2D Human Pose Estimation: New Benchmark and State of the Art Analysis

Mykhaylo Andriluka[1,3], Leonid Pishchulin[1], Peter Gehler[2] and Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Germany [2]Max Planck Institute for Intelligent Systems, Germany [3]Stanford University, USA

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Randomly chosen activities and images from 18 top level categories of our MPII Human Pose dataset. One image per activity is shown. The full dataset is available at `human-pose.mpi-inf.mpg.de`.

Overview

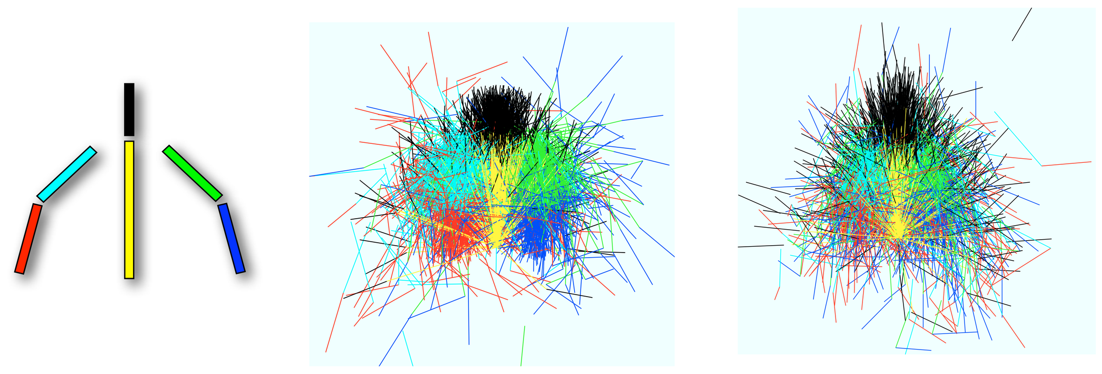
We analyse the state of the art in articulated human pose estimation using a new large-scale benchmark dataset.

- Our **MPII Human Pose** benchmark includes more than **40,000** images systematically collected using an established taxonomy of human activities [1]. The collected images cover **410 human activities** in total.
- Our analysis is based on the **rich annotations** that include body pose, body part occlusion, torso and head viewpoint, and person activity.

- Dataset and **web-based evaluation toolkit** are available at **human-pose.mpi-inf.mpg.de**.

Related Datasets

Dataset	#training	#test	img. type
Full body pose datasets			
Parse [Ramanan, NIPS'06]	100	205	diverse
LSP [Johnson&Everingham, BMVC'10]	1,000	1,000	sports (8 types)
PASCAL Person Layout [Everingham et al., IJCV'10]	850	849	everyday
Sport [Wang et al., CVPR'11]	649	650	sports
UIUC people [Wang et al., CVPR'11]	346	247	sports (2 types)
LSP extended [Johnson&Everingham, CVPR'11]	10,000	-	sports (3 types)
FashionPose [Dantone et al., CVPR'13]	6,530	775	fashion blogs
J-HMDB [Jhuang et al., ICCV'13]	31,838	-	diverse (21 act.)
Upper body pose datasets			
Buffy Stickmen [Ferrari et al, CVPR'08]	472	276	TV show (Buffy)
ETHZ PASCAL Stickmen [Eichner&Ferrari, BMVC'09]	-	549	PASCAL VOC
Human Obj. Int. (HOI) [Yao&Fei-Fei, CVPR'10]	180	120	sports (6 types)
We Are Family [Eichner&Ferrari, ECCV'10]	350 imgs.	175 imgs.	group photos
Video Pose 2 [Sapp et al., CVPR'11]	766	519	TV show (Friends)
FLIC [Sapp&Taskar, CVPR'13]	6,543	1,016	feature movies
Sync. Activities [Eichner&Ferrari, PAMI'12]	-	357 imgs.	dance / aerobics
Armlets [Gkioxari et al., CVPR'13]	9,593	2,996	PASCAL VOC/Flickr
MPII Human Pose (this paper)	**28,821**	**11,701**	diverse (491 act.)



(a) (b) (c)

Visualization of the upper body pose variability. (a) color coding of the body parts, (b) annotations from the Armlets dataset [Gkioxari et al. CVPR'13], and (c) annotations from our MPII Human Pose dataset.

