PoseTrack: A Benchmark for Human Pose Estimation and Tracking

Mykhaylo Andriluka[4,**] Umar Iqbal[2] Eldar Insafutdinov[1] Leonid Pishchulin[1]
Anton Milan[3,*] Juergen Gall[2] Bernt Schiele[1]

[1]MPI for Informatics, Saarbrücken, Germany
[2]Computer Vision Group, University of Bonn, Germany
[3]Amazon Research
[4]Google AI Perception

Figure 1: Sample video from our benchmark. We select sequences that represent crowded scenes with multiple articulated people engaging in various dynamic activities and provide dense annotations of person tracks, body joints and ignore regions.

Abstract

Existing systems for video-based pose estimation and tracking struggle to perform well on realistic videos with multiple people and often fail to output body-pose trajectories consistent over time. To address this shortcoming this paper introduces PoseTrack which is a new large-scale benchmark for video-based human pose estimation and articulated tracking. Our new benchmark encompasses three tasks focusing on i) single-frame multi-person pose estimation, ii) multi-person pose estimation in videos, and iii) multi-person articulated tracking. To establish the benchmark, we collect, annotate and release a new dataset that features videos with multiple people labeled with person tracks and articulated pose. A public centralized evaluation server is provided to allow the research community to evaluate on a held-out test set. Furthermore, we conduct an extensive experimental study on recent approaches to articulated pose tracking and provide analysis of the strengths and weaknesses of the state of the art. We envision that the proposed benchmark will stimulate productive research both by providing a large and representative training dataset as well as providing a platform to objectively evaluate and compare the proposed methods. The benchmark is freely accessible at `https://posetrack.net/`.

1. Introduction

Human pose estimation has recently made significant progress on the tasks of single person pose estimation in individual frames [46, 45, 44, 4, 49, 15, 18, 31, 2, 36] and videos [34, 6, 21, 12] as well as multi-person pose estimation in monocular images [35, 18, 20, 3, 32]. This progress has been facilitated by the use of deep learning-based architectures [41, 14] and by the availability of large-scale benchmark datasets such as "MPII Human Pose" [1] and "MS COCO" [28]. Importantly, these benchmark datasets not only have provided extensive training sets required for training of deep learning based approaches, but also established detailed metrics for direct and fair performance comparison across numerous competing approaches.

Despite significant progress of single frame based multi-person pose estimation, the problem of *articulated multi-person body joint tracking* in monocular video remains largely unaddressed. Although there exist training sets for special scenarios, such as sports [51, 23] and upright frontal people [6], these benchmarks focus on *single isolated individuals* and are still limited in their scope and variability of represented activities and body motions. In this work, we aim to fill this gap by establishing a new large-scale, high-quality benchmark for video-based multi-person pose estimation and articulated tracking.

Our benchmark is organized around three related tasks focusing on single-frame multi-person pose estimation, multi-person pose estimation in video, and multi-person articulated tracking. While the main focus of the dataset is

*This work was done prior to joining Amazon.
**This work was done prior to joining Google.

Figure 2: Example frames and annotations from our dataset.

on multi-person articulated tracking, progress in the single-frame setting will inevitably improve overall tracking quality. We thus make the single frame multi-person setting part of our evaluation procedure. In order to enable timely and scalable evaluation on the held-out test set, we provide a centralized evaluation server. We strongly believe that the proposed benchmark will prove highly useful to drive the research forward by focusing on remaining limitations of the state of the art.

To sample the initial interest of the computer vision community and to obtain early feedback we have organized a competition based on our benchmark at one of the recent computer vision meetings. We obtained largely positive feedback from the twelve teams that participated in the competition. We incorporate some of this feedback into this paper. In addition we analyze the currently best performing approaches and highlight the common difficulties for pose estimation and articulated tracking.

2. Related Datasets

The commonly used publicly available datasets for evaluation of 2D human pose estimation are summarized in Tab. 1. The table is split into blocks of single-person single-frame, single-person video, multi-person single-frame, and multi-person video data.

