POSETRACKREID: DATASET DESCRIPTION

A PREPRINT

Andreas Doering   **Di Chen**   **Shanshan Zhang**   **Bernt Schiele**   **Juergen Gall**

November 13, 2020

ABSTRACT

Current datasets for video-based person re-identification (re-ID) do not include structural knowledge in form of human pose annotations for the persons of interest. Nonetheless, pose information is very helpful to disentangle useful feature information from background or occlusion noise. Especially real-world scenarios, such as surveillance, contain a lot of occlusions in human crowds or by obstacles. On the other hand, video-based person re-ID can benefit other tasks such as multi-person pose tracking in terms of robust feature matching. For that reason, we present PoseTrackReID, a large-scale dataset for multi-person pose tracking and video-based person re-ID. With PoseTrackReID, we want to bridge the gap between person re-ID and multi-person pose tracking. Additionally, this dataset provides a good benchmark for current state-of-the-art methods on multi-frame person re-ID.

*K*eywords Dataset · Pose Estimation · Multi-Person Pose Tracking · Person re-identification

1 Introduction

Multi-person pose tracking [1, 2, 3] and video-based person re-identification (re-ID) [4, 5, 6, 7] are two very active research fields in computer vision, relevant for many applications like sports, autonomous driving and security. Both tasks are challenging due to different viewpoints, varying (low) image resolutions, illumination changes, occlusions, motion blur and crowded scenes.

In multi-person pose tracking, the target is to estimate all human poses for all frames of a video and link them over time. Person re-identification, on the other hand, aims to find a person of interest across multiple sequences and non-overlapping cameras. Although both tasks are mainly studied separately, they can benefit from each other. For example, person re-identification can benefit from human pose estimation by utilizing the semantic structure of the person of interest [8]. Person re-ID, on the other hand, can be utilized to link all instances of the same person across time [9].

Common datasets for person re-ID [10, 11, 12, 13] provide tight person crops and mostly contain the person of interest only. Though, in many real world scenarios such as surveillance or sports, persons are often occluded by obstacles. Further, these scenarios contain a lot of crowded scenes in which persons frequently occlude each other. This leads to a lot of ambiguities, especially if a person is partially visible or multiple persons are present within a single crop. To overcome problems such as occlusion, Miao et al. [8] propose the Occluded-DukeMTMC dataset as an extension of the DukeMTMC-ReID dataset [12] and provide keypoint information obtained from an off-the-shelf pose estimator [14]. However, off-the-shelf pose estimators such as [14] perform poorly in crowded scenes with severe occlusions, resulting in misplaced keypoints.

Video keypoint annotations, on the other hand, are costly and time consuming. Consequently, a limited amount of training data is publicly available, especially for the task of multi-person pose tracking. We argue that pose tracking and person re-ID are two related tasks which should be solved jointly. For that reason, we extended a large-scale multi-person pose estimation and tracking dataset (PoseTrack [15]) and propose PoseTrackReID. PoseTrackReID combines both tasks of multi-person pose tracking and video-based person re-ID with the aim to improve pose-based person re-ID methods in videos. To the best of our knowledge, this is the first dataset combining person re-ID and multi-person pose tracking.

2 Dataset

We briefly describe the PoseTrack 2018 dataset [15] in Section 2.1 and introduce PoseTrackReID as well as its annotation process in Section 2.2.

2.1 PoseTrack

The PoseTrack 2018 dataset[1] [15] is a large-scale dataset for multi-person pose estimation and tracking, containing 823 sequences. Following the split of the MPII Human Pose dataset [16], the sequences are split into 293, 170 and 375 videos for training, validation and testing. The majority of sequences range from 41 to 151 frames and contain 30 densely annotated frames around the middle of each sequence. On top, validation and test sequences are annotated with a step size of four frames. In total, PoseTrack 2018 consists of 46,933 labeled frames.

