A Local Basis Representation for Estimating Human Pose from Cluttered Images

Ankur Agarwal and Bill Triggs

GRAVIR-INRIA-CNRS, Grenoble, France

Asian Conference on Computer Vision, January 2006

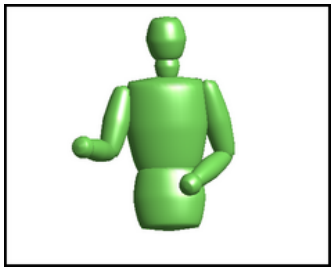
Introduction

Goal

- Recovering 3D human body pose from *raw* images

Applications

- Human computer interaction, gesture recognition ...



Existing Methods

- Optimize projection error of an explicit body model, **OR**
- Learning-based recognition from segmented images

  E.g. silhouettes

Present Focus

- Learning-based bottom up pose estimation
  - no explicit body and camera models
  - easily adaptable to different appearances/people
- Pose reconstruction in the **presence of background clutter**
  - need robust image features sensitive human pose but not to background, lighting etc.
- Assume approximate localization (i.e. a 'detection'), but no segmentation
- Work with upper body frontal poses

Approach

- Use a dense overlapping grid of SIFT-like descriptors
- Recode each descriptor using a learned local ***Nonnegative Matrix Factorization*** basis
  - suppresses background and stabilizes estimation
- Regression for obtaining 3D pose directly from these descriptors
  - train a regressor using motion capture / synthetic data

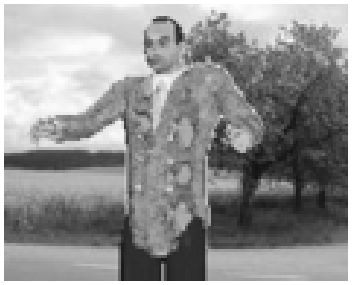
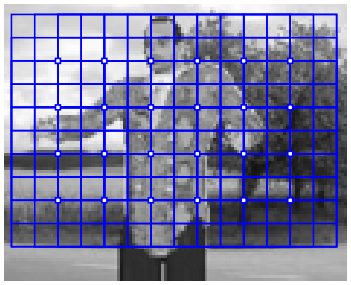
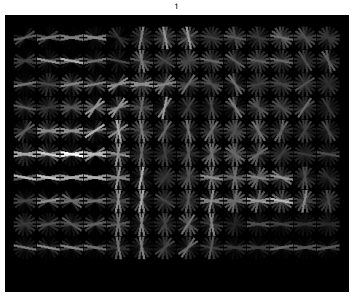