The most popular benchmarks to date for evaluation of single person pose estimation are "LSP" [25] (+ "LSP Extended" [26]) and "MPII Human Pose (Single Person)" [1]. LSP and LSP Extended datasets focus on sports scenes featuring a few sport types. Although a combination of both datasets results in 11,000 training poses, the evaluation set of 1000 is rather small. FLIC [38] targets a simpler task of upper body pose estimation of frontal upright individuals in feature movies. In contrast to LSP and FLIC datasets, MPII Single-Person benchmark covers a much wider variety of everyday human activities including various recreational, occupational and household activities and consists of over 26,000 annotated poses with 7000 poses held out for

Dataset	# Poses	Multi-person	Video-labeled poses	Data type
LSP [25]	2,000			sports (8 act.)
LSP Extended [26]	10,000			sports (11 act.)
MPII Single Person [1]	26,429			diverse (491 act.)
FLIC [38]	5,003			feature movies
FashionPose [9]	7,305			fashion blogs
We are family [10]	3,131	✓		group photos
MPII Multi-Person [1]	14,993	✓		diverse (491 act.)
MS COCO Keypoints [28]	105,698	✓		diverse
Penn Action [51]	159,633		✓	sports (15 act.)
JHMDB [23]	31,838		✓	diverse (21 act.)
YouTube Pose [6]	5,000		✓	diverse
Video Pose 2.0 [39]	1,286		✓	TV series
Multi-Person PoseTrack [22]	16,219	✓	✓	diverse
Proposed	**153,615**	✓	✓	**diverse**

Table 1: Overview of publicly available datasets for articulated human pose estimation in single frames and video. For each dataset we report the number of annotated poses, availability of video pose labels and multiple annotated persons per frame, as well as types of data.

evaluation. Both benchmarks focus on single person pose estimation and provide rough location scale of a person in question. In contrast, our dataset addresses a much more challenging task of body tracking of multiple highly articulated individuals where neither the number of people, nor their locations or scales are known.

The single-frame multi-person pose estimation setting was introduced in [10] along with "We Are Family (WAF)" dataset. While this benchmark is an important step towards more challenging multi-person scenarios, it focuses on a simplified setting of upper body pose estimation of multiple upright individuals in group photo collections. The "MPII Human Pose (Multi-Person)" dataset [1] has significantly advanced the multi-person pose estimation task in terms of diversity and difficulty of multi-person scenes that show highly-articulated people involved in hundreds of every day activities. More recently, MS COCO Keypoints Challenge [28] has been introduced to provide a new large-scale benchmark for single frame based multi-person pose estimation. All these datasets are only limited to single-frame based body pose estimation. In contrast, our dataset also focuses on a more challenging task of multi-person pose estimation in video sequences containing highly articulated people in dense crowds. This not only requires annotations of body keypoints, but also a unique identity for every person appearing in the video. Our dataset is based on the MPII Multi-Person benchmark, from which we select a subset of key frames and for each key frame include about five seconds of video footage centered on the key frame. We provide dense annotations of video sequences

with person tracking and body pose annotations. Furthermore, we adapt a completely unconstrained evaluation setup where the scale and location of the persons is completely unknown. This is in contrast to MPII dataset that is restricted to evaluation on group crops and provides rough group location and scale. Additionally, we provide ignore regions to identify the regions containing very large crowds of people that are unreasonably complex to annotate.

Recently, [22] and [17] also provided datasets for multi-person pose estimation in videos. However, both are at a very small scale. [22] provides only 60 videos with most sequences containing only 41 frames, and [17] provides 30 videos containing only 20 frames each. While these datasets make a first step toward solving the problem at hand, they are certainly not enough to cover a large range of real-world scenarios and to learn stronger pose estimation models. We on the other hand establish a large-scale benchmark with a much broader variety and an open evaluation setup. The proposed dataset contains over $150{,}000$ annotated poses and over $22{,}000$ labeled frames.

Our dataset is complementary to recent video datasets, such as J-HMDB [23], Penn Action [51] and YouTube Pose [6]. Similar to these datasets, we provide dense annotations of video sequences. However, in contrast to [23, 51, 6] that focus on single isolated individuals we target a much more challenging task of multiple people in dynamic crowded scenarios. In contrast to YouTube Pose that focus on frontal upright people, our dataset includes a wide variety of body poses and motions, and captures people at different scales from a wide range of viewpoints. In contrast to sports-focused Penn Action and J-HMDB that focuses on a few simple actions, the proposed dataset captures a wide variety of everyday human activities while being at least 3x larger compared to J-HMDB.

Our dataset also addresses a different set of challenges compared to the datasets such as "HumanEva" [40] and "Human3.6M" [19] that include images and 3D poses of people but are captured in controlled indoor environments, whereas our dataset includes real-world video sequences but provides 2D poses only.

3. The PoseTrack Dataset and Challenge

We will now provide the details on data collection and the annotation process, as well as the established evaluation procedure. We build on and extend the newly introduced datasets for pose tracking in the wild [17, 22]. To that end, we use the raw videos provided by the popular MPII Human Pose dataset. For each frame in MPII Human Pose dataset we include $41 - 298$ neighboring frames from the corresponding raw videos, and then select sequences that represent crowded scenes with multiple articulated people engaging in various dynamic activities. The video sequences are chosen such that they contain a large amount of body motion and body pose and appearance variations. They also contain severe body part occlusion and truncation, *i.e.*, due to occlusions with other people or objects, persons often disappear partially or completely and re-appear again. The scale of the persons also varies across the video due to the movement of persons and/or camera zooming. Therefore, the number of visible persons and body parts also varies across the video.

