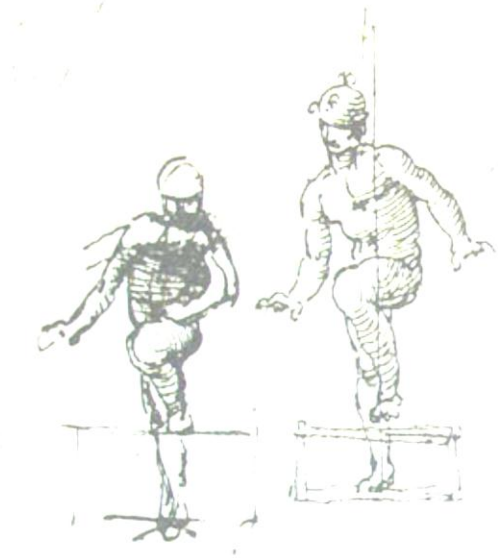
Reconnaissance d'objets et vision artificielle 2012



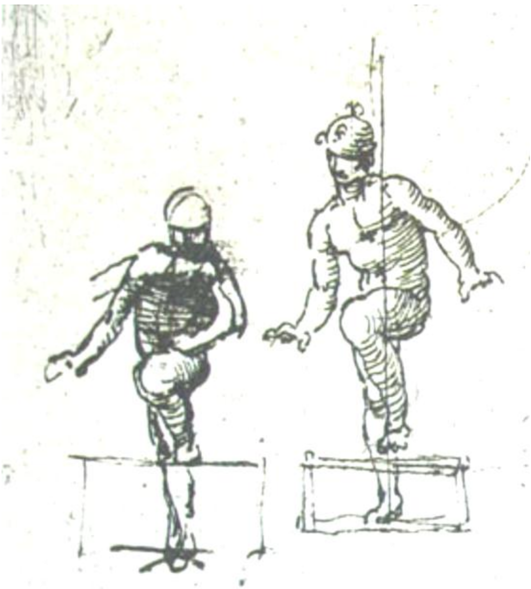
Motion and Human Actions

Ivan Laptev

ivan.laptev@inria.fr

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

Class overview



Motivation

Historic review
Modern applications

Appearance-based methods

Motion history images
Active shape models
Tracking and motion priors

Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Local space-time features
Action classification and detection
Weakly-supervised action learning

Motivation I: Artistic Representation

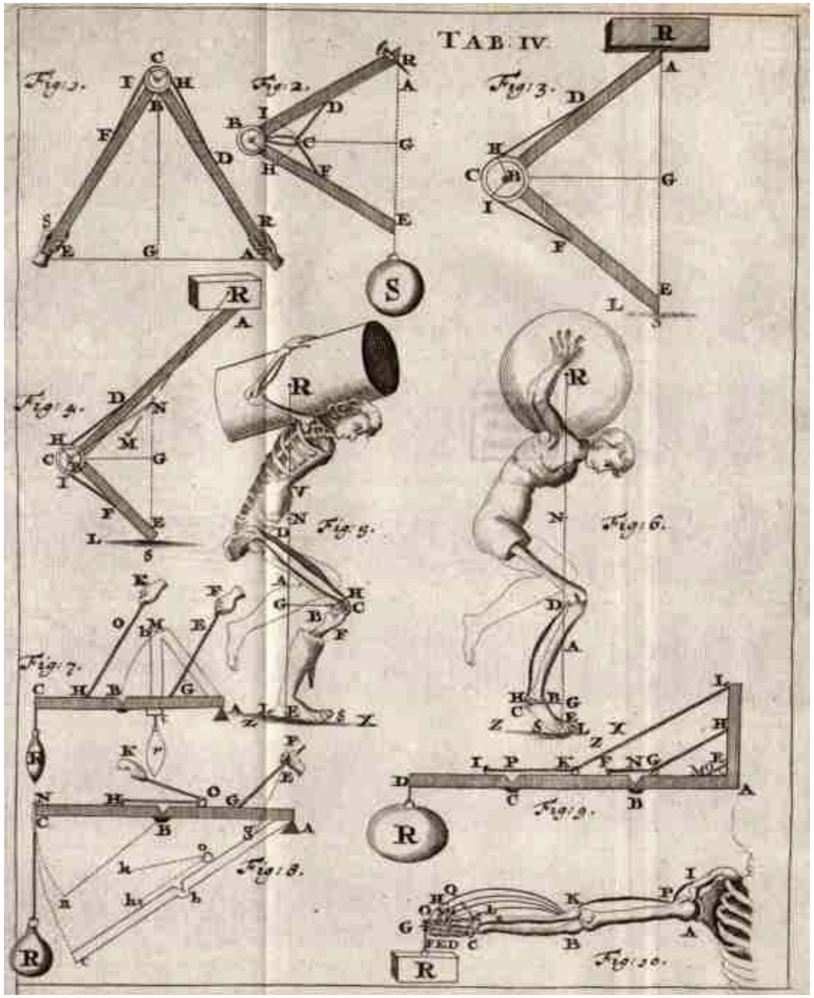
Early studies were motivated by human representations in Arts

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

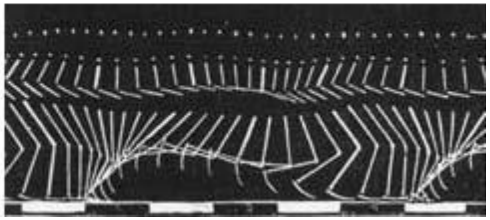


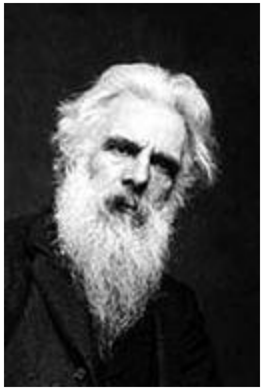
Giovanni Alfonso Borelli (1608–1679)

- 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

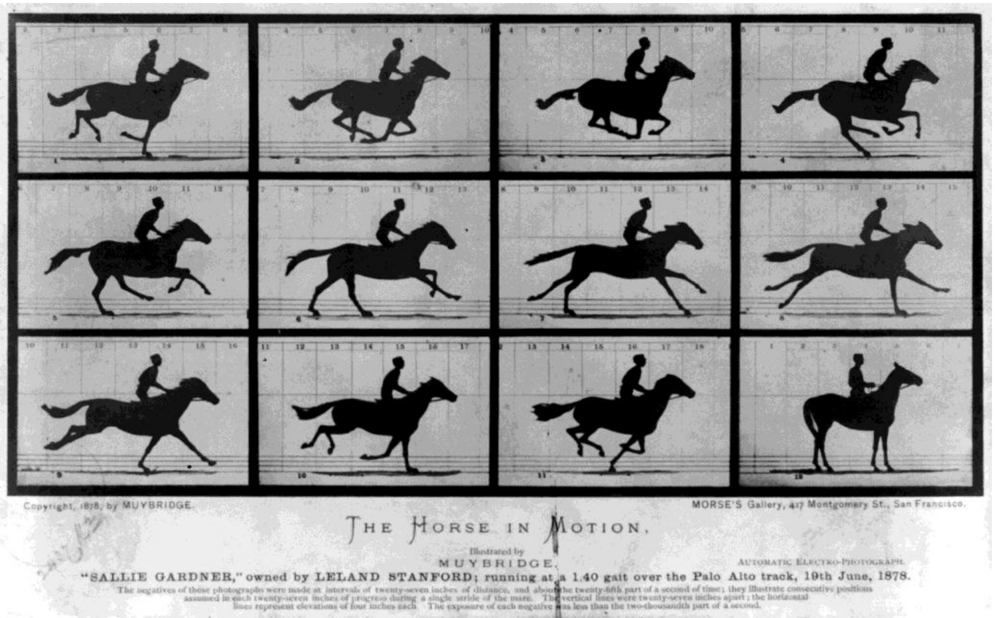
Motivation III: Motion perception

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



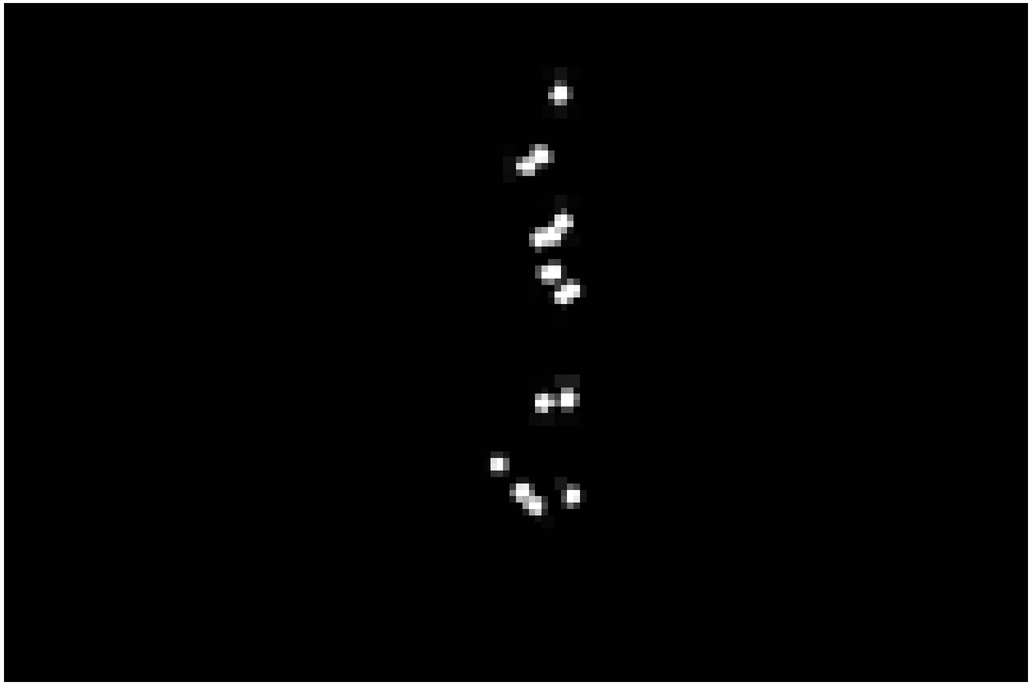


Eadweard Muybridge (1830–1904) invented a machine for displaying the recorded series of images. He pioneered motion pictures and applied his technique to movement studies



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

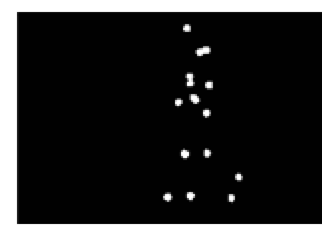
Human actions: Historic overview

15th century
studies of
anatomy

17th century
emergence of
biomechanics

19th century
emergence of
cinematography

1973
studies of human
motion perception

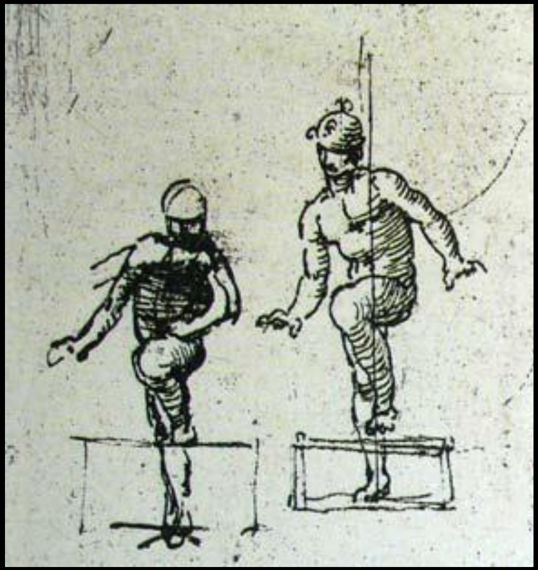


Modern computer vision

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

Why automatic video understanding?

- Huge amount of video is available and growing

BBC Motion Gallery

TV-channels recorded
since 60’s

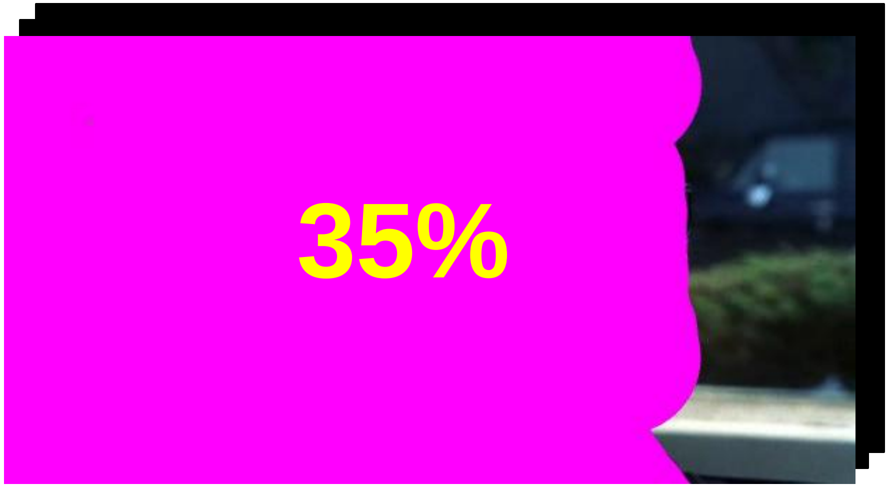
>34K hours of video
upload every day

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

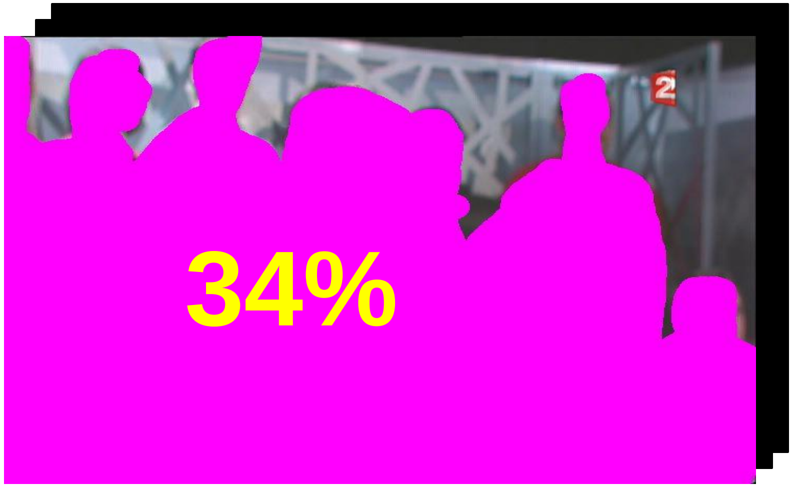
Movies

TV

YouTube



Movies



TV

YouTube

Why action recognition

- Analyzing video archives

First appearance of
N. Sarkozy on TV

Sociology research:
Influence of character
smoking in movies

Education: How do I
make a pizza?

