A Joint Model for 2D and 3D Pose Estimation from a Single Image

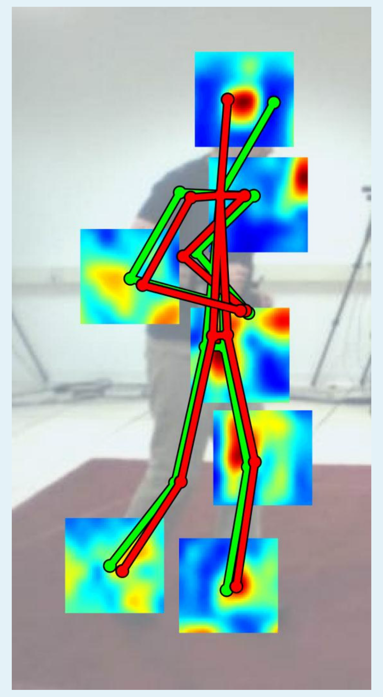
Institut de Robòtica i Informàtica Industrial · CSIC · UPC

E. Simo-Serra[1] A.Quattoni[2] C. Torras[1] F. Moreno-Noguer[1]

[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, 08028 Barcelona, Spain

[2]Universitat Politècnica de Catalunya, UPC, 08028 Barcelona, Spain

CVPR 2013 Portland, Oregon June 23-28



PROBLEM:

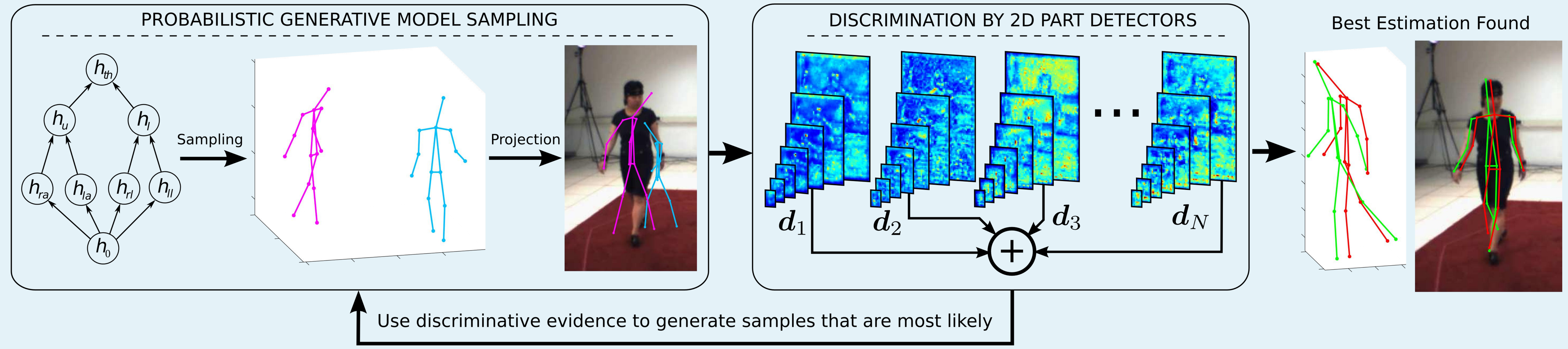
Retrieval of a 2D and 3D Human Pose from a single image

STATE-OF-THE ART LIMITATIONS:

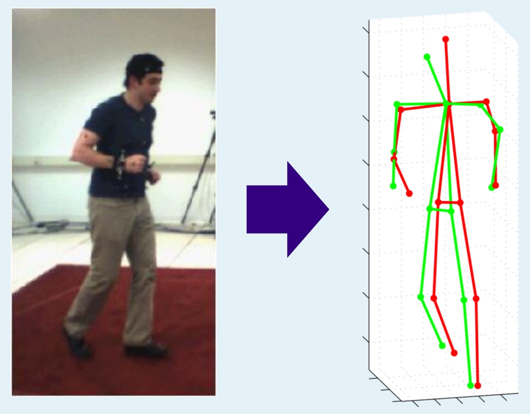
- Use of temporal information or background subtraction
- Unrealistic assumption of good 2D input

CONTRIBUTIONS:

- Novel probabilistic generative model for 3D Human Motion
- Bayesian framework for joint inference of 2D and 3D pose



