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

Abstract

Human pose estimation has made significant progress during the last years. However current datasets are limited in their coverage of the overall pose estimation challenges. Still these serve as the common sources to evaluate, train and compare different models on. In this paper we introduce a novel benchmark "MPII Human Pose"[1] that makes a significant advance in terms of diversity and difficulty, a contribution that we feel is required for future developments in human body models. This comprehensive dataset was collected using an established taxonomy of over 800 *human activities [1]. The collected images cover a wider variety of human activities than previous datasets including various recreational, occupational and householding activities, and capture people from a wider range of viewpoints. We provide a rich set of labels including positions of body joints, full 3D torso and head orientation, occlusion labels for joints and body parts, and activity labels. For each image we provide adjacent video frames to facilitate the use of motion information. Given these rich annotations we perform a detailed analysis of the leading human pose estimation approaches gaining insights for the success and failures of these methods.*

1. Introduction

Recent pose estimation methods employ complex appearance models [2, 9, 15] and rely on learning algorithms to estimate model parameters from the training data. The performance of these approaches crucially depends on the availability of annotated training images that are representative for the appearance of people clothing, strong articulation, partial (self-)occlusions and truncation at image borders. Although there exists training sets for special scenarios such as sport scenes [12, 13] and upright people [17, 2], these benchmarks are still limited in their scope and variability of represented activities. Sport scene datasets typically include highly articulated poses, but are limited with respect to variability of appearance since people are typically wearing tight sports outfits. In turn, datasets such as "FashionPose" [2] and "Armlets" [9] aim to collect images of people wearing a variety of different clothing types, and include occlusions and truncation but are dominated by people in simple upright standing poses.

To the best of our knowledge no attempt has been made to establish a more representative benchmark aiming to cover a wide pallet of challenges for human pose estimation. We believe that this hinders further development on this topic and propose a new benchmark "MPII Human Pose". Our benchmark significantly advances state of the art in terms of appearance variability and complexity, and includes more than 40,000 images of people. We used YouTube as a data source and collected images and image sequences using queries based on the descriptions of more than 800 activities. This results in a diverse set of images covering not only different activities, but indoor and outdoor scenes, a variety of imaging conditions, as well as both amateur and professional recordings (*c.f.* Fig. 1). This allows us to study existing body pose estimation techniques and identify their individual failure modes.

Related work The commonly used publicly available datasets for evaluation of 2D human pose estimation are summarized in Tab. 1 according to the year of the corresponding publication. Both full body and upper body datasets are included.

Existing benchmarks cover aspects of the human pose estimation task such as sport scenes [12, 21], frontal-facing people [8, 3, 17], people interacting with objects [23], pose estimation in group photos [5] and pose estimation of people performing synchronized activities [4].

Earlier datasets such as "Parse" [16] and "Buffy" [8] are still commonly found in evaluations [22, 15]. However the small training sets included in these datasets make them unsuitable for training models with complex appearance representations and multiple components [13, 17, 2], which have been shown to perform best.

[1]Available at human-pose.mpi-inf.mpg.de.

Figure 1. Randomly chosen images from each of 20 activity categories of the proposed "MPII Human Pose" dataset. Image captions indicate activity category (1st row) and activity (2nd row). To view the full dataset visit human-pose.mpi-inf.mpg.de.

Some efforts have been made to collect larger sets of images. For example [13] extends the LSP dataset to 10,000 images of people performing gymnastics, athletics and parkour. [2] proposes a large "FashionPose" dataset collected from fashion blogs. This dataset aims to cover a wide variety in people clothing. The LSP and FashionPose datasets are complementary and focus on two different challenges for human pose estimation: pose variability and variability of people appearance. However since they are collected with a specific focus in mind, these datasets do not cover real-life challenges such as truncation, occlusions by scene objects and variability of imaging conditions.

The works of [6] and [9] propose a challenging dataset building on the PASCAL VOC image collection. Results reported in [9] indicate that the best performing approaches for pose estimation of people in the presence of occlusion and complex appearance are under-performing on sport-oriented datasets such as LSP [12] and vice versa. There are qualitative differences between methods that work well for LSP and "Armlets" datasets. On LSP the best performing methods are typically based on flexible part-based models that are well suited for capturing pose variability. In contrary on the "Armlets" dataset the best performing approach [9] uses a set of rigid detectors for groups of parts, that are more robust to the variability in appearance.

