PoseTrack: A Benchmark for Human Pose Estimation and Tracking

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

[1]MPI for Informatics, Saarbrücken, Germany
[2]Department of Computer Science, University of Bonn, Germany
[3]School of Computer Science, University of Adelaide, Australia

Abstract

Human poses and motions are important cues for analysis of videos with people and there is strong evidence that representations based on body pose are highly effective for a variety of tasks such as activity recognition, content retrieval and social signal processing. In this work, we aim to further advance the state of the art by establishing "PoseTrack" , a new large-scale benchmark for video-based human pose estimation and articulated tracking, and bringing together the community of researchers working on visual human analysis. The benchmark encompasses three competition tracks focusing on i) single-frame multi-person pose estimation, ii) multi-person pose estimation in videos, and iii) multi-person articulated tracking. To facilitate the benchmark and challenge we collect, annotate and release a new dataset that features videos with multiple people labeled with person tracks and articulated pose. A centralized evaluation server is provided to allow participants to evaluate on a held-out test set. 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 [34, 33, 32, 4, 35, 11, 13, 23, 2, 27] and videos [25, 5, 16, 9] and multi-person pose estimation in monocular images [26, 13, 15, 3, 24]. This progress has been facilitated by the use of deep learning-based architectures [31, 10] and by the availability of large-scale benchmark datasets such as "MPII Human Pose" [1] and "MS COCO" [21]. 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 [37, 18] and upright frontal people [5], 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. We collect, annotate and release a new large-scale dataset that features videos with multiple people labeled with person tracks as well as articulated body pose.

The dataset is organized around a challenge with three competition tracks 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 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 the overall challenge. In order to enable timely and scalable evaluation on the held-out test set, we provide a centralized evaluation server. 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.

2. Related Datasets

The commonly used publicly available datasets for evaluation of 2D human pose estimation are summarized in Tab. 2. We note the total number of annotated body poses, availability of video pose labels and multiple annotated persons per frame, and types of data.

Dataset	# Poses	Multi-person	Video-labeled poses	Data type
LSP [19]	2,000			sports (8 act.)
LSP Extended [20]	10,000			sports (11 act.)
MPII Single Person [1]	26,429			diverse (491 act.)
FLIC [28]	5,003			feature movies
FashionPose [7]	7,305			fashion blogs
We are family [8]	3,131	✓		group photos
MPII Multi-Person [1]	14,993	✓		diverse (491 act.)
MS COCO Keypoints [21]	105,698	✓		diverse
Penn Action [37]	159,633		✓	sports (15 act.)
JHMDB [18]	31,838		✓	diverse (21 act.)
YouTube Pose [5]	5,000		✓	diverse
Video Pose 2.0 [29]	1,286		✓	TV series
Multi-Person PoseTrack [17]	16,219	✓	✓	diverse
Proposed	**153,615**	✓	✓	**diverse**

Table 1. Overview of the 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, and types of data.

The most popular benchmarks to date for evaluation of single person pose estimation are "LSP" [19] (+ "LSP Extended" [20]) 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 $1,000$ is rather small. FLIC [28] 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 householding activities and consists of over $26,000$ annotated poses with $7,000$ poses hold out for evaluation. Both benchmarks focus on single person pose estimation task 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 number of people, nor their locations and scales are known.

The single-frame multi-person pose estimation setting was introduced in [8] 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. "MPII Human Pose (Multi-Person)" dataset [1] has significantly advanced 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 [21] has been introduced as an attempt to provide a new large-scale benchmark for single frame based multi-person pose estimation. While it contains over $100,000$ annotated poses it includes many images of non-challenging isolated individuals. All these datasets are only limited to single-frame based body pose estimation. In contrast, our datasets 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 the video. Our dataset is based on the MPII Multi-Person benchmark, from which we carefully select challenging key frames and for each central key frame include up to 150 neighboring frames from the corresponding publicly available video sequences. 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 extremely hard to annotate.

