Articulated People Detection and Pose Estimation: Reshaping the Future

Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen and Bernt Schiele
Max Planck Institute for Informatics, Saarbrücken
{leonid, ajain, andriluk, thormae, schiele}@mpi-inf.mpg.de

Goal

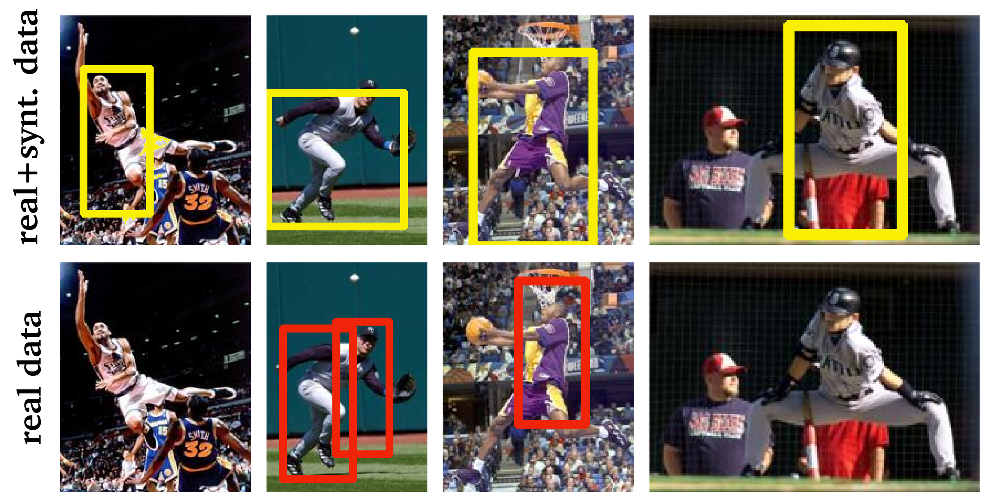
- Detection and pose estimation of highly articulated people in sport scenes

- Address the lack-of-training-data problem by automatically generating novel training samples
- Improve pose estimation by leveraging the strong evidence from people detector

Contributions

- Novel method for automatic generation of multiple training examples from a single image

⇒ training from generated and real data improves the performance over real data alone

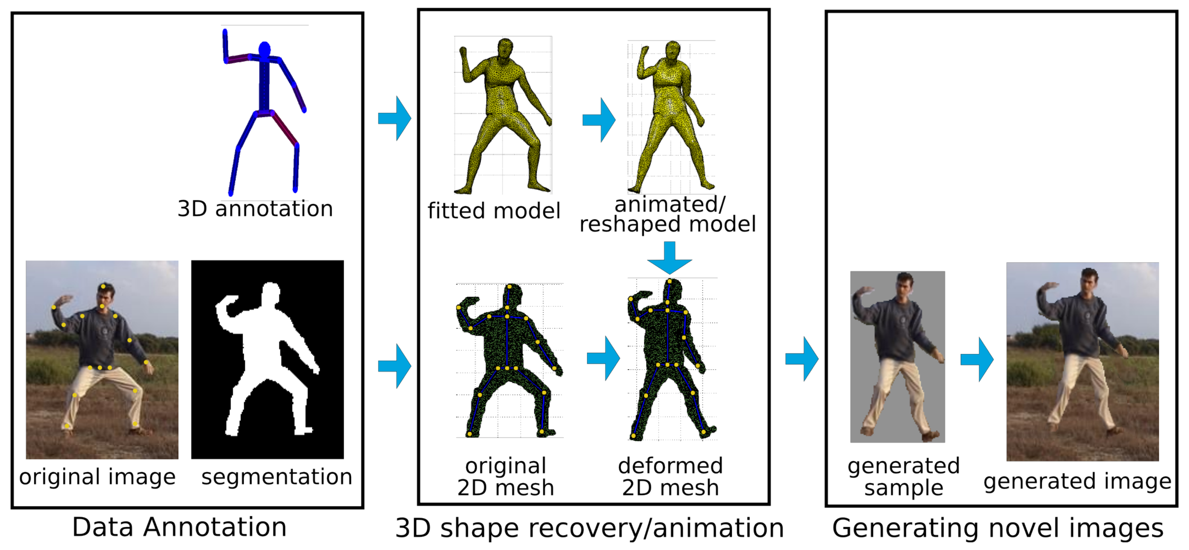


- Propose a joint model: integrate the evidence from DPM [4] into the Pictorial Structures (PS) [3, 5]

- Define a new challenge: joint detection and pose estimation of multiple people *"in the wild"*

⇒ **Data and pose estimation software available**
`https://www.d2.mpi-inf.mpg.de/articulated-data`

