Single Image 3D Human Pose Estimation from Noisy Observations

Institut de Robòtica i Informàtica Industrial

E. Simo-Serra A. Ramisa G. Alenyà C. Torras F. Moreno-Noguer

Institut de Robòtica i Informàtica Industrial, CSIC-UPC

08028 Barcelona, Spain

PROBLEM:

Retrieval of a 3D Human Pose from a single image.

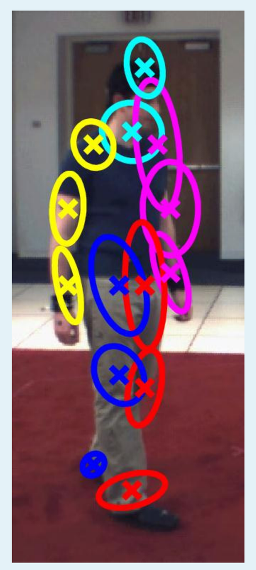
STATE-OF-THE ART LIMITATIONS:

- Use of temporal information or background subtraction
- Unable to handle large amounts of 2D noise

CONTRIBUTIONS:

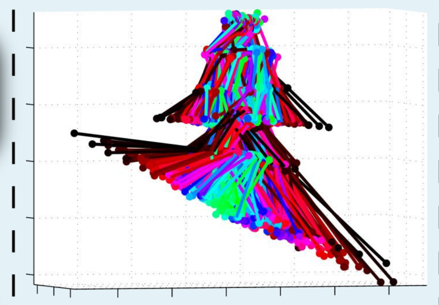
- Proposal of an approach to efficiently explore the space of possible 3D solutions given noisy 2D input
- Coarse to fine approach to constrain the solution space, until a single solution is obtained

Method Overview

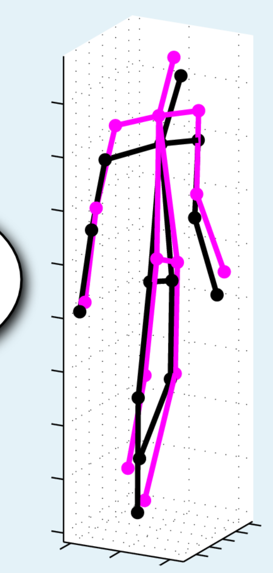


Single Monocular 2D Image → Raw Image → 2D Part Detector → 2D Gaussians U

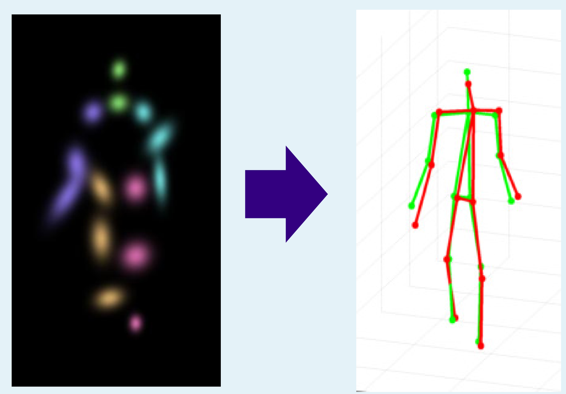
EXPLORATION OF AMBIGUOUS HYPOTHESES: Linear Projection U → X → α → Gaussian α → CMA Sampling → Set of α samples → Linear Projection α → X → Set of X samples → Gmeans Clustering → Subset of X samples



HYPOTHESES DISAMBIGUATION: Calculate distance vectors → Set of (X,d) vectors → Linear Projection d → β → Set of (X,β) samples → Sort by OCSVM run on β vectors → Sorted set of X samples → Choose best sample → 3D Pose X → 3D Pose Estimation



Problem Definition



2D Gaussians 3D Pose

GIVEN:

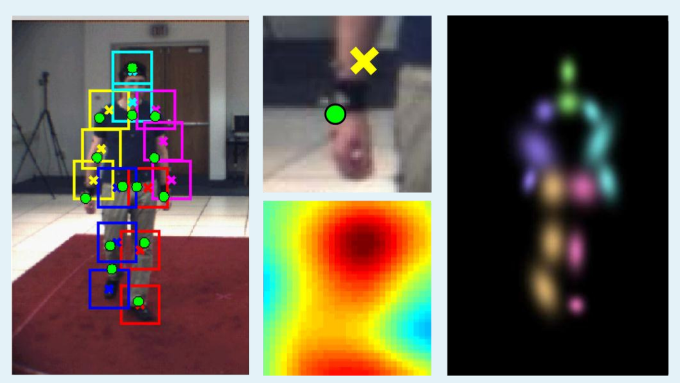
- 2D noisy detections of the parts as a set of Gaussians $\mathbf{u}_i \sim N(\mathbf{u}_i, \Sigma_u)$
- Internal Calibration Matrix $\mathbf{A}$