Recently, [17] and [12] also provide datasets for multi-person pose estimation in videos. However, both are at a very small scale. [17] provides only 60 videos with most sequences containing only 41 frames, and [12] provides 30 videos containing only 20 frames each. While these datasets provide a first step toward solving the problem at hand, they are certainly not enough to cover a large range of testing scenarios and to learn stronger pose estimation models. We on the other hand provide a large-scale benchmark with a much broader variety and well defined 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 [18], Penn Action [37] and YouTube Pose [5]. Similar to these datasets, we provide dense annotations of video sequences. However, in contrast to [18, 37, 5] 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" [30] and "Human3.6M" [14] that include images and 3D poses of people but are captured in the controlled indoor environments, whereas our dataset includes real-world video se-

quences but provides 2D poses only.

3. The PoseTrack Dataset and Challenge

In this work we introduce a new large-scale video dataset and challenge for articulated human pose tracking in the wild. We build on and extend the newly introduced datasets for pose tracking in the wild [17, 12]. 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 selected 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 vary 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 annotated the ignore regions to identify people the are extremely hard to annotate. This allows us to not penalize methods on estimating poses in those regions during training as well as evaluation. Afterwards, the head bounding boxes for each person across the videos were annotated and a track-id was assigned to every 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. Since a video can contain multiple video shots, we found that person re-identification between different shots can sometimes be very difficult even for the human annotators. We, therefore, assign a new ID to a person if they reappear in the video or appear in multiple shots. Since, in this work we do not target person re-identifcation, having different ID for the same person in different shots is not very crucial. Afterwards, the pose for every person with a track-id is separately annotated in the complete 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 that allows one to speed-up the annotations by interpolating annotations between two frames. All interpolation errors were manually removed by visualizing the annotations multiple times. Finally, we performed two additional iterations of data cleaning to remove any annotation errors and obtain high quality annotations. All annotations were performed by in-house workers. The annotators were provided with a clearly defined protocol, detailed instructions containing several examples and video tutorials, and the authors were always accessible to resolve any annotation ambiguity. Figure 1 shows example frames from the dataset.

Overall, the dataset contains 514 video sequences including $66,374$ frames. We split them into 300, 50, 208 videos for training, validation and testing, respectively. In order to reduce the annotation effort while also including more diverse scenarios, we annotated training and validation/testing videos in different manners. The length of the training videos ranges between 41-151 frames and we densely annotate 30 frames from the center of the video. Whereas, the number of frames in validation/testing videos ranges between 65 to 298 frames. In this case, we densely annotated 30 frames around the keyframe from MPII Pose dataset and afterwards annotate every fourth frame. This strategy allows us to evaluate more diverse body pose articulations and also long range articulated tracking, while having significantly fewer annotations. In total, this constitutes roughly $23,000$ labeled frames and $153,615$ pose annotations which to-this-date is the highest number of pose annotations for any multi-person pose dataset. The already available data in [17, 12] comprises approximately 100 video sequences totaling around $4,500$ labeled frames and $20,000$ annotated humans. However, in order to be consistent across the dataset and have same skeleton structure for all annotations, we re-annotated all videos. The testing set remains unpublished to avoid over-fitting, and no information about the test set is revealed including the information about which frames are labeled.

3.2. Challenges

The dataset accompanies three different challenges.

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

2. **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.

3. **Pose Tracking in the Wild.** Finally, this new challenge will require 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.

Figure 1. Example frames and annotations from the new PoseTrack dataset.

3.3. Evaluation Server

We provide an online evaluation server to evaluate the performance of different methods on thr 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 [26] and multi-target tracking [22]. 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 the 60% of the diagonal of 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 [26], and one from the multi-target tracking literature [36, 6, 22] to evaluate multi-person pose tracking.

Per-frame multi-person pose estimation. For measuring frame-wise multi-person pose accuracy, we use *Mean Average Precision* (mAP) as is done in [26]. The protocol to evaluate multi-person pose estimation in [26] requires that the location of a group of persons and their rough scale is known during evaluation [26]. 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.

Video-based pose tracking. To evaluate multi-person pose tracking, we use Multiple Object Tracking (MOT) metrics [22]. 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 not exceeding 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. We report MOTA metric for each body joint class and average over all body joints, while for MOTP, Precision, and Recall we report averages only.

The source code for evaluation metrics is available publicly on the benchmark website.

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

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

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

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

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

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

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

[12] 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*, Dec. 2017. arXiv: 1612.01465.

[13] 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.

[14] 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.

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

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

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

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

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

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

[21] 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.

[22] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, 2016.

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

[24] 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.

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

[26] 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.

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

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

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

[30] 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.

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

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

[33] J. 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.

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

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

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

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