Monocular Human Motion Capture with a Mixture of Regressors

Ankur Agarwal and Bill Triggs
GRAVIR-INRIA-CNRS, Grenoble, France

IEEE Workshop on Vision for Human-Computer Interaction, 21 June 2005

Introduction

Goal

- To recover 3D human body pose from **monocular** image silhouettes

  - 3D pose = joint angles

  - Use either individual images or video sequences

Applications

- Human computer interaction
- Gesture interpretation / activity recognition
- Markerless motion capture
- Visual surveillance

“Model Free” Learning Based Approach

- No explicit 3D model — recovers 3D pose directly from robust silhouette descriptors
- Human motion capture data used to train mixture of kernel regressors

Advantages

- No need to build an explicit 3D model
- Easily adaptable to different appearances/people; possibly more robust
- Motion capture data captures **typical human movements**, not just *kinematically possible* ones

Disadvantages

- Harder to interpret than explicit model, and may be less accurate

Regression based pose estimation from Silhouettes

- Robustly encode silhouette shape using vector quantized local Shape Context Histograms ($\mathbf{z}$)
- Denote the output (3D pose) as a vector ($\mathbf{x}$)
- Learn a regressive mapping $\mathbf{x} \sim \mathbf{r}(\mathbf{z}) \equiv \mathbf{A}\,\phi(\mathbf{z}) + \mathbf{b}$

  $\mathbf{A}$: matrix of weight vectors, $\phi(\mathbf{z})$: vector of scalar basis functions
- Can allow for robustness/sparseness with the use of SVM/RVM ...

Training data

- Real silhouettes from motion capture videos supplemented with synthetic silhouettes from several human body models

Ambiguities in static pose reconstruction

- The silhouette ($\mathbf{z}$) to pose ($\mathbf{x}$) problem is inherently **multi-valued**.
- Treating it as a function can lead to averaging or zig-zagging between different solutions.

Multimodal Pose Estimation

- Introduce a **discrete latent variable** $k \in \{1, 2 \ldots K\}$ to encode the information missing in the silhouette.

- Assume a **mixture of experts** model based on K underlying functional regression rules $\mathbf{x} \sim \mathbf{r}_k(\mathbf{z})$:

$$p(\mathbf{x} \,|\, \mathbf{z}) = \sum_{k=1}^{K} p(\mathbf{x} \,|\, \mathbf{z}, k)\, p(k \,|\, \mathbf{z})\,, \qquad p(\mathbf{x} \,|\, \mathbf{z}, k) = \mathcal{N}(\mathbf{r}_k(\mathbf{z}), \mathbf{\Lambda}_k)$$