3.1. Data Annotation

We annotated the selected video sequences with person locations, identities, body pose and ignore regions. The annotations were performed in four steps. First, we labeled ignore regions to exclude crowds and people for which pose can not be reliably determined due to poor visibility. Afterwards, the head bounding boxes for each person across the videos were annotated and a track ID was assigned to each person. The head bounding boxes provide an estimate of the absolute scale of the person required for evaluation. We assign a unique track ID to each person appearing in the video until the person moves out of the camera field-of-view. Note that each video in our dataset might contain several shots. We do not maintain track ID between shots and same person might get different track ID if it reappears in another shot. Poses for each person track are then annotated in the entire video. We annotate 15 body parts for each body pose including *head, nose, neck, shoulders, elbows, wrists, hips, knees and ankles*. All pose annotations were performed using the VATIC tool [48] that allows to speed-up annotation by interpolating between frames. We chose to skip annotation of the body joints that can not be reliably localized by the annotator due to strong occlusion or difficult imaging conditions. This has proven the be a faster alternative to requiring annotators to guess the location of the joint and/or marking it as occluded.

Fig. 2 shows example frames from the dataset. Note the variability in appearance and scale, and complexity due to substantial number of people in close proximity.

Overall, the dataset contains 550 video sequences with $66{,}374$ frames. We split them into 292, 50, 208 videos for training, validation and testing, respectively. The split follows the original split of the MPII Human Pose dataset makeing it possible to train a model on the MPII Human Pose and evaluate on our test and validation sets.

The length of the majority of the sequences in our dataset ranges between 41 and 151 frames. The sequences correspond to about 5 seconds of video. Differences in the sequence length are due to variation in the frame rate of the videos. A few sequences in our dataset are longer than five seconds with the longest sequence having 298 frames. For each sequence in our benchmark we annotate the 30 frames in the middle of the sequence. In addition, we densely annotate validation and test sequences with a step of four frames.

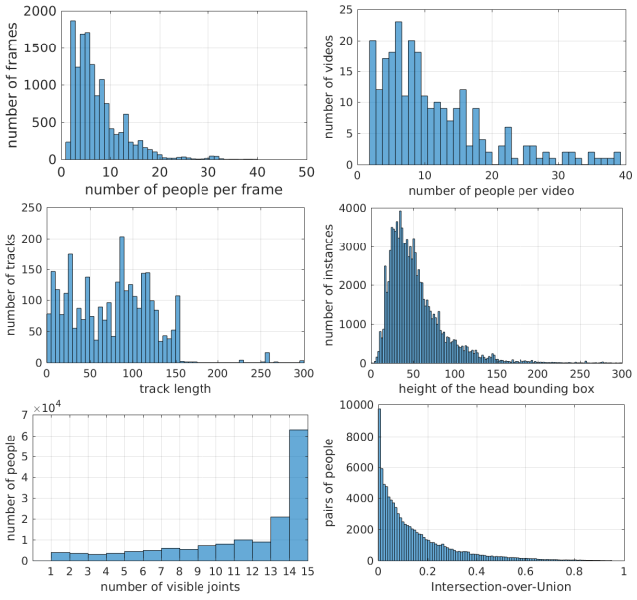

Figure 3: Various statistics of the PoseTrack benchmark.

The rationale behind this annotation strategy is that we aim to evaluate both smoothness of body joint tracks as well as ability to track body joints over longer number of frames. We did not densely annotate the training set to save the annotation resources for the annoation of test and validation set. In total, we provide around 23,000 labeled frames with 153,615 pose annotations. To the best of our knowledge this makes PoseTrack the largest multi-person pose estimation and tracking dataset released to date. In Fig. 3 we show additional statistics of the validation and test sets of our dataset. The plots illustrate the distribution of 'crowdness' per frame and per video, the track length and people size measured by the head bounding box. Note that substantial portion of the videos has a large number of people as shown in the plot on the top-right. The abrupt fall off in the plot of the track length in the bottom-left is due to fixed length of the sequences included in the dataset.

3.2. Challenges

The benchmark consists of the following challenges:

Single-frame pose estimation. This task is similar to the ones covered by existing datasets like MPII Pose and MS COCO Keypoints, but on our new large-scale dataset.