- Surveillance

Where is my cat?

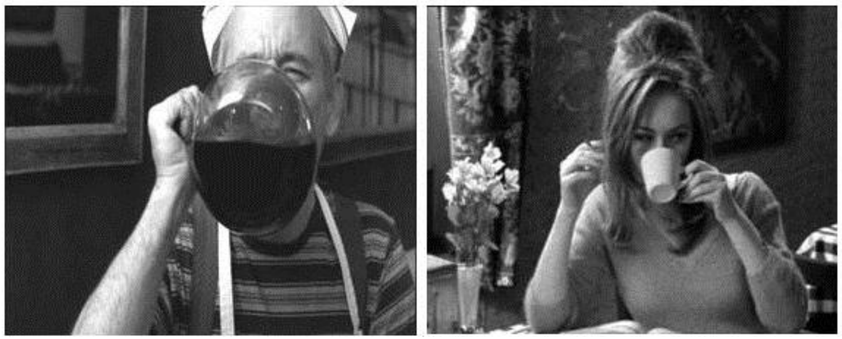
Predicting crowd behavior
Counting people

- Graphics

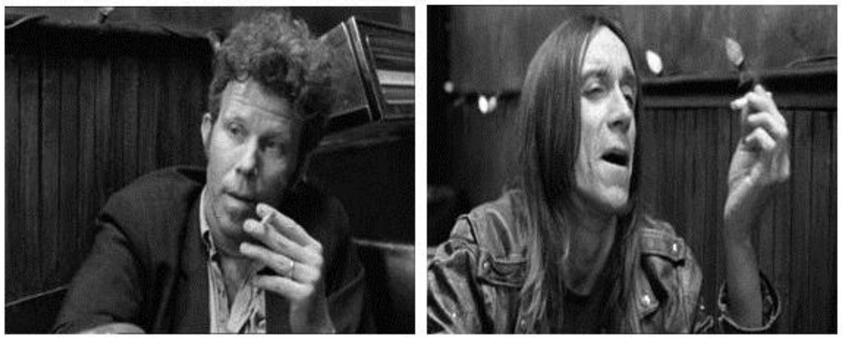
Motion capture and animation

Problem 1: Variability

- Need to deal with large appearance variations



Drinking



Smoking

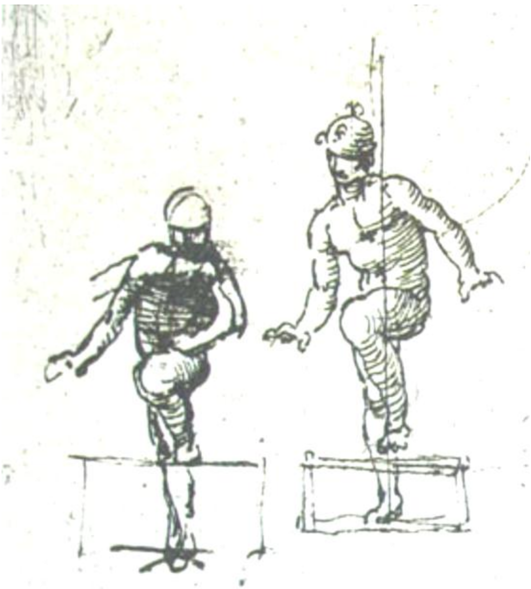
- Large number of classes

Problem 2: Granularity

Source: http://www.youtube.com/watch?v=eYdUZdan5i8

Do we want to learn *person-throws-cat-into-trash-bin* classifier?

Class overview



Motivation

- Historic review
- Modern applications

Appearance-based methods

- Motion history images
- Active shape models
- Tracking and motion priors

Motion-based methods

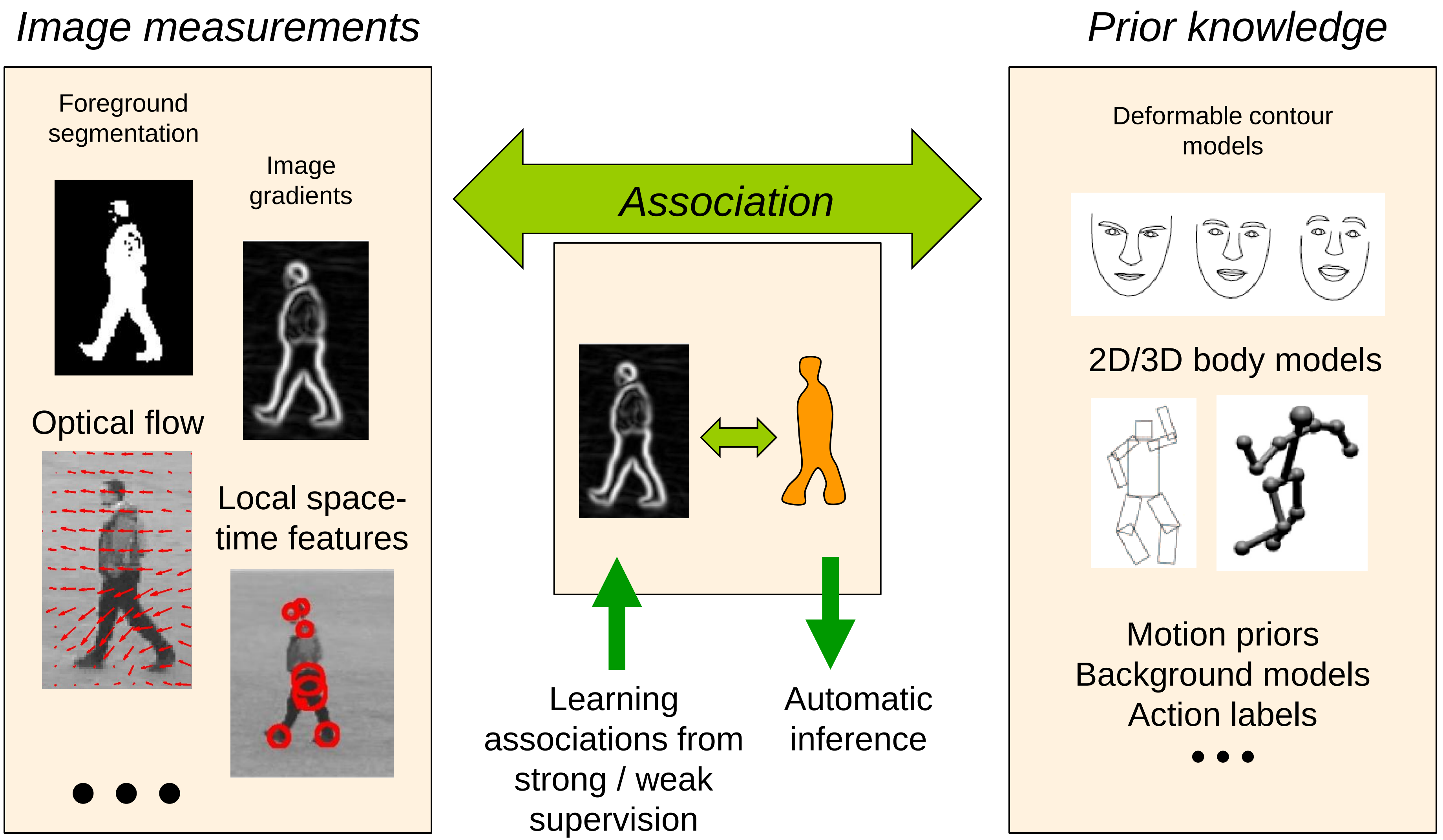
- Generic and parametric Optical Flow
- Motion templates

Space-time methods

- Local space-time features
- Action classification and detection
- Weakly-supervised action learning

How to recognize actions?

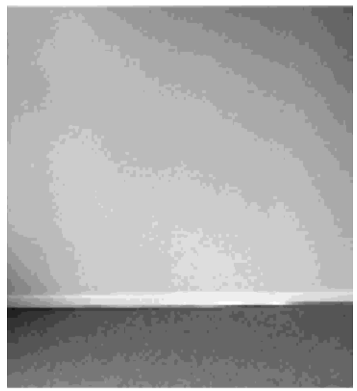
Action understanding: Key components



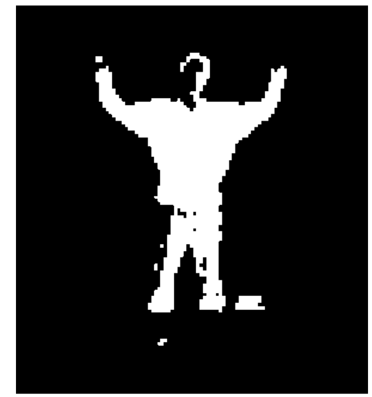
Foreground segmentation

Image differencing: a simple way to measure motion / temporal change

-



> Const

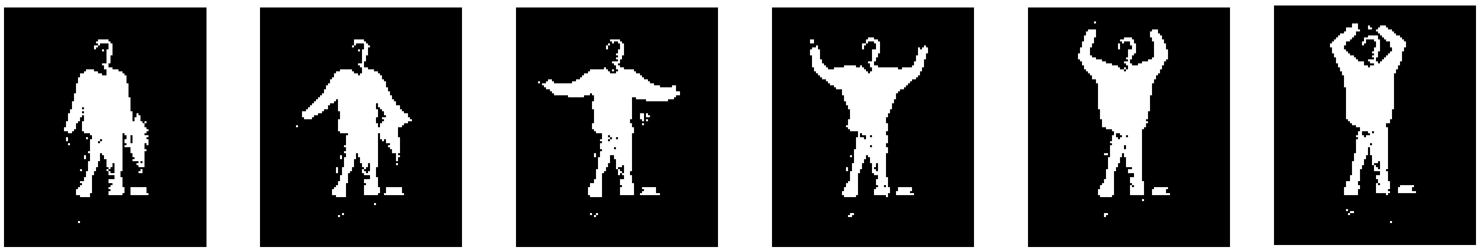


Better Background / Foreground separation methods exist:

- 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

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



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

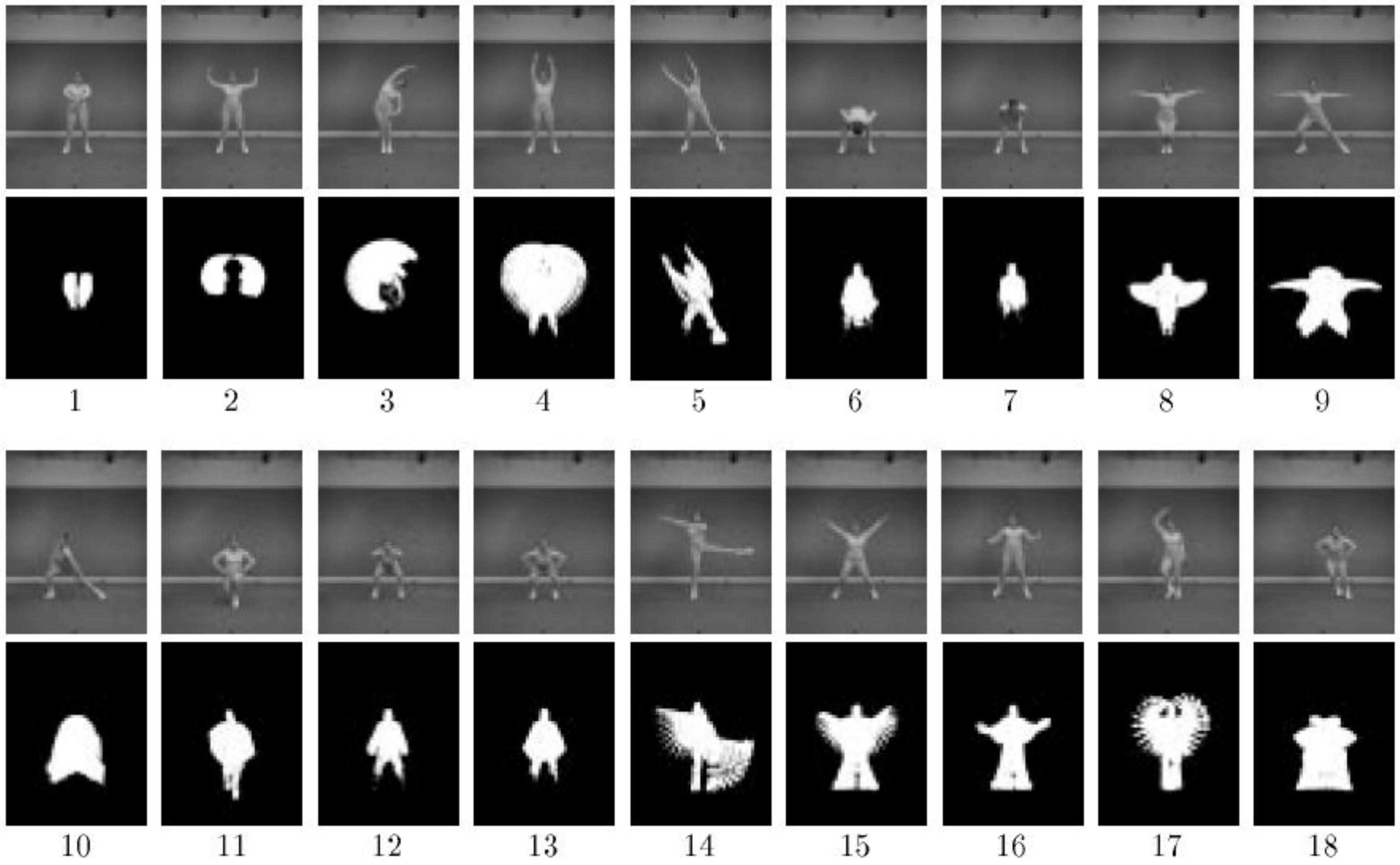
$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max\ (0, H_\tau(x, y, t-1) - 1) & \\ \quad \text{otherwise} & \end{cases}$$