Generation of Novel Training Samples



1. Fit a 3D shape model [6] to the annotated 3D pose
2. Reshape and animate the 3D shape model
3. Compute 2D mesh with linear blending weights [7] for the pre-segmented person
4. Morph 2D mesh: use projected 3D shape model joints
5. Render the original appearance to the changed mesh

Statistical 3D human shape model [6]

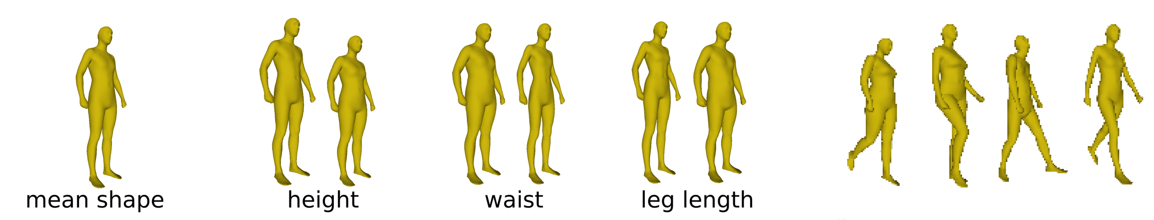
- Shape learned from 3D laser scans of humans
- Represent shape variations via PCA
- Embed kinematic skeleton with linear blend skinning

Model fitting

- Retarget the 3D shape model skeleton to the annotated 3D pose
  - compute inverse kinematics: optimize 3D shape/pose

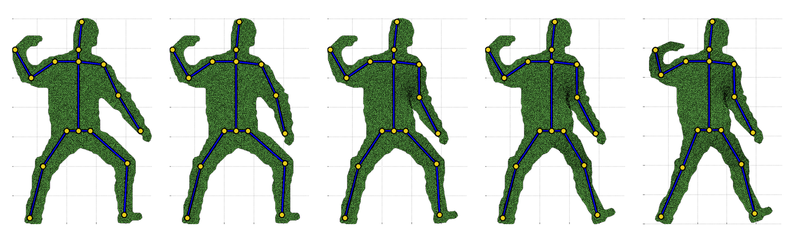
Varying model shape and pose

- Change shape: sample from the 3D shape distribution
- Change pose: use the motion capture data [1]



Generation of novel images

- Use projected 3D model joints to reshape/animate the 2D mesh having the pre-computed skinning weights [7]



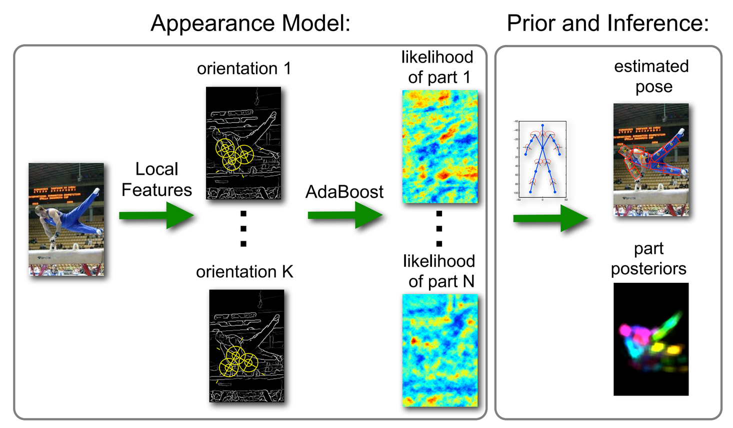
- Render the appearance and project the novel sample into the background image

Models

Pose estimation: Pictorial Structures (PS) [3, 5]

- Flexible configuration of body parts with pose prior

$$p(L|E) \propto \prod_{i=0}^{N} p(e_i(l_i)) \prod_{(i,j)\in G} p(l_i|l_j)$$



People detection: Deformable Part Model (DPM) [4]

⇒ robust *articulated* people detection when trained on our novel samples

Joint PS-DPM model

- Adapt the DPM: train linear regression to predict the torso endpoints from the DMP model parts
- New torso likelihood: $p(e_i(l_i)) = p_{ps}(e_i(l_i))p_{dpm}(e_i(l_i))$

Data

- Example novel training images

- Proposed *"in the wild"* challenge
  - Multi-scale Leeds Sport Poses (LSP)