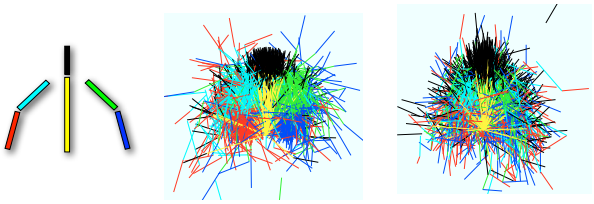
Our dataset is complementary to the J-HMDB dataset [11] and provides more images and a wider coverage of activities (491 in our dataset vs. 21 in J-HMDB), whereas J-HMDB provides densely annotated image sequences and larger number of videos for each activity. Our dataset also addresses a different set of challenges compared to the datasets such as "HumanEva" [19] and "Human3.6M" [10] that include images and 3D poses of people but are captured in the controlled indoor environments, whereas our dataset includes real-world images but provides 2D poses only.

Dataset	#training	#test	img. type
Full body pose datasets			
Parse [16]	100	205	diverse
LSP [12]	1,000	1,000	sports (8 types)
PASCAL Person Layout [6]	850	849	everyday
Sport [21]	649	650	sports
UIUC people [21]	346	247	sports (2 types)
LSP extended [13]	10,000	-	sports (3 types)
FashionPose [2]	6,530	775	fashion blogs
J-HMDB [11]	31,838	-	diverse (21 act.)
Upper body pose datasets			
Buffy Stickmen [8]	472	276	TV show (Buffy)
ETHZ PASCAL Stickmen [3]	-	549	PASCAL VOC
Human Obj. Int. (HOI) [23]	180	120	sports (6 types)
We Are Family [5]	350 imgs.	175 imgs.	group photos
Video Pose 2 [18]	766	519	TV show (Friends)
FLIC [17]	6,543	1,016	feature movies
Sync. Activities [4]	-	357 imgs.	dance / aerobics
Armlets [9]	9,593	2,996	PASCAL VOC/Flickr
MPII Human Pose (this paper)	**28,821**	**11,701**	diverse (491 act.)

Table 1. Overview of the publicly available datasets for articulated human pose estimation. For each dataset we report the number of annotated people in training and test sets and the type of images the set include. The numbers indicate the number of unique annotated people without mirroring.

2. Dataset

In this paper we introduce a large dataset of images that covers a wide variety of human poses and clothing types and includes people interacting with various objects and environments. The key rationale behind our data collection strategy is that we want to represent both common and rare human poses that might be missed when simply collecting more images without aiming for good coverage. To this end, we use a two-level hierarchy of human activities proposed in [1] to guide the collection process. This hierarchy was developed for the assignment of standardized energy levels during physical activity surveys and includes 823 activities in total of 21 different activity categories. The activities at the first level of the hierarchy correspond to thematically related groups of activities such as "Home Activities", "Lawn and Garden" or "Sports". The activities at the second level then correspond to individual activities such as "Washing windows", "Picking fruit" or "Rock climbing". Note that using the activity hierarchy for collection has an additional advantage that all images have an associated activity label. As a result one can assess and analyze any performance measure also on subsets of activities or activity categories.



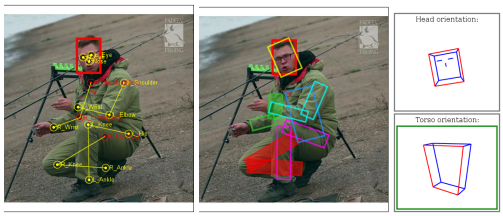
(a) (b) (c)

Figure 2. Visualization of upper body pose variability. From left to right we show, (a) color coding of the body parts (b) annotations of the "Armlets" dataset [9], and (c) annotations of this dataset.

Due to the coverage of the hierarchy the images in our dataset are representative of the diversity of human poses, overcoming one of the main limitations of previous collections. In Fig. 2 we visualize this diversity by comparing upper body annotations of the "Armlets" dataset Fig. 2(b) and our proposed dataset (c). Note that although "Armlets" contain about 13,500 images, the annotations resemble a person with arms down along the torso (distribution of red, cyan, green, and blue sticks).

We collect images from YouTube using queries based on the activity descriptions. Using YouTube allows us to access a rich collection of videos originating from various sources, including amateur and professional recordings and capturing a variety of public events and performances. In Fig. 2 (c) we show the distribution of upper body poses on our dataset. Note the variability in the location of hands and the absence of distinctive peaks for the upper and lower arms that are present in the case of the "Armlets" dataset.