- Linear regressor within each component k

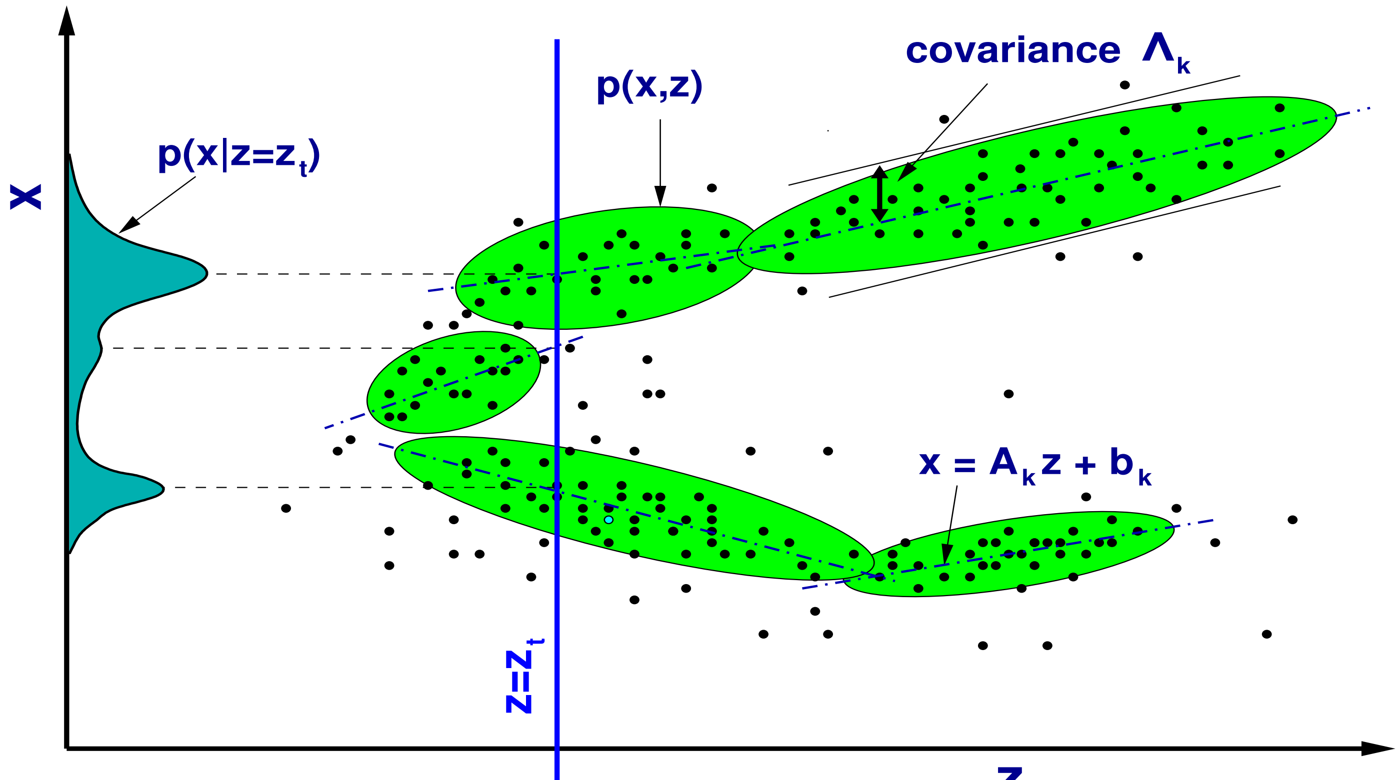
$$\mathbf{r}_k(\mathbf{z}) \;\equiv\; \mathbf{A}_k\, \boldsymbol{\phi}(\mathbf{z}) + \mathbf{b}_k$$

- Obtain multimodal probabilistic solutions in 3D pose space

Mixture of Regressors

- Fit a **mixture of regressive Gaussians** to the joint density $(\phi(\mathbf{z}), \mathbf{x})$:

$$\begin{pmatrix} \phi(\mathbf{z}) \\ \mathbf{x} \end{pmatrix} \simeq \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Gamma}_k)$$



Mixture of Regressors (contd.)