PoseTrack provides ignore regions to exclude crowds and small people, keypoint annotations, head bounding boxes to estimate the scale of a person and track ids. Unfortunately, track ids are not unique throughout the dataset, not even within a single sequence. If a person leaves a scene and re-enters, for instance, it is assigned a new track id. This can result in ambiguities for appearance-based similarity approaches.

2.2 PoseTrackReID

The PoseTrack 2018 dataset contains multiple sequences sampled from the same video. For that reason, the dataset is well suited to be extended by person IDs as the same person appears in multiple sequences. In addition, PoseTrack contains a lot of challenging sequences with a high degree of occlusion and motion blur. In contrast to related person re-ID datasets [10, 11, 12, 13], this results in person bounding boxes with a high degree of occlusion and multiple persons within a single person crop. Due to its challenging nature, our proposed dataset can provide a new measure of robustness for current state-of-the-art approaches. The provided keypoint annotations further allow to incorporate structural pose information paving the way for pose-based person re-ID methods. On the other hand, our dataset opens ways for re-ID based pose tracking methods. .

2.3 Data Annotation

We developed a new tool to annotate the PoseTrack 2018 dataset and adopt the same annotation format. At the moment, we only provide annotations for the training and validation sets.

The annotation process was performed in four steps. First, we re-annotated the existing bounding boxes on the training and validation set as the bounding boxes provided by PoseTrack have been an artifact of the annotation process and were therefore not accurately annotated. Specifically, most bounding boxes only contained the visible part of a person. We extended all bounding boxes to cover the entire portion of a person, including occluded body parts. To increase the difficulty of person re-ID, we further annotated bounding boxes for small persons and persons in crowds. Then, we interpolated and revised the bounding boxes between non-keyframes[2]. This results in additional **25,478** and **11,238** annotated frames on the training and validation sets, respectively. Second, we adapted the ignore regions on the training set such that the newly added persons are not anymore excluded. In a third step, we annotated unique person identities throughout the training and validation sets. Additionally, we added head bounding boxes for newly added persons for all frames on the validation set. On the training set, we only provide head bounding boxes from PoseTrack 2018.

In a final step, we adapt and annotate person keypoints for all keyframes on the training and validation sets. The original annotations of the PoseTrack 2018 dataset consist of 15 keypoints where each keypoint contains a flag, whether it is annotated or not. Unfortunately, there exist several cases in which these flags are not set reliably. In our dataset, we re-define the purpose of the keypoint flags and include occluded keypoints. In that way, pose estimation and re-ID approaches can utilize occlusion information within their training pipelines. We define a joint $j = (x, y, v)$ as occluded, if $x \geq 0, y \geq 0$ and $v = 0$. A joint is truncated if $x = -1, y = -1$ and $v = 0$. Otherwise, a joint is defined as visible if $v = 1$. After refining the original keypoints, we ran an off-the-shelf pose estimator [3] to estimate the poses for all newly added bounding boxes on keyframes only, which were then refined manually afterwards.

---

[1]We refer to the final version of PoseTrack 2018 which was released in 2019, containing more sequences as described in [15].
[2]We denote as keyframe all frames which were annotated in PoseTrack 2018.

References

[1] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In *CVPR*, 2020.

[2] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *CVPR*, 2020.

[3] Umer Rafi, Andreas Doering, Bastian Leibe, and Juergen Gall. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *ECCV*, 2020.

[4] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, 2020.

[5] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *CVPR*, 2020.

[6] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, 2020.

[7] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, 2020.

[8] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019.

[9] Rui Zhang, Zheng Zhu, Peng Li, Rui Wu, Chaoxu Guo, Guan Huang, and Hailun Xia. Exploiting offset-guided network for pose estimation and tracking. In *CVPRW*, 2019.

[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[11] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.

[12] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[13] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[15] M. Andriluka, U. Iqbal, E. Ensafutdinov, L. Pishchulin, A. Milan, J. Gall, and Schiele B. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018.

[16] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.