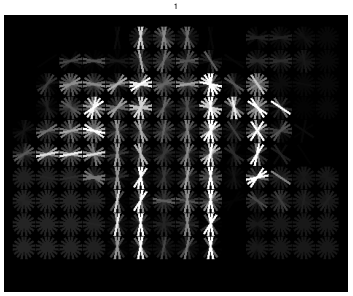
image window



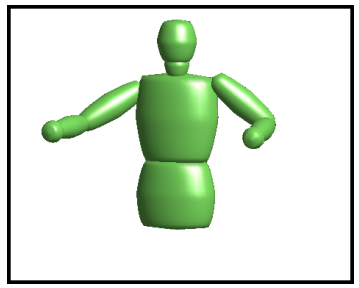
overlapping SIFT grid



window descriptors



descriptors after NMF reduction

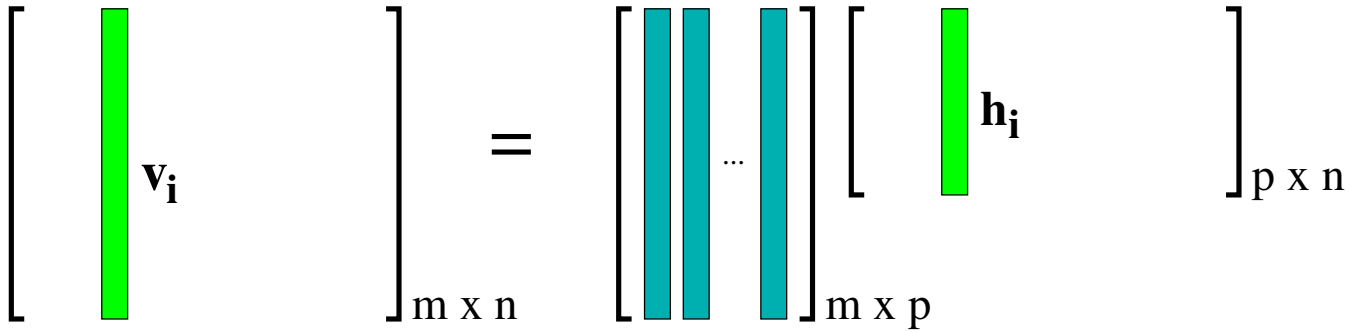


final pose

Nonnegative Matrix Factorization

- Approximates a nonnegative matrix as lower rank product of nonnegative factors:

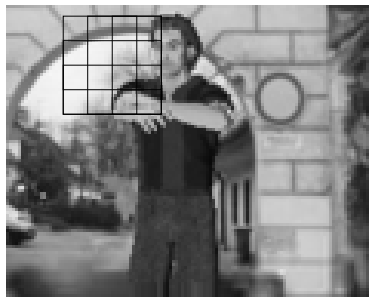
$$\mathbf{V}_{m\times n} = \mathbf{W}_{m\times p}\,\mathbf{H}_{p\times n} \qquad p \le m, n \quad \mathbf{V}_{ij}, \mathbf{W}_{ij}, \mathbf{H}_{ij} \ge 0$$



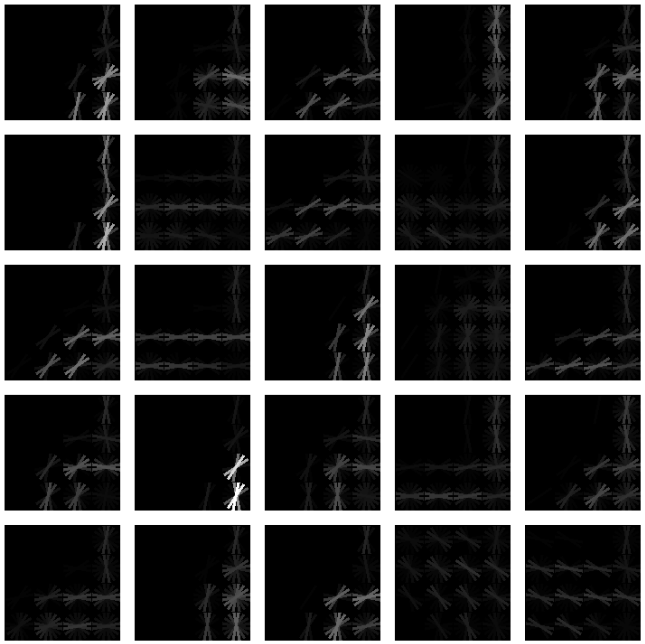
- Calculate factorization with simple iterative updating algorithms – least squares or entropy (logarithmic barrier) cost functions
- Hoyer's algorithm: add terms to encourage sparsity of one/both factors

