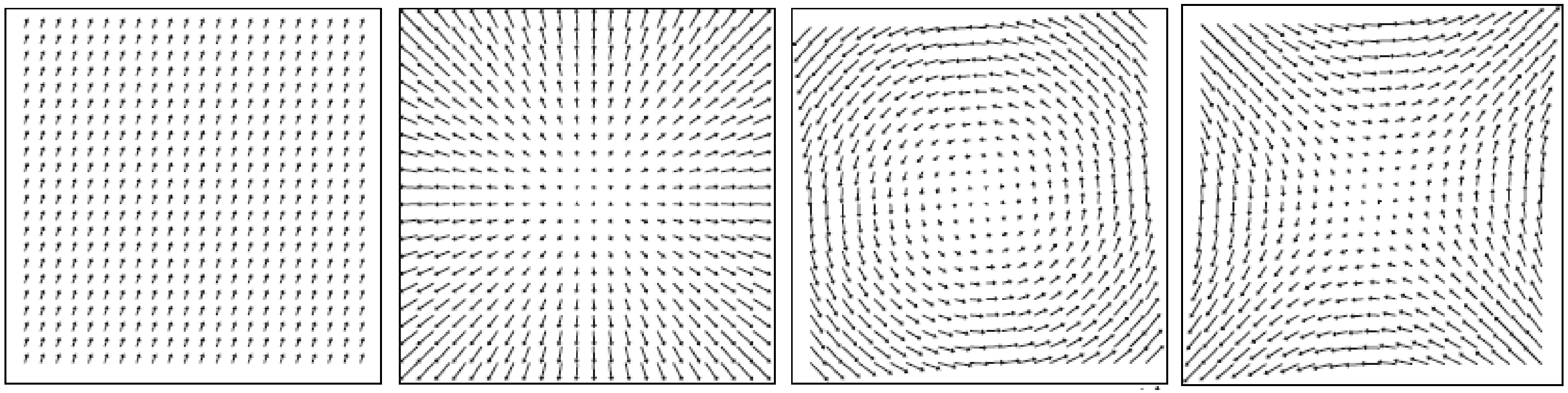
  ➡ Use more sophisticated motion model

Parameterized Optical Flow

- Constant velocity model: $\mathbf{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix}$
- Upgrade to affine motion model: $\mathbf{v} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} v_x \\ v_y \end{pmatrix}$

  Now motion depends on the position $(x, y)^\top$ inside the neighborhood

Examples of Affine motion models for different parameters:



- Can be formulated as Least Squares approach to estimate $\mathbf{v}$ as before!

Parameterized Optical Flow

- Another extension of the constant motion model is to compute PCA basis flow fields from training examples

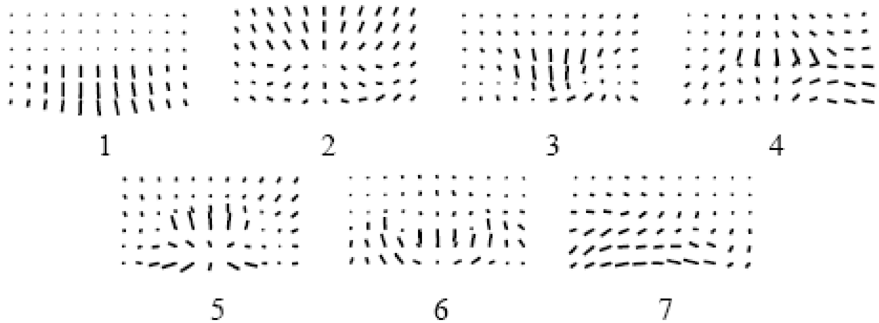
1. Compute standard Optical Flow for many examples
2. Put velocity components into one vector