Data collection. As a first step of the data collection we manually query YouTube using descriptions of activities from [1]. We select up to 10 videos for each activity filtering out videos of low quality and those that do not include people. This resulted in $3,913$ videos spanning 491 different activities. Note that we merged a number of the original 823 activities due to high similarity between them, such as cycling at different speeds. In the second step we manually pick several frames with people from each video. As the focus of our benchmark is pose estimation we do not include video frames in which people are severely truncated or in which pose is not recognizable due to poor image quality or small scale. We aim to select frames that either depict different people present in the video or the same person in a substantially different pose. In addition we restrict the selected frames to be at least 5 seconds apart. This step resulted to a total of $24,920$ extracted frames from all collected videos. Next, we annotate all people present in the collected images, but ignore dense people crowds in which significant number of people are almost fully occluded. Following this procedure we collect images of $40,522$ people. We allocate roughly tree quaters of the collected images for training and use the rest for testing. Images from the same video are either all in the training or all in the test set. This results in a training/test set split of $28,821$ to $11,701$.



(a) (b) (c)

Figure 3. Example of the provided annotations. Annotated are (a) positions and visibility of the main body joints, locations of the eyes and nose and the head bounding box (occluded joints are shown in red), (b) occlusion of the main body parts (occluded parts are shown with filled rectangles), and (c) 3D viewpoints of the head and torso. On the illustration the viewpoint is shown using a simplified body model, the front face of the model is shown in red.

Data annotation. We provide rich annotations for the collected images, an example can be seen in Fig. 3. Annotated are the body joints, 3D viewpoint of the head and torso, and position of the eyes and nose. Additionally for all body joints and parts visibility is annotated. Following [13, 9] we annotate joints in a "person centric" way, meaning that the left/right joints refer to the left/right limbs of the person. At test time this requires pose estimation with both a correct localization of the limbs of a person along with the correct match to the left/right limb. The annotations are performed by in-house workers and via Amazon Mechanical Turk (AMT). In our annotation process we build and extend the annotation tools described in [14]. Similarly to [13, 20] we found that effective use of AMT requires careful selection of qualified workforce. We pre-select AMT workers based on a qualification task, and then maintain data quality by manually inspecting the annotated data.

Experimental protocol and evaluation metrics. We define the baseline evaluation protocol on our dataset following the current practices in the literature [13, 9, 15]. We assume that at test time the rough location and scale of a person are known, and we exclude the cases with multiple people in close proximity to each other from the evaluation. We feel that these simplifications are necessary for the rapid adoption of the dataset as the majority of the current approaches does not address multiple people pose estimation and does not search over people positions and scales.

We consider three metrics as indicators for the pose estimation performance. The widely adopted "PCP" metric [8] that considers a body part to be localized correctly if the estimated body segment endpoints are within 50% of the ground-truth segment length from their true locations. The "PCP" metric has a drawback that foreshortened body parts should be localized with higher precision to be considered correct. We define a new metric denoted as "PCPm" that uses 50% of the mean ground-truth segment length over the

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Upper body	Full body
Gkioxari et al. [9]	51.3	-	-	28.0	12.4	-	26.4	-
Sapp&Taskar [17]	51.3	-	-	27.4	16.3	-	27.8	-
Yang&Ramanan [22]	61.0	36.6	36.5	34.8	17.4	70.2	33.1	38.3
Pishchulin et al. [15]	63.8	39.6	37.3	39.0	26.8	70.7	39.1	42.3
Gkioxari et al. [9] + loc	65.1	-	-	33.7	14.9	-	32.4	-
Sapp&Taskar [17] + loc	65.1	-	-	32.6	19.2	-	33.7	-
Yang&Ramanan [22] + loc	**67.2**	39.7	39.4	37.4	18.6	**75.7**	35.8	41.4
Pishchulin et al. [15] + loc	66.6	**40.5**	**38.2**	**40.4**	**27.7**	74.5	**40.6**	**43.9**

Table 2. Pose estimation results (PCPm) on the proposed dataset without and with using rough body location ("+ loc" in the table).

entire test set as a matching threshold, but otherwise follows the definition of "PCP". Finally, we consider the "PCK" metric from [22] that measures accuracy of the localization of the body joints. In [22] the threshold for matching of the joint position to the ground-truth is defined as a fraction of the person bounding box size. We use a slight modification of the "PCK" and define the matching threshold as 50% of the head segment length. We denote this metric as "PCKh". We choose to use head size because we would like to make the metric articulation independent.

3. Analysis of the state of the art

In this section we analyse the performance of leading human pose estimation approaches on our benchmark. We take advantage of our rich annotations and conduct a detailed analysis of various factors influencing the results, such as foreshortening, activity and viewpoint, previously not possible in this detail. The goal of this analysis is to evaluate the robustness of the current approaches in various challenges for articulated pose estimation, identify the existing limitations and stimulate further research advances.