WE WANT TO RETRIEVE:

- The 3D pose of the input image

"Off-the-shelf" 2D Body Part Detector [1]

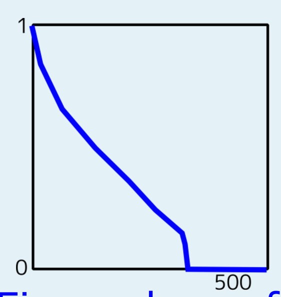
- Output modified to provide local Gaussian estimations
- Provides robust 2D detections at the cost of location noise



Projective Linear Deformation Model

- 2D-3D correspondences can be written as the rank-deficient linear system

$$\mathbf{Mx} = 0$$



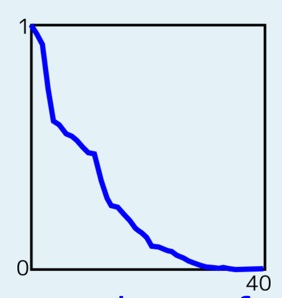
Eigenvalues of $\mathbf{M}$

- Shape = Linear combination of deformation modes:

$$\mathbf{x} = \underbrace{\mathbf{Q}}_{\text{Deformation Modes}}\underbrace{\boldsymbol{\alpha}}_{\text{Modal Weights}} + \underbrace{\mathbf{x}_0}_{\text{Mean Shape}}$$

- The correspondence problem becomes:

$$\mathbf{MQ\alpha} + \mathbf{Mx}_0 = \mathbf{0}$$



Eigenvalues of $\mathbf{MQ}$

- Still rank deficient, but much less.

Projective Linear Deformation Model

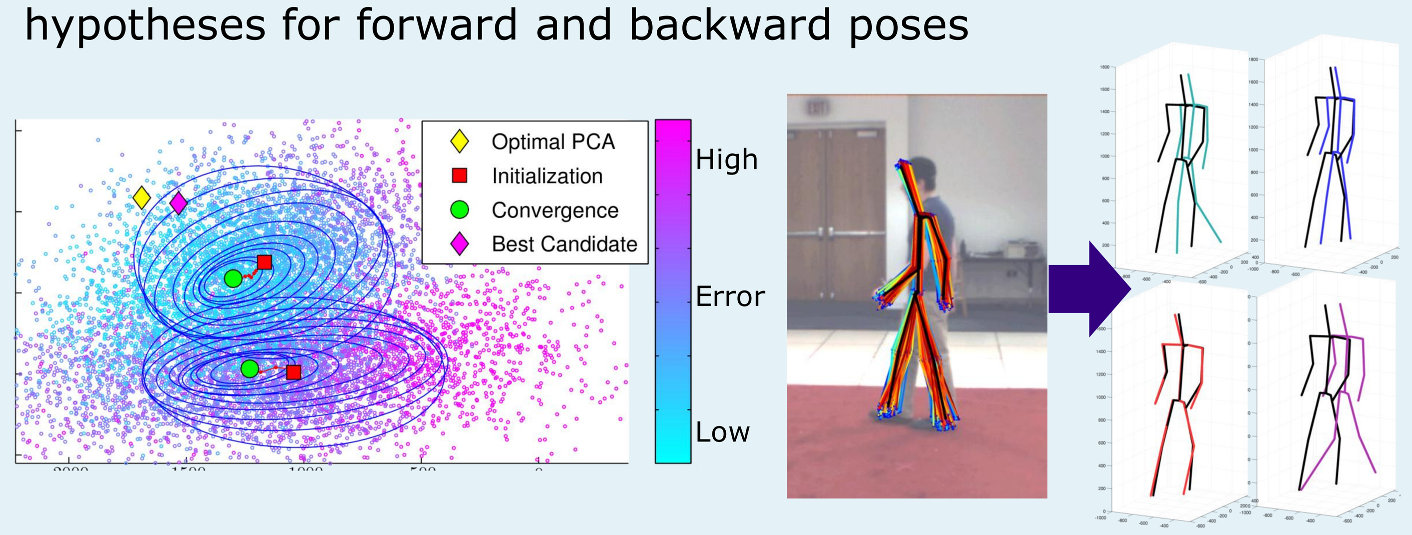
- Mean: $\boldsymbol{\mu}_\alpha = -(\mathbf{MQ})^\dagger \mathbf{Mx}_0$, where $\mathbf{M}$ function of $\mathbf{A}$, $\mathbf{u}_i$
- Covariance: $\Sigma_\alpha = \frac{\partial \boldsymbol{\alpha}}{\partial \mathbf{u}} \Sigma_u \left(\frac{\partial \boldsymbol{\alpha}}{\partial \mathbf{u}}\right)^T$