$$\mathbf{w} = (v_x^1, v_y^1, v_x^2, v_y^2, \ldots, v_x^n, v_y^n)^\top$$

3. Do PCA on $\mathbf{w}$ and obtain most informative PCA flow basis vectors

Training samples

PCA flow bases



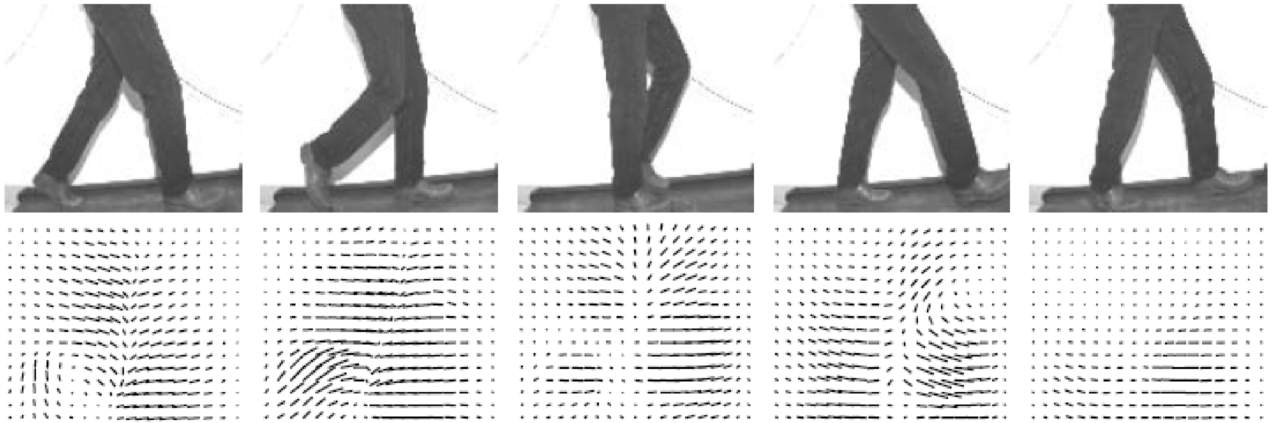
Learning Parameterized Models of Image Motion
M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, **CVPR 1997**

Parameterized Optical Flow

- Use PCA flow bases to *regularize* solution of motion estimation
- Motion estimation for test samples can be computed *without* explicit computation of optical flow!