In our analysis we consider two full body and two upper body pose estimation approaches. The full body approaches are the version 1.3 of the *flexible mixture of parts* (FMP) approach of Yang and Ramanan [22] and the *pictorial structures* (PS) approach of Pishchulin *et al*. [15]. The upper body pose estimation approaches are the *multimodal decomposable models* (MODEC) approach of Sapp *et al*. [17] and the *Armlets* approach of Gkioxari *et al*. [9]. In case of FMP and MODEC we use publicly available code and pre-trained models. The PS model used here corresponds to our best model published in [15]. In case of the Armlets model, the code and pre-trained model provided by the authors correspond to the version from [9] that includes the HOG features only. The performance of our version of Armlets on the "Armlets" dataset is 3.3 PCP lower than the version based on combination of all features.[2]

Note that the approaches considered in this evaluation are the best performing ones in their respective categories. The PS approach achieves the best results to date on LSP that is focused on the strongly articulated people [15]. The Armlets approach is best on the "Armlets" dataset [9] that includes large number of truncation and occlusions, and MODEC is the best on the recent upper body pose estimation dataset "FLIC" [17]. We include the FMP approach that is widely used in the literature and typically shows competitive performance for a variety of settings. In the following experiments we use "PCPm" as our working metric, while also providing results for "PCP" and "PCKh" in the supplementary material. While we observe little performance differences when using each metric, all conclusions obtained during "PCPm"-based evaluation are valid for "PCP" and "PCKh"-based evaluations as well.

Overall performance evaluation. We begin our analysis by reporting the overall pose estimation performance of each approach and summarize the results in Tab. 2. We include both upper- and full body results to enable comparison across different models. The PS approach achieves the best result of 42.3% PCPm, followed by the FMP approach with 38.3% PCPm. On the upper body evaluation, PS performs best with 39.1%, while both MODEC (27.8% PCPm) and Armlets (26.4% PCPm) perform significantly worse.

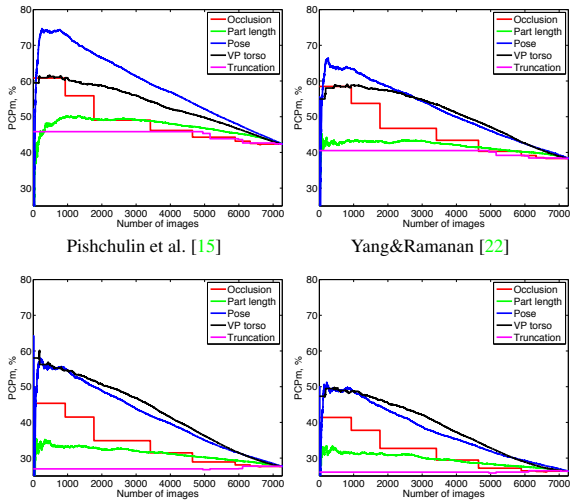
The interesting outcome of this comparison is that both upper body approaches MODEC and Armlets are outperformed by the full body approaches PS and FMP evaluated on upper body only. This is interesting because significant portion of the dataset (15 %) includes people that have only upper body visible. It appears that the PS and FMP approaches are sufficiently robust to missing parts to produce reliable estimates even in the case of lower body occlusion.

Lower part of Tab. 2 shows the results when using provided rough location of person during test time inference. We observe, that while the performance increases for all methods, upper body approaches profit at most, as they heavily depend on correct torso localization. For the sake of fair comparison among the methods, we *do not* use the rough location in the following experiments. Another interesting outcome is that the achieved performance is substantially lower than current best results on the sports-centric LSP dataset, but comparable to results on the "Armlets" dataset (42.2 PCP on our benchmark (see supplemental) vs. 69.2 on LSP [15] vs. 36.2 PCP on "Armlets"). This suggests that sport activities are not necessary the most difficult cases for pose estimation; challenges such as appearance variability, occlusion and truncation apparently deserve more attention in the future.

3.1. Analysis of pose estimation challenges

We now analyse the performance of each approach with respect to the following five factors: part occlusion, foreshortening, body pose, viewpoint, and activity of the person. For the purpose of this analysis we define quantitative

[2] See Tab.1 in [9] for the comparison.



Pishchulin et al. [15] Yang&Ramanan [22]

Sapp&Taskar [17] Gkioxari et al. [9]

Figure 4. Performance (PCPm) as a function of the five complexity measures.

