Motion and Human Actions

Ivan Laptev
ivan.laptev@ens.fr
Equipe-projet WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d’Informatique, Ecole Normale Supérieure, Paris
Computer vision grand challenge: Video understanding

Objects: outdoors, building, car, glasses, people, etc...

Scene categories: indoors, outdoors, street scene, etc...

Actions: drinking, running, door exit, car enter, etc...

Geometry: Street, wall, field, stair, etc...

Constraints:
# Class overview

## Motivation
- Historic review
- Modern applications

## Overview of methods
Role of image measurements, prior knowledge and data association

<table>
<thead>
<tr>
<th>Methods I</th>
<th>Methods II</th>
<th>Methods III</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Silhouette methods</strong></td>
<td><strong>Optical Flow</strong></td>
<td><strong>Discriminative models</strong></td>
</tr>
<tr>
<td>FG/BG separation;</td>
<td>general OF, parametric</td>
<td>Boosted ST feature</td>
</tr>
<tr>
<td>Motion history images,</td>
<td>dense OF models,</td>
<td>models, realistic action</td>
</tr>
<tr>
<td>Human interfaces</td>
<td>articulated models</td>
<td>detection in movies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Deformable models</strong></td>
<td><strong>Space-time methods</strong></td>
<td><strong>Local features</strong></td>
</tr>
<tr>
<td>Active shape models,</td>
<td>ST-OF models, ST</td>
<td>Detectors, descriptors,</td>
</tr>
<tr>
<td>motion priors, particle</td>
<td>correlation, ST self-</td>
<td>matching, Bag of</td>
</tr>
<tr>
<td>filters, gesture</td>
<td>similarity, irregular</td>
<td>Features representations,</td>
</tr>
<tr>
<td>recognition</td>
<td>behavior</td>
<td>recognition</td>
</tr>
</tbody>
</table>
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

- 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

Giovanni Alfonso Borelli (1608–1679)
Motivation III: Study of motion

**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: Study of motion


- “Moving Light Displays” (LED) enable identification of familiar people and the gender and inspired many works in computer vision.

Human actions: Historic review

- 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: Animation

Motion Synthesis from Annotations
Okan Arikan, David A. Forsyth, James O'Brien, SIGGRAPH 2003
Modern applications: Animation

Motion Synthesis from Annotations
Okan Arikan, David A. Forsyth, James O'Brien, SIGGRAPH 2003
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*
Modern applications: Video editing

Recognizing Action at a Distance
Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, ICCV 2003
Applications: Human-Machine Interfaces

http://vismod.media.mit.edu/vismod/demos/kidsroom/kidsroom.html
Applications: Unusual Activity Detection

e.g. for surveillance

Detecting Irregularities in Images and in Video
Boimana & Irani, ICCV 2005
Applications: Search & Indexing

- **Video search**
  - TV & Web: e.g. “Fight in a parliament”
  - Home videos: e.g. “My daughter climbing”
  - Surveillance: suspicious behavior

Useful for TV production, entertainment, social studies, security,

- **Video mining**
  - e.g. Discover age-smoking-gender correlations now vs. 20 years ago

- **Auto-scripting (video2text)**
  
  JANE
  I need a father who’s a role model, not some horny geek-boy who’s gonna spray his shorts whenever I bring a girlfriend home from school.
  (snorts)
  What a lame-o. Somebody really should put him out of his misery.
Applications: Video Annotation

for video search, indexing, etc…

Learning realistic human actions from movies
Laptev, Marszalek, Schmid and Rozenfeld, CVPR 2008
How to recognize actions?
Action understanding: Key components

Image measurements
- Foreground segmentation
- Image gradient
- Optical flow
- Local space-time features

Association

Prior knowledge
- Deformable contour models
- 2D/3D body models
- Motion priors
  - Background models
  - Space-time templates
  - SVM classifiers
Image differencing: one of the simplest ways to measure motion/change

Better Background (BG) / Foreground (FG) separation methods are available:

- Modeling of color variation at each pixel with Gaussian Mixture Models (GMMs).
- Dominant motion estimation and compensation for sequences with moving camera
- Motion layer separation for scenes with non-static backgrounds
Foreground regions segmentation

Pros:
+ Simple and fast
+ Gives acceptable results under restricted conditions

Cons:
- Often unreliable due to shadows, low image contrast, etc.
- Requires background model => not well suited for scenes with dynamic BG and/or motion parallax
Temporal Templates of Bobick & Davis

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

Idea: summarize motion in video in a

*Motion History Image (MHI)*:

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

*The Recognition of Human Movement Using Temporal Templates*  
Aaron F. Bobick and James W. Davis, *PAMI* 2001
Temporal Templates of Bobick & Davis

- Compute MHI for each action sequence
- Describe each sequence with the translation and scale invariant vector of 7 Hu moments

\[ d = (m_{20}, m_{11}, m_{02}, m_{30}, m_{21}, m_{12}, m_{03})^\top \]

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

- Nearest Neighbor action classification with Mahalanobis distance between training and test descriptors \( d \).
Aerobics Dataset
Temporal Templates: Summary

Pros:
- Simple
- Fast

Cons:
- Assumes static camera, static background
- Sensitive to segmentation errors
- Silhouettes do not capture interior motion/shape

Possible improvements:
- Not all shapes are validRestrict the space of admissible shapes to overcome segmentation errors
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

\[ \Phi = (\phi_1 \mid \phi_2 \mid \ldots \mid \phi_t) \]

such that

\[ \mathbf{x} \approx \bar{\mathbf{x}} + \Phi \mathbf{b} \]

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

\[ S = \frac{1}{s-1} \sum_{i=1}^{s} (x_i - \bar{x})(x_i - \bar{x})^T \]

Eigenvectors

\[ \Phi = (\phi_1 | \phi_2 | \ldots | \phi_t) \]

Eigenvalues \( \lambda_1, \ldots, \lambda_t \)
Active Shape Models of Cootes et al.

