Monocular Human Motion Capture with a Mixture of Regressors

Ankur Agarwal and Bill Triggs

GRAVIR-INRIA-CNRS, Grenoble, France http://lear.inrialpes.fr {Ankur.Agarwal,Bill.Triggs}@inrialpes.fr

1 In Brief

Goal

- Recover **3D human body pose** from monocular **image silhouettes**
  - 3D pose = joint angles
  - use either individual images or video sequences

Applications

- Human computer interaction
- Markerless motion capture
- Gesture recognition
- Visual surveillance

Contributions

- "Model-free" learning based approach – no explicit 3D model
- Mixture of kernel regressors trained using human motion capture data
- Multimodal probabilistic solutions in 3D pose space
- Temporal fusion using a particle filter style tracker

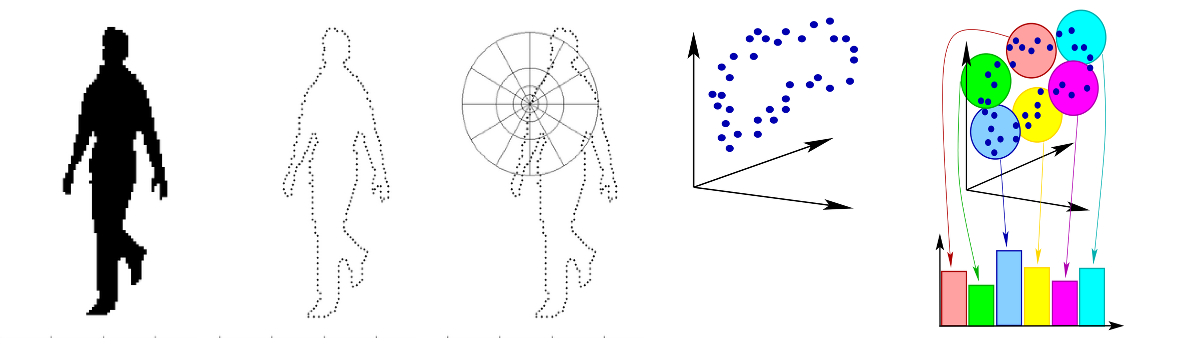
2 Silhouette Descriptors

Why Silhouettes

\+
- Relatively simple and low-level
- Capture most of the available pose information
- Insensitive to surface attributes (clothing colour, texture..)

−
- Distortions caused by background subtraction, shadows
- Ambiguity: hides internal details and depth ordering

Robust encoding of local shape — Shape Context Histograms



(a) extract silhouette (b) sample edge points (c) find local shape contexts (d) distribution in SC space (e) vector quantize to get histogram

3 Training and Test Data

- Capture **typical** human movements, not just *kinematically possible* ones, using real human motion capture data
- Use both real silhouettes from motion capture and synthetic silhouettes from several human body models (POSER from Curious Labs)

4 Multimodal Pose Estimation

- 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.
- Introduce a **discrete latent variable** $k \in \{1, 2 \dots K\}$ to encode the information missing in the silhouette.
- Assume a **mixture of experts** model based on K underlying functional regression rules $\mathbf{x} \sim \mathbf{r}_k(\mathbf{z})$:

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

5 Mixture of Regressors by E-M

- Reduce dimensionality of silhouette data using kernel PCA: $\mathbf{z} \to \boldsymbol{\phi}(\mathbf{z})$
- Initialize clustering with local connected components analysis
- Fit a **mixture of regressive Gaussians** to the joint density $(\boldsymbol{\phi}(\mathbf{z}), \mathbf{x})$:

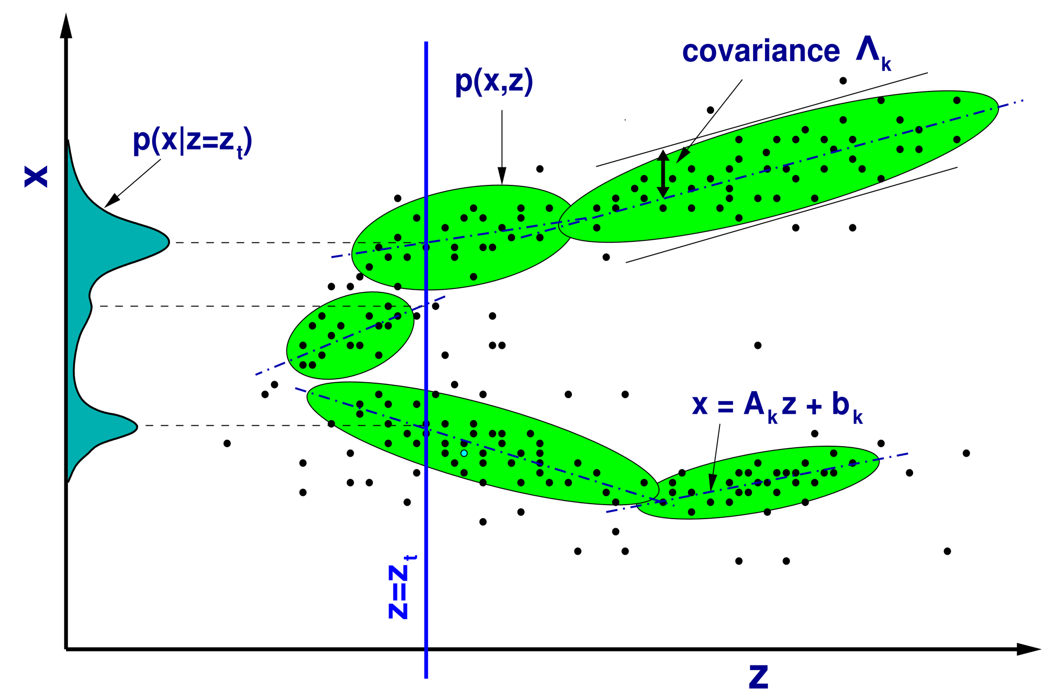
$$\begin{pmatrix} \boldsymbol{\phi}(\mathbf{z}) \\ \mathbf{x} \end{pmatrix} \simeq \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Gamma}_k)$$

- Linear regressor within each component k

$$p(\mathbf{x}\,|\,\mathbf{z}) \;=\; \mathcal{N}(\mathbf{r}_k(\mathbf{z}), \mathbf{\Lambda}_k) \qquad \mathbf{r}_k(\mathbf{z}) \;\equiv\; \mathbf{A}_k\,\boldsymbol{\phi}(\mathbf{z}) + \mathbf{b}_k$$