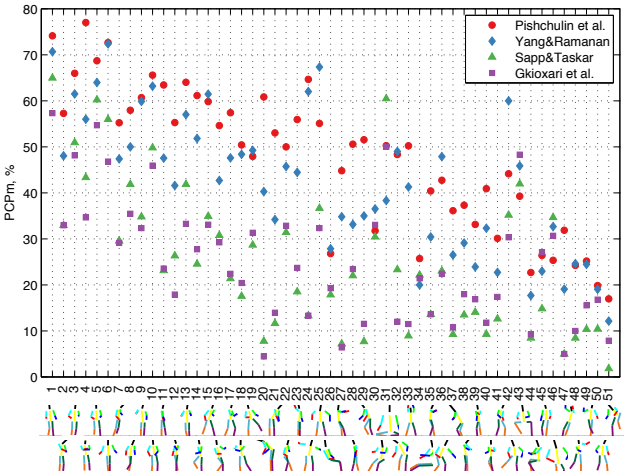


Figure 5. Performance (PCPm) on images clustered by full body pose. Clusters are ordered by increasing mean pose complexity and representatives are shown beneath. Results using upper body and lower body clusters can be found in supplementary material.

complexity measures that map body image annotations to a real value that relates to the complexity of the image with respect to each factor.

Let us denote the annotation of the person by $L = \{L^{pose}, L^{view}, L^{vis}\}$, where $L^{pose} = \{l_i, i = 1, \ldots, N\}$ corresponds to the positions of body parts, $L^{view} = \{\alpha_1, \alpha_2, \alpha_3\}$ are the Euler angles representation of the torso rotation, and $L^{vis} = \{(\rho_i, \theta_i), i = 1, \ldots, N\}$ encodes body part visibility via a set of occlusion labels $\rho_i \in \{0, 1\}$ and truncation labels $\theta_i \in \{0, 1\}$.

We define the following complexity measures. Pose complexity is measured as the deviation from the mean pose on the entire dataset. We define $\mathrm{m}_{pose}(L) = \prod_{(i,j)\in E} p_{ps}(l_i|l_j)$, where E is a set of body joints and $p_{ps}(l_i|l_j)$ is a Gaussian distribution measuring relative position of the two adjacent body parts using the transformed state-space representation introduced in [7]. Note that $\mathrm{m}_{pose}(L)$ corresponds to the likelihood of the pose under the tree structured pictorial structures model [7]. The amount of foreshortening is measured by $\mathrm{m}_f(L) = \sum_{i=1}^{N} |d(l_i) - m_i|/m_i$, where $d(l_i)$ is the length of the body part i, and m_i is the mean length over the entire dataset. The viewpoint complexity is measured by the deviation from the frontal viewpoint: $\mathrm{m}_v(L) = \sum_{i=1}^{3} \alpha_i$. Finally, the amount of occlusion and truncation correspond to the number of occluded and truncated body parts: $\mathrm{m}_{occ} = \sum_{i=1}^{N} \rho_i$, and $\mathrm{m}_t = \sum_{i=1}^{N} \tau_i$.

Performance as a function of the complexity measures To visualize the influence of the various factors on pose estimation performance we plot PCPm scores for the images sorted in the order of increasing complexity (see Fig. 4). In general and as expected, the performance drops for all measures as the complexity increases. There are interesting differences however. Body pose complexity clearly influences the performance of all approaches the most. The second most influential factor is the viewpoint of the torso. For upper body pose estimation approaches this factor is equally influential as body pose. The third most influential factors is occlusion while for the full body estimation approaches this is equally influential as the torso orientation. Contrary to our expectation we found that the part length is less influential. Part length and in particular foreshortening effects are considered to be the key difficulties for both pose estimation. Based on this analysis the above mentioned factors have a higher influence on the performance. The least influential factor is truncation having the smallest effect. In the case of upper body estimation the performance even slightly increases as the amount of truncation increases due to two factors. As truncation if more likely for the lower body these approaches suffer less from truncation and also truncated poses are biased towards frontal views for which the methods are more suited. We now discuss and analyze each factor in more detail.

Body pose performance. As stated above the complexity of the pose is a dominating factor for the performance of all considered approaches. For example the PS approach achieves 72.8% PCPm on the 1000 images with lowest pose complexity, compared to 42.3% for the entire dataset. The same is true for the FMP model, 63.4% PCPm on 1000 least pose complex images vs. 38.3% overall.

To highlight variations in performance across different body configurations we cluster the test images according to the body pose and measure performance for each cluster. We repeat this three times, clustering all body joints, only the upper body joints, and finally the lower body joints.