- Special covariance structure enforces "regressive" noise model

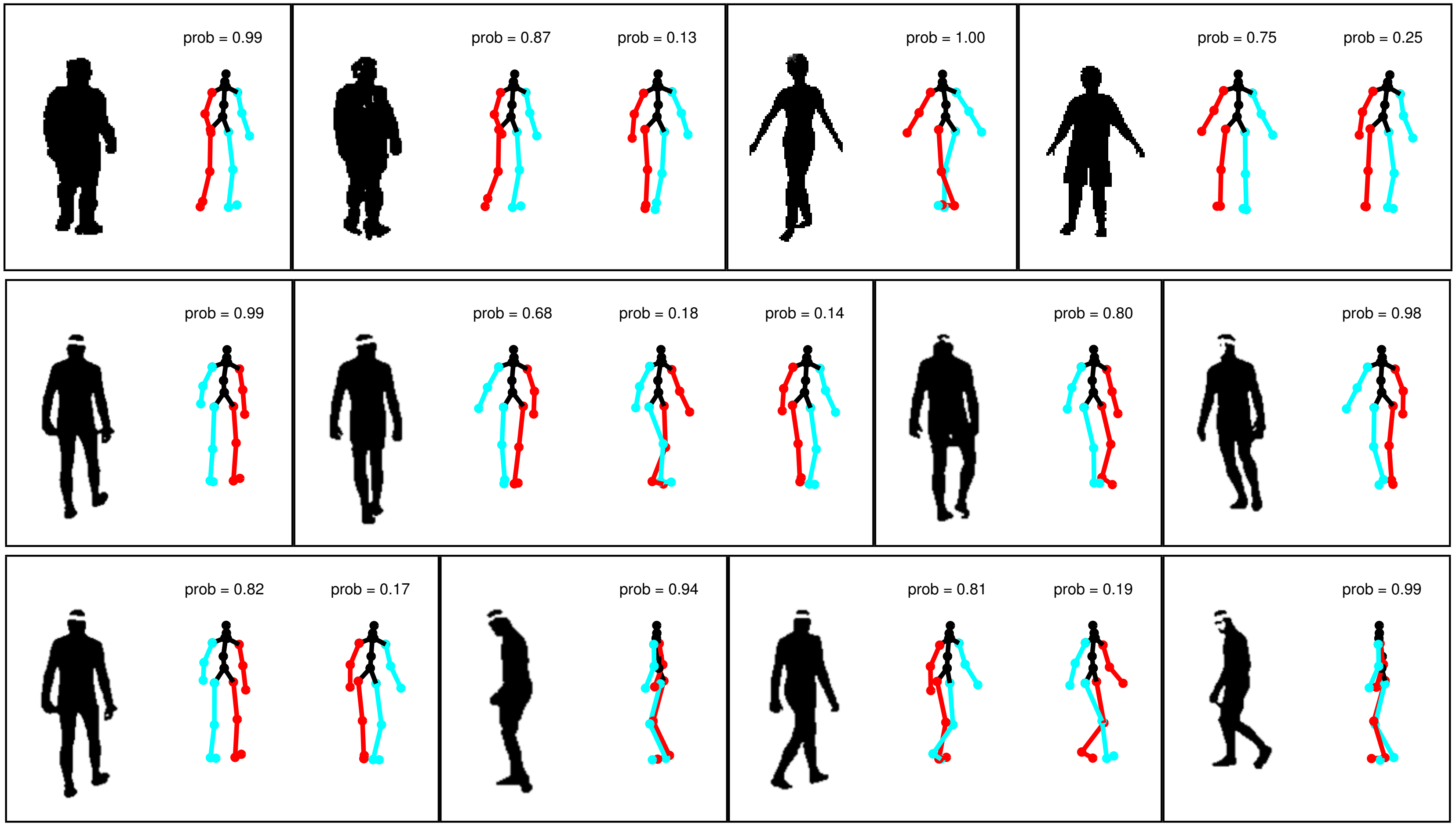
$$\boldsymbol{\mu}_k = \begin{pmatrix} \phi(\bar{\mathbf{z}}_k) \\ \mathbf{r}_k(\bar{\mathbf{z}}_k) \end{pmatrix}, \boldsymbol{\Gamma}_k = \begin{pmatrix} \boldsymbol{\Sigma}_k & \boldsymbol{\Sigma}_k \mathbf{A}_k^\top \\ \mathbf{A}_k \boldsymbol{\Sigma}_k & \mathbf{A}_k \boldsymbol{\Sigma}_k \mathbf{A}_k^\top + \boldsymbol{\Lambda}_k \end{pmatrix}$$

- Parameters learned using **Expectation Maximization**

M-step: Estimate $\mathbf{A}_k, \mathbf{b}_k$ by weighted least squares regression, Λ_k from residual errors. Compute μ_k, Σ_k, π_k for each class.

E-step: Reestimate class membership weights for each point.

Multimodal Pose Estimation from Static Images



- Provides multiple solutions for pose, with corresponding probabilities
- Most cases of ambiguity are identified

	% of frames with m solutions			Error in the top solution	Error in best of top 4 solutions
	$m = 1$	$m = 2$	$m \geq 3$		
Test person	62	28	10	6.14°	4.84°
Test motion	65	28	6	7.40°	5.37°
Train subset	72	23	5	6.14°	4.55°

Numbers of solutions and RMS joint angle reconstruction errors for 3 test sequences.