Solution formulation e.g. in terms of Least Squares

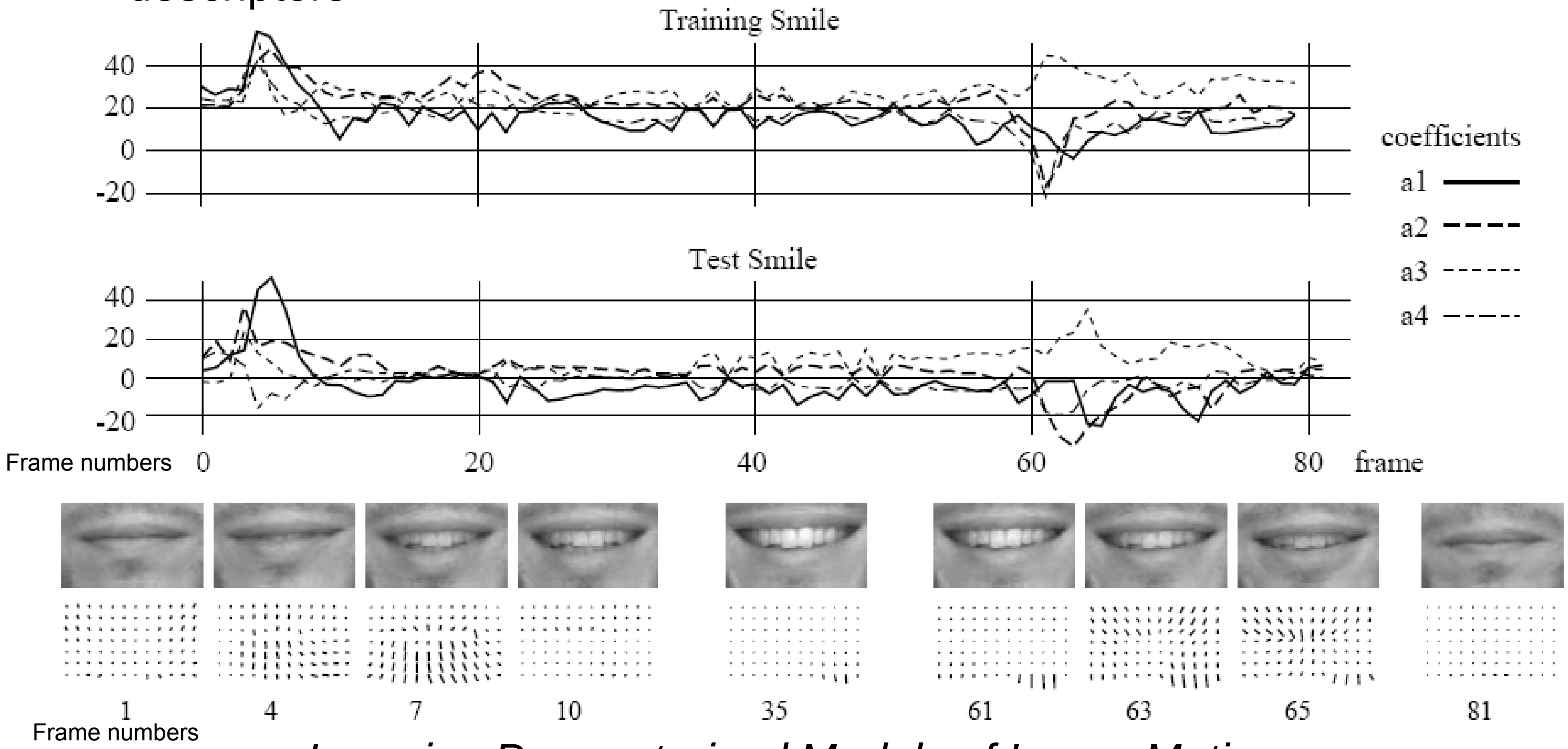
Direct flow recovery:



Learning Parameterized Models of Image Motion
M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, **CVPR 1997**