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

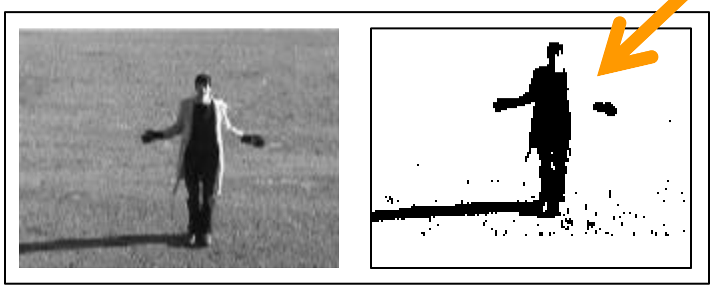
Pros:

+ 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*

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

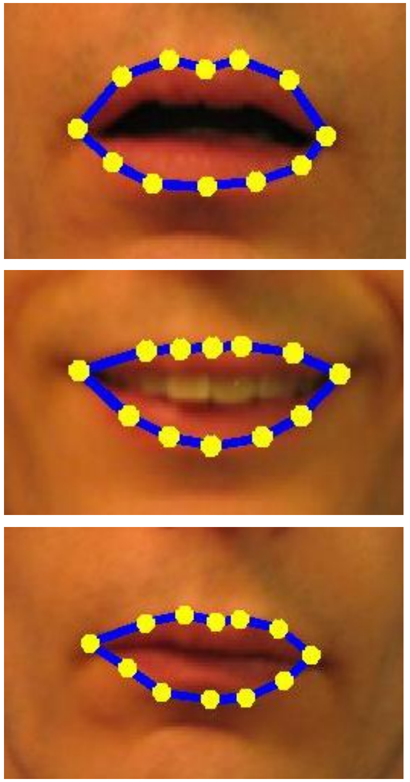
- Assume each shape can be represented by the linear combination of basis shapes

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

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 $\mathbf{b}$

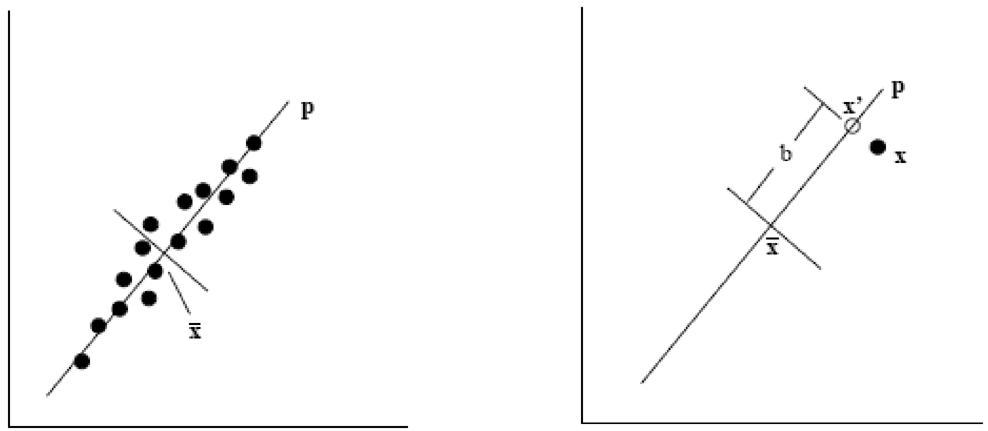


Active Shape Models of Cootes et al.

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

2D Example:

(each point can be thought as a shape in N-Dim space)



Principle Component Analysis (PCA):

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

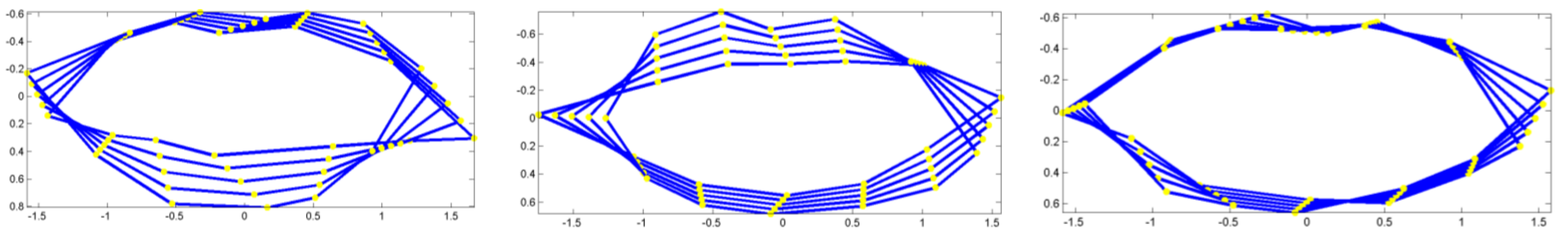
Eigenvectors $\mathbf{\Phi} = (\phi_1|\phi_2|\ldots|\phi_t)$ eigenvalues $\lambda_1, \ldots, \lambda_t$

Active Shape Models of Cootes et al.

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

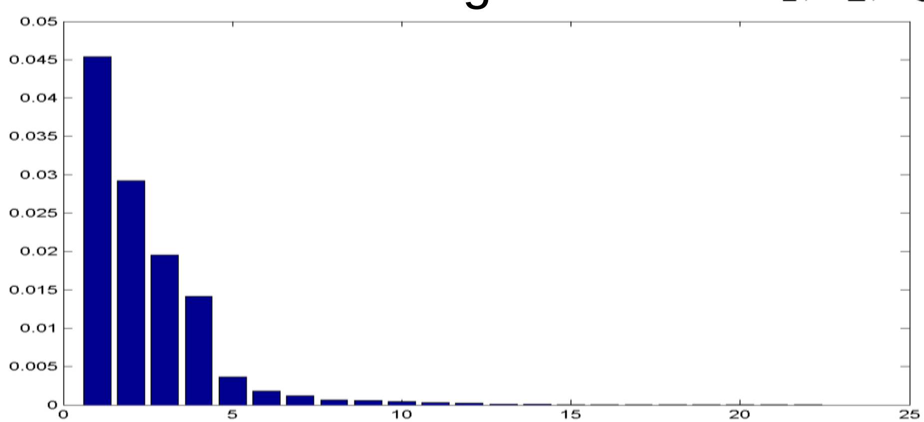
➡ Three main modes of lips-shape variation:

$\mathbf{b} = (\mu\lambda_1, 0, 0, ...)^\top$ $\mathbf{b} = (0, \mu\lambda_2, 0, 0, ...)^\top$ $\mathbf{b} = (0, 0, \mu\lambda_3, 0, 0, ...)^\top$



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

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.

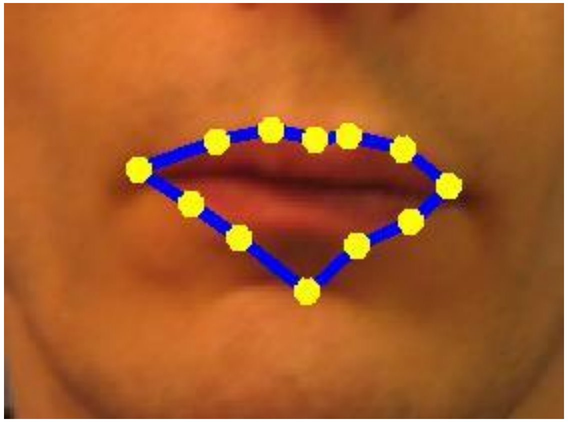
- $\mathbf{\Phi}$ is orthonormal basis, therefore $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^\top$

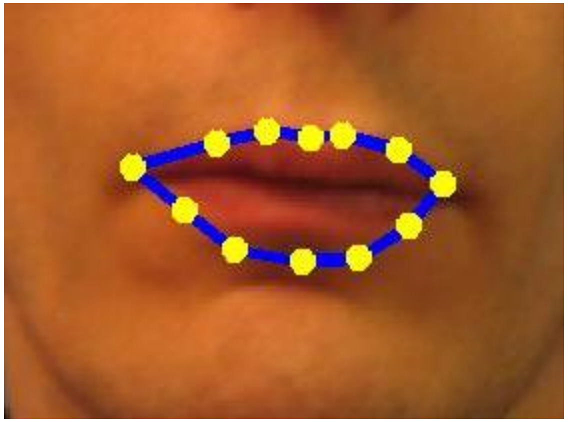
  ➡ Given estimate of $\mathbf{x}$ we can recover shape parameters $\mathbf{b}$

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

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

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

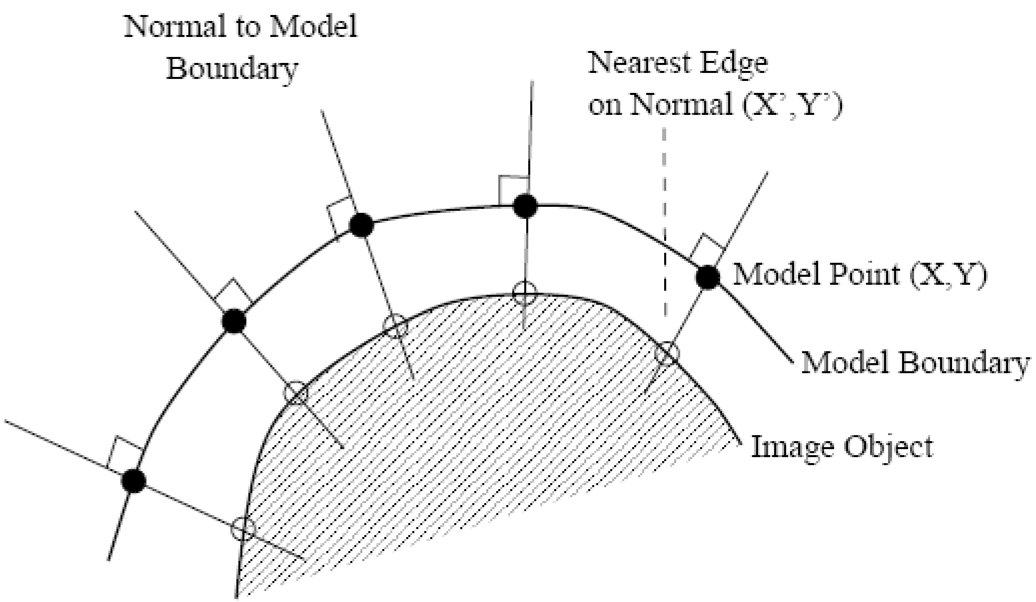




Active Shape Models of Cootes et al.