Tracking with automatic (re)initialization

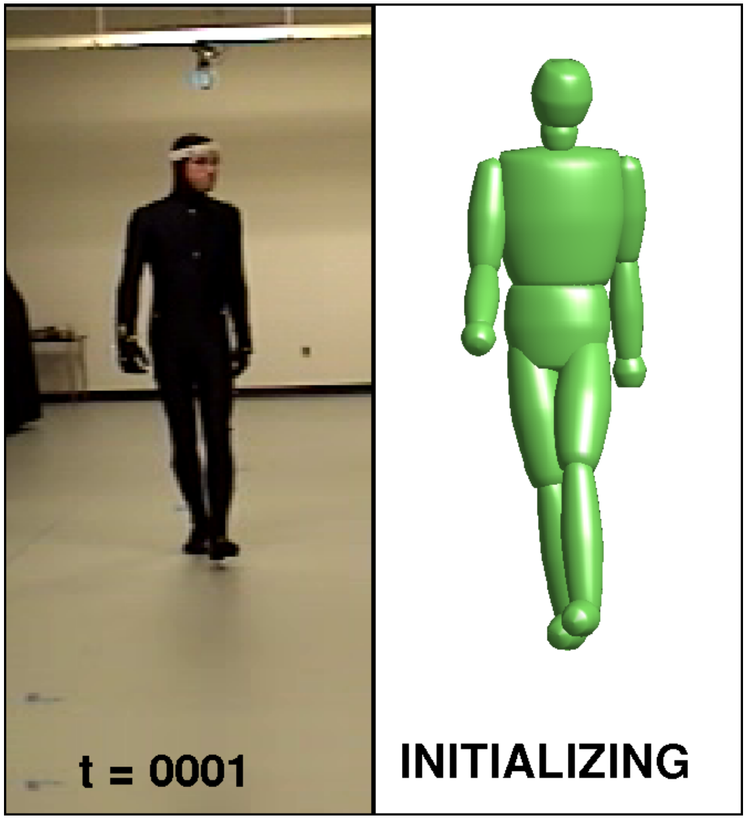
- Particle filter tracker, samples from dynamics $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ as usual
- Uses regressive mixture $p(\mathbf{x}_t \mid \mathbf{z}_t)$ to assign posterior particle weights
- (Re)initializes by sampling from full mixture

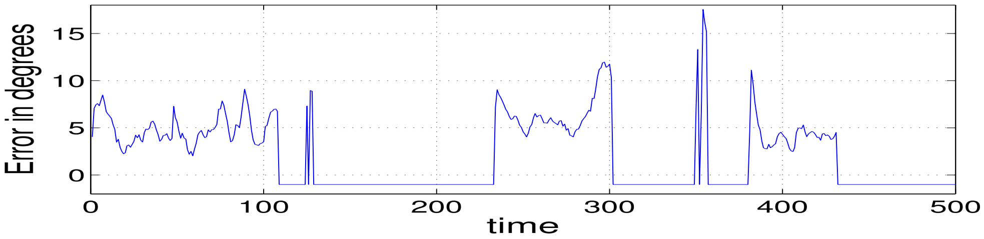
$$p(\mathbf{x}_0 \mid \mathbf{z}_0) = \sum_{k=1}^{K} p(k \mid \mathbf{z}_0).\ \mathcal{N}(\mathbf{r}_k(\mathbf{z}_0), \mathbf{\Lambda}_k)$$

- Potentially real time owing to closed form solution for posterior.

Self-Initializing 3D Tracking

Detects the presence of a person and decides whether to wait, initialize or track using observed silhouette shape.





