2d-3D Relations

Cristian Sminchisescu, TTI-C/ UToronto

People Tracking Tutorial, CVPR 2006
David Forsyth, Deva Ramanan, Cristian Sminchisescu
Inferring 3D from 2D

• History
• Monocular vs. multi-view analysis
• Difficulties
  – structure of the solution and ambiguities
  – static and dynamic ambiguities
• Modeling frameworks for inference and learning
  – top-down (generative, alignment-based)
  – bottom-up (discriminative, predictive, exemplar-based)
  – Learning joint models
• Take-home points
History of Analyzing Humans in Motion

• Markers (*Etienne Jules Marey*, 1882)

• Multiple Cameras (*Eadweard Muybridge*, 1884)
Human motion capture today
120 years and still fighting …

• VICON ~ 100,000 $
  – Excellent performance,
    *de-facto* standard for special
effects, animation, *etc*

• But heavily instrumented
  – Multiple cameras
  – Markers in order to simplify the image correspondence
  – Special room, simple background

**Major challenge:** Move from the laboratory to the real world
What is so different between multi-view and single-view analysis?

• Different emphasis on the relative importance of measurement and prior knowledge
  – Depth ambiguities
  – Self-occluded body parts

• Similar techniques at least one-way
  – Transition monocular->multiview straightforward

• Monocular as the `robust limit’ of multi-view
  – Multiple cameras unavailable, or less effective in real-world environments due to occlusion from other people, objects, etc.
3D Human Motion Capture Difficulties

- General poses
- Self-occlusions
- Difficult to segment the individual limbs
- Loss of 3D information in the monocular projection
- Partial Views
- Several people, occlusions
- Reduced observability of body parts due to loose fitting clothing
- Accidental alignments
- Motion blur

- Different body sizes
Levels of 3d Modeling

This section

- Coarse body model
  - 30 - 35 d.o.f
  - Simple appearance (implicit texture map)

- Complex body model
  - 50 - 60 d.o.f
  - Simple appearance (edge histograms)

- Complex body model
  - ? (hundreds) d.o.f
  - Sophisticated modeling of clothing and lighting
Difficulties

• High-dimensional state space (30-60 dof)

• Complex appearance due to articulation, deformation, clothing, body proportions

• Depth ambiguities and self-occlusion

• Fast motions, only vaguely known a-priori
  – External factors, objects, sudden intentions…

• Data association (what is a human part, what is background – see the Data Association section)
Difficulties, more concretely

- **Depth ambiguities**
- **Occlusions** (missing data)
- **Left arm**
- **Preservation of physical constraints**
- **Data association ambiguities**
- **Left / right leg ?**
Articulated 3d from 2d Joint Positions

Structure of the monocular solution: Lee and Chen, CVGIP 1985 (!)

• Characterizes the space of solutions, assuming
  – 2d joint positions + limb lengths
  – internal camera parameters

• Builds an interpretation tree of projection-consistent hypotheses (3d joint positions)
  – obtained by forward-backward flips in-depth
  – $O(2^{\# \text{ of body parts}})$ solutions
  – In principle, can prune some by physical reasoning
  – But no procedure to compute joint angles, hence difficult to reason about physical constraints

• Not an automatic 3d reconstruction method
  – select the true solution (out of many) manually

• Adapted for orthographic cameras (Taylor 2000)
Why is 3D-from-monocular hard? <v>

Static, Kinematic, Pose Ambiguities

- Monocular static pose optima
  - \( \sim 2^{\text{Nr of Joints}} \), some pruned by physical constraints
  - Temporally persistent

Sminchisescu and Triggs '02
Trajectory Ambiguities $<v>$

General smooth dynamics

Model / image | Filtered | Smoothed

Sminchisescu and Jepson '04

2 (out of several) plausible trajectories
Visual Inference in a 12d Space

*6d rigid motion + 6d learned latent coordinate*

**Interpretation #1**
Points at camera when conversation ends (before the turn)

**Interpretation #2**
Says `salut' when conversation ends (before the turn)
Trajectory Ambiguities <v>

Learned latent space and smooth dynamics

Interpretation #1
Says `salut’ when conversation ends
(before the turn)

Interpretation #2
Points at camera when conversation ends
(before the turn)

• Image consistent, smooth, typically human…

Sminchisescu and Jepson ‘04
The Nature of 3D Ambiguities

- Persistent over long time-scales (each S-branch)
- Loops (a, b, c) have limited time-scale support, hence ambiguity cannot be resolved by extending it
Generative vs. Discriminative Modelling

- $x$ is the model state
- $r$ are image observations

Goal: $p_\theta(x | r)$

$\theta$ are parameters to learn

given training set of $(r,x)$ pairs

\[ p_\theta(x | r) \propto p_\theta(r | x) \cdot p(x) \]

