3D Human Pose from Silhouettes by Relevance Vector Regression

Ankur Agarwal and Bill Triggs


1 In Brief

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

Contributions
- “Model-free” learning based approach
- no explicit 3D model — recovers 3D pose by direct regression against robust silhouette descriptors
- Sparse kernel regressor trained using human motion capture data
- Regression based filtering for resolving reconstruction ambiguities
- Mean errors of only 4-6° per joint angle on test sequences

Method

Silhouette Descriptor → RVM regressor → 3D pose vector

2 Silhouette Descriptors

Why Silhouettes
+ Relatively simple and low-level
- Capture most of the available pose information
- Insensitive to surface attributes (cloth colour, texture...)--Frequently distorted by background subtraction / shadows
- Ambiguity: internal details and depth ordering are hidden

Robust encoding of local shape — Shape Context Histograms
(a) extract silhouette  (b) sample edge points  (c) find local shape contexts  (d) distribution of points  (e) vector quantize to get histogram

3 Training and Test Data

For the movements, we use real human motion capture data — captures typical human movements, not just possible ones
- Synthesize silhouettes with Poseer human modeller (Curious Labs) — somewhat artificial, but gives ground truth for testing, allows a wide range of training viewpoints.
- Also tested on real sequences of other people (without ground truth)

4 Nonlinear Regression Model

Given input: shape context histogram vector x
Desired output: 3D human pose vector y

\[ y = \hat{A} f(x) + \epsilon = \sum_{i=1}^{n} a_i \phi_i(x) + \epsilon \]

\[ f(x) = (\phi_1(x) \cdots \phi_k(x))^T \quad \text{vector of scalar basis functions} \quad \phi_i(x) \]

\[ A = (a_1 \cdots a_k) \quad \text{matrix of weight vectors} \quad a_i \text{ to be learned} \]

\[ \epsilon \quad \text{residual error vector} \]

4.1 Penalized Least Squares

Estimate \( \hat{A} \), given a set of training pairs \( \{(y_i, x_i) | i = 1 \ldots n\} \):

\[ \hat{A} = \arg \min_{A} \sum_{i=1}^{n} (A f(x_i) - y_i)^2 + R(A) \]

- \( x_i \) enter only via feature vectors \( f(x_i) = \phi_i(x) \)
- \( R(A) \) is a regularizer on \( A \) to control overfitting

Ridge Regression:
\[ R(A) = \lambda \|A\|^2 \]

4.2 The Relevance Vector Machine

- A Bayesian-motivated approach to regression and classification
- Uses a singular power-law prior to aggressively prune unneeded weights, giving sparse solutions \( A \)
- Regularizer:
\[ R(A) = \epsilon \sum_{a_i} \log |a_i| \]

- \( \epsilon \) is the pruning / shrinkage strength
- \( a_i \) can be the components, the columns, or the rows of \( A \)

RVM Training Algorithm
0. Initialize \( A \) with ridge regression. Initialize the running scale estimates \( u_{\text{scale}} = |a_i| \) for the components or vectors \( a_i \)
1. Approximate the \( \log |a_i| \) penalty terms with “quadratic bridges”:
2. Solve the resulting linear least squares problem in \( A \):
3. Remove any components \( a_i \) that have become zero, update the scale estimates \( u_{\text{scale}} = |a_i| \) and continue from 1 until convergence.

Linear bases
- \( f(x) \propto x \rightarrow \text{the RVM selects relevant silhouette features.} \)

Kernel bases
- \( f(x) = (K(x, x_1) \cdots K(x, x_n))^T \) where \( K(x, x_i) \) is a kernel function instantiated at training examples \( x_i \)
- RVM selects relevant training examples, here only 6%

5 Pose from Static Images

Reconstruction on test sequence

Spiral walking sequence not included in training data. Gaussian kernel regressor. Mean angular error per d.o.f = 6.0°

6 Pose from Video Sequences

Tracking Framework
- Tracking reduces glitches caused by silhouette ambiguities
- Regression based filtering for dynamical prediction and observation update (\( x_i \): 3D pose state, \( x \): Silhouette descriptor)

Dynamics
- A second order global dynamical model suffices:
\[ \dot{x} = A x + B z + w \]
\[ z = C x + s \]

State-Sensitive Observation Update
- Nonlinear kernel regressor “selects” observation update to apply using state prediction
- Our full regression model also includes an explicit \( s_i \) term to represent the direct contribution of the dynamics
\[ s_i = C x_i + \sum_{j=1}^{n} \lambda_j \phi_j(\hat{x}_j) \]

Results

Dynamics

Observations

Angles (in degrees) vs. Time.  --- Ground truth  —— Estimate

Sample reconstructions from real images

Original (middle) and new (bottom) viewpoints

7 Conclusion

- “Model free” methods for recovering 3D human pose from monocular silhouettes
- Direct nonlinear regression of pose again robust shape descriptors
- Tested different regression methods: ridge regression, RVM, SVM
- Pose recovery from static images and image sequences

Work supported by European projects VIBES, LAVA and PASCAL