Pose estimation in videos. The evaluation of this challenge is performed on single frames, however, the data will also include video frames before and after the annotated ones, allowing methods to exploit video information for a more robust single-frame pose estimation.

Pose tracking. This task requires to provide temporally consistent poses for all people visible in the videos. Our evaluation include both individual pose accuracy as well as temporal consistency measured by identity switches.

3.3. Evaluation Server

We provide an online evaluation server to quantify the performance of different methods on the held-out test set. This will not only prevent over-fitting to the test data but also ensures that all methods are evaluated in the exact same way, using the same ground truth and evaluation scripts, making the quantitative comparison meaningful. Additionally, it can also serve as a central directory of all available results and methods.

3.4. Experimental Setup and Evaluation Metrics

Since we need to evaluate both the accuracy of multi-person pose estimation in individual frames and articulated tracking in videos, we follow the best practices followed in both multi-person pose estimation [35] and multi-target tracking [30]. In order to evaluate whether a body part is predicted correctly, we use the PCKh (head-normalized probability of correct keypoint) metric [1], which considers a body joint to be correctly localized if the predicted location of the joint is within a certain threshold from the true location. Due to the large scale variation of people across videos and even within a frame, this threshold needs to be selected adaptively, based on the person's size. To that end, we follow [1] and use 50% of the head length where the head length corresponds to 60% of the diagonal length of the ground-truth head bounding box. Given the joint localization threshold for each person, we compute two sets of evaluation metrics, one which is commonly used for evaluating multi-person pose estimation [35], and one from the multi-target tracking literature [50, 8, 30] to evaluate multi-person pose tracking. During evalution we ignore all person detections that overlap with the ignore regions.

Multi-person pose estimation. For measuring frame-wise multi-person pose accuracy, we use *mean Average Precision* (mAP) as is done in [35]. The protocol to evaluate multi-person pose estimation in [35] requires that the location of a group of persons and their rough scale is known during evaluation [35]. This information, however, is almost never available in realistic scenarios, particularly for videos. We therefore, propose not to use any ground-truth information during testing and evaluate the predictions without rescaling or selecting a specific group of people for evaluation.

Articulated multi-person pose tracking. To evaluate multi-person pose tracking, we use Multiple Object Tracking (MOT) metrics [30]. The metrics require predicted body poses with tracklet IDs. First, for each frame, for each body joint class, distances between predicted locations and GT locations are computed. Then, predicted tracklet IDs and GT tracklet IDs are taken into account and all (prediction, GT) pairs with distances below the PCKh threshold are considered during global matching of predicted tracklets to GT tracklets for each particular body joint. Global matching

minimizes the total assignment distance. Finally, Multiple Object Tracker Accuracy (MOTA), Multiple Object Tracker Precision (MOTP), Precision, and Recall metrics are computed. Evaluation server reports MOTA metric for each body joint class and average over all body joints, while for MOTP, Precision, and Recall we report averages only. In the following experimental evaluation MOTA is used as our main tracking metric. The source code for evaluation metrics is available publicly on the benchmark website.

4. Analysis of the State of the Art

Articulated pose tracking in unconstrained videos is a relatively new topic in computer vision research. To the best of our knowledge only few approaches for this task have been proposed in the literature [17, 22]. Therefore, to analyze the performance of the state of the art on our new dataset, we proceed in two ways.

First, we propose two baseline methods based on the state of the art approaches [17, 22]. Note that our benchmark includes an order of magnitude more sequences compared to the datasets used in [17, 22] and the sequences in our benchmark are about five times longer, which makes it computationally expensive to run the graph partitioning on the full sequences as in [17, 22]. We, therefore, modify these methods to make them applicable on the proposed dataset. The baselines and corresponding modifications are explained in Sec. 4.1.

Second, in order to broaden the scope of our evaluation we organized a *PoseTrack Challenge* in conjunction with ICCV'17 on our dataset by establishing an online evaluation server and inviting submissions from the research community. In the following we consider the top five methods submitted to the online evaluation server both for the pose estimation and pose tracking tasks. In Tab. 2 and 3 we list the best performing methods on each task sorted by MOTA and mAP, respectively. In the following we first describe our baselines based on [17, 22] and then summarize the main observations made in this evaluation.