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

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

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

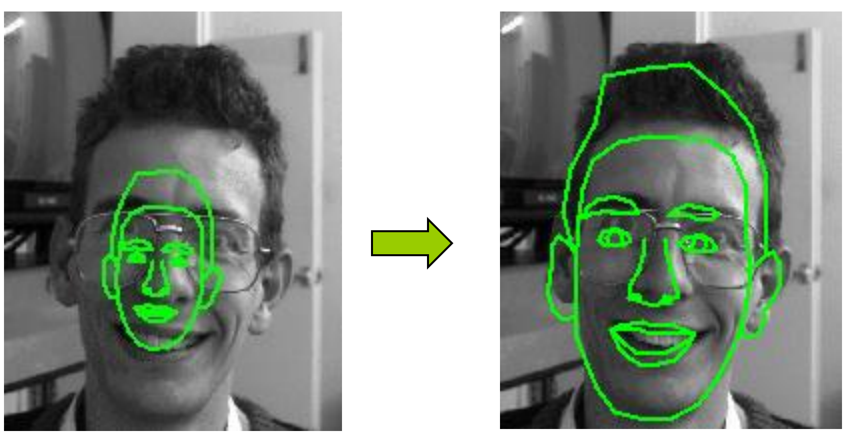
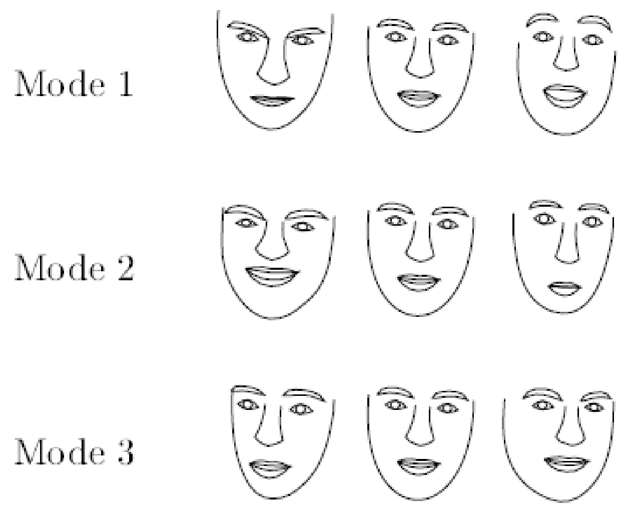


Illustration of face shape space



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

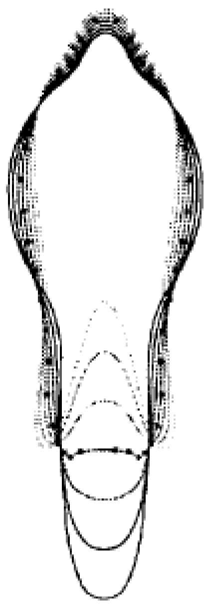
Active Shape Model tracking

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

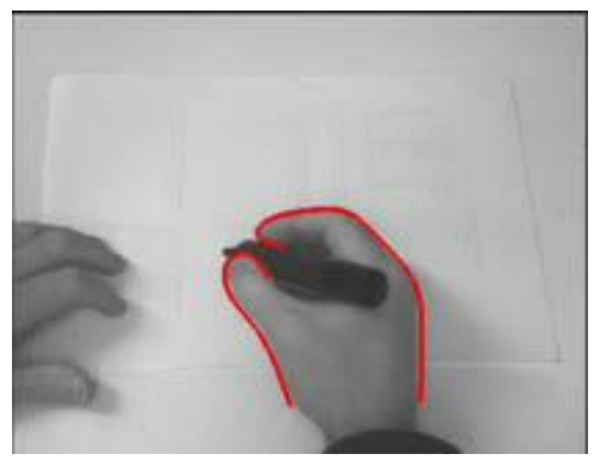
Example:

Drawing with 3 action modes

line drawing

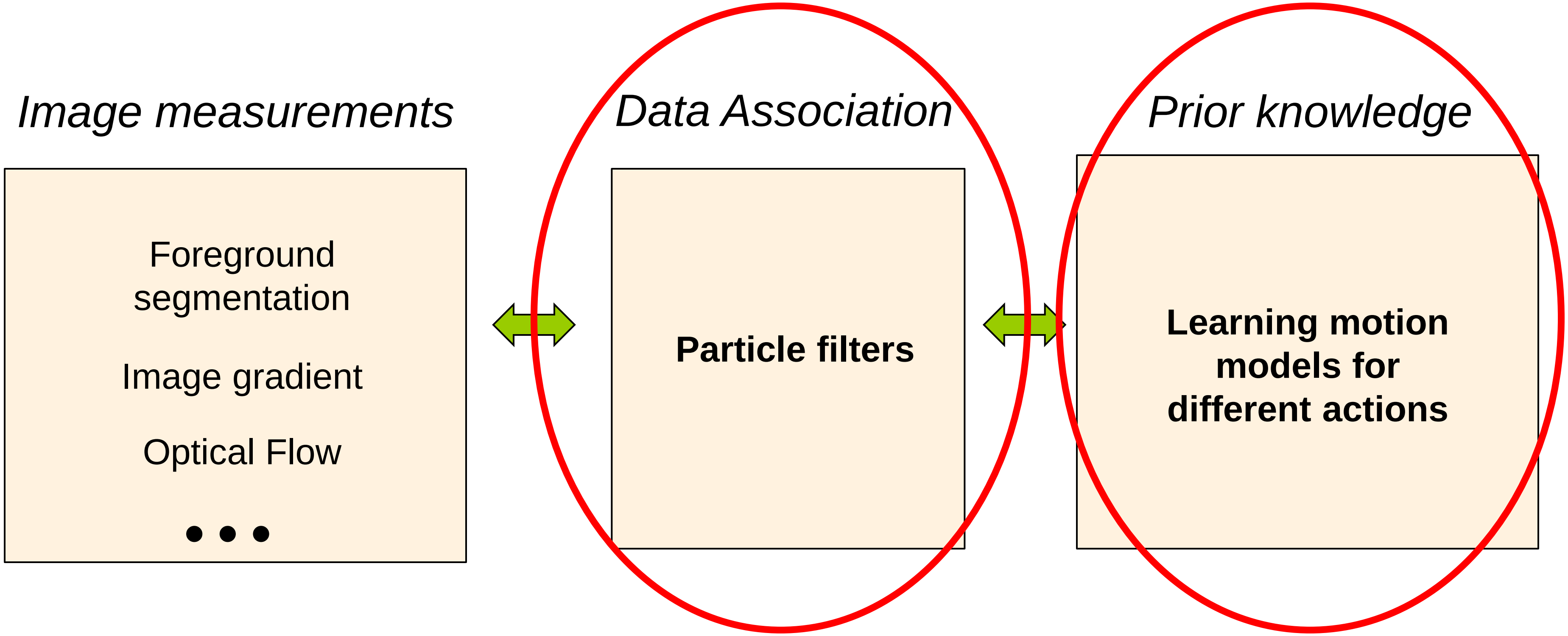
scribbling

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

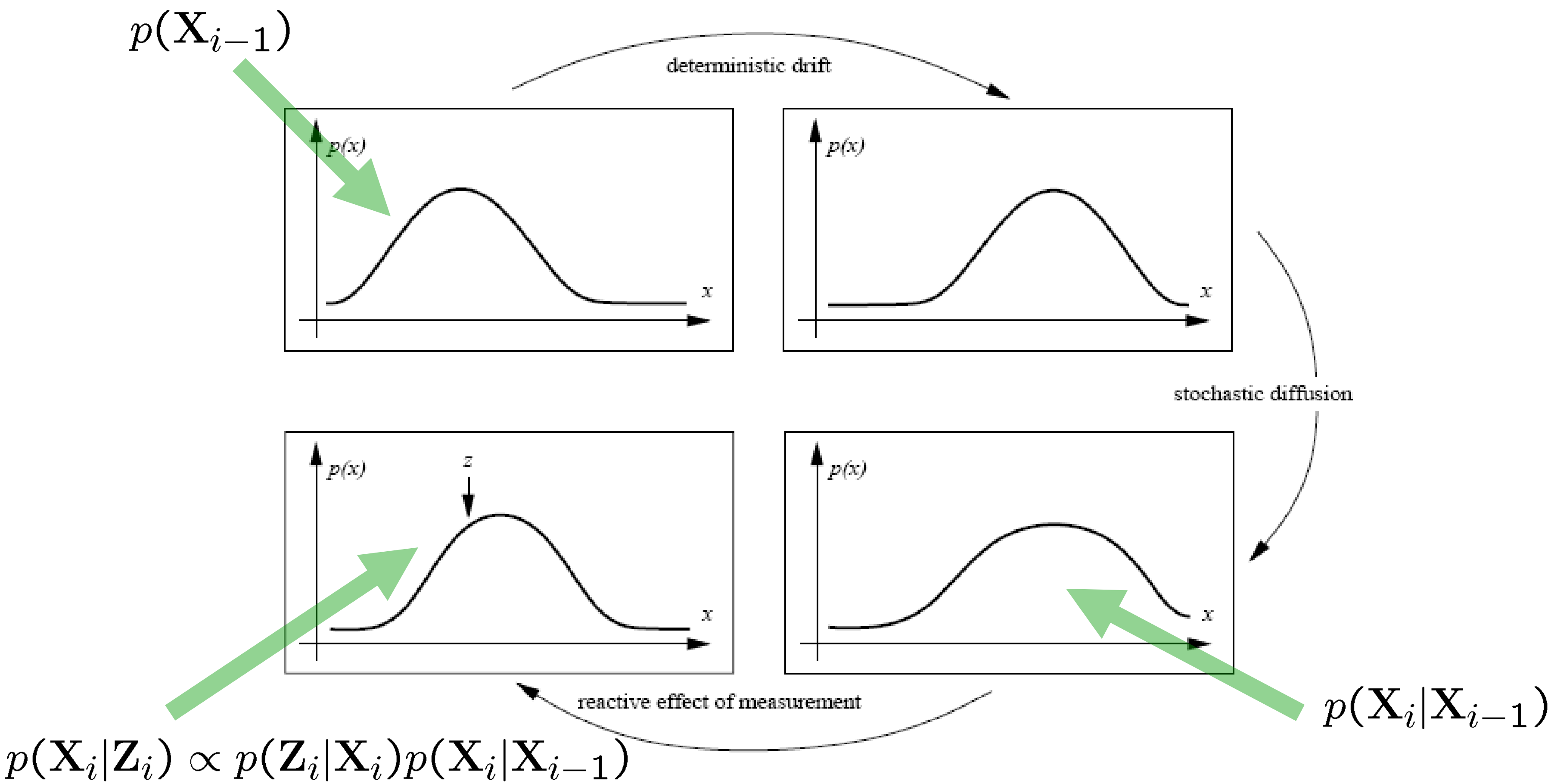
$$p(\mathbf{X}|\mathbf{Z}) \propto \mathbf{p}(\mathbf{Z}|\mathbf{X})\mathbf{p}(\mathbf{X})$$

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

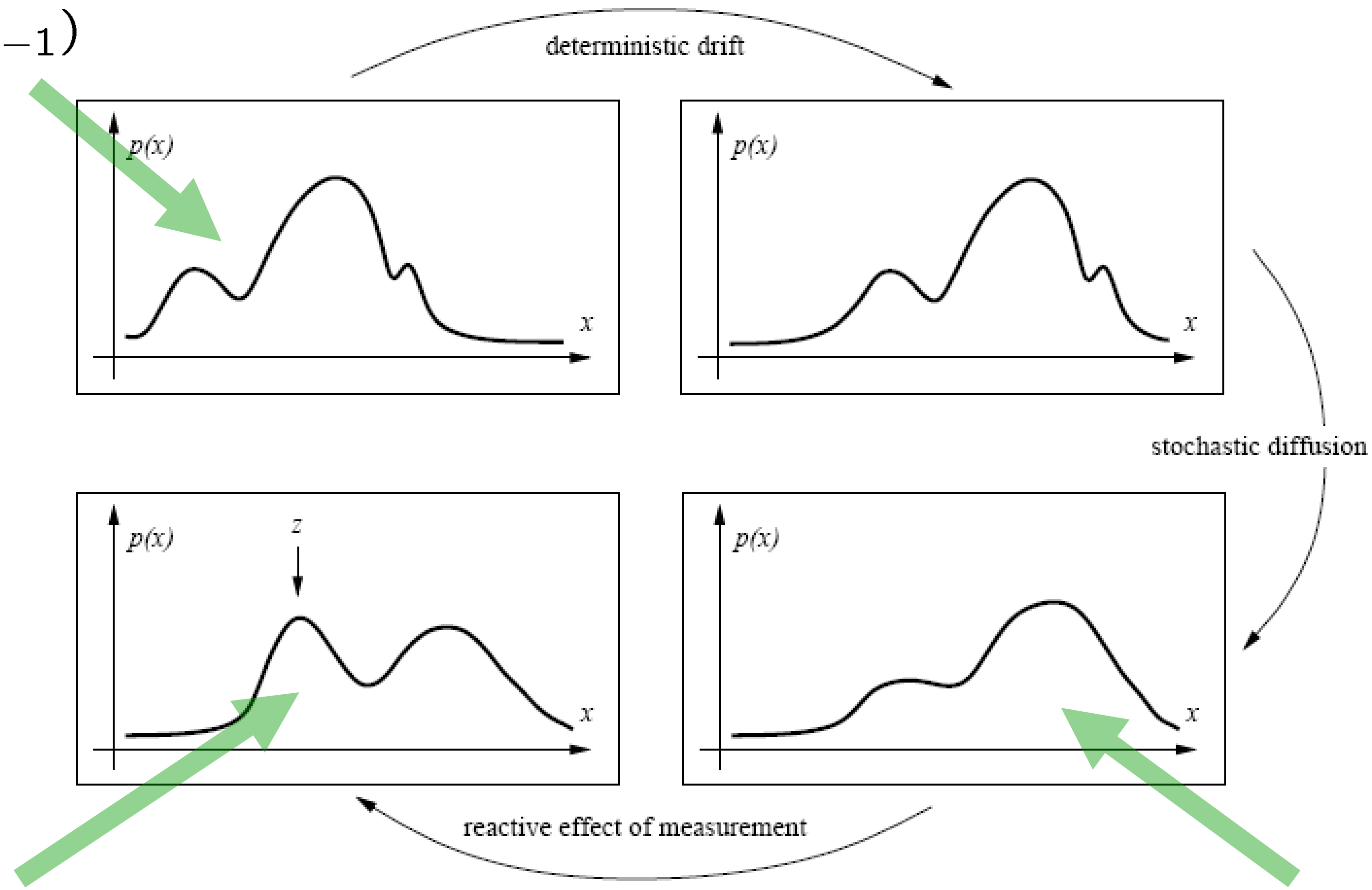
If all probability densities are uni-modal, specifically Gussians, the posterior can be evaluated in the closed form