Problem Definition



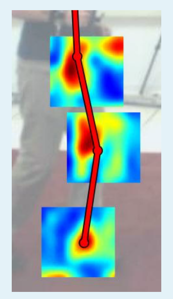
GIVEN:

- Input Image
- Camera Focal Length α

WE WANT TO RETRIEVE:

- Both the 3D and 3D pose of the subject in the input image

Bayesian Formulation



- Image evidence given body configuration

$$\underbrace{p(D \mid L)}_{\text{Image Evidence},\ \text{2D Pose}} = \prod_{i=1}^{N} p(\boldsymbol{d}_i \mid \boldsymbol{l}_i)$$

- Consider 2D to be projection of true 3D model generated by smaller latent model

Latent Space

3D Pose

$$p(X \mid D) \propto \underbrace{p(H)\, p(X \mid H)}_{\text{generative}} \underbrace{\prod_{i=1}^{N} \left(p(\boldsymbol{d}_i \mid \boldsymbol{l}_i)\, p(\boldsymbol{l}_i \mid \boldsymbol{x}_i)\right)}_{\text{discriminative}}$$

Generative model reduces search space during inference

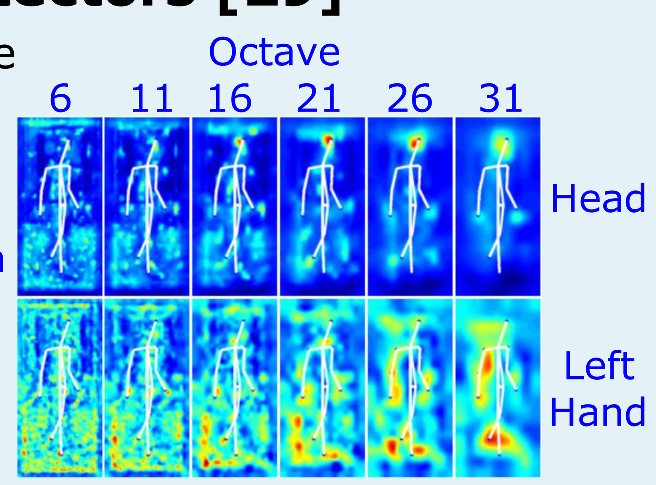
Discriminative 2D detectors enforce consistency of the 3D pose with the image evidence

Discriminative 2D Part Detectors [29]

- Smooth response good for inference
- Scale estimated from depth with β:

$$s_i^{-1} = \alpha^{-1} \beta z_i$$

Part scale · Part depth · Focal length



- Weighted based on usefulness for 3D pose estimation
- Score interpreted as log-likelihood

$$\log p(L \mid D) \approx \mathrm{score}(L) = \sum_{i=1}^{N} k_i \boldsymbol{d}_i(u_i, v_i, s_i)$$

Detector at scale space coordinates

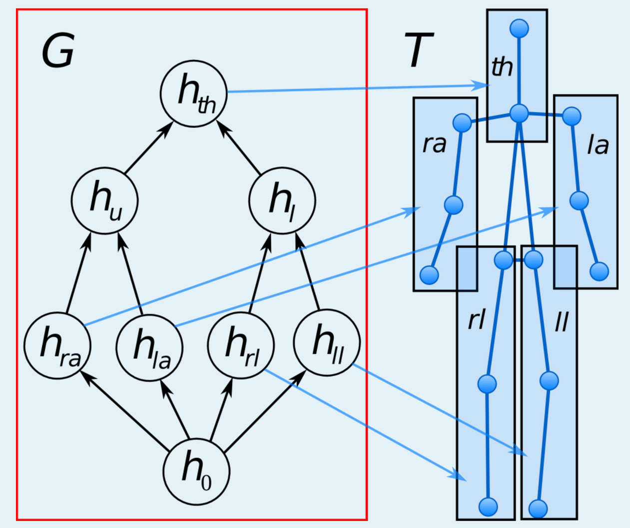
Relative weighting

Latent Generative Model

- Learns compression function:

$$\phi(X^L) : \mathcal{X}^L \rightarrow \mathcal{H}$$

3D Poses · Latent Space



- 3D Poses are discretized
- Directed Acyclic Graph allows efficient dynamic programming:

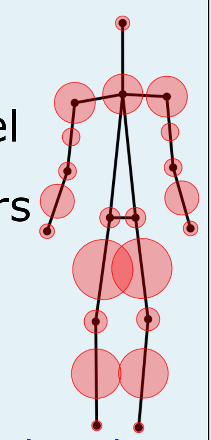
Compression Function: $\phi(X^L) = \arg\max_{H}\ p(X^L, H)$

Decompression Function: $\phi^{-1}(H) = \arg\max_{X^L}\ p(X^L, H)$

Parameter Learning (k_i, β)

- Parameters serve to combine detectors with latent model
- Human symmetry exploited to reduce needed parameters
- Optimized on randomly generated negatives

$$\arg\max_{\boldsymbol{k},\beta}\ \log \mathbb{E}\left(\mathrm{score}(L^{+})\right) - \log \mathbb{E}\left(\mathrm{score}(L^{-})\right)$$



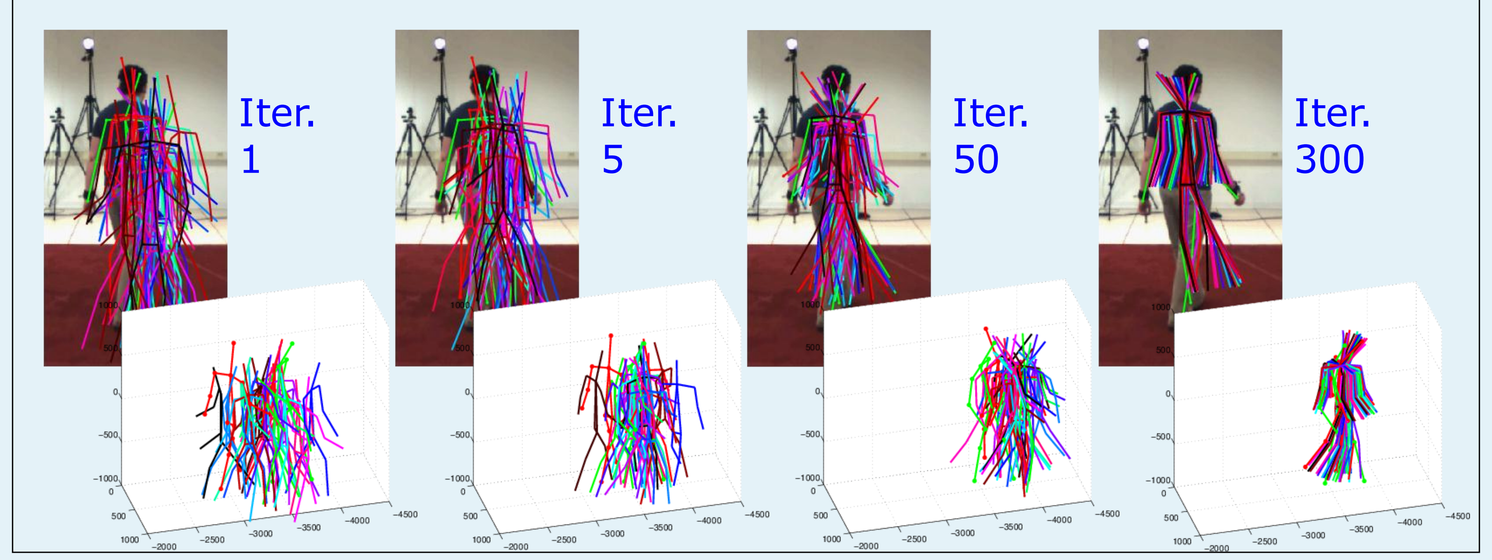
k_i values

Inference

$$<X^{*}> = \arg\max_{X} \prod_{i=1}^{N} p(\boldsymbol{d}_i \mid \boldsymbol{l}_i)\, p(\boldsymbol{l}_i \mid \boldsymbol{x}_i)\, p(X \mid H)\, p(H)$$