Parameterized Optical Flow

- Estimated coefficients of PCA flow bases can be used as action descriptors

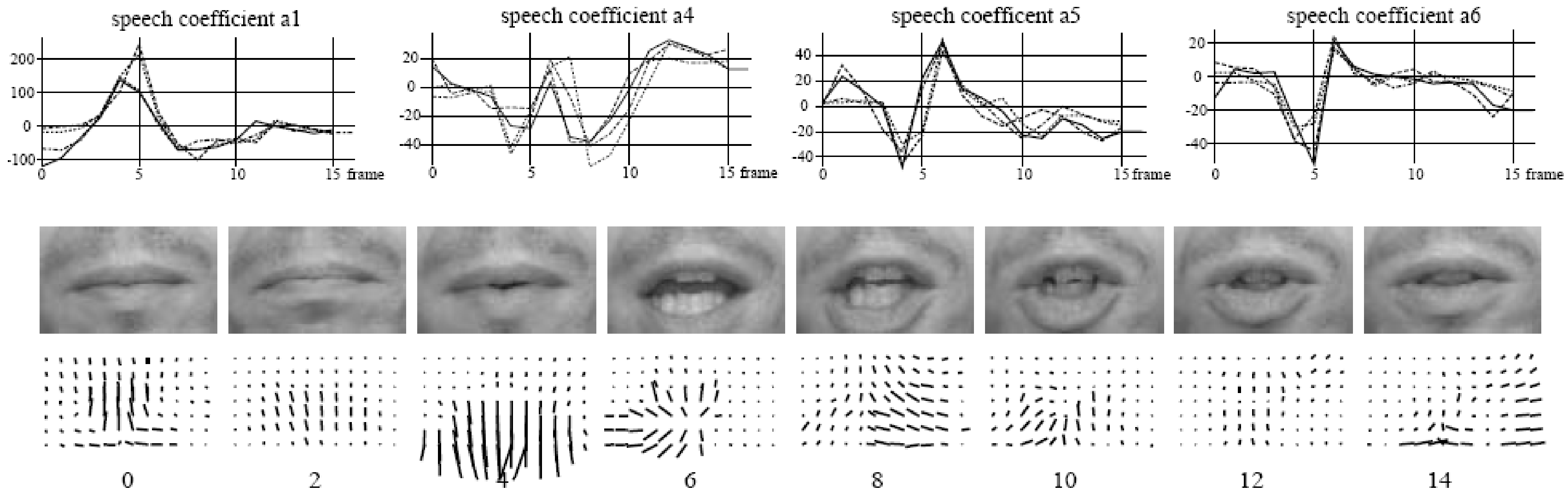


Learning Parameterized Models of Image Motion

M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, **CVPR 1997**

Parameterized Optical Flow

- Estimated coefficients of PCA flow bases can be used as action descriptors



Frame numbers

➡ Optical flow seems to be an interesting descriptor for motion/action recognition

Spatial Motion Descriptor

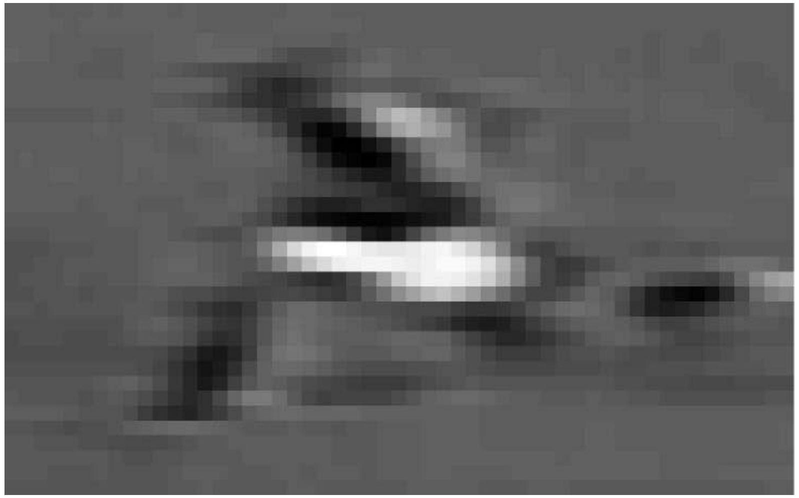
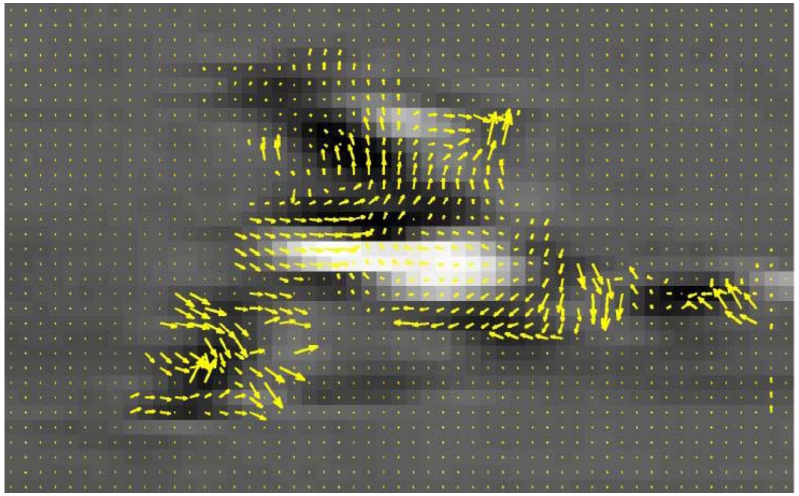
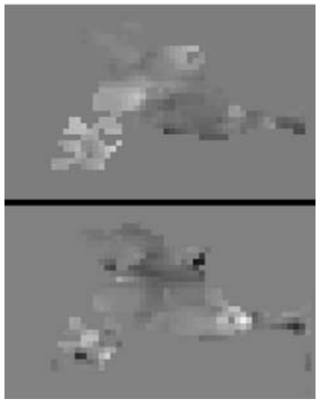