Upper Body Gesture Recognition

- Associate (by hand) different mixture components with gestures
- Use posterior class probabilities to identify action

Training gestures (Basketball signals)

—

Traveling

Illegal dribble

Illegal defense

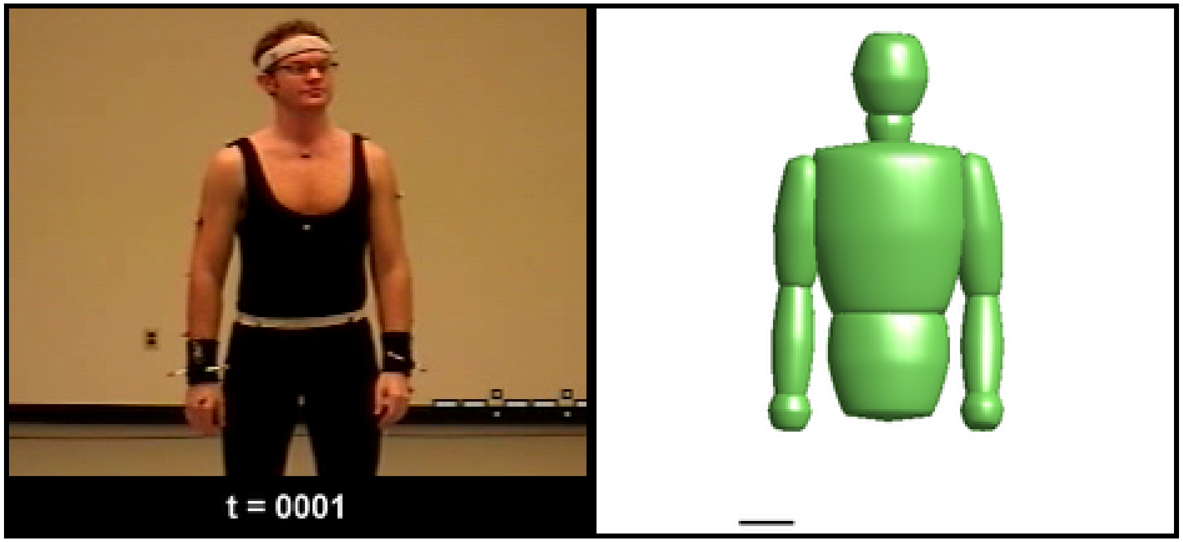
No score

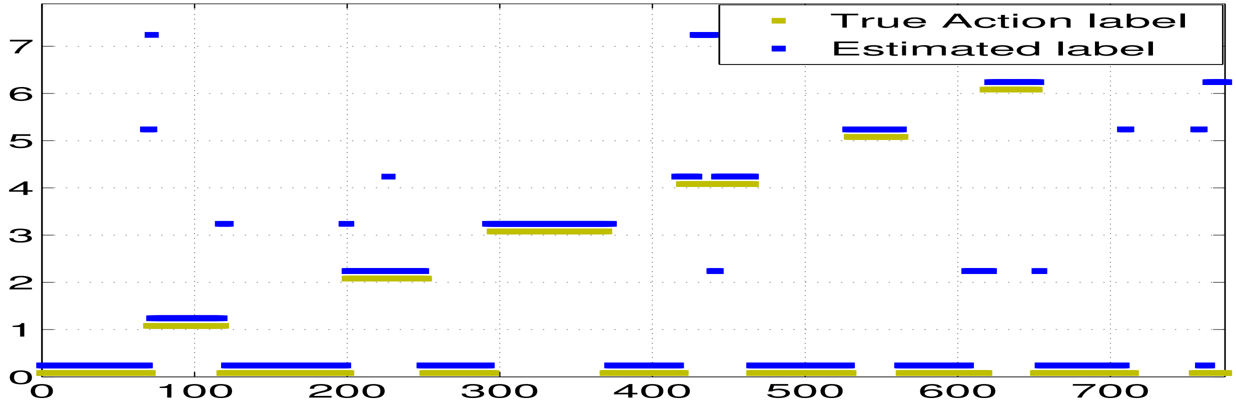
Stop clock

Hand check

Technical foul

Tracking and labelling gestures





Conclusion

- “Model free” methods for recovering 3D human pose from monocular silhouettes
- Multiple hypothesis pose estimates with associated probabilities
- Stable pose recovery from static images and image sequences
- Action recognition using mixture components