Local NMF Basis

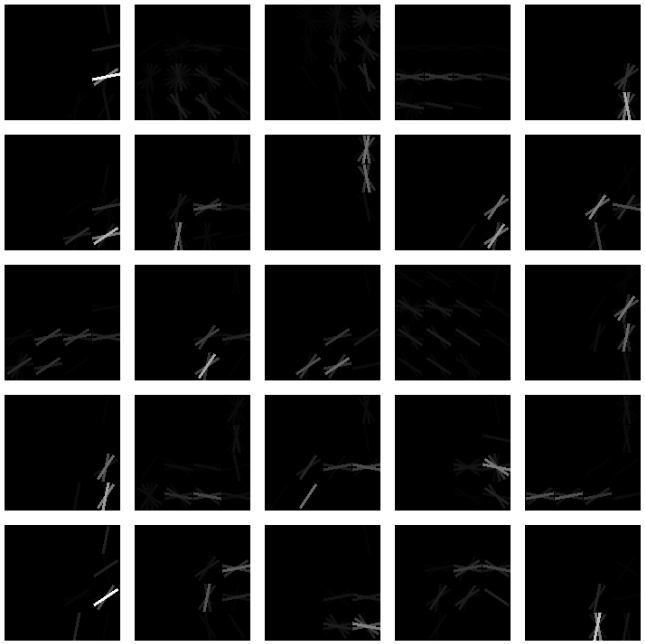
- Each SIFT patch has its own NMF basis.
- Learn bases by NMF on matrix of training examples
  - either empty background or varying background
  - captures consistent structure (significant edges)
  - suppresses inconsistent structure (varying backgrounds).
- 30–40 basis elements per 128-d SIFT patch suffice.
- For larger bases, sparse coding may help
  - sparsifying coefficients (but not basis elements)