Analysis of the state of the art

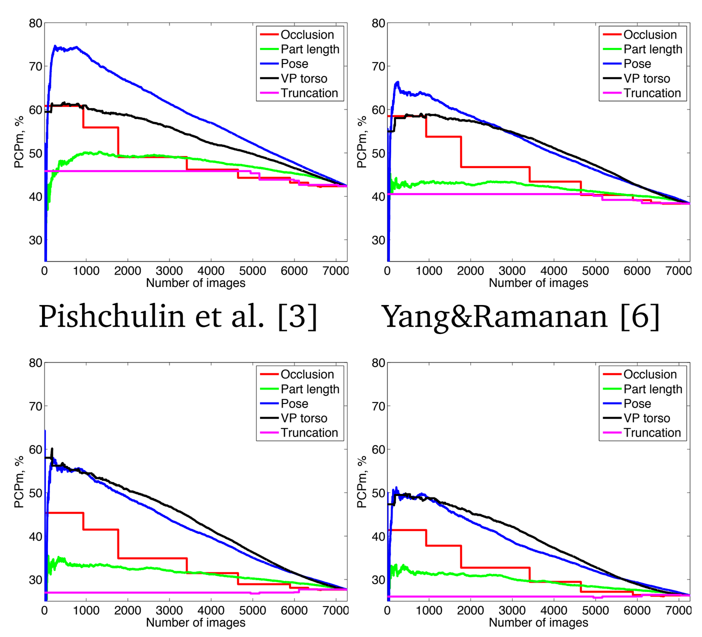
- Performance of the upper-body and full-body state-of-the-art approaches:

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Upper body	Full body
Gkioxari et al. [2]	51.3	-	-	28.0	12.4	-	26.4	-
Sapp&Taskar [5]	51.3	-	-	27.4	16.3	-	27.8	-
Yang&Ramanan [6]	61.0	36.6	36.5	34.8	17.4	70.2	33.1	38.3
Pishchulin et al. [3]	63.8	39.6	37.3	39.0	26.8	70.7	39.1	42.3
Gkioxari et al. [2] + loc	65.1	-	-	33.7	14.9	-	32.4	-
Sapp&Taskar [5] + loc	65.1	-	-	32.6	19.2	-	33.7	-
Yang&Ramanan [6] + loc	**67.2**	39.7	39.4	37.4	18.6	**75.7**	35.8	41.4
Pishchulin et al. [3] + loc	66.6	**40.5**	**38.2**	**40.4**	**27.7**	74.5	**40.6**	**43.9**
Yang&Ramanan [6] retrained	**69.3**	39.5	38.8	**43.4**	27.7	74.6	**42.3**	44.7
Pishchulin et al. [3] retrained	68.4	**42.7**	**42.8**	42.0	**29.2**	**76.3**	42.1	**46.1**

⇒ PS and FMP models improve significantly after retraining.

⇒ Full-body approaches perform better than upper-body ones.

- We define complexity measures with respect to *body pose*, *viewpoint of the torso*, *body part length*, *occlusion*, and *truncation*, and analyze sensitivity of the state-of-the-art approaches to each of these factors:



Pishchulin et al. [3] Yang&Ramanan [6]

Sapp&Taskar [5] Gkioxari et al. [2]

Complexity measures:

Occlusion: number of occluded body parts
$m_{occ}(L) = \sum_{i=1}^{N} \rho_i$

Part length: average deviation from the mean part length
$m_f(L) = \sum_{i=1}^{N} |d(l_i) - m_i|/m_i$