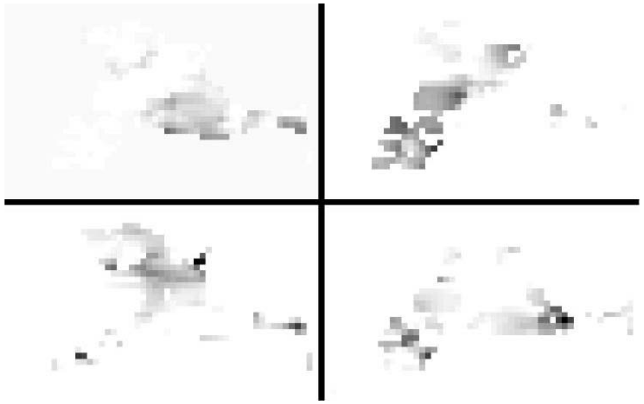
Image frame



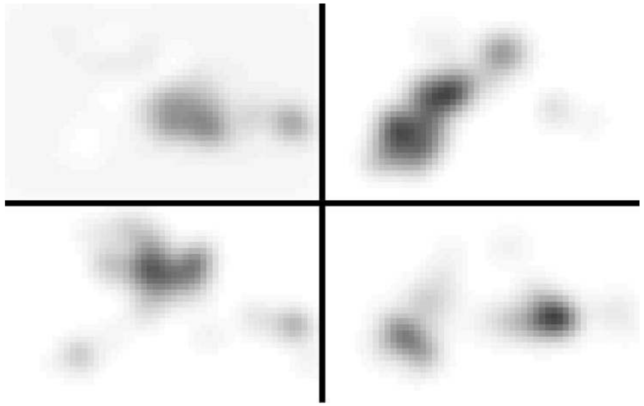
Optical flow $F_{x,y}$



F_x, F_y

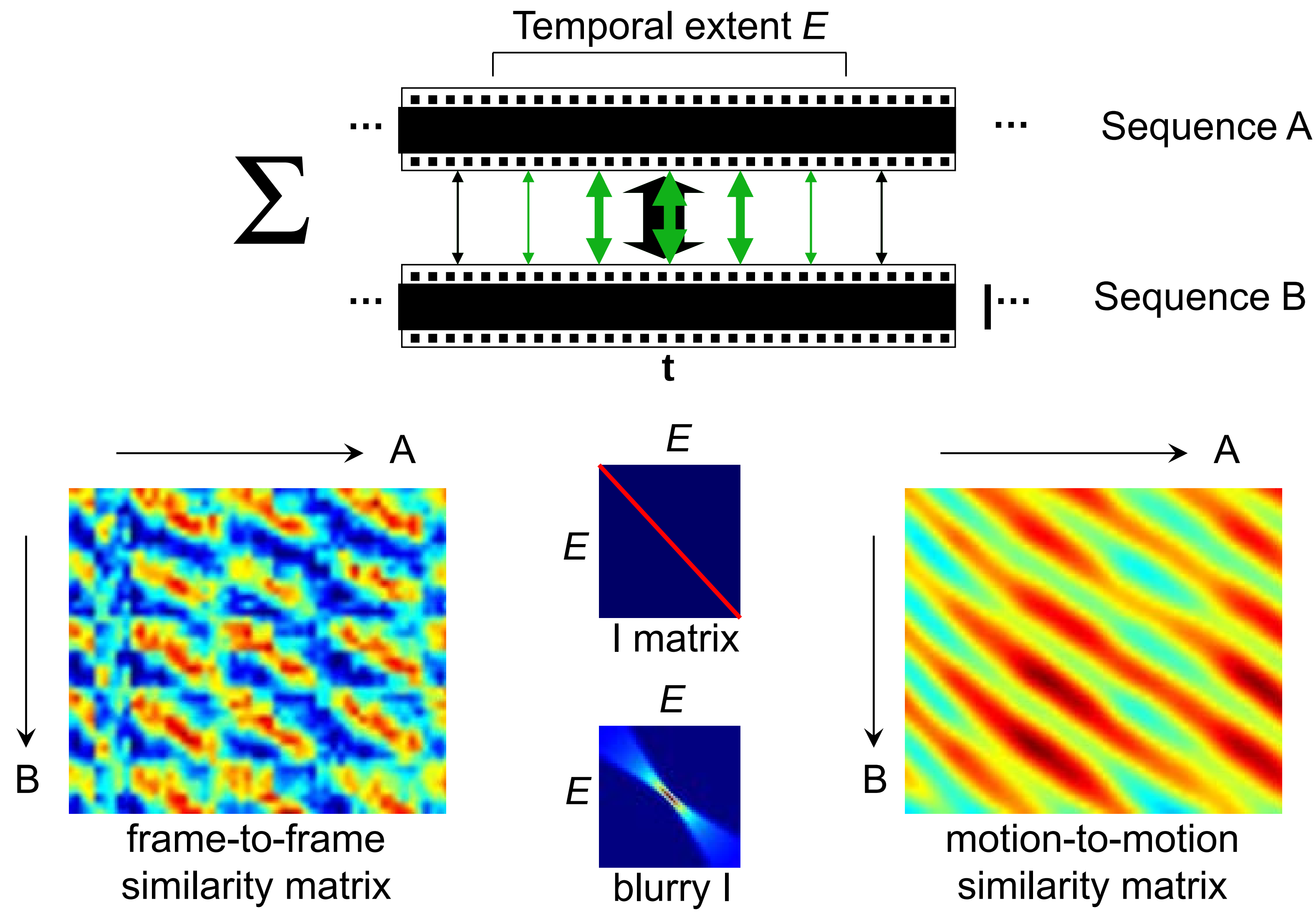


$F_x^-, F_x^+, F_y^-, F_y^+$



blurred $F_x^-, F_x^+, F_y^-, F_y^+$

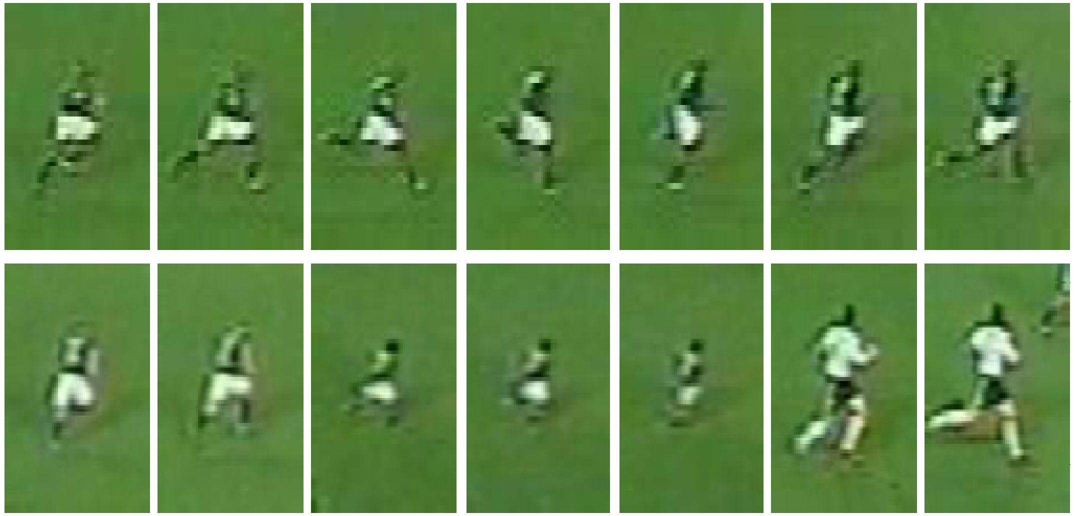
Spatio-Temporal Motion Descriptor

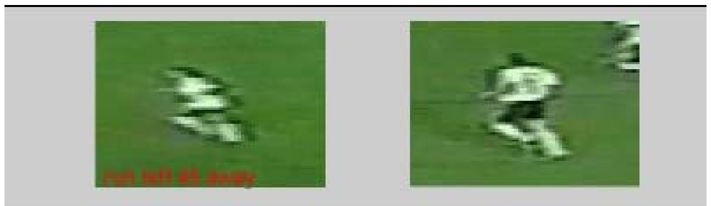