In the latter two cases we measure performance on the upper/lower body parts only. These three clusterings correspond both to different types of challenges as well as applications. Furthermore, this allows to directly compare full vs. upper body techniques. We show the average PCPm for all full body clusters with more than 25 examples in Fig. 5 ordering the results from left to right by increasing mean pose complexity. Note the significant variations in performance across different clusters. For example, results on full body clusters vary between 77% and 2% PCPm. The best performance is achieved on clusters with poses similar to the mean pose *e.g*. clusters 1 and 5 (see Figure 5). Examining clusters with poor performance we immediately discover several failure modes of PS and FMP approaches. Consider the clusters 42 and 43 that correspond to people with slightly foreshortened torso. FMP improves over PS by 14% PCPm on cluster 25 (54% PCPm for PS vs. 68% PCPm for FMP) and by 16% PCMm on cluster 42 (44% PCPm for PS vs. 60% PCPm for FMP), as it can better model torso foreshortening by representing torso as configuration of multiple flexible parts, whereas PS models torso as a single rigid part. Also, the flexibility of FMP model accounts for its better performance on frontal sitting people (cluster 43) where FMP improves over PS by 7% PCPm (46% PCPm for FMP vs. 39% PCPm for PS), mainly due to better modeling of the foreshortened upper legs. However, performance on the sideview sitting people (*e.g*. clusters 26, 30, 34, 44) is poor for all methods. Another prominent failure mode for all approaches are people facing away from the camera, e.g. cluster 50. Such part configurations are commonly mistaken for the frontal view which leads to a mismatch between left and right body parts resulting in incorrect estimation. These findings demonstrate inability of current methods to reliably discriminate between frontal and backward views of people. Interestingly, upper body approaches outperform full body methods on the full body cluster 31. This is an easy case for the former group of methods due to frontal upright upper body, but is a challenging task for the full body upproaches as legs are hard to estimate in this case. However, both MODEC und Armlets fail on examples when torso start deviating from canonical orientation (*e.g*. clusters 20, 27, 37). At the same time both full body methods perform better, as they are more robust to the viewpoint changes. Surprisingly, full body methods outperform upper body approaches on "easy" examples (c.f. cluster 1, 3 and 5). We attribute this effect to the correct integration of signals from the legs into a more reliable upper body estimate.

Occlusion and truncation performance. In Fig. 4 we clearly see difference in how occlusion and truncation influences the results. As expected we observe that the performance is best for fully visible people, but full visibility does not result in success rate similar to the one we observed for the images with simple poses, e.g. PS approach achieves 72.8% PCPm for 1000 most simple poses vs. 60% PCPm for same amount of people with least occlusion. We observe that occlusion results in significant performance drop on the order of 10% PCPm, e.g. in the case of PS approach 19.3% vs. 31.2% PCPm for the forearm with and without occlusion.

As mentioned above, truncation showed the least influence overall among the discussed factors. There are at least two reasons. First, the number of images with truncation is limited in our dataset (about 30% of the test data contain truncated people). Second, and more importantly, for truncation one cannot annotate positions of body parts outside of the image. Therefore the standard procedure is to exclude truncated body parts from the evaluation. In that sense approaches that wrongly estimate the position of a truncated body part are not punished for that. This limitation could be addressed by requiring that models have to also report which parts of the body are truncated.

Viewpoint performance. We evaluate the pose estimation for various torso viewpoints in two ways. In Fig. 4 we show results using our standard analysis method based on images ordered by deviation from the frontal viewpoint. For a more detailed analysis we quantize the space of viewpoints by clustering training examples according to their 3D torso orientations. We show results for the viewpoint clusters in Fig. 6 ordering them by the number of examples corresponding to each cluster. The number of examples per cluster ranges between 1453 examples for the largest cluster corresponding to the frontal viewpoint, and 53 examples for the viewpoint with extreme torso tilt.

We observe that in contrast to the full body approaches, viewpoint has profound influence on the performance of the upper body approaches considered in our evaluation. The performance of both Armlets and MODEC approaches drops significantly for non-frontal views.

A per viewpoint evaluation reveals significant performance differences across viewpoints. In Fig. 6 we show the results for the "person centric" annotations that we use throughout experiments in this paper and in addition for the "observer centric" (OC) annotations, in which body limbs are labeled as left/right based on their image location with respect to the torso. Frontal and near-frontal viewpoints are performing best. We observe a large drop in performance for backward facing people when performance is measured in "person centric" manner, which suggests that large portion of incorrect pose estimates for backward views is due to incorrect matching of left/right limbs.

We observe that all approaches handle extreme viewpoints poorly. PS approach is the only one in our evaluation that gracefully handles in-plane rotations (cluster 12), whereas performance of other approaches significantly degrades in that case. Also, PS outperforms other methods