- Predict state distributions from image features
- Learning to `invert` perspective projection and kinematics is difficult and produces multiple solutions
  - Multivalued mappings $\equiv$ multimodal conditional state distributions
- Temporal extensions necessary

- Optimize alignment with image features
- Can learn state representations, dynamics, observation models; but difficult to model human appearance
- State inference is expensive, need effective optimization
Temporal Inference (tracking)

- Generative (top-down) chain models
  (Kalman Filter, Extended KF, Condensation)

- Discriminative (bottom-up) chain models
  (Conditional Bayesian Mixture Of Experts Markov Model - BM³E, Conditional Random Fields - CRF, Max. Entropy Models - MEMM)
Temporal Inference

- $x_t$: state at time $t$
- $O_t = (o_1, o_2, ..., o_t)$: observations up to time $t$
- $p(x_t | O_t)$

$p(x_{t+1} | O_t)$
$p(o_{t+1} | x_{t+1})$
$p(x_{t+1} | O_{t+1})$

cf. CONDENSATION, Isard and Blake, 1996
Generative / Alignment Methods

- Modeling
- Methods for temporal inference
- Learning low-dimensional representations and parameters
Model-based Multiview Reconstruction

Kehl, Bray and van Gool ‘05

- Body represented as a textured 3D mesh
- Tracking by minimizing distance between 3d points on the mesh and volumetric reconstruction obtained from multiple cameras
Generative 3D Reconstruction
Annealed Particle Filter
(Deutscher, Blake and Reid, ‘99-01)

Improved results (complex motions) when multiple cameras (3-6) were used

Careful design

- Dynamics
- Observation likelihood
  - edge + silhouettes
- Annealing-based search procedure, improves over particle filtering
- Simple background and clothing
Generative 3D Reconstruction

Sidenbladh, Black and Fleet, ’00-02; Sigal et al ‘04

**Monocular**
- Condensation-based filter
- Dynamical models
  - walking, snippets
- Careful learning of observation likelihood distributions

**Multi-camera**
- Non-parametric belief propagation, initialization by limb detection and triangulation
Kinematic Jump Sampling
Sminchisescu & Triggs ’01,’03

\[ C \rightarrow C=SelectSamplingChain(m_i) \rightarrow \]
Candidate Sampling Chains

\[ C \rightarrow C=SelectSamplingChain(m_i) \rightarrow \]

\[ s=CovarianceScaledSampling(m_i) \]

\[ T=BuildInterpretationTree(s,C) \]

\[ E=InverseKinematics(T) \]

Prune and locally optimize \( E \)

\[ m_i=(\mu_i, \Sigma_i) \]

\[ p^{t-1} \]
Kinematic Jump Sampling <v>
What can we learn?

• Low-dimensional perceptual representations, dynamics (unsupervised)
  – What is the intrinsic model dimensionality?
  – How to preserve physical constraints?
  – How to optimize efficiently?

• Parameters (typically supervised)
  – Observation likelihood (noise variance, feature weighting)
  – Can learn separately (easier) but how well we do?
  – Best to learn by doing (i.e. inference)
    • Maximize the probability of the right answer on the training data, hence learning = inference in a loop
    • Need efficient inference methods
Intrinsic Dimension Estimation <v> and Latent Representation for Walking

- 2500 samples from motion capture
- The Hausdorff dimension ($d$) is effectively 1, lift to 3 for more flexibility
- Use non-linear embedding to learn the latent 3d space embedded in an ambient 30d human joint angle space

\[ d = \lim_{r \to 0} \frac{\log N(r)}{\log(1/r)} \]

Sminchisescu and Jepson ‘04
3D Model-Based Reconstruction

(Urtasun, Fleet, Hertzmann and Fua’05)

- Track human joints using the \textit{WSL} tracker (Jepson et al’01)
- Optimize model joint re-projection error in a low-dimensional space obtained using probabilistic PCA (Lawrence’04)
Learning Empirical Distribution of Edge Filter Responses

(Original slide courtesy of Michael Black)

\[ p_{on}(F) \]

\[ p_{off}(F) \]

Likelihood ratio, \( p_{on}/p_{off} \), used for edge detection

Geman & Jedynak and Konishi, Yuille, & Coughlan
Learning Dependencies

(original slide courtesy of Michael Black); Roth, Sigal and Black ’04

Filter responses are not conditionally independent
Leaning by Maximum Entropy
The effect of learning on the trajectory distribution

**Before**
- Learn body proportions + parameters of the observation model (weighting of different feature types, variances, *etc*)
- Notice reduction in uncertainty
- The ambiguity diminishes significantly but does not disappear

**After**

*Sminchisescu, Welling and Hinton ‘03*
Conditional /Discriminative/ Indexing Methods