- Special covariance structure enforces "regressive" noise model

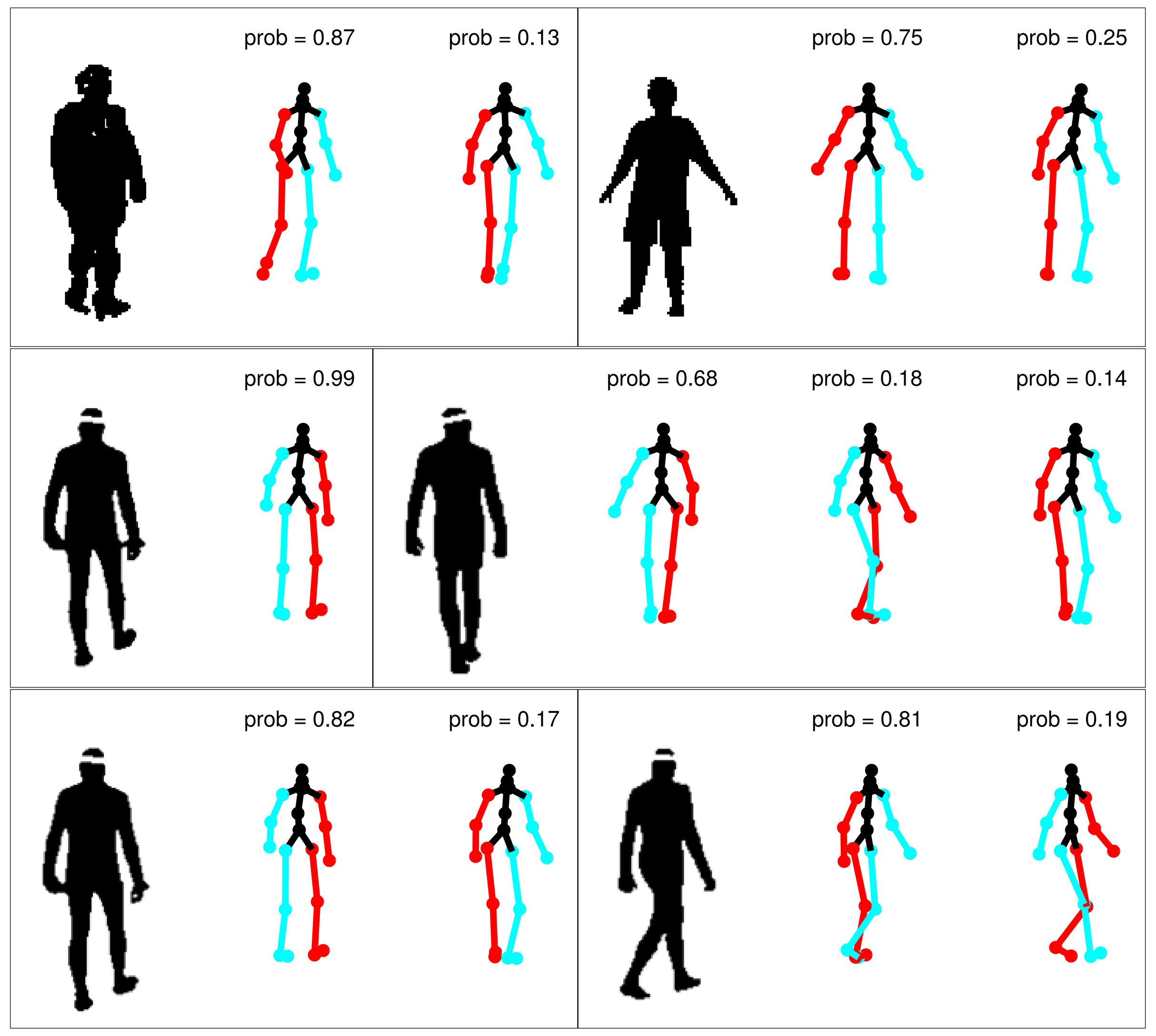
$$\boldsymbol{\mu}_k = \begin{pmatrix} \boldsymbol{\phi}(\bar{\mathbf{z}}_k) \\ \mathbf{r}_k(\bar{\mathbf{z}}_k) \end{pmatrix}, \mathbf{\Gamma}_k = \begin{pmatrix} \mathbf{\Sigma}_k & \mathbf{\Sigma}_k \mathbf{A}_k^\top \\ \mathbf{A}_k \mathbf{\Sigma}_k & \mathbf{A}_k \mathbf{\Sigma}_k \mathbf{A}_k^\top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



- **M-step**: Estimate $\mathbf{A}_k, \mathbf{b}_k$ by weighted least squares regression, $\mathbf{\Lambda}_k$ from residual errors. Compute $\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \pi_k$ for each class.
- **E-step**: Reestimate class membership weights for each point.