4.1. Baseline Methods

We build the first baseline model following the graph partitioning formulation for articulated tracking introduced in [17], but introduce two simplifications that follow [32]. First, we rely on a person detector to establish locations of people in the image and run pose estimation independently for each person detection. This allows us to deal with large variation in scale present in our dataset by cropping and rescaling images to canonical scale prior to pose estimation. In addition, this also allows us to group together the body-part estimates inferred for a given detection bounding box. As a second simplification we apply the model on the level of full body poses and not on the level of individual body parts as in [17, 22]. We use a publicly available Faster-RCNN [37] detector from the TensorFlow Object Detection API [16] for people detection. This detector has been trained on the "MS COCO" dataset and uses Inception-ResNet-V2 [42] for image encoding. We adopt the DeeperCut CNN architecture from [18] as our pose estimation method. This architecture is based on the ResNet-101 converted to a fully convolutional network by removing the global pooling layer and utilizing atrous (or dilated) convolutions [7] to increase the resolution of the output scoremaps. Once all poses are extracted, we perform non-maximum suppression based on pose similarity criteria [32] to filter out redundant person detections. We follow the cropping procedure of [32] with the crop size 336x336px. Tracking is implemented as in [17] by forming the graph that connects body-part hypotheses in adjacent frames and partitioning this graph into connected components using an approach from [27]. We use Euclidean distance between body joints to derive costs for graph edges. Such distance based features were found to be already effective in [17] with additional features adding minimal improvements at the cost of substantially slower inference.

For the second baseline, we use the publicly available source code of [22] and replace the pose estimation model with [3]. We emprically found that the pose estimation model of [3] is better at handling large scale variations as compared to DeeperCut [18] used in the original paper, in particular, when performing bottom-up multi-person pose estimation. We do not make any changes in the graph partitioning algorithm, but reduce the window size to 21 as compared to 31 used in the original model. We refer the readers to [22] for more details. The goal of constructing these strong baseline is to validate the results submitted to our evaluation server and to allow us to perform additional experiments presented in Sec. 5. In the rest of this paper, we refer to these baselines as ArtTrack [17] and PoseTrack [22], respectively.

4.2. Main Observations

Two-stage design. The first observation is that all submissions follow a two-stage tracking-by-detection design. In the first stage, a combination of person detector and single-frame pose estimation method is used to estimate poses of people in each frame. The exact implementation of single-frame pose estimation method varies. Each of the top three articulated tracking methods builds on a different pose estimation approach (Mask-RCNN [13], PAF [3] and DeeperCut [18]). On the other hand, when evaluating methods according to pose estimation metric (see Tab. 3) three of the top four approaches build on PAF [3]. The performance still varies considerably among these PAF-based methods (70.3 for submission ML-LAB [52] vs. 62.5 for submission SOPT-PT [43]) indicating that large gains can be achieved within the PAF framework by introducing incremental im-

Submission	Pose model	Tracking model	Tracking granularity	Additional training data	mAP	MOTA
ProTracker [11]	Mask R-CNN [13]	Hungarian	pose-level	COCO	59.6	**51.8**
BUTD [24]	PAF [3]	graph partitioning	person-level and part-level	COCO	59.2	50.6
SOPT-PT [43]	PAF [3]	Hungarian	pose-level	MPII-Pose + COCO	62.5	44.6
ML-LAB [52]	modification of PAF [3]	frame-to-frame assign.	pose-level	MPII-Pose + COCO	**70.3**	41.8
ICG [33]	novel single-/multi-person CNN	frame-to-frame assign.	pose-level	-	51.2	32.0
ArtTrack-baseline	Faster-RCNN [16] + DeeperCut [18]	graph partitioning	pose-level	MPII-Pose + COCO	59.4	48.1
PoseTrack-baseline	PAF [3]	graph partitioning	part-level	COCO	59.4	48.4

Table 2: Results of the top five pose tracking models submitted to our evaluation server and of our baselines based on [17] and [22]. Note that mAP for some of the methods might be intentionally reduced to achieve higher MOTA (see discussion in text).

Submission	Pose model	Additional training data	mAP
ML-LAB [52]	modification of PAF [3]	COCO	**70.3**
BUTDS [24]	PAF [3]	MPII-Pose + COCO	64.5
ProTracker [11]	Mask R-CNN [13]	COCO	64.1
SOPT-PT [43]	PAF [3]	MPII-Pose + COCO	62.5
SSDHG	SSD [29] + Hourglass [31]	MPII-Pose + COCO	60.0
ArtTrack-baseline	DeeperCut	MPII-Pose + COCO	65.1
PoseTrack-baseline	PAF [3]	COCO	59.4

Table 3: Results of the top five pose estimation models submitted to our evaluation server and of our baselines. The methods are ordered according to mAP. Note that the mAP of ArtTrack and submission ProTracker [11] is different from Tab. 2 because the evaluation in this table does not threshold detections by the score.