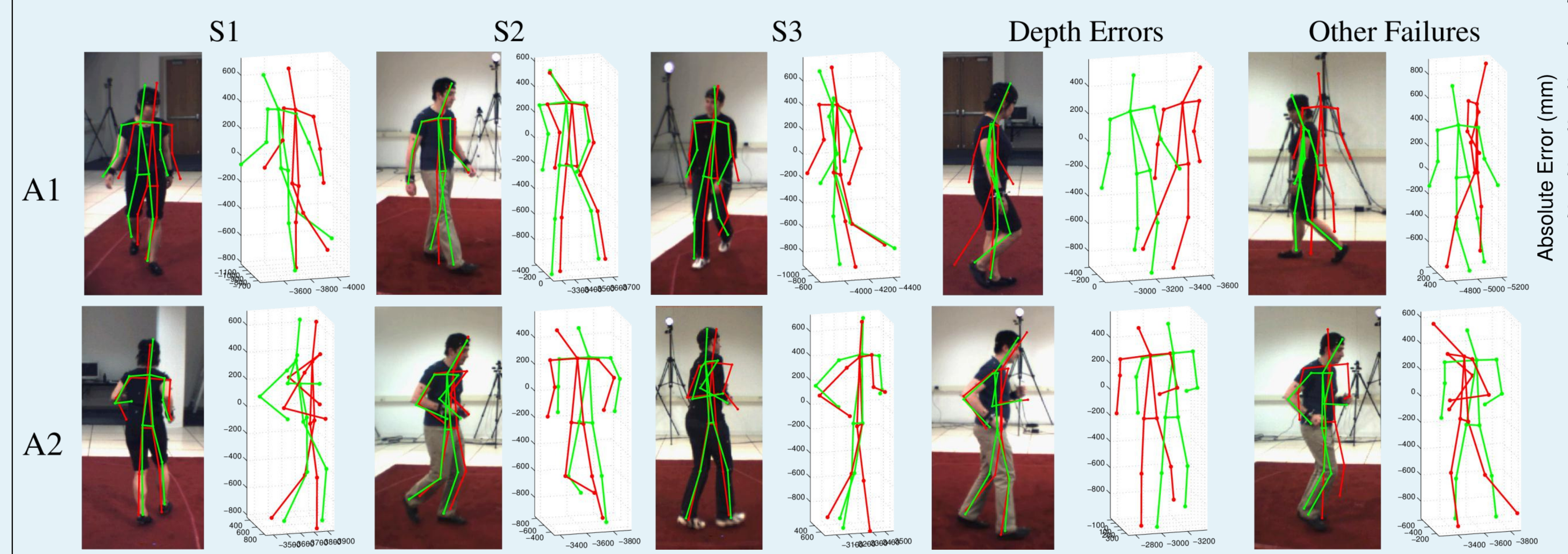
- 3D Pose consists of global transformation and local deformation
- Treated as global optimization problem (using CMA-ES [10]):

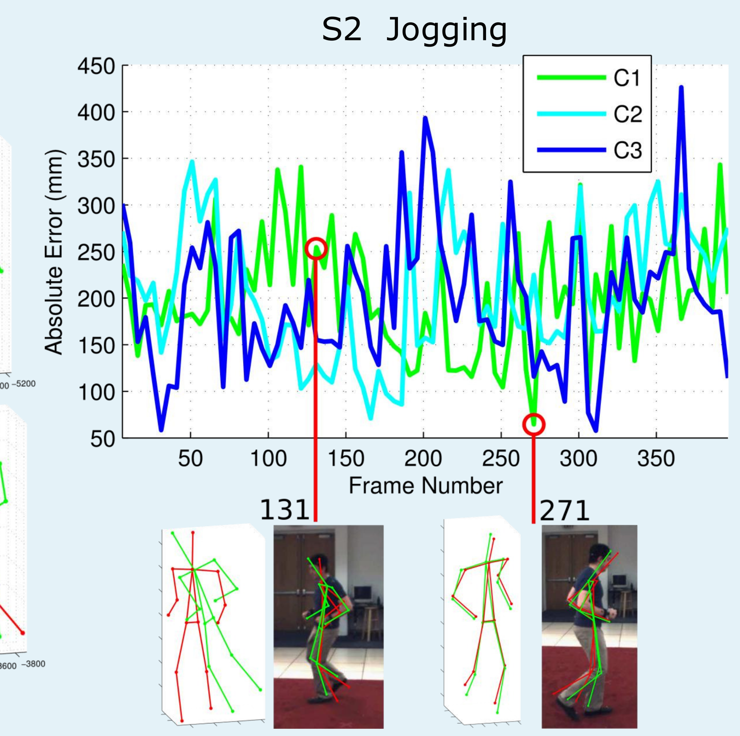
$$\arg\max_{R,\boldsymbol{t},H}\ \mathrm{score}\left(\mathrm{proj}_{R,\boldsymbol{t}}(\phi^{-1}(H))\right) + \log\left(p\left(\phi^{-1}(H), H\right)\right)$$