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

Three main modes of lips-shape variation:

- $b = (\mu \lambda_1, 0, 0, ...)^\top$
- $b = (0, \mu \lambda_2, 0, 0, ...)^\top$
- $b = (0, 0, \mu \lambda_3, 0, 0, ...)^\top$

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

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

A small fraction of basis shapes (eigenvectors) accounts for the most of shape variation ($\Rightarrow$ landmarks are redundant)
Active Shape Models of Cootes et al.

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

  Given estimate of $x$, we can recover shape parameters $b$

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

- Projection onto the shape-space serves as a regularization

  $$x \quad \Rightarrow \quad b = \Phi^\top(x - \bar{x}) \quad \Rightarrow \quad x_{\text{reg}} = \bar{x} + \Phi b$$
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.

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

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

using similarity transformation

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

A simple way to estimate $T$ is to assign $(t_x, t_y)$ and $a$ to the mean position and the standard deviation of points in $x$ 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 $\bar{x}$, $\Phi$ have to be computed using normalized image point coordinates $x_{\text{norm}} = T^{-1}(x)$.
Active Shape Models of Cootes et al.

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

Example: face alignment

Illustration of face shape space

Active Shape Models: Their Training and Application
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 $b$:

\[ b_i^{(k)} = b_i^{(k-1)} + w_i^{k-1} \]

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

For similarity transformation $T$

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

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

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 apply specific motion priors for 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

From M. Isard and A. Blake, *ICCV* 1998
Incorporating motion priors

**Image measurements**
- Foreground segmentation
- Image gradient
- Optical Flow

**Data Association**
- Particle filters

**Prior knowledge**
- Learning motion models for different actions

---

Incorporating motion priors
Bayesian Tracking

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

Notation:
- $Z_i$: image data at time $i$
- $X_i$: model parameters at time $i$ (e.g. shape and its dynamics)
- $p(X_i)$: prior density for $X_i$
- $p(Z_i|X_i)$: likelihood of data for the given model configuration

We search posterior defined by the Bayes’ rule

$$p(X|Z) \propto p(Z|X)p(X)$$

For tracking the Markov assumption gives the prior

$$p(X_i|X_{i-1})$$

Temporal update rule:

$$p(X_i|Z_i) \propto p(Z_i|X_i)p(X_i|X_{i-1})$$
Kalman Filtering

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

\[ p(X_{i-1}) \]

\[ p(x) \]

\[ p(z) \]

\[ p(x) \]

\[ p(z|x) \]

\[ p(x|z) \]

\[ p(x|x_{i-1}) \]

\[ p(X_i|Z_i) \propto p(Z_i|X_i)p(X_i|X_{i-1}) \]
Particle Filtering

In reality probability densities are almost always multi-modal

\[ p(X_{i-1}) \]

deterministic drift

\[ p(z) \]

\[ p(x) \]

stochastic diffusion

\[ p(z|X_i) \propto p(Z_i|X_i)p(X_i|X_{i-1}) \]

reactive effect of measurement

\[ p(X_i|X_{i-1}) \]
Particle Filtering

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

Approximate distributions with weighted particles
Particle Filtering

Tracking examples:

\[ X \text{ describes leave shape} \quad X \text{ describes head shape} \]

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

- Dynamic model: 2\textsuperscript{nd} order Auto-Regressive Process

\[
\mathcal{X}_k = \begin{pmatrix} \mathbf{X}_{k-1} \\ \mathbf{X}_k \end{pmatrix}
\]

State

Update rule:

\[
\mathcal{X}_k - \bar{\mathcal{X}} = A(\mathcal{X}_{k-1} - \bar{\mathcal{X}}) + B\mathbf{w}_k
\]

Model parameters:

\[
A = \begin{pmatrix} 0 & I \\ A_2 & A_1 \end{pmatrix}, \quad \bar{\mathcal{X}} = \begin{pmatrix} \bar{\mathbf{X}} \\ \bar{\mathbf{X}} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ B_0 \end{pmatrix}
\]

Learning scheme:

1. **Shape Space**
2. **Hand-built dynamics**
3. **Training sequence slow, clutter-free**
4. **Fast test sequences**
5. **Faster training sequence**
6. **Infer dynamical model**
7. **Iterate**
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 \( \chi_k^+ \) = \( \begin{pmatrix} \chi_k \\ y_k \end{pmatrix} \)

Transition probability matrix
\[
P(y_k = j | y_{k-1} = i) = T_{i,j},
\]
or more generally
\[
P(y_k = j | y_{k-1} = i, \chi_{k-1}) = T_{i,j}(\chi_{k-1})
\]

Incorporation of the mixed-state model into a particle filter is straightforward, simply use \( \chi_k^+ \) instead of \( \chi_k \) and the corresponding update rules.
Dynamics with discrete states

Example: Drawing

Transition probability matrix

\[
T = \begin{pmatrix}
0.9800 & 0.0015 & 0.0185 \\
0.0850 & 0.9000 & 0.0150 \\
0.0050 & 0.0150 & 0.9800 \\
\end{pmatrix}
\]

Result: simultaneously improved tracking and gesture recognition

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

A probabilistic framework for matching temporal trajectories: CONDENSATION-based recognition of gestures and expressions
M.J. Black and A.D. Jepson, ECCV 1998
So far...

**Image measurements**
- Foreground segmentation
- Hu moments and Fourier descriptors
- Image edges

**Data Association**
- Particle filters
- NN classifiers

**Prior knowledge**
- Background models
- Temporal templates
- Deformable shape models
- Motion priors