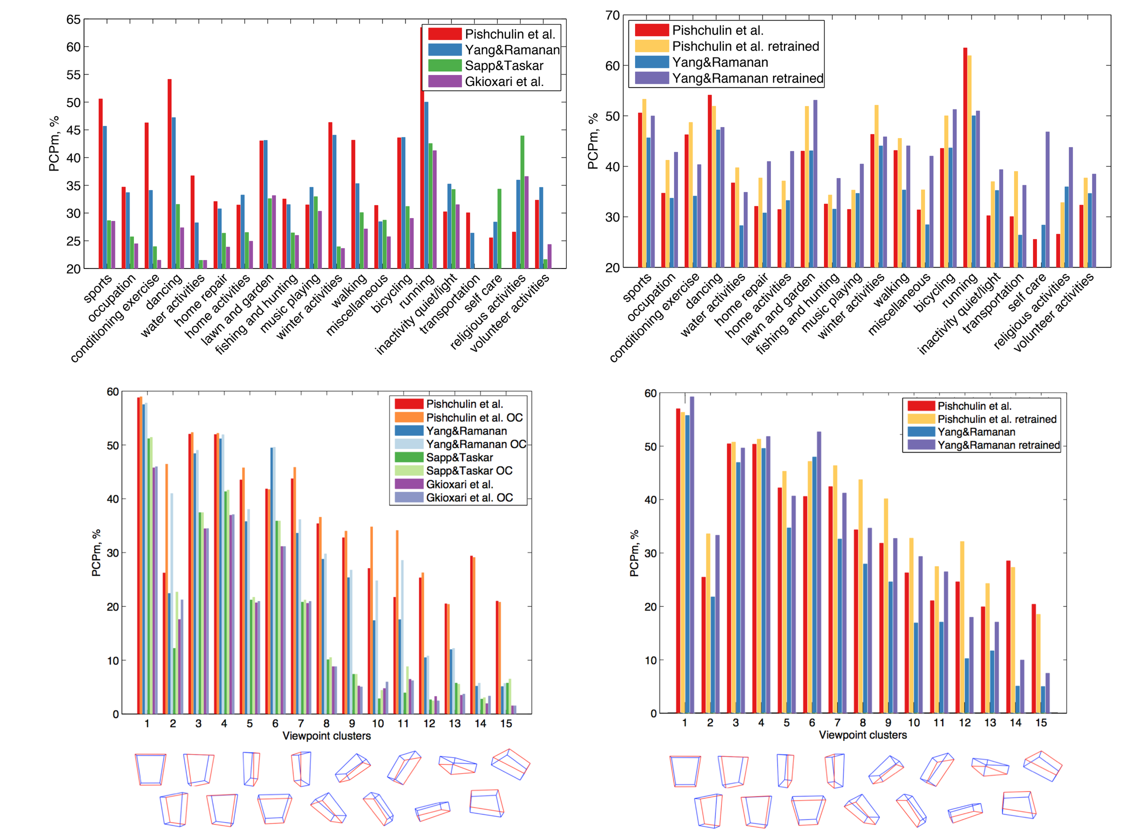
Pose: deviation from the mean pose represented via PS prior distribution
$m_{pose}(L) = \prod_{(i,j)\in E} p_{ps}(l_i|l_j)$

VP torso: deviation of the torso from the frontal viewpoint
$m_v(L) = \sum_{i=1}^{3} \alpha_i$

Truncation: number of truncated body parts
$m_t(L) = \sum_{i=1}^{N} \tau_i$

⇒ Body pose has a significantly more profound effect on performance than occlusion and truncation.

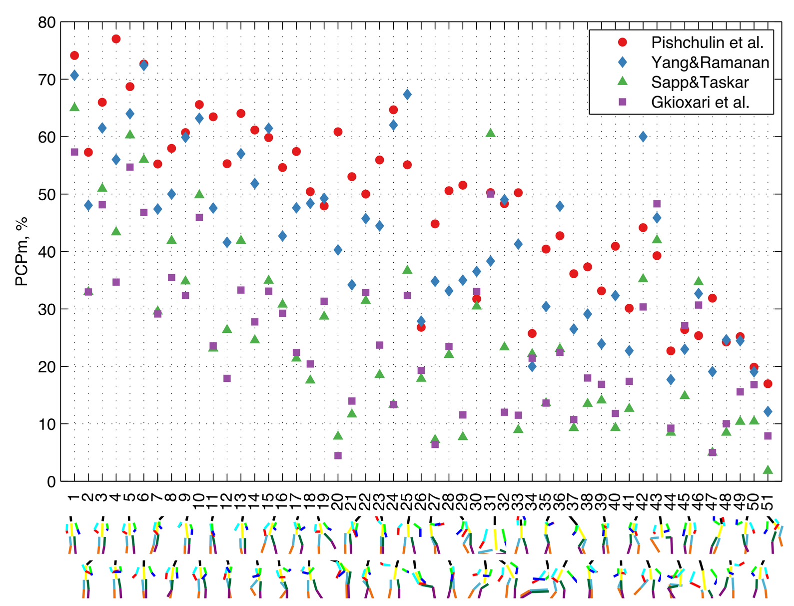
- Detailed performance analysis for activity, viewpoint and body poses types:



⇒ Striking performance differences for all approaches across activities and viewpoints even after retraining.