6 Pose 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.

7 Self-Initializing 3D Tracking

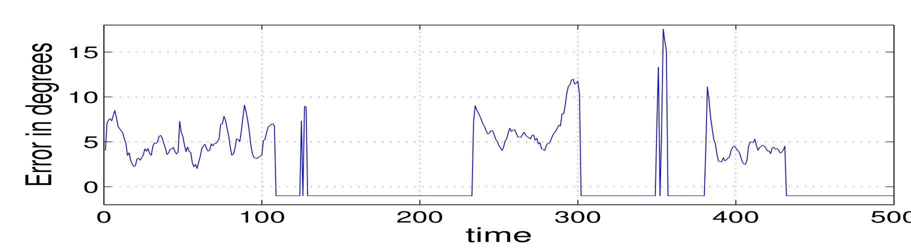
- Particle filter tracker, samples from dynamics $p(\mathbf{x}_t\,|\,\mathbf{x}_{t-1})$ as usual
- Uses regressive mixture $p(\mathbf{x}_t\,|\,\mathbf{z}_t)$ to assign posterior particle weights
- (Re)initializes by sampling from full mixture

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

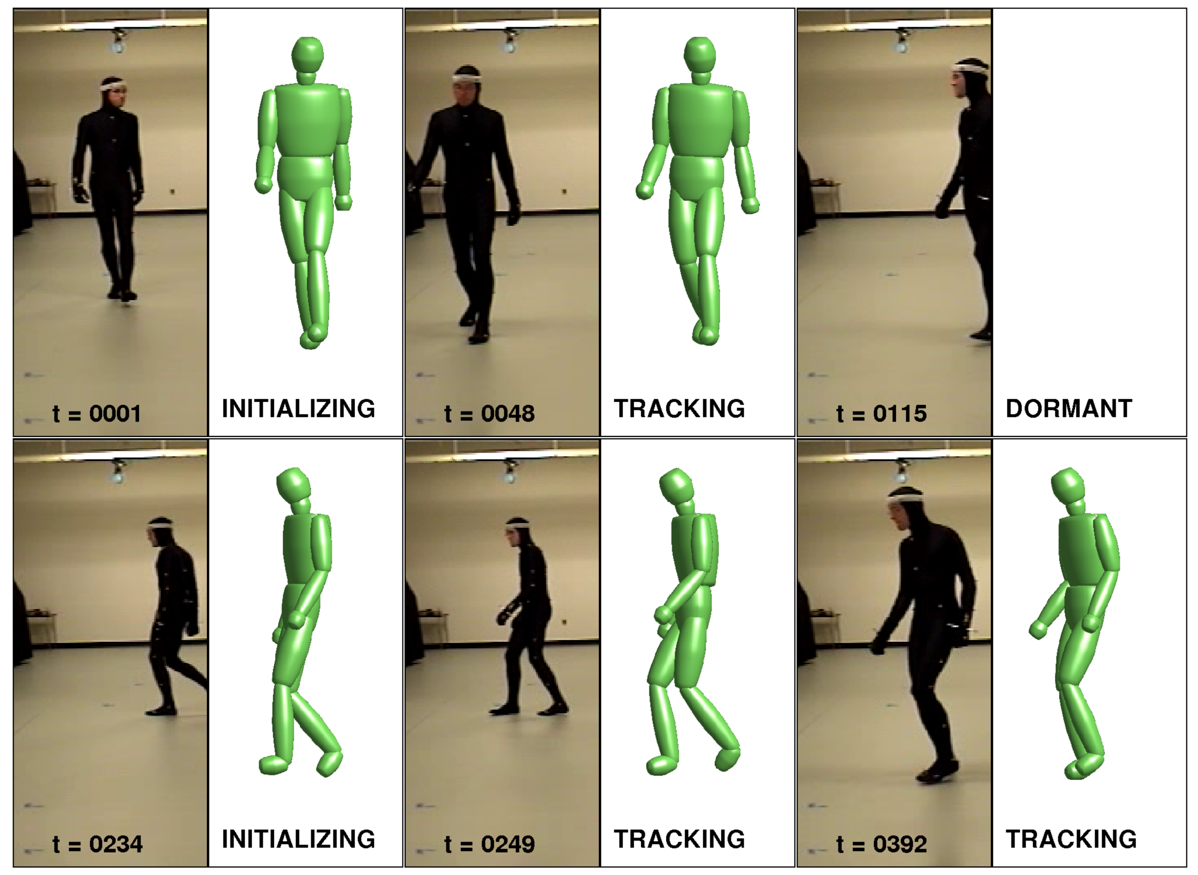
- Potentially real time owing to closed form solution for posterior.