Particle Filtering

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

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



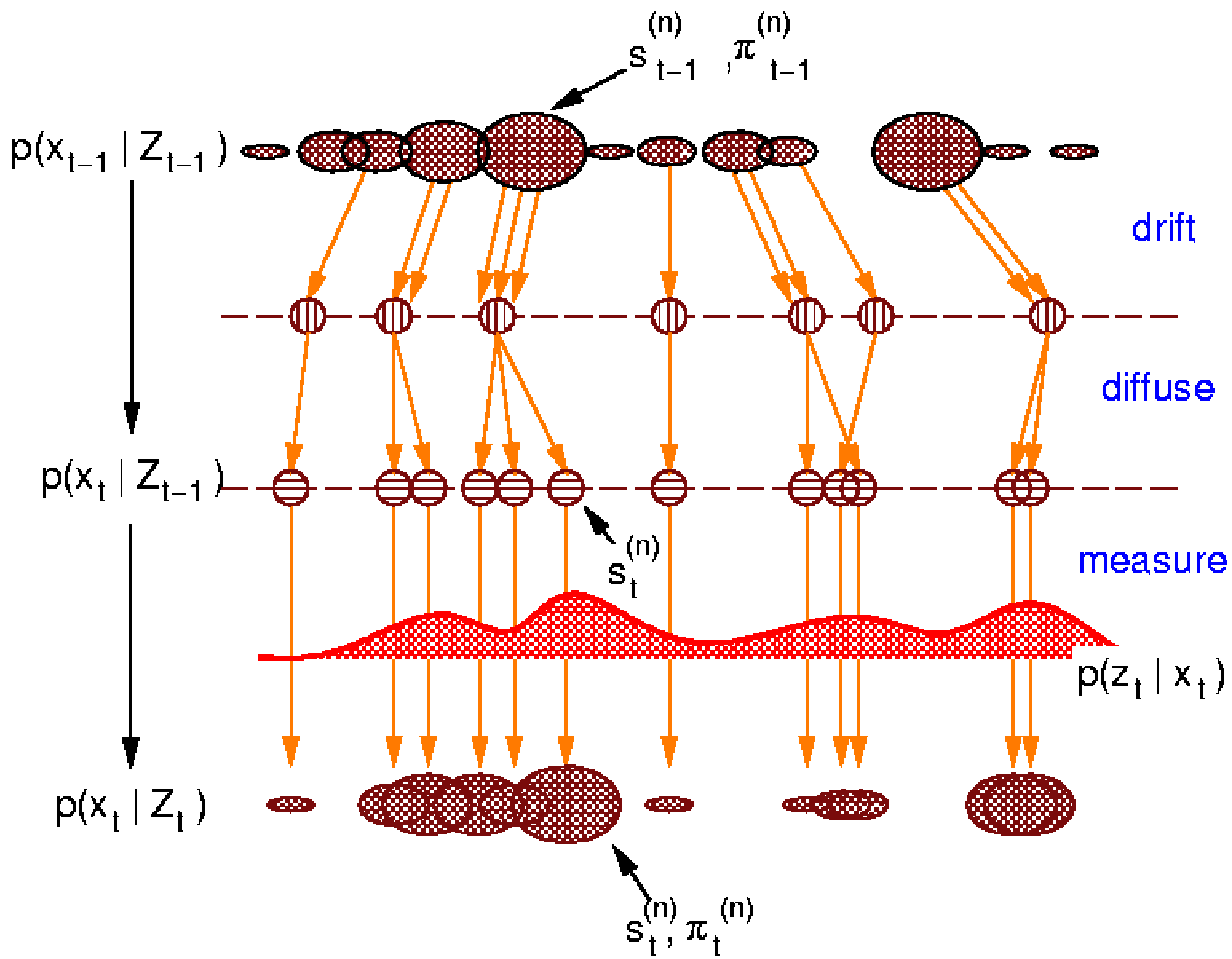
$p(\mathbf{X}_i|\mathbf{Z}_i) \propto p(\mathbf{Z}_i|\mathbf{X}_i)p(\mathbf{X}_i|\mathbf{X}_{i-1})$

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

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

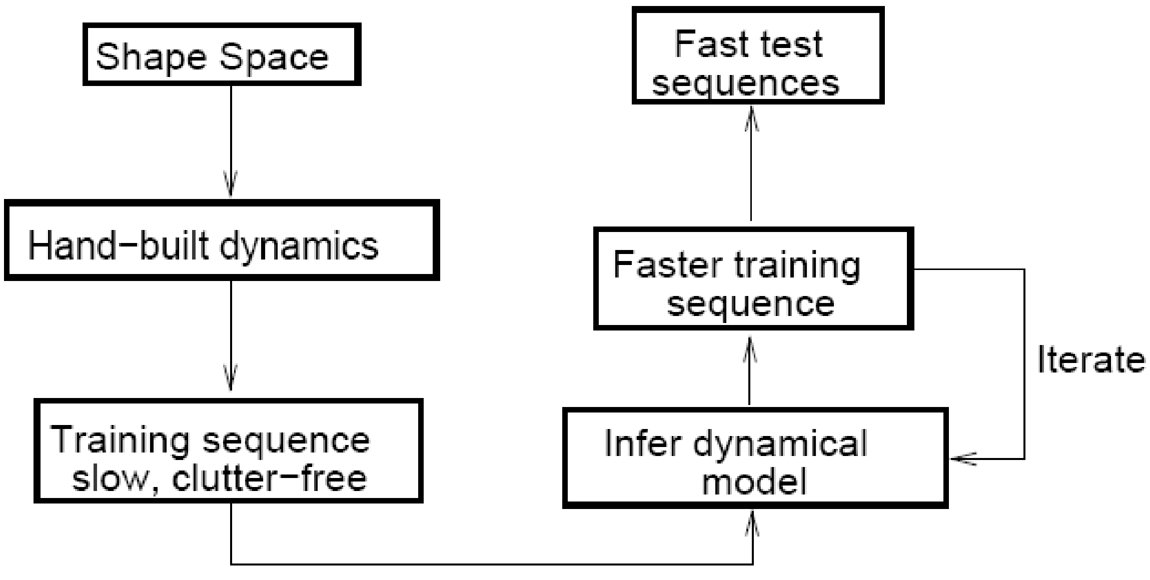
- Dynamic model: 2nd order Auto-Regressive Process

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

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

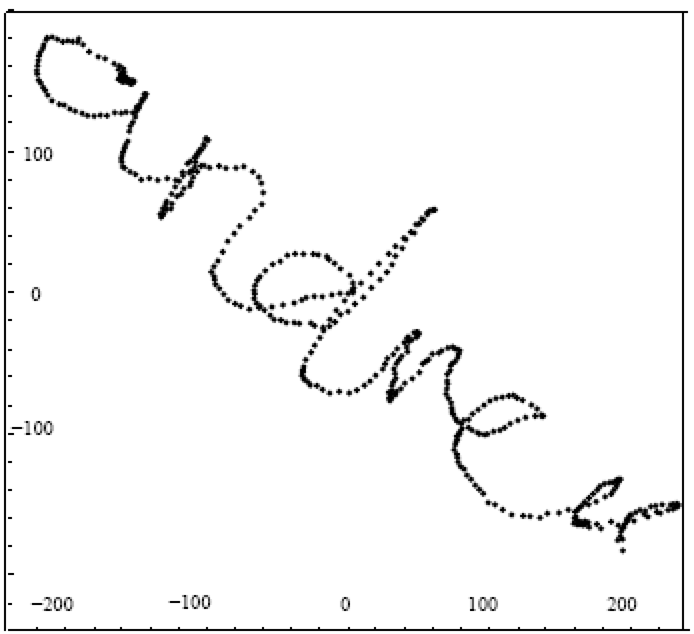
Model parameters: $$A = \begin{pmatrix} 0 & I \\ A_2 & A_1 \end{pmatrix}, \quad \overline{\mathcal{X}} = \begin{pmatrix} \overline{\mathbf{X}} \\ \overline{\mathbf{X}} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ B_0 \end{pmatrix}$$

Learning scheme:

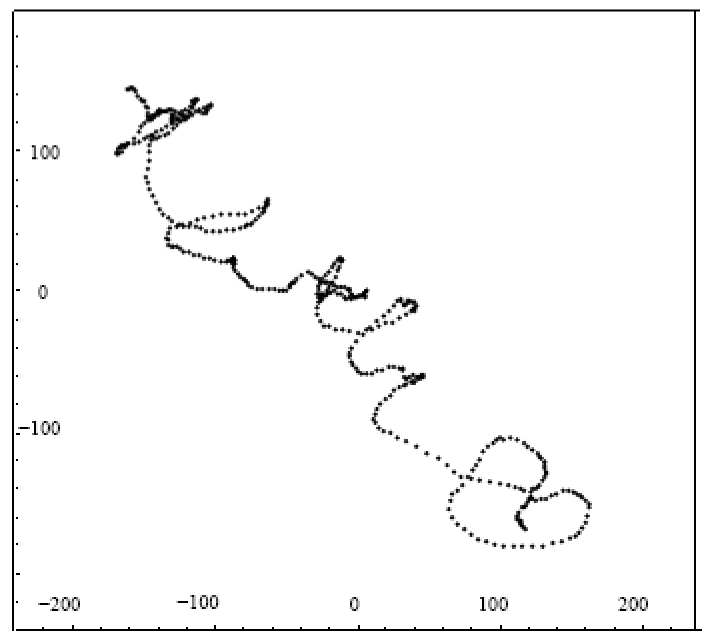


Learning dynamic prior

Learning point sequence



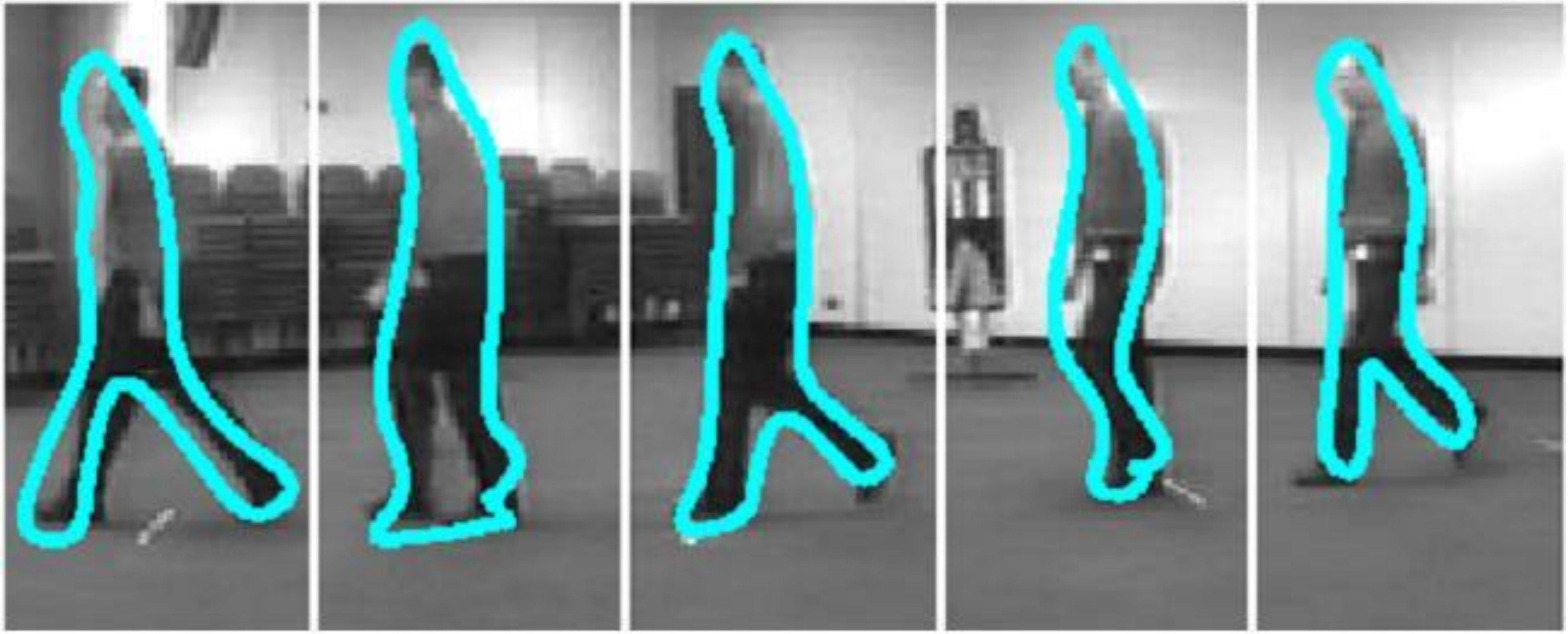
Random simulation of the learned dynamical model



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

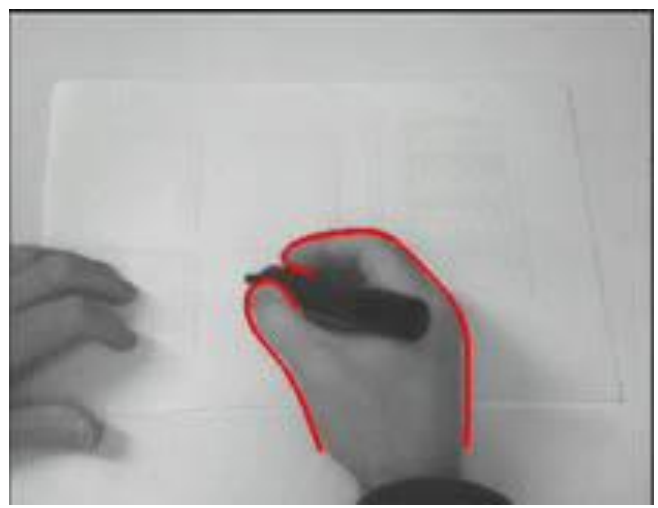
Dynamics with discrete states

Example: Drawing

Transition probability matrix

$$
T = \begin{array}{cccc}
 & \text{line} & \text{idle} & \text{scribbling} \\
 & \left(\begin{matrix} 0.9800 & 0.0015 & 0.0185 \\ 0.0850 & 0.9000 & 0.0150 \\ 0.0050 & 0.0150 & 0.9800 \end{matrix}\right) & & \begin{matrix}\text{line} \\ \text{idle} \\ \text{scribbling}\end{matrix}
\end{array}
$$