Football Actions: matching

Input Sequence

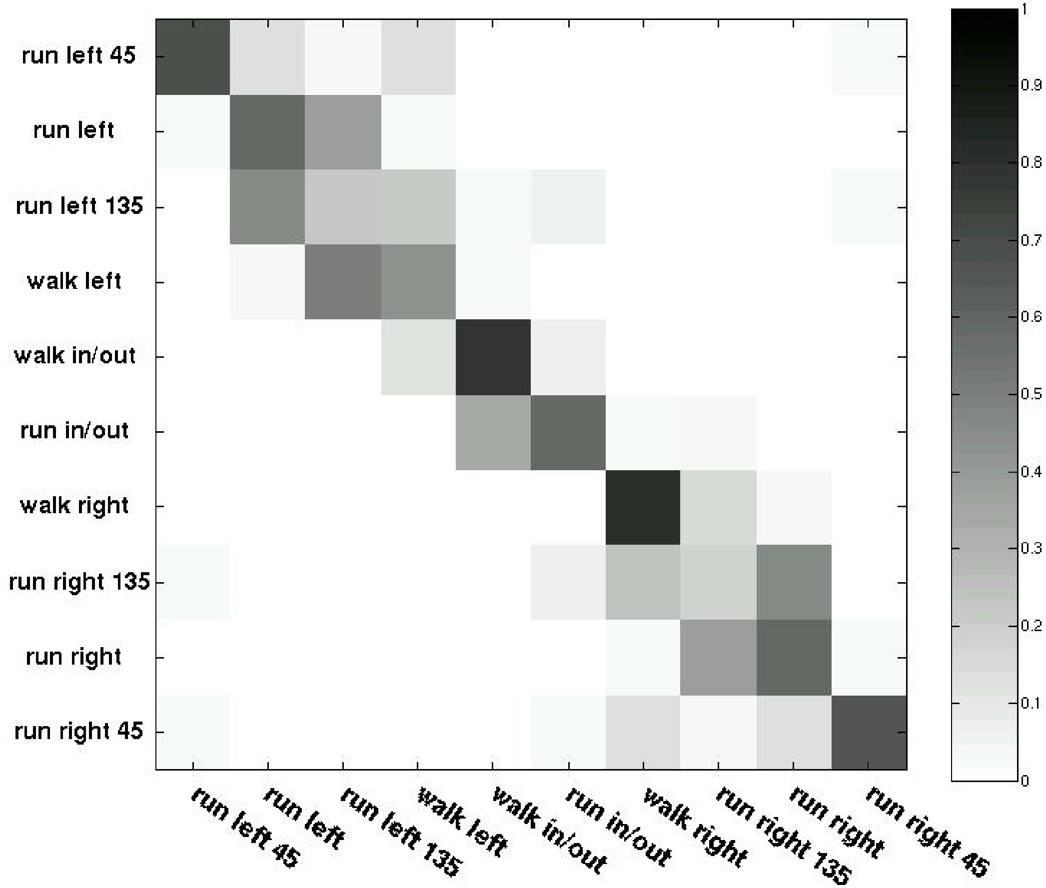
Matched Frames





input matched

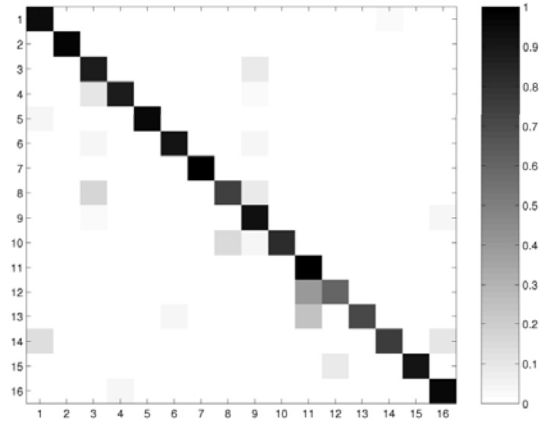
Football Actions: classification



10 actions; 4500 total frames; 13-frame motion descriptor

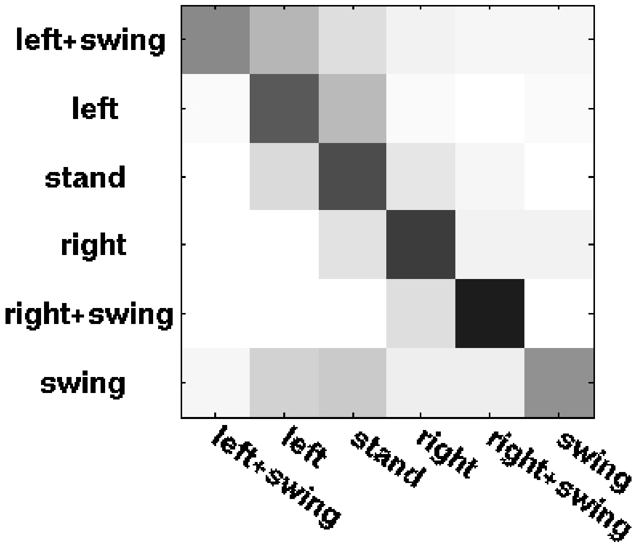
Classifying Ballet Actions