7.1 Automatic (re)initialization

- Errors stabilize rapidly on (re)initialization



RMS tracking error of individual joint angles on a 500 frame sequence (-1 when no person is detected).



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

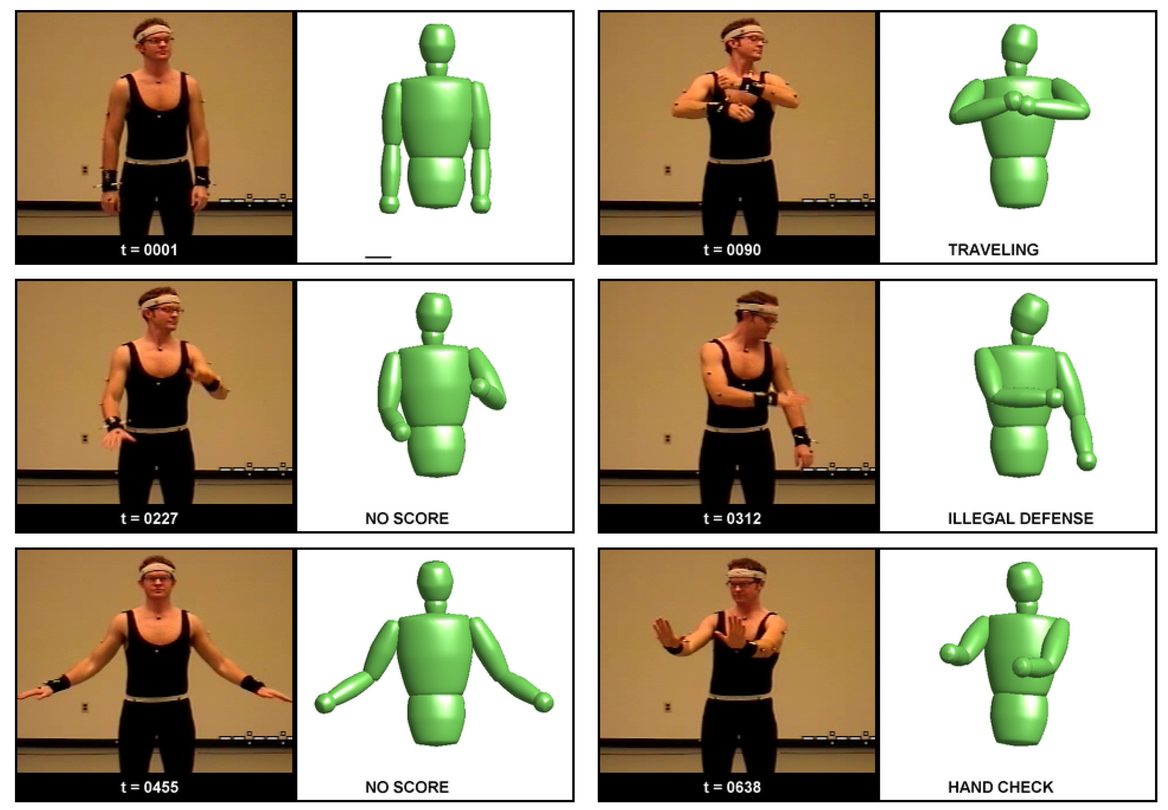
7.2 Upper Body Gesture Recognition

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

Training gestures (Basketball signals)

Traveling Illegal defense No score Stop clock Hand check Technical foul

Test sequence labelling



8 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

Work supported by an MENRT Doctoral Fellowship (French Education Ministry) and the European project LAVA