NMF gives sparse basis vectors

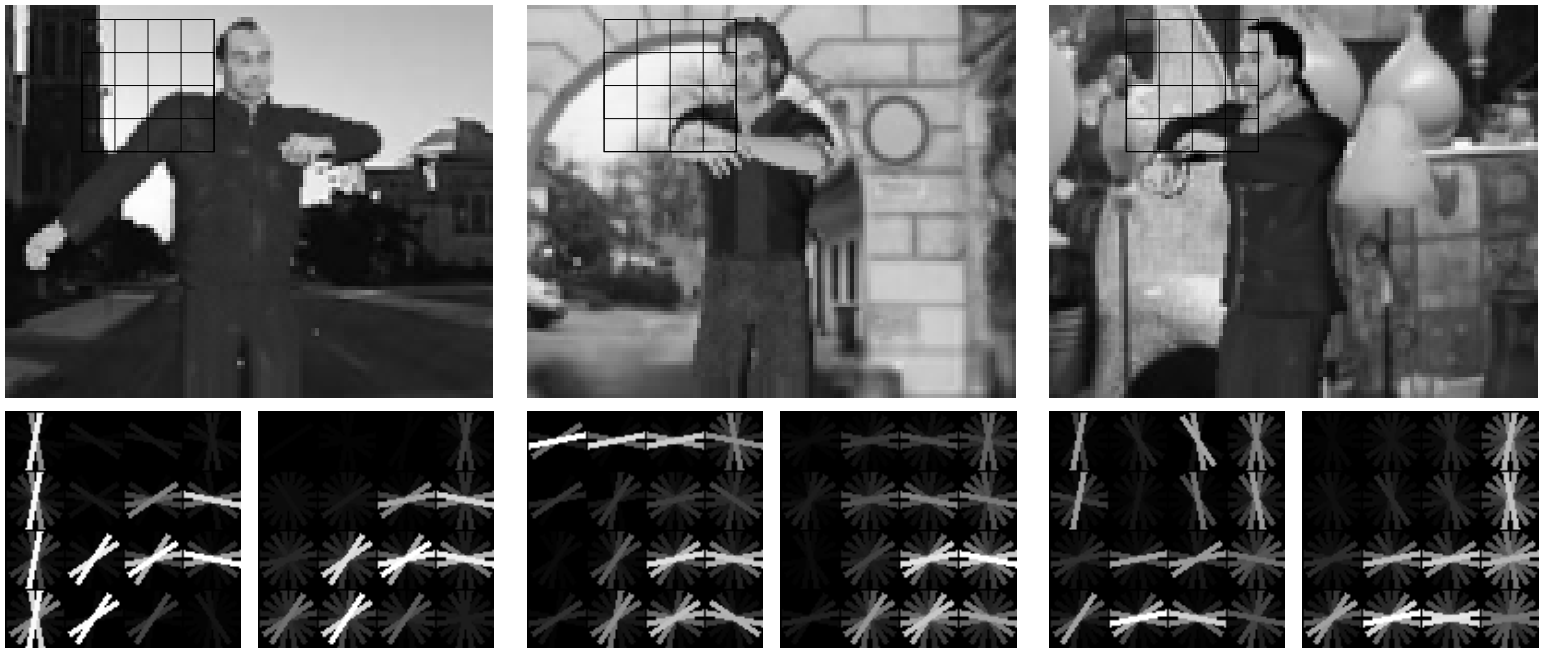


input SIFT patches



patches after NMF coding

Examples of NMF Coding



input patch | after NMF | input patch | after NMF | input patch | after NMF

Note the good suppression of background edges

Obtaining Pose using Regression

- Initial feature vector $\mathbf{x} \equiv (\mathbf{v}^{1\top}, \mathbf{v}^{2\top}, \ldots \mathbf{v}^{L\top})^\top$.

  $\mathbf{v}^k$: 128-d SIFT vector at patch k.

- $\mathbf{v}^k = \sum_j \mathbf{w}_j^k h_j^k$

  $\mathbf{w}^k$: basis vectors at patch k

  h_j^k : coefficient of j^{th} basis vector

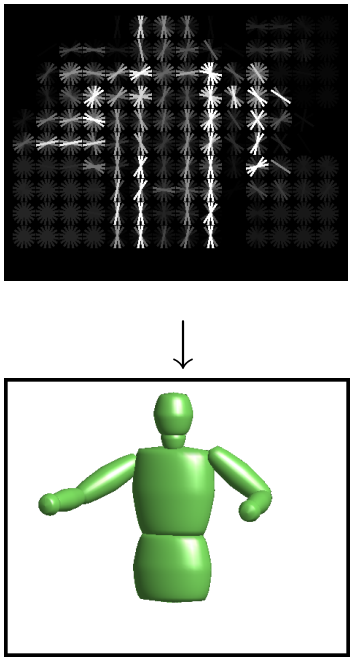
- Nonlinearly coded feature vector:

  $\boldsymbol{\phi}(\mathbf{x}) \equiv (\mathbf{h}^{1\top}, \mathbf{h}^{2\top}, \ldots \mathbf{h}^{L\top})^\top$

- Pose vector $\mathbf{y}$ obtained by regression:

  $\mathbf{y} \;=\; \mathbf{A}\,\boldsymbol{\phi}(\mathbf{x}) + \boldsymbol{\epsilon}$