Result: simultaneously improved tracking and gesture recognition



line drawing

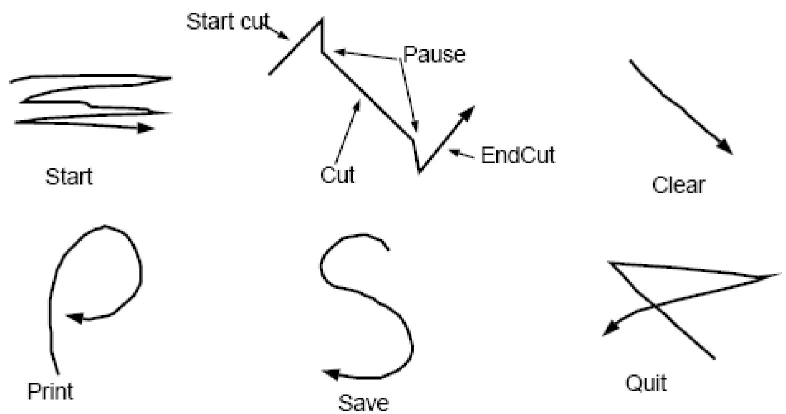
scribbling

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

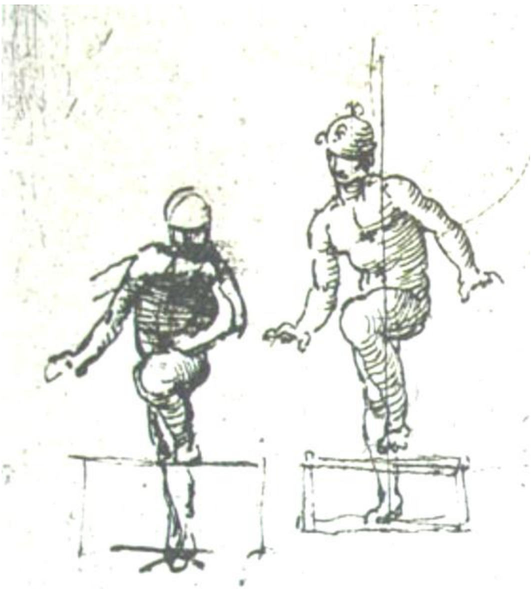
Pros:

+ 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

Appearance-based methods

Motion history images
Active shape models
Tracking and motion priors

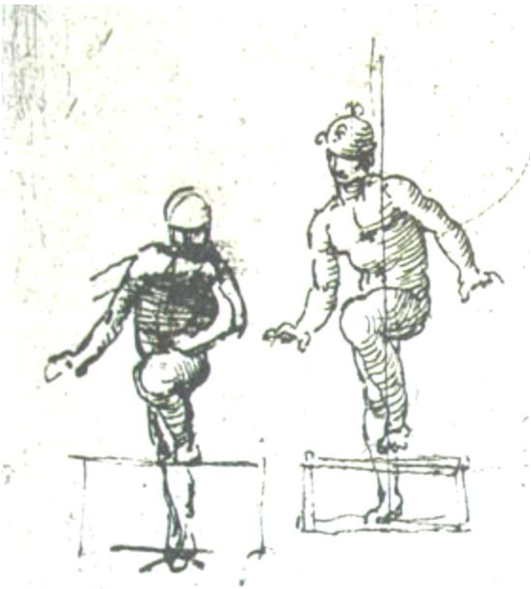
Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Local space-time features
Action classification and detection
Weakly-supervised action learning

Class overview



Motivation

Historic review
Modern applications

Appearance-based methods

Motion history images
Active shape models
Tracking and motion priors

Motion-based methods

Generic and parametric Optical Flow
Motion templates

Space-time methods

Local space-time features
Action classification and detection
Weakly-supervised action learning

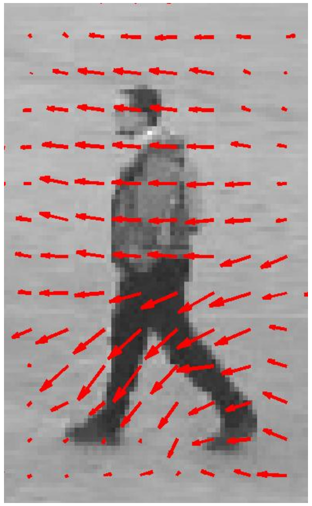
Shape and Appearance vs. Motion

- 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

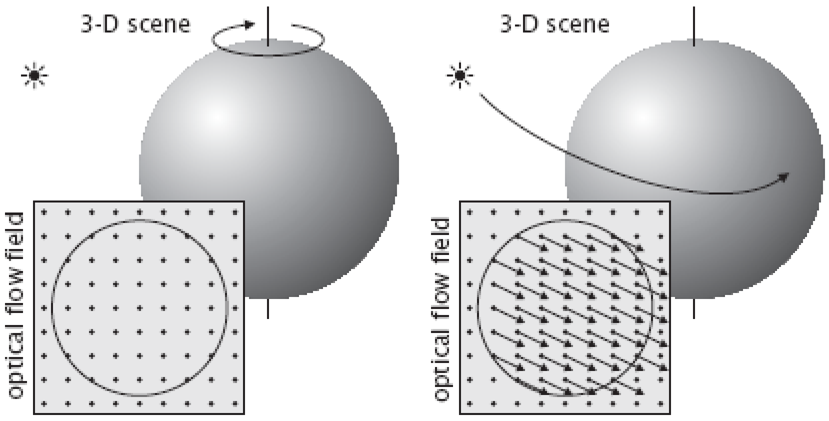
  How? We only have access to image pixels
  ⇨ Estimate pixel-wise correspondence between frames = Optical Flow

- 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)

$$(\nabla I)^\top \mathbf{v} + I_t = 0 \qquad \begin{aligned} \mathbf{v} &= (v_x, v_y)^\top \ \text{Optical flow} \\ \nabla I &= (I_x, I_y)^\top \ \text{Image gradient} \end{aligned}$$

One equation, two unknowns => cannot be solved directly

$\Rightarrow$ 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

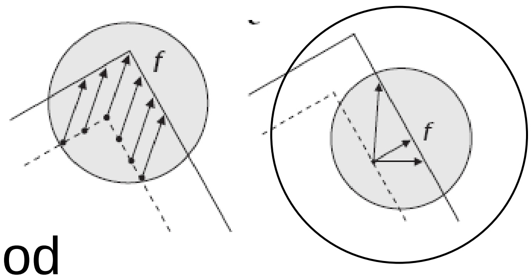
- The solution of $< \nabla I (\nabla I)^\top > \mathbf{v} = - < \nabla I I_t >$ assumes

  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 $< \cdot >$

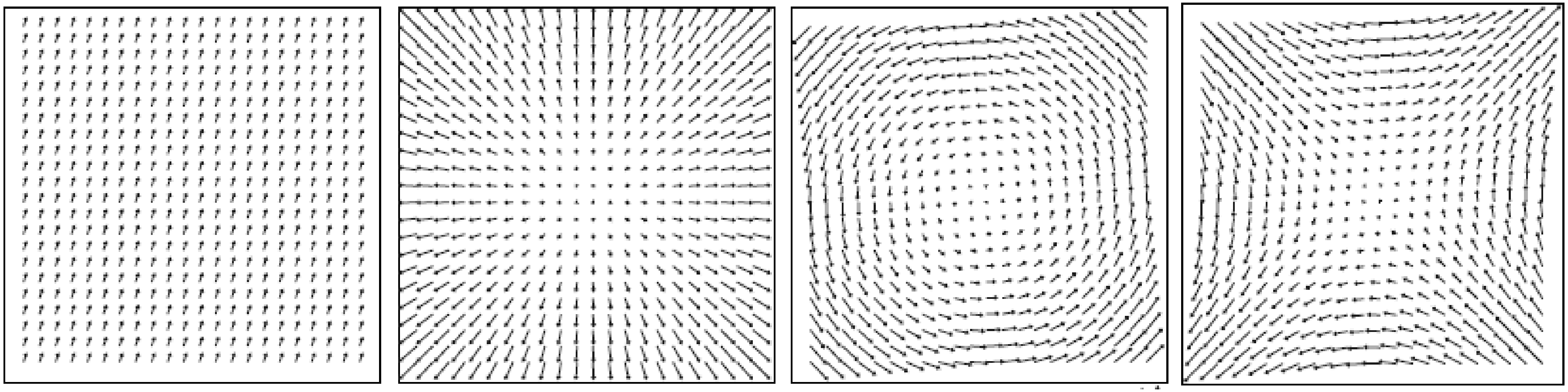
  ➡ Use more sophisticated motion model

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

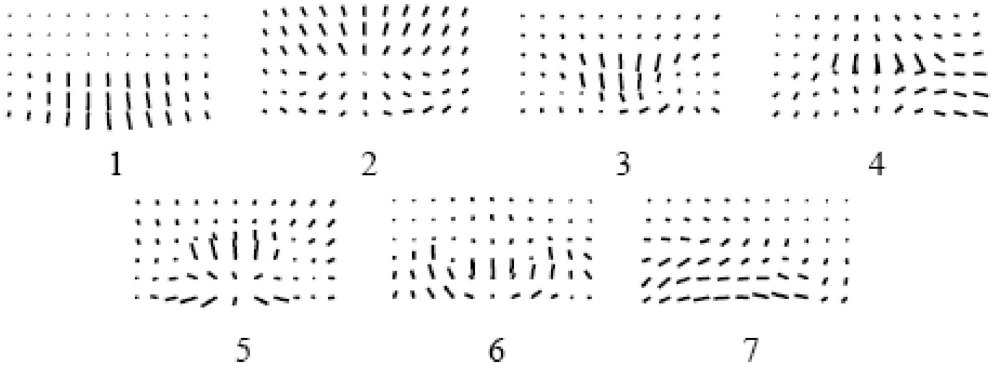
1. Compute standard Optical Flow for many examples
2. Put velocity components into one vector

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

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

Training samples

PCA flow bases



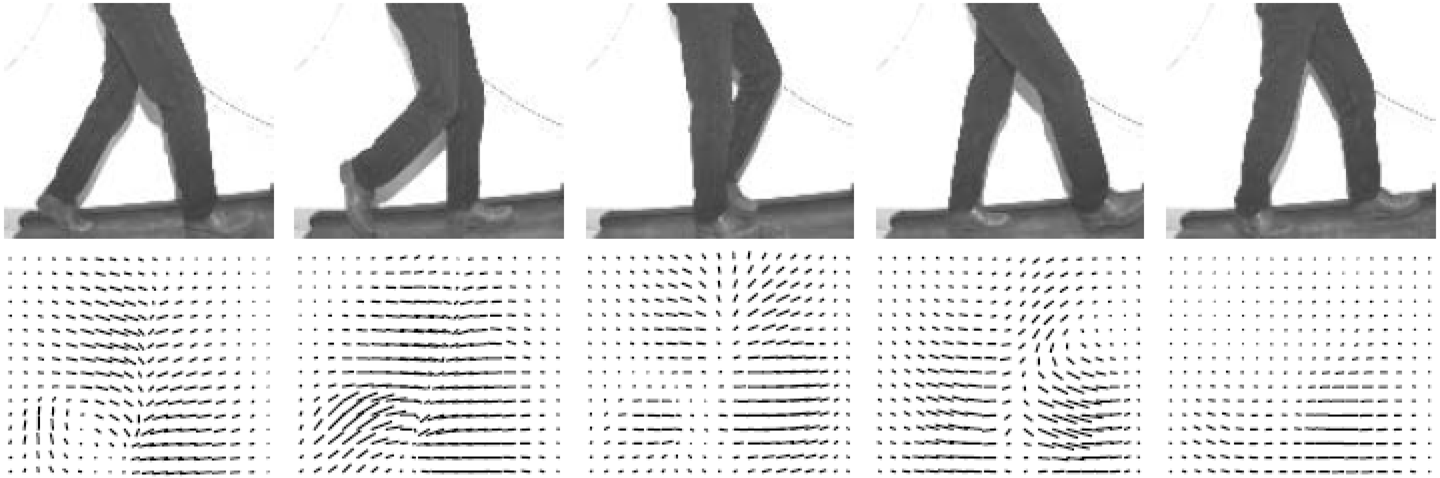
Learning Parameterized Models of Image Motion
M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, **CVPR 1997**