Exploration of Ambiguous Hypotheses

- Error metric based on reprojection and learnt limb length.

$$\varepsilon_{lr} = \varepsilon_l \cdot \varepsilon_r \left[\begin{array}{l} \varepsilon_l = \sum_{i,j \in N} \| \tilde{l}_{ij} - l_{ij}^{train} \| \, \sigma_{ij}^{-1} \quad \leftarrow \text{Length Error} \\ \varepsilon_r = \sum_{i}^{n_v} \sqrt{(\tilde{\mathbf{u}}_i - \mathbf{u}_i)^T \Sigma_{\mathbf{u}_i}^{-1} (\tilde{\mathbf{u}}_i - \mathbf{u}_i)} \quad \leftarrow \text{Reprojection Error} \end{array}\right.$$

- Sampling solution space using Covariance Matrix Adaptation [2]
- Ambiguity in orientation facing solved by proposing two hypotheses for forward and backward poses

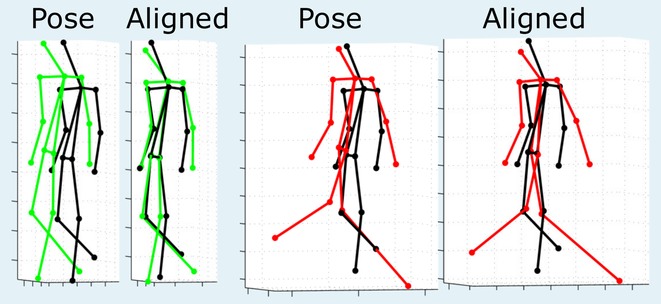


- Potential solutions (~10,000) are clustered to reduce their number (~300 clusters)
- Error function is not discriminative enough for a single solution

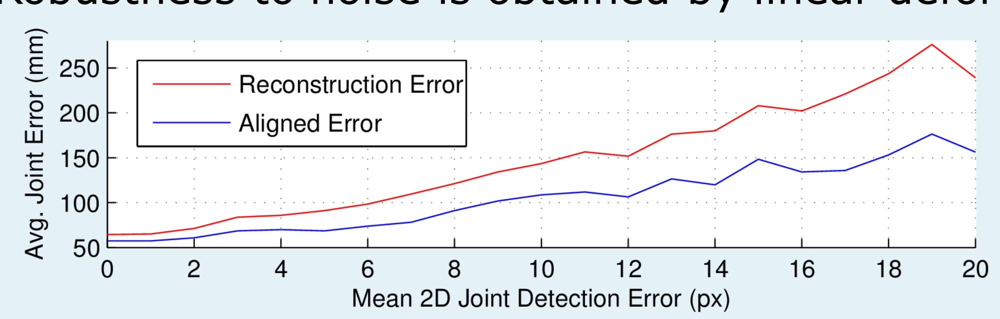
Hypotheses Disambiguation

- One-Class Support Vector Machine is used to obtain a value indicative of pose anthropomorphism from distance matrices

	Sample #1		Sample #2
Error Value (ε_{lr})	6.883	≈	6.885
SVM Output	2.8e-04	>	-7.6e-03
Reconst. Err. (mm)	199.9	≈	214.9
Aligned Err. (mm)	56.4	<	167.7