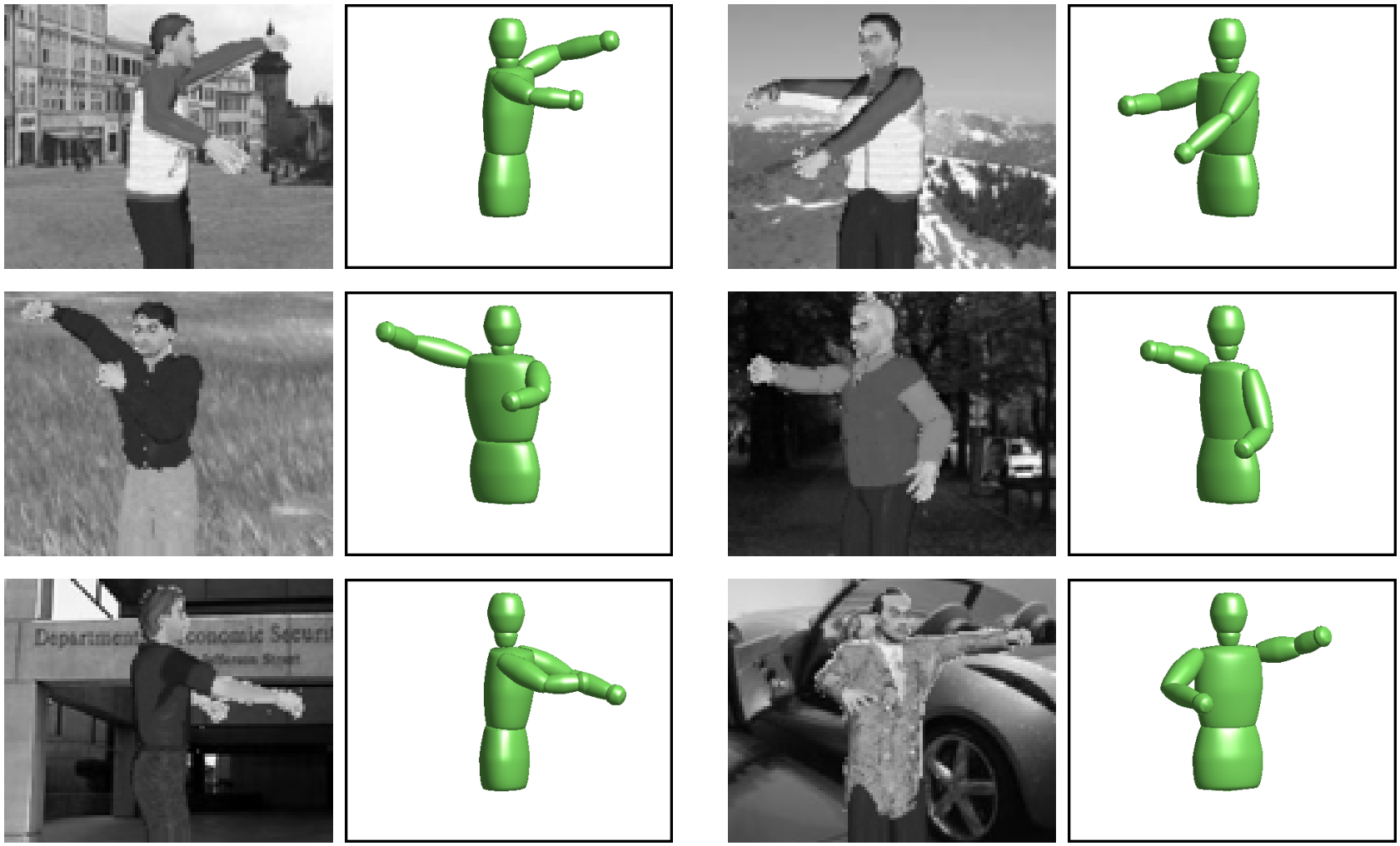
  $\mathbf{y}$: 24-d vector encoding position of 8 upper body joints



Experiments – Training and Test Sets

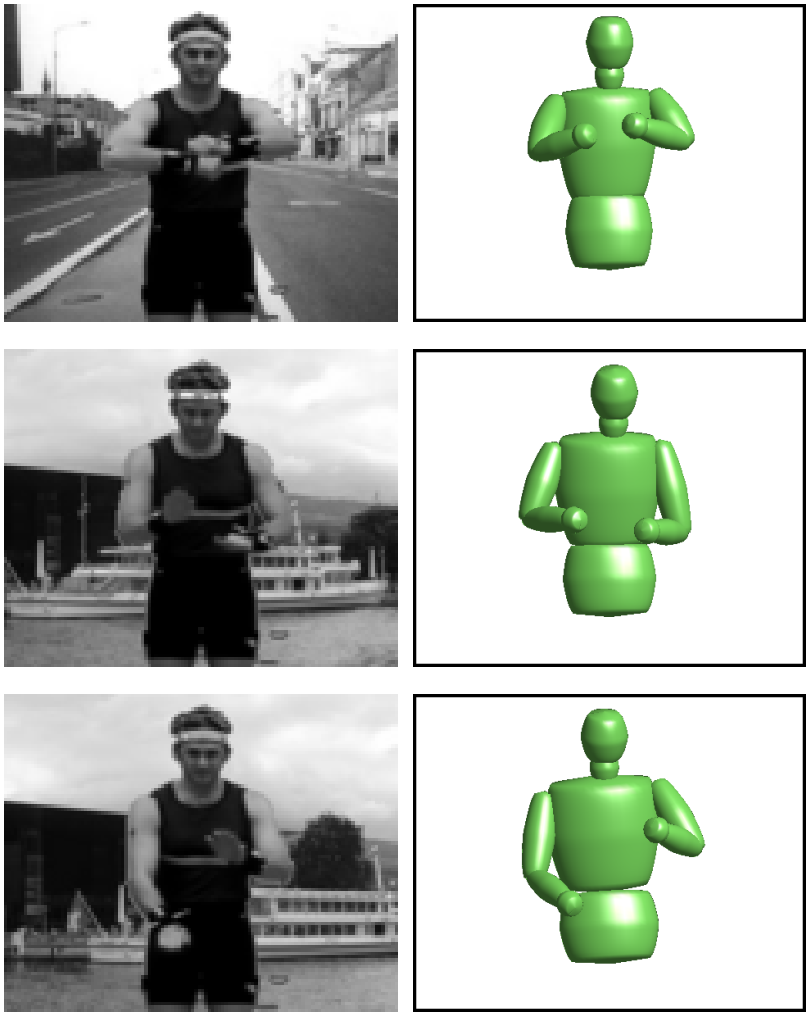
1. Synthetic human images (Poser) on clean or random backgrounds
   - random poses: widely varying but not very natural
   - below we use 4000 training examples, 1000 test examples
2. Real dataset – CMU basketball signals with motion capture data
   - 1600 training images (9 sequences)
   - 1 test sequence of another person
   - natural poses but limited range
   - we also add artificial backgrounds