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

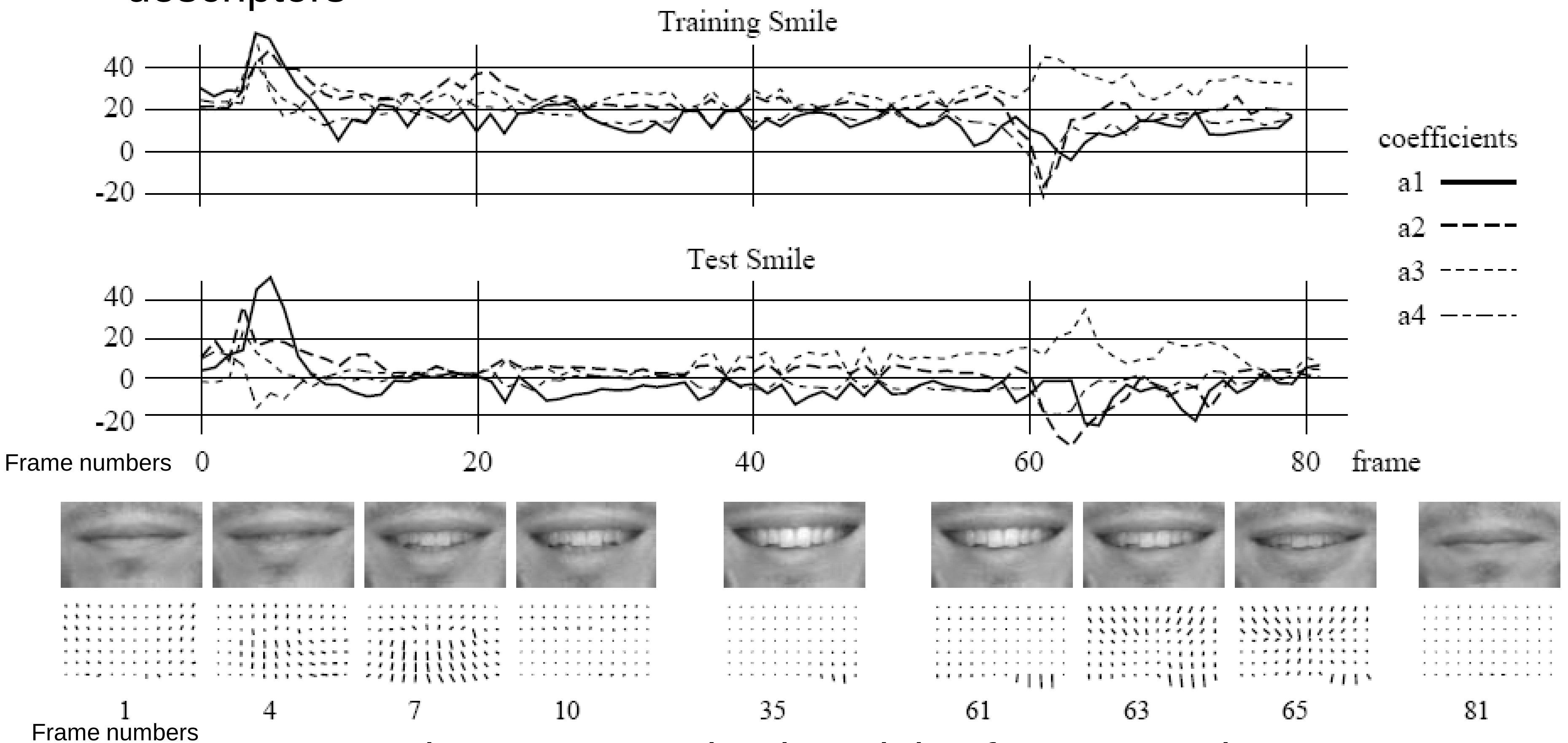
Direct flow recovery:



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

Parameterized Optical Flow

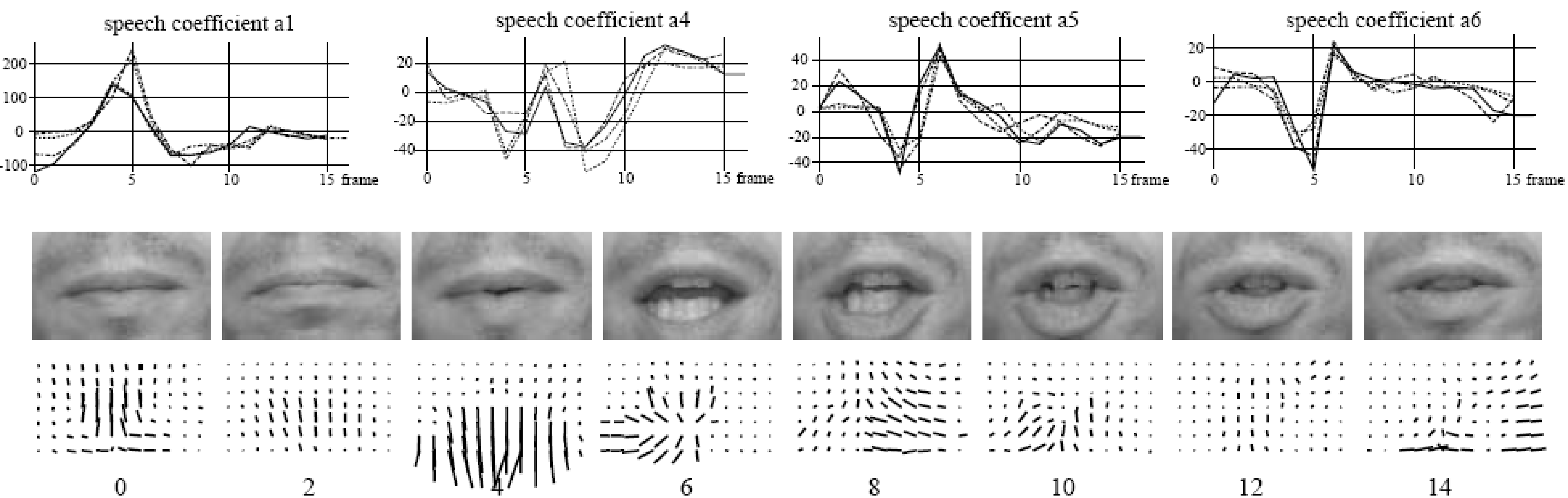
- Estimated coefficients of PCA flow bases can be used as action descriptors



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

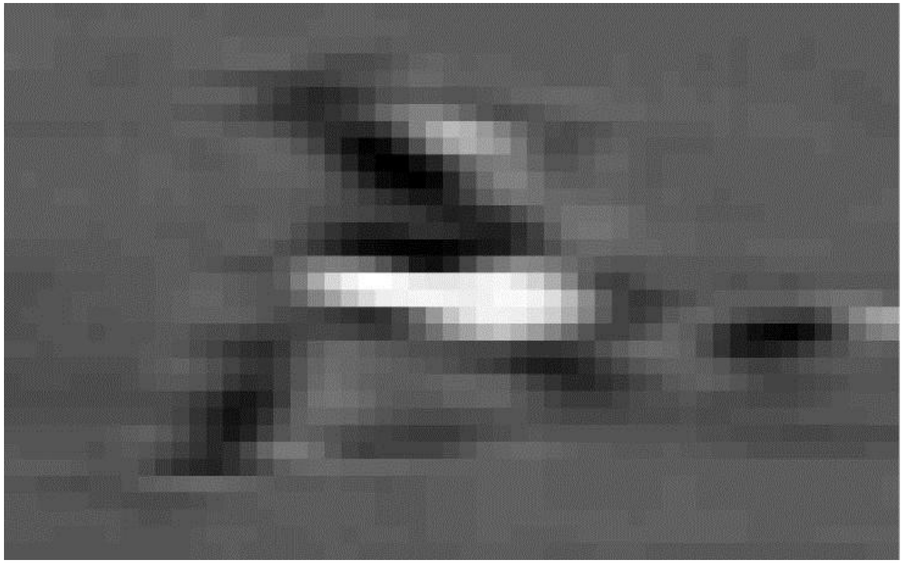
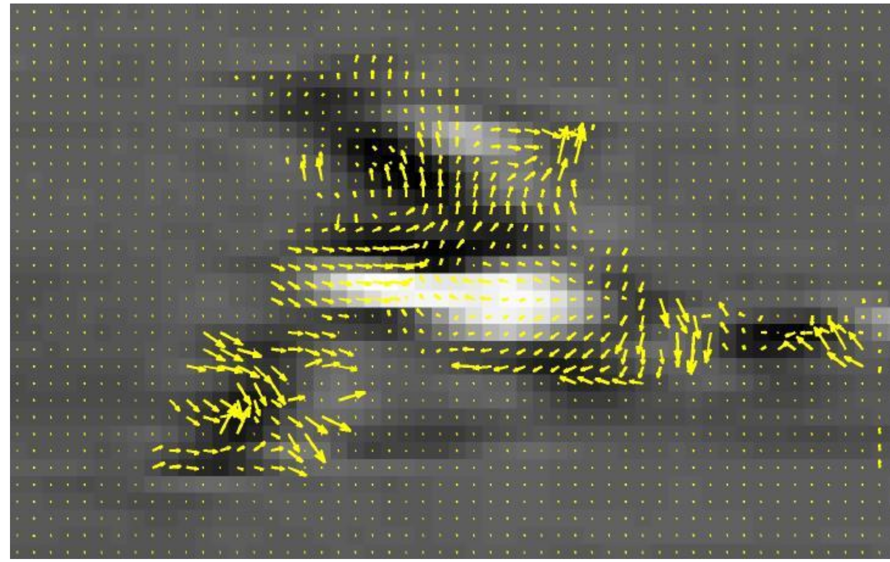
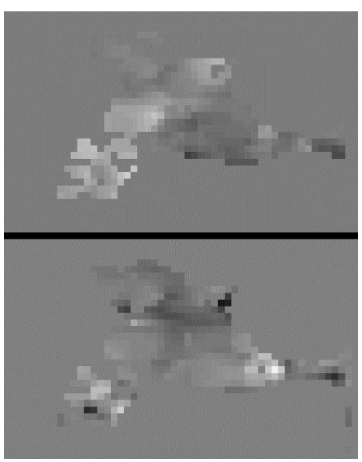