⇒ Results on sports activities are the best, even though they have been previously considered among the most difficult for pose estimation.

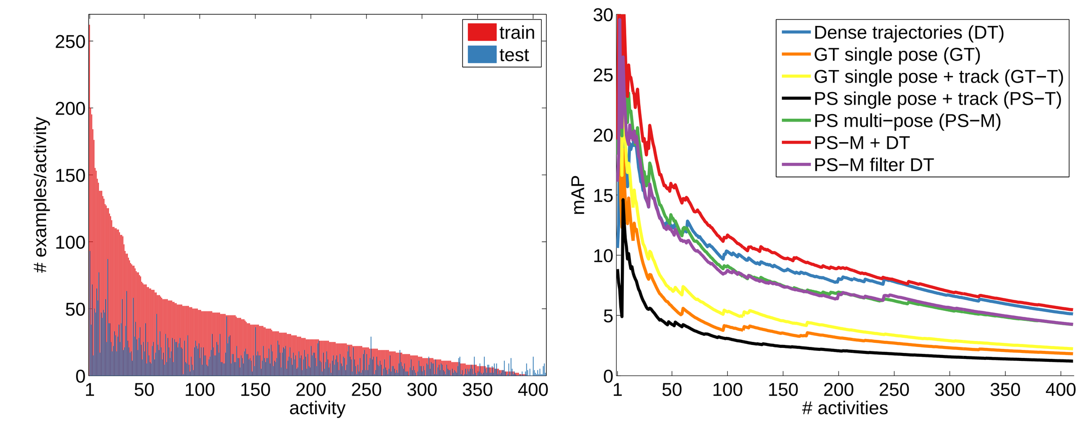
⇒ Detailed analysis reveals differences between FMP and PS, even though overall performance is comparable.



Performance (mPCP) for different pose types.

Activity recognition

We evaluate the performance of the state-of-the-art activity recognition methods on our benchmark in [4].



(a) (b)

(a) number of examples per activity. (b) Comparison of the activity recognition performance of [Wang et al., ICCV'13] (DT) and variants of posed-based approach of [Jhuang et al.,ICCV'13]. The plot shows performance (mAP) as a function of the number of activities. Activities are sorted by the number of training examples.

	yoga, power	bicycling, mountain	skiing, downhill	cooking or food prep.	skate-boarding	rope skipping	softball, general	forestry
Dense trajectories (DT)	10.6	14.5	51.9	0.5	11.4	36.0	**12.7**	8.4
PS multi-pose (PS-M)	18.3	34.0	27.3	2.6	17.2	**90.5**	3.0	5.2
PS-M + DT	19.6	**40.7**	32.9	2.2	**19.5**	88.7	3.9	7.2
PS-M filter DT	16.1	20.4	**52.2**	0.8	13.5	55.7	4.2	**10.6**

	carpentry, general	bicycling, racing	golf	rock climbing	ballet, modern	aerobic step	resistance training	total
Dense trajectories (DT)	5.5	5.5	33.0	**41.5**	12.7	24.5	**16.5**	19.0
PS multi-pose (PS-M)	3.4	8.6	47.9	4.7	22.9	10.4	7.2	20.2
PS-M + DT	5.0	12.1	**51.9**	14.4	**23.7**	17.1	14.4	**23.5**
PS-M filter DT	**6.1**	**15.5**	15.9	38.6	7.1	**25.8**	9.6	19.5

Activity recognition results (mAP) on the 15 largest classes.

References

[1] B. Ainsworth, W. Haskell, S. Herrmann, N. Meckes, D. Bassett, C. Tudor-Locke, J. Greer, J. Vezina, M. Whitt-Glover, and A. Leon. 2011 compendium of physical activities: a second update of codes and MET values. *MSSE'11*.

[2] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *CVPR'13*.

[3] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV'13*.

[4] L. Pishchulin, M. Andriluka, and B. Schiele. Fine-grained activity recognition with holistic and pose based features. *arXiv:1406.1881*, 2014.

[5] B. Sapp and B. Taskar. Multimodal decomposable models for human pose estimation. In *CVPR'13*.

[6] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI'13*.