Experimental Results

HumanEva I Dataset

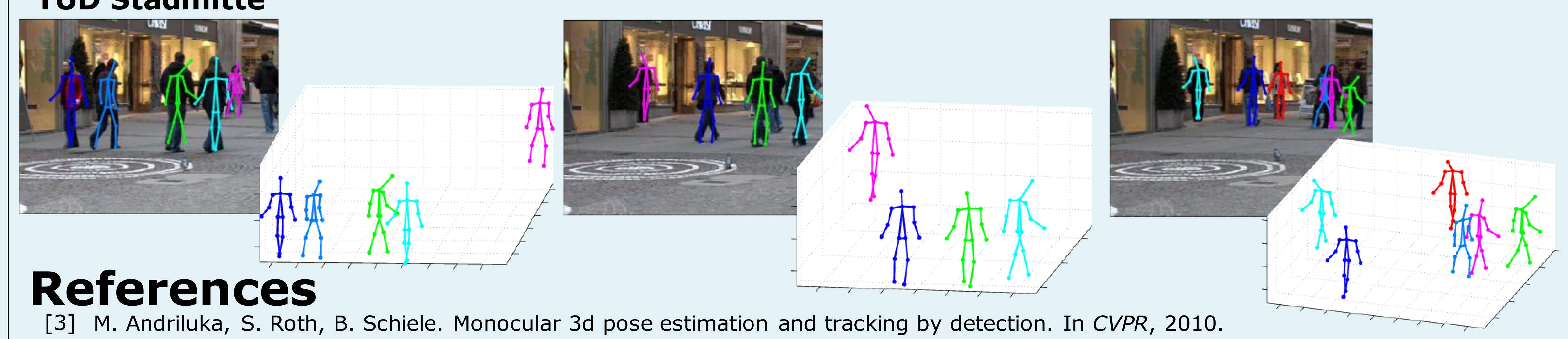




	Walking (A1,C1)		
	S1	S2	S3
Ours	65.1 (17.4)	48.6 (29.0)	73.5 (21.4)
[29] (evaluates fewer frames)	99.6 (42.6)	108.3 (42.3)	127.4 (24.0)
[3] (tracking)	89.3	108.7	113.5
[7] (tracking)	-	107 (15)	-
[6] (background subtraction)	38.2 (21.4)	32.8 (23.1)	40.2 (23.2)
	Jogging (A2,C1)		
	S1	S2	S3
Ours	74.2 (22.3)	46.6 (24.7)	32.2 (17.5)
[29] (evaluates fewer frames)	109.2 (41.5)	93.1 (41.1)	115.8 (40.6)
[6] (background subtraction)	42.0 (12.9)	34.7 (16.6)	46.4 (28.9)

	[29]	Ideal Detector			Our Approach		
Err.	2D	2D	3D	Pose	2D	3D	Pose
All	21.7	11.0	106.6	51.6	19.5	237.3	55.3
C1	19.5	11.1	113.8	52.3	18.9	239.1	55.2
C2	22.9	11.1	109.7	51.2	19.6	245.8	55.4
C3	22.8	10.8	96.2	51.2	20.0	227.1	55.4
S1	21.8	10.2	96.8	63.4	19.9	277.2	69.3
S2	21.8	10.8	108.0	44.8	18.6	206.6	46.8
S3	21.6	12.3	119.0	43.7	20.1	221.4	46.6
A1	20.9	10.7	106.0	56.2	19.3	254.4	60.3
A2	22.7	11.3	107.2	46.6	19.7	219.0	50.0

TUD Stadmitte



References

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

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

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

[10] 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.

[24] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, F. Moreno-Noguer. Single Image 3D Human Pose Estimation from Noisy Observations. In *CVPR*, 2012.

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