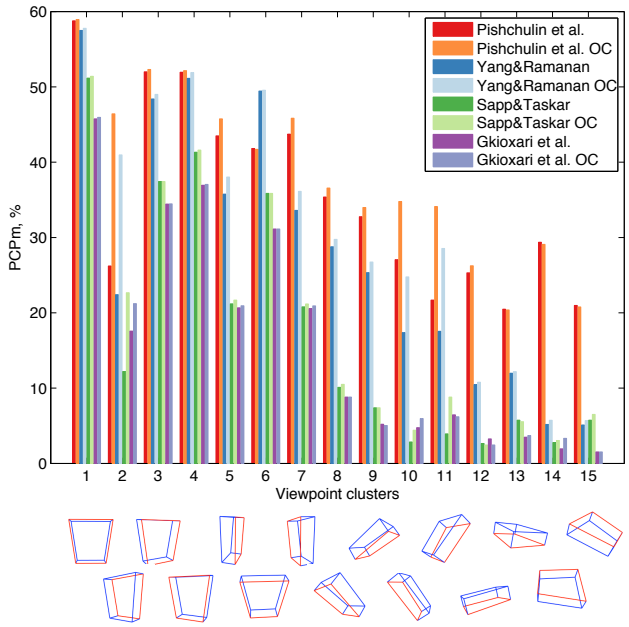


Figure 6. Pose estimation results (PCPm) grouped by viewpoint. Viewpoint clusters ordered decreasingly w.r.t. number of images. Each cluster is visualized in bottom row using 3D model of the torso corresponding to the cluster medoid.

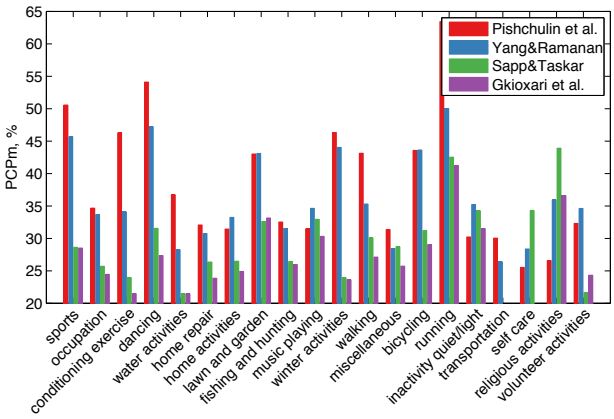


Figure 7. Pose estimation results (PCPm) grouped by activity categories shown in decreasing order w.r.t. number of images.

in case of extreme torso tilts (e.g. cluster 14). The performance for clusters with extreme torso rotation is on the level of 20 - 30% PCPm for the best method, corresponding to only 2 - 3 out of 10 body parts being localized correctly for such viewpoints.

Part length performance. Fig. 4 also shows the influence of part length on the performance of each approach. In this context, foreshortening is the most influential aspect and considered an important challenge for articulated pose estimation. The key observation is that the presence or absence of foreshortening has relatively little influence on the result compared to the other factors such as pose and occlusion. The best performing PS model is the most robust

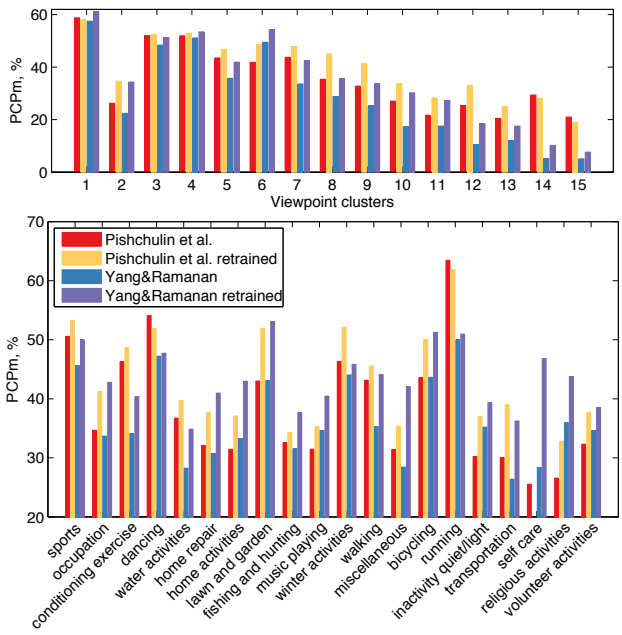


Figure 8. Comparison of performance (PCPm) on viewpoint (top) and activity category clusters (bottom) before and after retraining. See Fig. 6 for visualization of the viewpoint clusters.

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Upper body	Full body
Yang&Ramanan [22]	61.0	36.6	36.5	34.8	17.4	70.2	33.1	38.3
Yang&Ramanan [22] retrained	**69.3**	39.5	38.8	**43.4**	27.7	74.6	**42.3**	44.7
Pishchulin et al. [15]	63.8	39.6	37.3	39.0	26.8	70.7	39.1	42.3
Pishchulin et al. [15] retrained	68.4	**42.7**	**42.8**	42.0	**29.2**	**76.3**	42.1	**46.1**

Table 3. Comparison of performance (PCPm) before and after retraining. For PCKh results see supplementary material.

to foreshortening compared to other three approaches. For example the performance for the first 4000 images ordered by increasing foreshortening remains nearly constant.