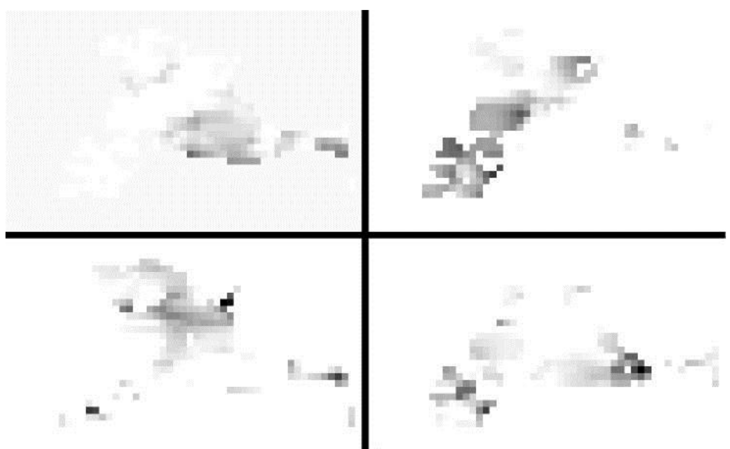
Image frame



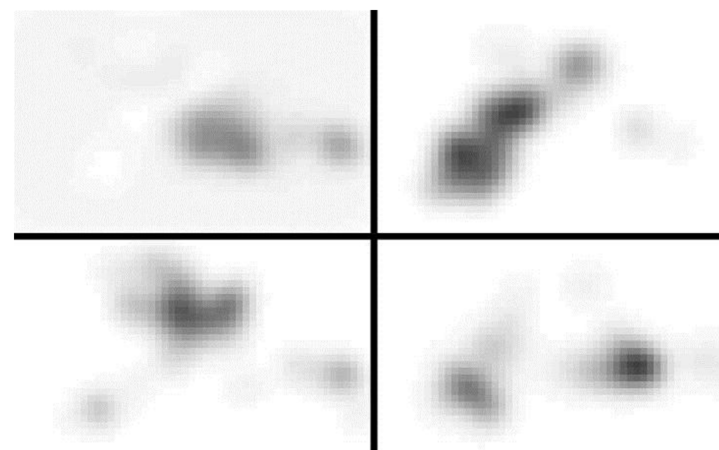
Optical flow $F_{x,y}$



F_x, F_y

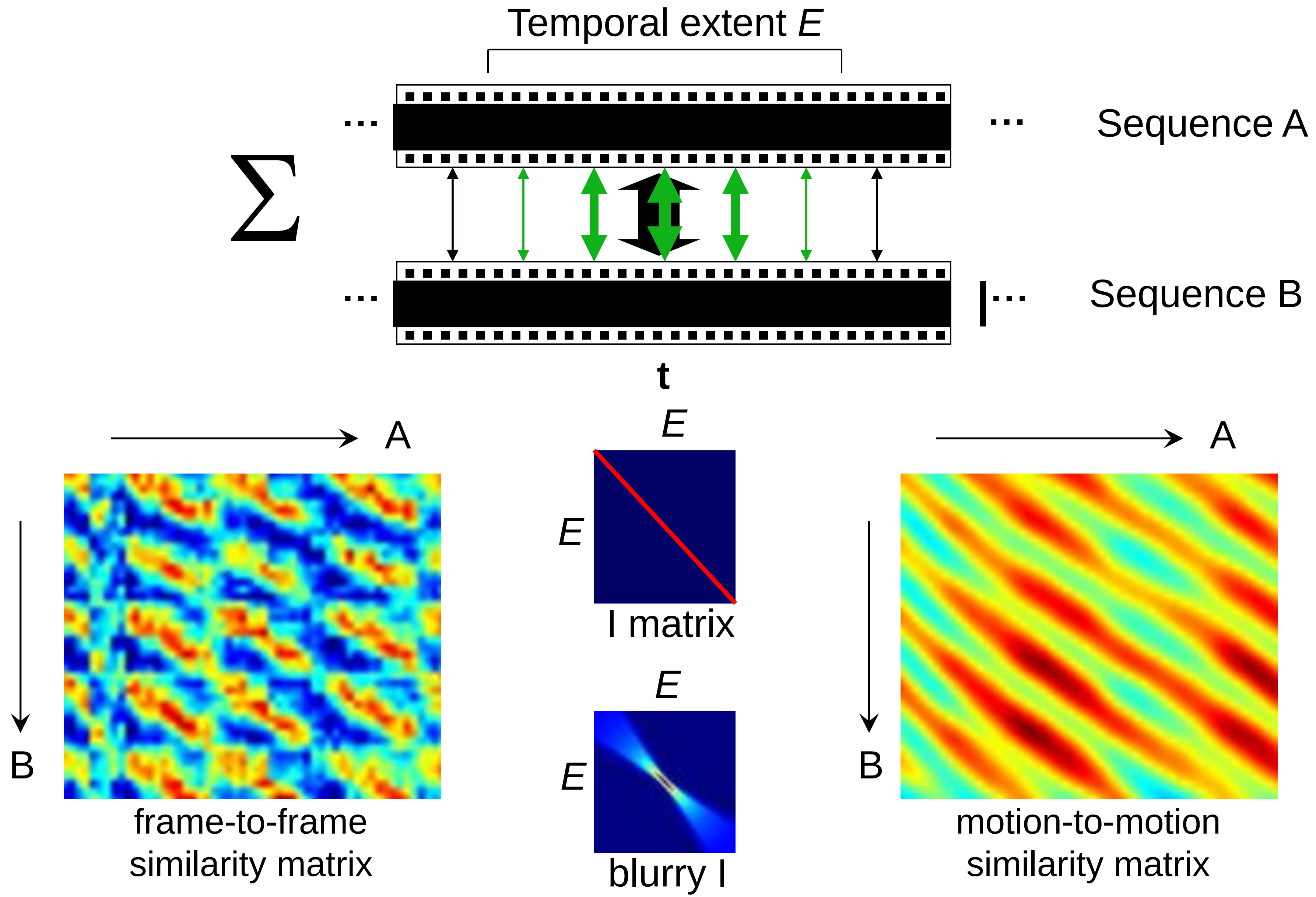


$F_x^-, F_x^+, F_y^-, F_y^+$



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

Spatio-Temporal Motion Descriptor



Slide credit: A. Efros

Football Actions: matching

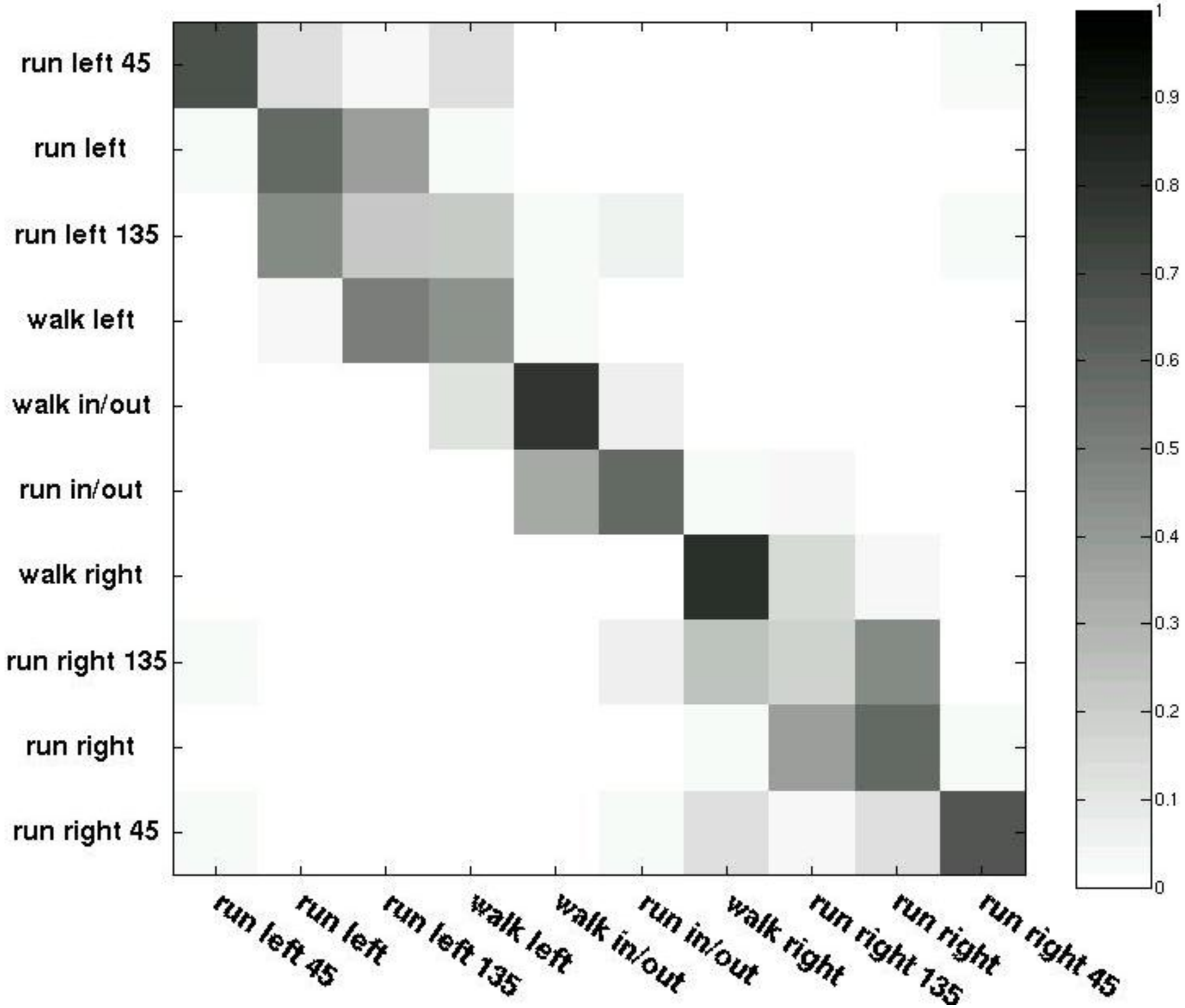
Input Sequence

Matched Frames

input matched

Slide credit: A. Efros

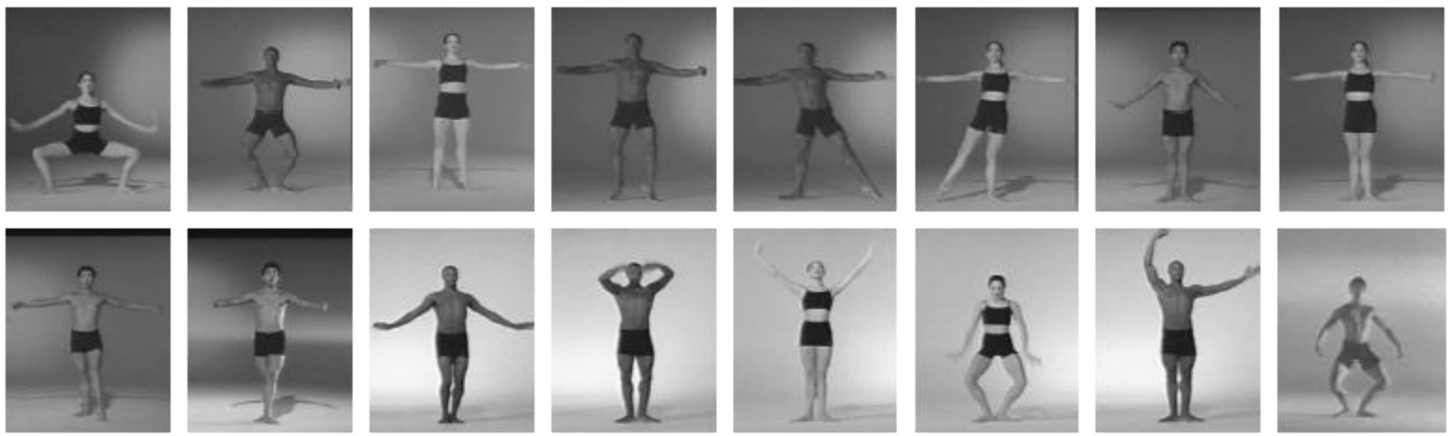
Football Actions: classification

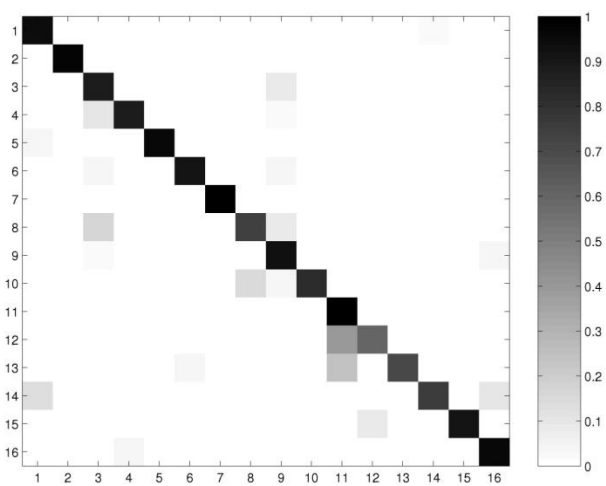


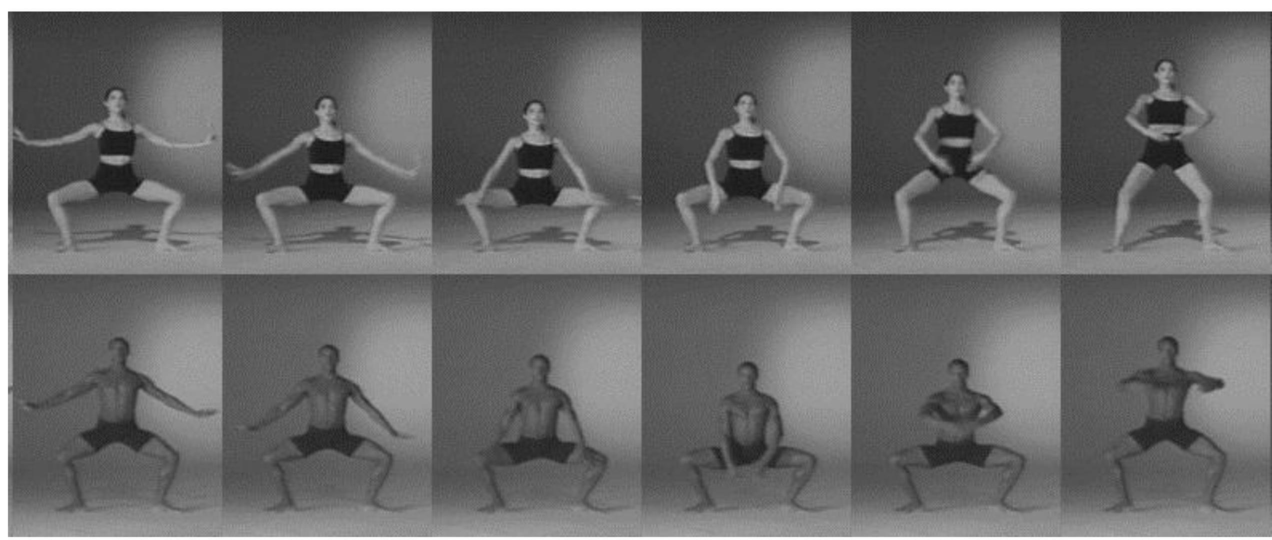
10 actions; 4500 total frames; 13-frame motion descriptor

Classifying Ballet Actions

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

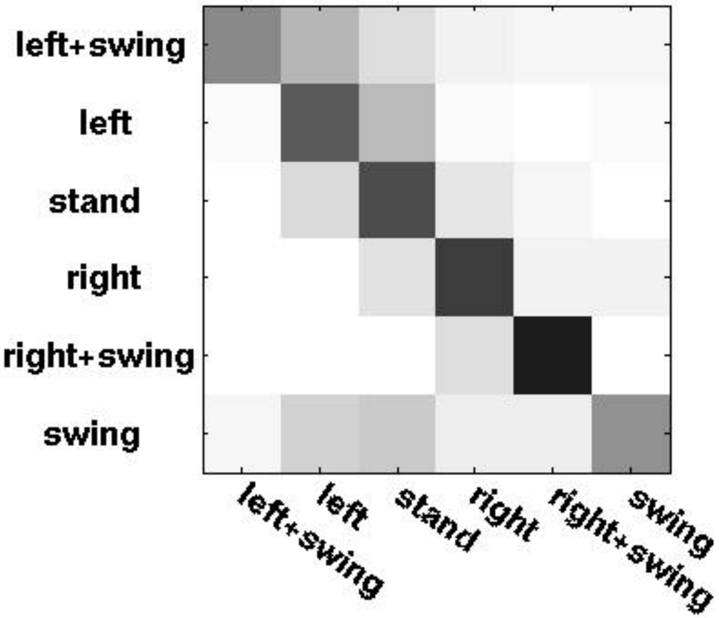






Classifying Tennis Actions

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

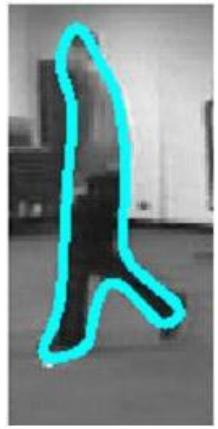


[Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, ICCV 2003]

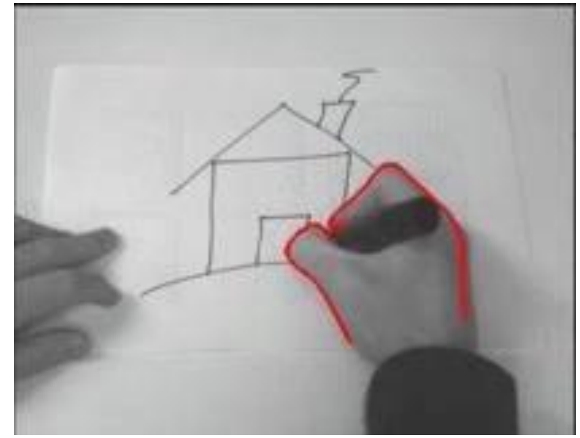
Where are we so far ?

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