Poser Dataset – Examples

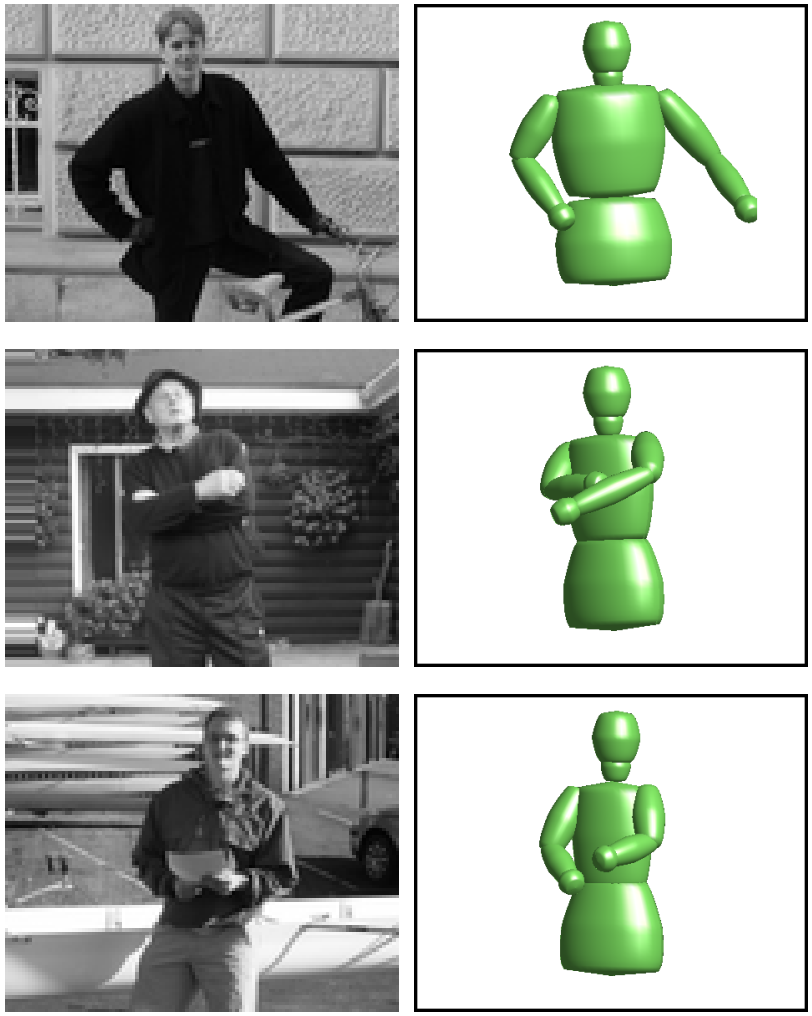


Mocap Dataset – Examples

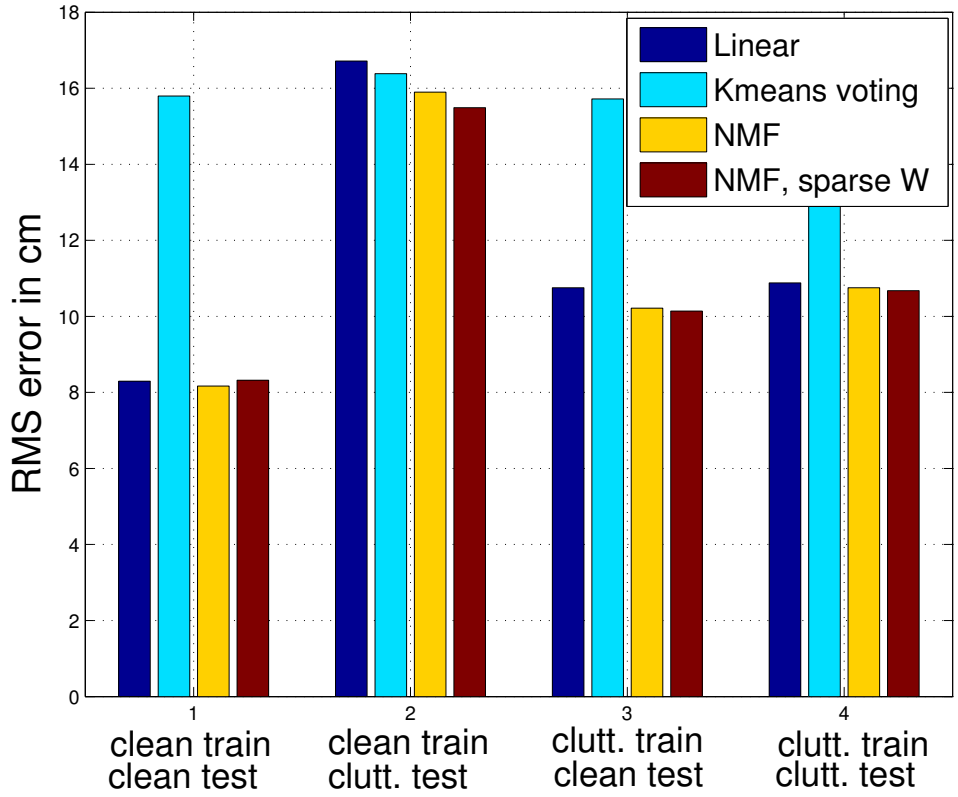
Mocap test



Web images

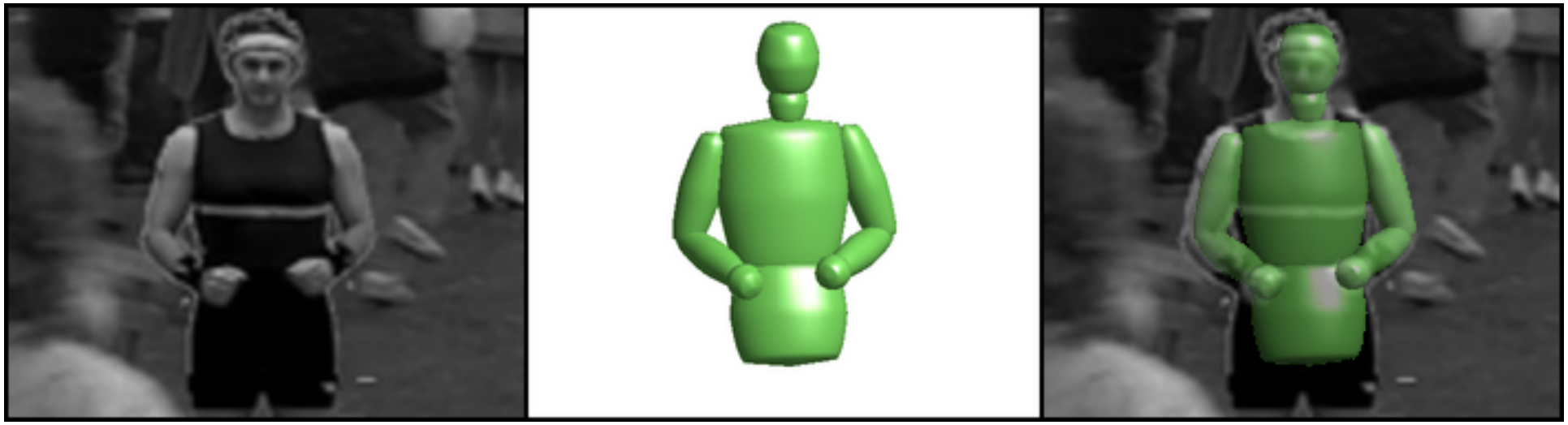


Poser Dataset – Performance



Precision in clutter ~ 10 cm

Mocap Dataset – Video



Test illustrating stability w.r.t. varrying background.

Conclusion

- Regression-based recovery of human pose from monocular images
  - needs rough localization but no prior segmentation
  - uses nonnegative factorization for resistance to cluttered backgrounds
- Faster and probably more robust than most model based methods

Future Work

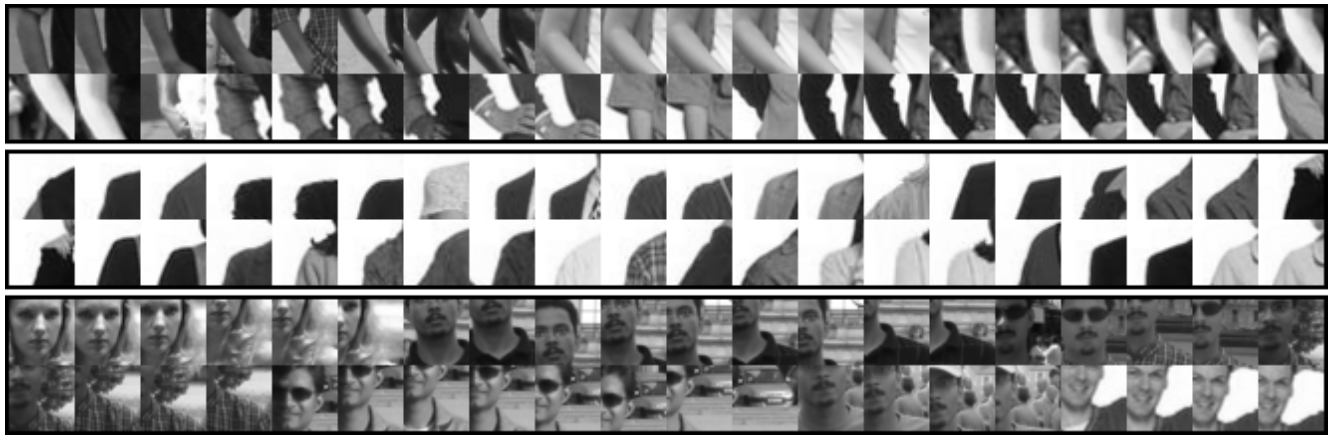
- More robustness to incorrect localization; scale invariance ...
- Handling a wider range of movements

Thank you.

Parameter Settings

- 3072-D image descriptor
  - window contains 4×6 grid of SIFT blocks
  - blocks overlap by 50% in both x & y
  - each block is 128-D: a 4×4 grid of 4×4 pixel spatial cells with 8 orientations (image size: 118×95 pixels).
  - reduced to 720-D using NMF
- Pose regressor is simple ridge regression (regularized linear least squares)
  - regularization parameter set using cross-validation

Bag of body-parts model



- K-means to obtain representative body parts
  - **– loses spatial information important for pose reconstruction**