Activity performance. Finally, we evaluate pose estimation performance as a function of the person activity. To that end we group test images by the activity categories in the hierarchy used for the image collection [1] and compute PCPm for each category. The results are shown in Fig. 7, where we order categories from left to right according to the number of test examples.

We observe strong variation of performance for different activity types. Best results are obtained on the sports- and dancing-centric activities (*e.g.* "Sports", "Running", "Winter Activities" and "Dancing"). Most difficult turn out to be activities that are performed in bulky clothing and involve use of tools (*e.g.* "Home Repair") and activities performed in cluttered scenes (*e.g.* "Fishing and Hunting"). MODEC outperforms all other approaches on the "Self care" activities (examples of activities from this category are "Eating, sitting", "Hairstyling", "Grooming" etc. with "Eating, sitting" containing by far the largest number of images.)

Retrained models. To showcase the usefulness of the benchmark as an analysis tool we retrain the PS and FMP models on the training set from our benchmark. To speed up training we consider a subset of 4000 images, which is 4 times as many images as in the LSP and 40 times as many as in the PARSE datasets used by the publicly available PS and FMP models. The results are shown in Tab. 3. FMP significantly benefits from retraining (44.7 PCPm for retrained vs. 38.3 for original). PS achieves slightly better result, although overall improvement due to retraining is smaller (46.1 PCPm for retrained vs. 42.3 PCPm the original).

Although performances for FMP and PS are close overall, we observe interesting differences when examining performance at the level of individual activities and viewpoints (thereby exploiting the rich annotations of our benchmark). Results are shown in Fig. 8. We observe that our publicly available PS model is winning by a large margin on the highly articulated categories, such as "Dancing" and "Running". Retraining the model boosts performance on activities with less articulation but more complex appearance (e.g. "Home Activities", "Lawn and Garden", "Bicycling", and "Occupation"). Our results show that training on the larger amount of more variable data significantly improved robustness of FMP to viewpoint changes. Performance of FMP improves on the difficult viewpoints by a large margin (e.g. for viewpoint cluster 10 improvement is from 17 to 31% PCPm). Retraining improves the performance of PS model on difficult viewpoints as well, although not as dramatically as for FMP, likely because PS already models in-plane rotations explicitly.

4. Conclusion

In this work we advance the state of the art in human pose estimation by establishing new qualitatively higher standards for evaluation and analysis of pose estimation methods and demonstrate the most promising research directions for the next years. To that end we propose a novel "MPII Human Pose" benchmark that we collected by leveraging a taxonomy of activities established in the literature. Compared to current datasets our benchmark covers significantly wider range of human poses spanning from householding to recreational activities and sports. Rich labeling of the collected data and a set of developed evaluation tools enable comprehesive analysis which we perform to demonstrate the strengths and weaknesses of the current methods.

Our findings indicate that current methods are significantly challenged by cases outside their comfort zone, such as large torso rotation and loose clothing. From all other factors, pose complexity has the most profound effect on the pose estimation performance. Current methods perform best on activities with simple tight clothing (*e.g.* in sport scenes), and are challenged by images with complex clothing and background clutter that are typical for many occupational and outdoor activities.

We will make the data, rich annotations for training images and evaluation tools publicly available in order to enable detailed analysis of future pose estimation methods. To prevent accidentally tuning on the test set, the annotations for the test images will be withheld and made accessible through an online evaluation tool. In the future we plan to extend our benchmark to joint pose estimation of multiple people and pose estimation in image sequences.

Acknowledgements. This work has been supported by the Max Planck Center for Visual Computing & Communication. The authors are thankful to Steve Hillyer and numerous anonymous Mechanical Turk workers for the help with preparation of the dataset.

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] M. Dantone, J. Gall, C. Leistner, and L. V. Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR'13*.

[3] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC'09*.

[4] M. Eichner and V. Ferrari. Human pose co-estimation and applications. *PAMI'12*.

[5] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV'10*.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV'10*.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV'05*.

[8] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR'08*.

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

[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI'13*.

[11] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV'13*.

[12] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC'10*.

[13] S. Johnson and M. Everingham. Learning Effective Human Pose Estimation from Inaccurate Annotation. In *CVPR'11*.

[14] S. Maji. Large scale image annotations on amazon mechanical turk. Technical report, EECS UC Berkeley, 2011.

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

[16] D. Ramanan. Learning to parse images of articulated objects. In *NIPS'06*.

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

[18] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR'11*.

[19] L. Sigal, A. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87, 2010.

[20] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV'12*.

[21] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR'11*.

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

[23] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.