- Nearest-neighbor, snippets
- Regression
- Mixture of neural networks
- Conditional mixtures of experts
- Probabilistic methods for temporal Integration
**Discriminative 3d: Nearest Neighbor Parameter Sensitive Hashing (PSH)**

Shakhnarovich, Viola and Darell '03

- Relies on database of (observation, state) pairs rendered artificially
  - Locates samples that have observation components similar to the current image data (nearest neighbors) and use their state as putative estimates

- Extension to multiple cameras and tracking by non-linear model optimization (PSH used for initialization Demirdjian et al, ICCV05)
  - Foreground / background segmentation from stereo
Discriminative 3d: Nearest Neighbor Matching

2D->3D Pose + Annotation

Ramanan and Forsyth '03

Annotations

{run, walk, wave, etc.}

Motion Synthesizer

match 1/2 second clips of motion

original video

model build

detect

2D track

3D pose and annotation

3D motion library

StandWave

Discriminative 3d: Nearest Neighbor Matching

2D->3D Pose + Annotation

Ramanan and Forsyth '03

Annotations

{run, walk, wave, etc.}

Motion Synthesizer

match 1/2 second clips of motion

original video

model build

detect

2D track

3D pose and annotation

3D motion library

StandWave
2D->3D pose + annotation <v>

Ramanan and Forsyth’03
**Discriminative 3d: Regression Methods**

*Aggarwal and Triggs ‘04, Elgammal & Lee ‘04*

- *(A&T)* 3d pose recovery by non-linear regression against silhouette observations represented as shape context histograms
  - Emphasis on sparse, efficient predictions, good generalization

- *(A&T)* Careful study of dynamical regression-based predictors for walking and extensions to mixture of regressors *(HCI’05)*

- *(E&L)* pose from silhouette regression where the dimensionality of the input is reduced using non-linear embedding
  - Latent (input) to joint angle (output) state space map based on RBF networks
Discriminative 3d:
Specialized Mappings Architecture
Rosales and Sclaroff ‘01

- Static 3D human pose estimation from silhouettes (Hu moments)
- Approximates the observation-pose mapping from training data
  - Mixture of neural networks
  - Models the joint distribution
- Uses the forward model (graphics rendering) to verify solutions
**Conditional Bayesian Mixtures of Experts**

- A single expert cannot represent multi-valued relations.
- Multiple experts can focus on representing parts of the data.
- But the expert contribution (importance) is contextual.
  - Disregarding context introduces systematic error (invalid extrapolation).
- The experts need observation-sensitive mixing proportions.
Discriminative Temporal Inference

BM³E = Conditional Bayesian Mixture of Experts Markov Model

• `Bottom-up’ chain

\[ p(x_t | R_t) = \int p(x_t | x_{t-1}, r_t) p(x_{t-1} | R_{t-1}) \, dx_{t-1}, \quad \text{where } R_t = (r_1, ..., r_t) \]

Local conditional \hspace{3cm} Temporal (filtered) prior

Conditions on the observation

• The temporal prior is a Gaussian mixture
• The local conditional is a Bayesian mixture of Gaussian experts
• Integrate pair-wise products of Gaussians analytically

Sminchisescu, Kanaujia, Li, Metaxas ‘05
Turn during Dancing \(<v>\)

Notice imperfect silhouettes

Sminchisescu, Kanaujia, Li, Metaxas ‘05
Washing a Window "v"

Sminchisescu, Kanaujia, Li, Metaxas ‘05
The pose prediction problem is highly structured
– Human joint angles are correlated, not independent
– Learn conditional mixtures between low-dimensional spaces decorrelated using kernel PCA (kBME)

RVM – Relevance Vector Machine
KDE – Kernel Dependency Estimator
Low-dimensional Discriminative Inference

(translation removed for better comparison)
Evaluation on artificially generated silhouettes with 3d ground truth

*(average error / average maximum error, per joint angle, in degrees)*

| Sequence       | \( p(x_t|r_t) \) | \( p(x_t|x_{t-1},r_t) \) |
|----------------|-------------------|---------------------------|
|                | NN                | RVM                       | BME                        | NN    | RVM  | BME  |
| Normal walk    | 4 / 20            | 2.7 / 12                  | 2 / 10                     | 7 / 25 | 3.7 / 11.2 | 2.8 / 8.1 |
| Complex walk   | 11.3 / 88         | 9.5 / 60                  | 4.5 / 20                   | 7.5 / 78 | 5.67 / 20 | 2.77 / 9   |
| Running        | 7 / 91            | 6.5 / 84                  | 5 / 94                     | 5.5 / 91 | 5.1 / 108 | 4.5 / 76   |
| Conversation   | 7.3 / 26          | 5.5 / 21                  | 4.15 / 9.5                 | 8.14 / 29 | 4.07 / 16 | 3 / 9      |
| Pantomime      | 7 / 36            | 7.5 / 53                  | 6.5 / 25                   | 7.5 / 49 | 7.5 / 43  | 7 / 41     |