16 Actions; 24800 total frames; 51-frame motion descriptor. Men used to classify women and vice versa.



Classifying Tennis Actions

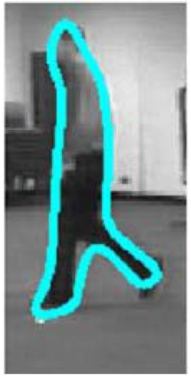
6 actions; 4600 frames; 7-frame motion descriptor
Woman player used as training, man as testing.



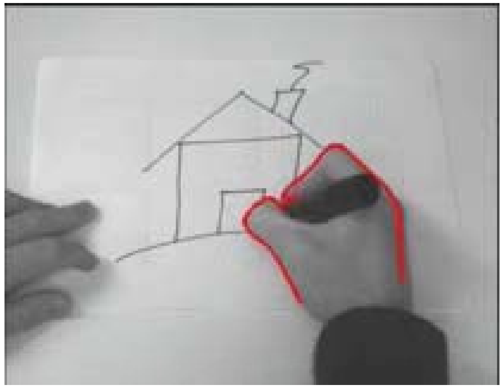
Where are we so far ?

Temporal templates:
+ simple, fast

- sensitive to segmentation errors



Active shape models:
+ shape regularization
- sensitive to initialization and tracking failures



Tracking with motion priors:
+ improved tracking and simultaneous action recognition
- sensitive to initialization and tracking failures

Motion-based recognition:
+ generic descriptors; less depends on appearance

- sensitive to localization/tracking errors