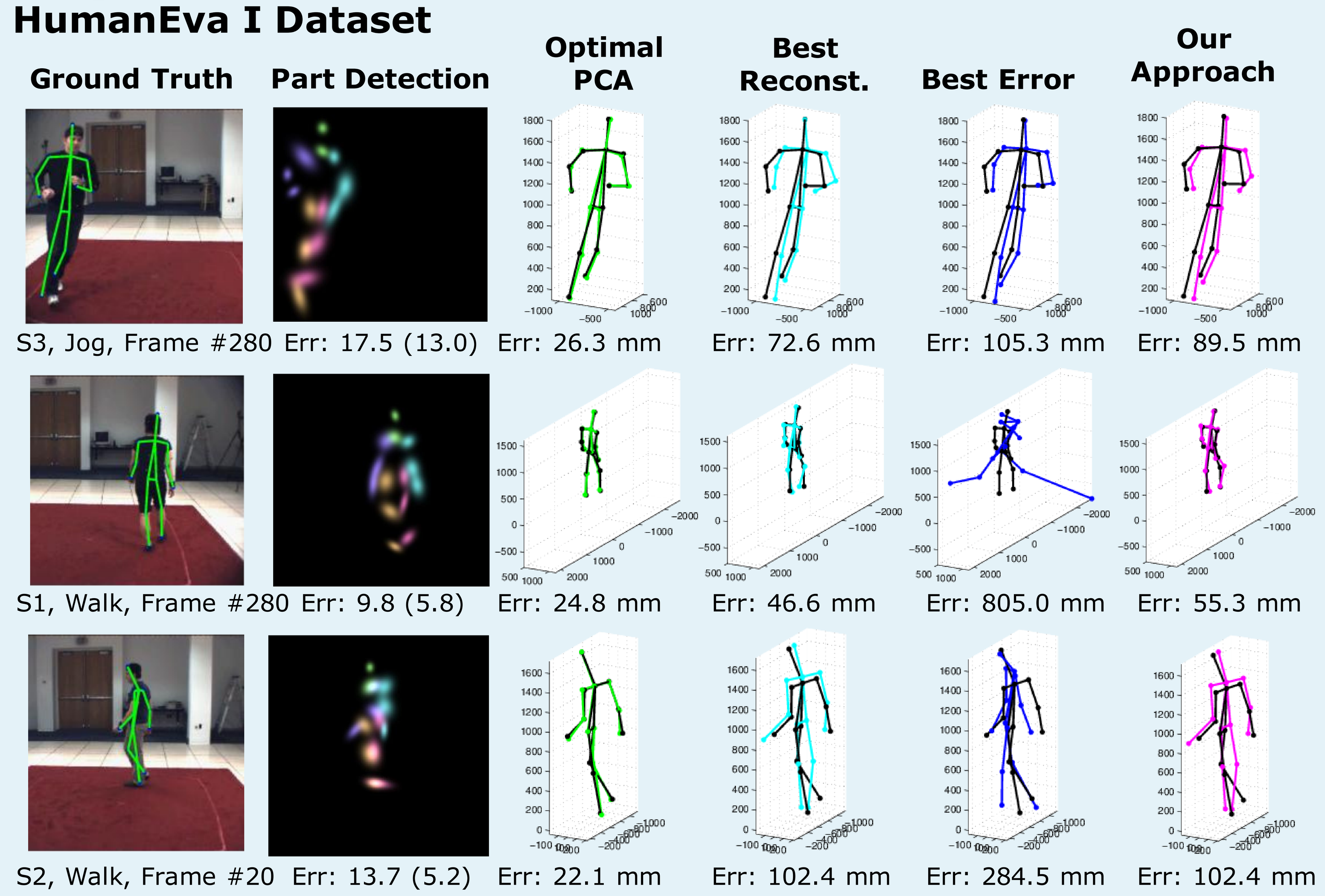
- Robustness to noise is obtained by linear deformation model