Model	Training Set	Head	Sho	Elb	Wri	Hip	Knee	Ank	mAP
ArtTrack-baseline	our dataset	73.1	65.8	55.6	47.2	52.6	50.1	44.1	55.5
ArtTrack-baseline	MPII	76.4	74.4	68.0	59.4	66.1	64.2	56.6	66.4
ArtTrack-baseline	MPII + our dataset	**78.7**	**76.2**	**70.4**	**62.3**	**68.1**	**66.7**	**58.4**	**68.7**

Table 4: Pose estimation performance (mAP) of our ArtTrack baseline for different training sets.

provements.

In the second stage the single-frame pose estimates are linked over time. For most of the methods the assignment is performed on the level of body poses, not individual parts. This is indicated in the "Tracking granularity" column in Tab. 2. Only submission BUTD [24] and our PoseTrack baseline track people on the level of individual body parts.

Model	Head	Sho	Elb	Wri	Hip	Knee	Ank	Total	mAP
ArtTrack-baseline, $\tau = 0.1$	58.0	56.4	34.0	19.2	44.1	35.9	19.0	38.1	**68.6**
ArtTrack-baseline, $\tau = 0.5$	63.5	62.8	48.0	37.8	52.9	48.7	36.6	50.0	66.7
ArtTrack-baseline, $\tau = 0.8$	**66.2**	**64.2**	**53.2**	**43.7**	**53.0**	**51.6**	**41.7**	**53.4**	62.1

Table 5: Pose tracking performance (MOTA) of ArtTrack baseline for different part detection cut-off thresholds τ.

Hence, most methods establish correspondence/assembly of parts into body poses on the per-frame level. In practice, this is implemented by supplying a bounding box of a person and running pose estimation just for this box, then declaring maxima of the heatmaps as belonging together. This is suboptimal as multiple people overlap significantly, yet most approaches choose to ignore such cases (possibly for inference speed/efficiency reasons). The best performing approach ProTracker [11] relies on simple matching between frames based on Hungarian algorithm and matching cost based on intersection-over-union score between person bounding boxes. None of the methods is end-to-end in the sense that it is able to directly infer articulated people tracks from video. We observe that the pose tracking performance of the top five submitted methods saturates at around 50 MOTA, with the top four approaches showing rather similar MOTA results (51.8 for submission ProTracker [11] vs. 50.6 for submission BUTD [24] vs. 48.4 for PoseTrack vs. 48.1 for ArtTrack), indicating room for improvement on this task.

Training data. Most submissions found it necessary to combine our training set with datasets of static images such as COCO and MPII-Pose to obtain a joint training set with larger appearance variability. The most common procedure was to pre-train on external data and then fine-tune on our training set. Our training set is composed of 2437 people tracks with 61,178 annotated body poses and is complementary to COCO and MPII-Pose which include an order of magnitude more individual people but do not provide motion information. We quantify the performance improvement due to training on additional data in Tab. 4 using our ArtTrack baseline. Extending the training data with the MPII-Pose dataset improves the performance considerably (55.5 vs. 68.7 mAP). The combination of our dataset and MPII-Pose still performs better than MPII-Pose alone (66.4 vs. 68.7) showing that datasets are indeed complementary.

None of the approaches in our evaluation employs any form of learning on the provided video sequences beyond simple cross-validation of a few hyperparameters. This can

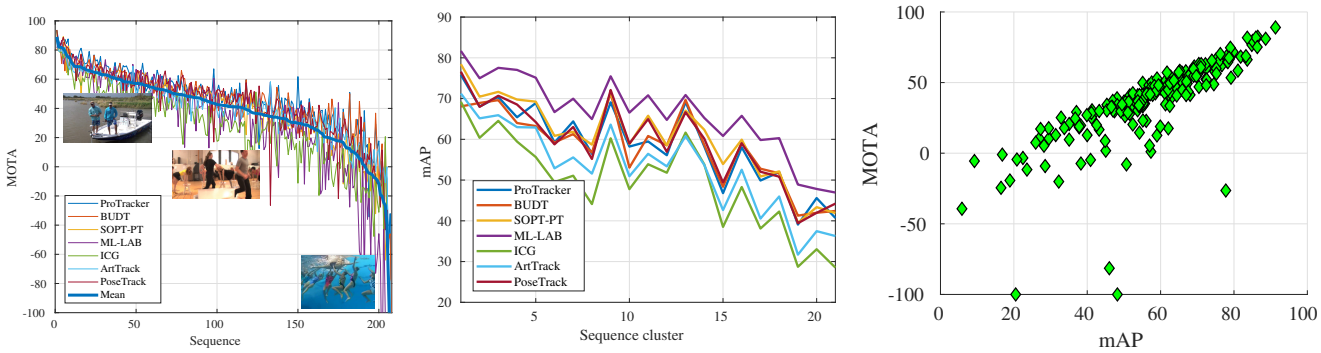

Figure 4: Sequences sorted by average MOTA (left). Pose estimation results sorted according to articulation complexity of the sequence (middle). Visualization of correlation between mAP and MOTA for each sequence (right). Note the outliers in right plot that correspond to sequences where pose estimation works well but tracking still fails.