- NN = nearest neighbor
- RVM = relevance vector machine
- BME = conditional Bayesian mixture of experts
Evaluation, low-dimensional models

(average error / joint angle in degrees)

<table>
<thead>
<tr>
<th>Activity</th>
<th>KDE-RR</th>
<th>RVM</th>
<th>KDE-RVM</th>
<th>BME</th>
<th>kBME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk and turn back</td>
<td>10.46</td>
<td>4.95</td>
<td>7.57</td>
<td>4.27</td>
<td>4.69</td>
</tr>
<tr>
<td>Conversation</td>
<td>7.95</td>
<td>4.96</td>
<td>6.31</td>
<td>4.15</td>
<td>4.79</td>
</tr>
<tr>
<td>Run and turn left</td>
<td>5.22</td>
<td>5.02</td>
<td>6.25</td>
<td>5.01</td>
<td>4.92</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Activity</th>
<th>KDE-RR</th>
<th>KDE-RVM</th>
<th>kBME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk and Turn</td>
<td>7.59</td>
<td>7.15</td>
<td>3.72</td>
</tr>
<tr>
<td>Run and Turn</td>
<td>17.7</td>
<td>16.08</td>
<td>8.01</td>
</tr>
</tbody>
</table>

- KDE-RR=ridge regressor between low-dimensional spaces
- KDE-RVM=RVM between low-dimensional spaces
  - Unimodal methods average competing solutions
- kBME=conditional Bayesian mixture between low-dimensional state and observation spaces
  - Training and inference is about 10 time faster
Self-supervised Learning of a Joint Generative-Recognition Model

- Maximize the probability of the (observed) evidence (e.g. images of humans)

\[
\log p_\theta(r) = \log \int_x Q_v(x | r) \frac{p_\theta(x, r)}{Q_v(x | r)} \geq \int_x Q_v(x | r) \log \frac{p_\theta(x, r)}{Q_v(x | r)} = KL(Q_v(x | r) \parallel p_\theta(x, r))
\]

\[
\log p_\theta(r) - KL(Q_v(x | r) \parallel p_\theta(x | r)) = KL(Q_v(x | r) \parallel p_\theta(x, r))
\]

- Hence, the KL divergence between what the generative model \( p \) infers and what the recognition model \( Q \) predicts, with tight bound at

\[
Q_v(x | r) = p_\theta(x | r)
\]
Self-supervised Learning of a Joint Generative-Recognition Model

Algorithm for Bidirectional Model Learning

E-step: $\nu^{k+1} = \arg \max_{\nu} \mathcal{L}(\nu, \theta^k)$
Train the recognition model using samples from the current generative model.

M-step: $\theta^{k+1} = \arg \max_{\theta} \mathcal{L}(\nu^{k+1}, \theta)$
Train the generative model to have state posterior close to the one predicted by the current recognition model.

- Local optimum for parameters
- Recognition model is a conditional mixture of data-driven mean field experts
  - Fast expectations, dimensions decouple
Generalization under clutter

Sminchisescu, Kanaujia, Metaxas '06
Take home points

• Multi-view 3d reconstruction reliable in the lab
  – Measurement-oriented
  – Geometric, marker-based
    • correspondence + triangulation
  – Optimize multi-view alignment
    • generative, model-based
  – Data-association in real-world (occlusions) open

• Monocular 3d as robust limit of multi-view
  – Difficulties: depth perception + self-occlusion
  – Stronger dependency on efficient non-convex optimization and good observation models
  – Increased emphasis on prior vs. measurement
Take home points (contd.)

• Top-down / Generative / Alignment Models
  – Flexible, but difficult to model human appearance
  – Difficult optimization problems, local optima
  – Can learn constrained representations and parameters
    • Can handle occlusion, faster search (low-d)
    • Fewer local optima -- the best more likely true solutions

• Discriminative / Conditional / Exemplar–based Models
  – Need to model complex multi-valued relations
  – Replace inference with indexing / prediction
  – Good for initialization, recovery from failure, on-line
  – Still need to deal with segmentation / data association
For more information

http://www.cs.toronto.edu/~crismin
http://ttic.uchicago.edu/~crismin

crismin@cs.toronto.edu
crismin@nagoya.uchicago.edu

For a tutorial on 3d human sensing from monocular images, see:

http://www.cs.toronto.edu/~crismin/PAPERS/tutorial_iccv05.pdf
Cristian Sminchisescu thanks the following collaborators on various elements of the research described in the 2d-3d relations section

- Geoffrey Hinton (Toronto)
- Allan Jepson (Toronto)
- Atul Kanaujia (Rutgers)
- Zhiguo Li (Rutgers)
- Dimitris Metaxas (Rutgers)
- Bill Triggs (INRIA)
- Max Welling (UCIrvine)