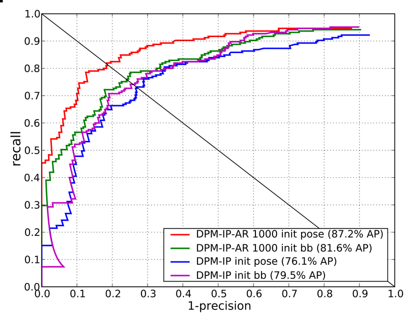
Results

People detection

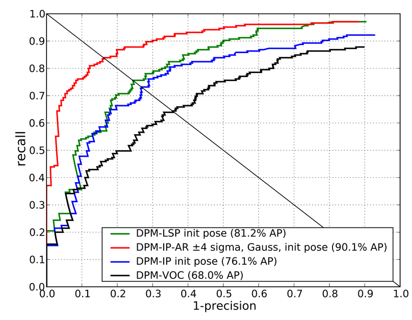
- Image Parsing (IP) [9]

real/synthetic	AP, [%]
100 IP/0	76.1
100 IP/400 R	83.9
100 IP/400 AR	87.2
100 IP/900 AR	**88.6**
100 IP/1900 AR	88.1

Data ratio



Initialization



Comparison to SOTA

Pose estimation

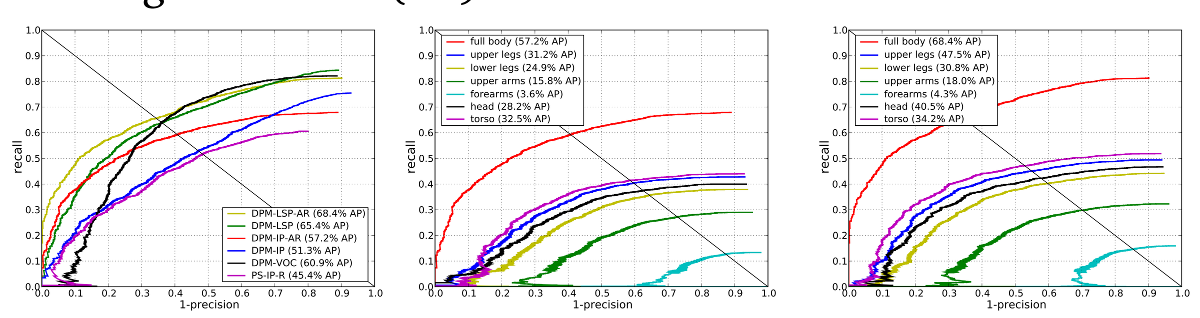
- Image Parsing (IP) [9]
- *Percentage Correct Parts (PCP)* criterion

Setting	Torso	Upper legs	Lower legs	Upper arms	Fore-arms	Head	Total
Image Parsing (IP)	84.9	71.5	61.5	50.2	36.6	71.2	59.6
+ Reshape (R)	87.8	75.1	65.9	52.4	36.1	71.7	61.9
+ Joint PS+DPM	**88.8**	**77.3**	**67.1**	53.7	36.1	73.7	63.1
Andriluka et al., [2] *	83.9	70.5	63.4	50.5	35.1	70.7	59.4
Yang&Ramanan, [10] *	82.9	69.0	63.9	55.1	35.4	**77.6**	60.7
Johnson&Everingham, [8]	87.6	74.7	**67.1**	**67.3**	**45.8**	76.8	**67.4**

* evaluated using our implementation of PCP criterion

People detection/pose estimations *"in the wild"*

- Multi-scale LSP
- *Average Precision (AP)* criterion



Full body | Body parts (DPM-IP-AR + PS) | Body parts (DPM-LSP-AR + J. PS+DPM)

Method	Torso	Upper legs	Lower legs	Upper arms	Fore-arms	Head	Total
DPM-LSP-AR + J. PS+DPM	**40.5**	**37.5**	**30.8**	**18.0**	**4.3**	**34.2**	**25.6**
DPM-LSP-AR	38.9	35.6	29.3	18.0	4.2	33.6	24.7
DPM-IP-AR + PS	32.5	31.2	24.9	15.8	3.6	28.2	21.2
DPM-VOC + PS	29.9	25.2	20.0	14.2	3.6	27.4	18.3
PS-IP-R + PS	29.1	28.7	23.5	14.7	4.0	24.5	19.5

References

[1] Cmu graphics lab motion capture database.
[2] M. Andriluka, S. Roth, and B. Schiele. Discriminative appearance models for pictorial structures. *IJCV'11*.
[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.
[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI'10*.
[5] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV'05*.
[6] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *CGF (Proc. Eurographics'08)*.
[7] A. Jacobson, I. Baran, J. Popović, and O. Sorkine. Bounded biharmonic weights for real-time deformation. In *SIGGRAPH'11*.
[8] S. Johnson and M. Everingham. Learning Effective Human Pose Estimation from Inaccurate Annotation. In *CVPR'11*.
[9] D. Ramanan. Learning to parse images of articulated objects. In *NIPS'06*.
[10] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR'11*.