be in part due to relatively small size of our training set. One of the lessons learned from our work on this benchmark is that creating truly large annotated datasets of articulated pose sequences is a major challenge. We envision that future work will combine manually labeled data with other techniques such as transfer learning from other datasets such as [5], inferring sequences of poses by propagating annotations from reliable keyframes [6], and leveraging synthetic training data as in [47].

Dataset difficulty. We composed our dataset by including videos around the keyframes from MPII Human Pose dataset that included several people and non-static scenes. The rationale was to create a dataset that would be non-trivial for tracking and require methods to correctly resolve effects such as person-person occlusions. In Fig. 4 we visualize performance of the evaluated approaches on each of the test sequences. We observe that test sequences vary greatly with respect to difficulty both for pose estimation as well as for tracking. *E.g.*, for the best performing submission ProTracker [11] the performance varies from nearly 80 MOTA to a score below zero[1]. Note that the approaches mostly agree with respect to the difficulty of the sequences. More difficult sequences are likely to require methods that are beyond simple tracking component based on frame-to-frame assignment used in the currently best performing approaches. To encourage submissions that explicitly address challenges in the difficult portions of the dataset we have defined easy/moderate/hard splits of the data and report results for each of the splits as well as the full set.

Evaluation metrics. The MOTA evaluation metric has a deficiency in that it does not take the confidence score of the predicted tracks into account. As a result achieving good MOTA score requires tuning of the pose detector threshold so that only confident track and pose hypothesis are supplied for evaluation. This in general degrades pose estimation performance as measured by mAP (*c.f.* performance of submission ProTracker [11] in Tab. 2 and 3). We quantify this in Fig. 5 for our ArtTrack baseline. Note that filtering the detections with score below $\tau = 0.8$ as compared to $\tau = 0.1$ improves MOTA from 38.1 to 53.4. One potential improvement to the evaluation metric would be to require that pose tracking methods assign confidence score to each predicted track as is common for pose estimation and object detection. This would allow one to compute a final score as an average of MOTA computed for a range of track scores. Current pose tracking methods typically do not provide such confidence scores. We believe that extending the evaluation protocol to include confidence scores is an important future direction.

[1]Note that MOTA metric can become negative for example when the number of false positives significantly exceeds the number of ground-truth targets.

5. Dataset Analysis

In order to better understand successes and failures of the current body pose tracking approaches, we analyze their performance across the range of sequences in the test set. To that end, for each sequence we compute an average over MOTA scores obtained by each of the seven evaluated methods. Such average score serves us as an estimate for the difficulty of the sequence for the current computer vision approaches. We then rank the sequences by the average MOTA. The resulting ranking is shown in Fig. 4 (left) along with the original MOTA scores of each of the approaches. First, we observe that all methods perform similarly well on easy sequences. Fig. 5 shows a few easy sequences with an average MOTA above 75%. Visual analysis reveals that easy sequences typically contain significantly separated individuals in upright standing poses with minimal changes of body articulation over time and no camera motion. Tracking accuracy drops with the increased complexity of video sequences. Fig. 6 shows a few hard sequences with average MOTA accuracy below 0. These sequences typically include strongly overlapping people, and fast motions of peo-

Figure 5: Selected frames from sample sequences with MOTA score above 75% with predictions of our ArtTrack-baseline overlaid in each frame. See text for further description.

1 2 3 4

5 6 7 8

Figure 6: Selected frames from sample sequences with negative average MOTA score. The predictions of our ArtTrack-baseline are overlaid in each frame. Challenges for current methods in such sequences include crowds (images 3 and 8), extreme proximity of people to each other (7), rare poses (4 and 6) and strong camera motions (3, 5, 6, and 8).

ple and camera.

We further analyze how tracking and pose estimation accuracy are affected by pose complexity. As a measure for the pose complexity of a sequence we employ an average deviation of each pose in a sequence from the mean pose. The computed complexity score is used to sort video sequences from low to high pose complexity and average mAP is reported for each sequence. The result of this evaluation is shown in Fig. 4 (middle). For visualization purposes, we partition the sorted video sequences into bins of size 10 based on pose complexity score and report average mAP for each bin. We observe that both body pose estimation and tracking performance significantly decrease with the increased pose complexity. Fig. 4 (right) shows a plot that highlights correlation between mAP and MOTA of the same sequence. We use the mean performance of all methods in this visualization. Note that in most cases more accurate pose estimation reflected by higher mAP indeed corresponds to higher MOTA. However, it is instructive to look at sequences where poses are estimated accurately (mAP is high), yet tracking results are particularly poor (MOTA near zero). One of such sequences is shown in Fig. 6 (8). This sequence features a large number of people and fast camera movement that is likely confusing simple frame-to-frame association tracking of the evaluated approaches. Please see supplemental material for additional examples and analyses of challenging sequences.