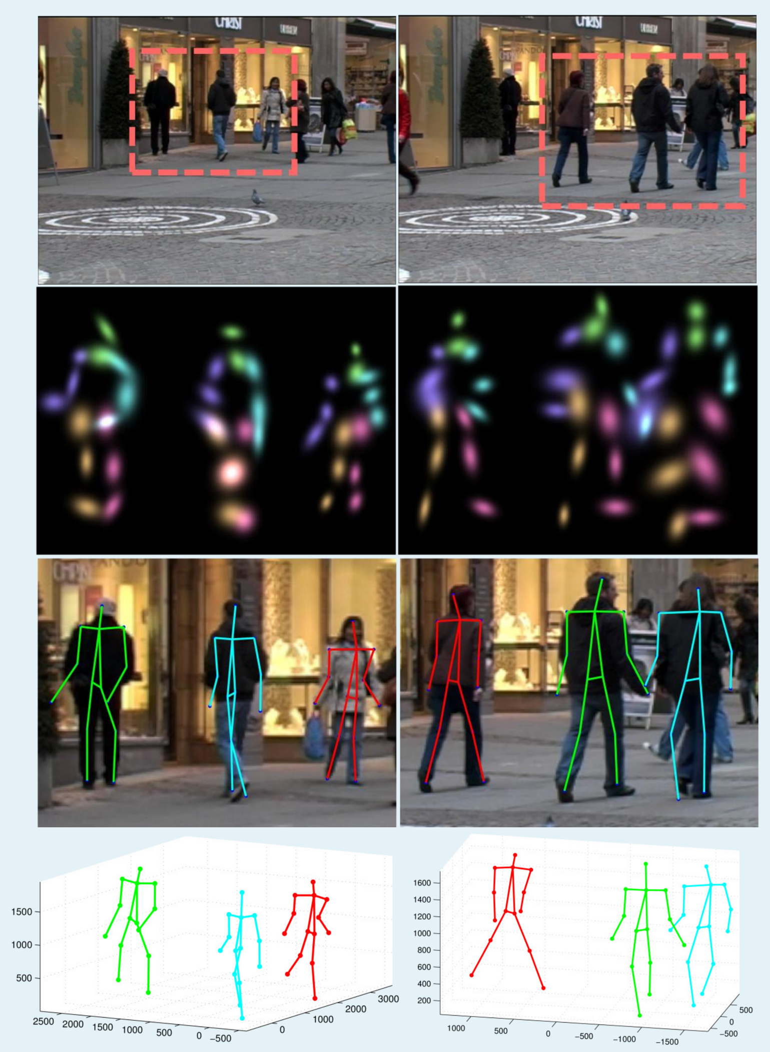
Propagation of synthetic noise from 2D estimations to 3D pose

Experimental Results

HumanEva I Dataset



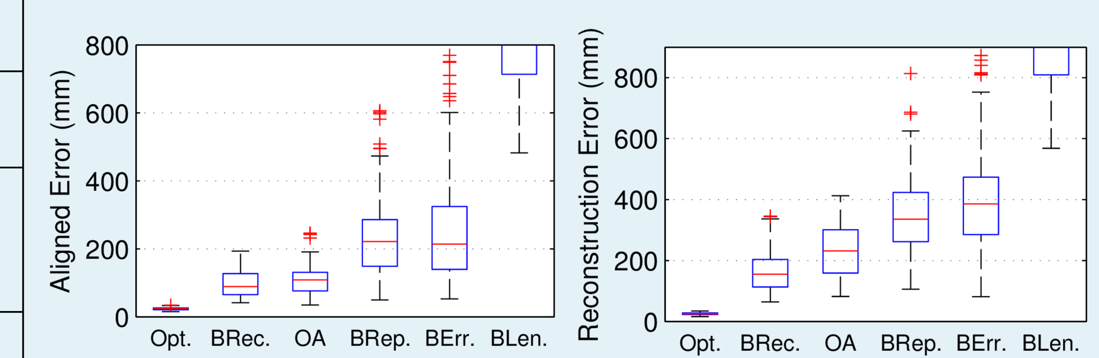
TUD Stadmitte



Numeric Results

	Walking S1	Walking S2	Walking S3
Our Approach	99.6 (42.7)	108.3 (42.3)	127.4 (24.0)
2D Input (px)	14.1 (7.5)	19.1 (8.1)	26.8 (8.0)
[3] (tracking)	-	107 (15)	-
[4] (tracking)	89.3	108	113.5
[5] (background subtraction)	38.2 (21.4)	32.8 (23.1)	40.2 (23.2)

	Jogging S1	Jogging S2	Jogging S3
Our Approach	109.2 (41.5)	93.1 (41.1)	115.8 (40.6)
2D Input (px)	18.3 (6.3)	18.1 (6.0)	20.9 (6.1)
[5] (background subtraction)	42.0 (12.9)	34.7 (16.6)	46.6 (28.9)



Our Approach (OA) outperforms using only error functions to obtain the 3D pose (BRep, BErr, BLen) and is generally close to finding the best reconstruction sample BRec

References

[1] Y. Yang, D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[2] N. Hansen. The CMA evolution strategy: a comparing review. In *Towards a new evolutionary computation, Adv. On estimation of distribution alg.*, pp 75-102. Springer, 2006.

[3] M. Andriluka, S. Roth, B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[4] B. Daubney, X. Xie. Tracking 3d human pose with large root node uncertainty. In *CVPR*, 2011.

[5] L. Bo, C. Sminchisescu. Twin Gaussian Processes for Structured Prediction. *IJCV*, 87(1-2): 28-52, 2010.