6. Conclusion

In this paper we proposed a new benchmark for human pose estimation and articulated tracking that is significantly larger and more diverse in terms of data variability and complexity compared to existing pose tracking benchmarks. Our benchmark enables objective comparison of different approaches for articulated people tracking in realistic scenes. We have set up an online evaluation server that permits evaluation on a held-out test set, and have measures in place to limit overfitting on the dataset. Finally, we conducted a rigorous survey of the state of the art. Due to the scale and complexity of the benchmark, most existing methods build on combinations of proven components: people detection, single-person pose estimation, and tracking based on simple association between neighboring frames. Our analysis shows that current methods perform well on easy sequences with well separated upright people, but are severely challenged in the presence of fast camera motions and complex articulations. Addressing these challenges remains an important direction for future work.

Acknowledgements. UI and JG have been supported by the DFG project GA 1927/5-1 (FOR 2535) and the ERC Starting Grant ARCA (677650).

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017.

[4] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[6] J. Charles, T. Pfister, D. Magee, and A. Hogg, D. Zisserman. Personalizing human video pose estimation. In *CVPR*, 2016.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *TPAMI*, 2017.

[8] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *ICCV 2015*.

[9] M. Dantone, J. Gall, C. Leistner, and L. V. Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, 2013.

[10] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV*, 2010.

[11] R. Girdhar, G. Gkioxari, L. Torresani, D. Ramanan, M. Paluri, and D. Tran. Simple, efficient and effective keypoint tracking. In *ICCV PoseTrack Workshop*, 2017.

[12] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *ECCV*, 2016.

[13] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] P. Hu and D. Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *CVPR*, 2016.

[16] J. Huang, V. Rathod, C. Sun, M. Zhu, A. K. Balan, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017.

[17] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. Arttrack: Articulated multi-person tracking in the wild. In *CVPR*, 2017.

[18] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.

[19] 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*, 2014.

[20] U. Iqbal and J. Gall. Multi-person pose estimation with local joint-to-person associations. In *ECCVw*, 2016.

[21] U. Iqbal, M. Garbade, and J. Gall. Pose for action - action for pose. In *FG*, 2017.

[22] U. Iqbal, A. Milan, and J. Gall. PoseTrack: Joint multi-person pose estimation and tracking. In *CVPR*, 2017.

[23] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013.

[24] S. Jin, X. Ma, Z. Han, Y. Wu, W. Yang, W. Liu, C. Qian, and W. Ouyang. Towards multi-person pose tracking: Bottom-up and top-down methods. In *ICCV PoseTrack Workshop*, 2017.

[25] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[26] S. Johnson and M. Everingham. Learning Effective Human Pose Estimation from Inaccurate Annotation. In *CVPR*, 2011.

[27] E. Levinkov, J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, A. Kirillov, C. Rother, T. Brox, B. Schiele, and B. Andres. Joint graph decomposition and node labeling: Problem, algorithms, applications. In *CVPR*, 2017.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[30] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, 2016.

[31] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[32] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017.

[33] C. Payer, T. Neff, H. Bischof, M. Urschler, and D. Stern. Simultaneous multi-person detection and single-person pose estimation with a single heatmap regression network. In *ICCV PoseTrack Workshop*, 2017.

[34] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015.

[35] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016.

[36] U. Rafi, I.Kostrikov, J. Gall, and B. Leibe. An efficient convolutional network for human pose estimation. In *BMVC*, 2016.

[37] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[38] B. Sapp and B. Taskar. Multimodal decomposable models for human pose estimation. In *CVPR*, 2013.

[39] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011.

[40] 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.

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

[43] TODO. Towards realtime 2d pose tracking: A simple online pose tracker. In *ICCV PoseTrack Workshop*, 2017.

[44] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015.

[45] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[46] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[47] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *CVPR*, 2017.

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

[49] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[50] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *CVPR 2012*, pages 2034–2041.

[51] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *CVPR*, 2013.

[52] X. Zhu, Y. Jiang, and Z. Luo. Multi-person pose estimation for posetrack with enhanced part affinity fields. In *ICCV PoseTrack Workshop*, 2017.