Interactive Labeling for Human Pose Estimation in Surveillance Videos

Mickael Cormier[1,2,3], Fabian Röpke[1], Thomas Golda[1,2,3], and Jürgen Beyerer[2,1,3]

[1]Karlsruhe Institute of Technology, [2]Fraunhofer IOSB, [3]Fraunhofer Center for Machine Learning

Abstract

Automatically detecting and estimating the movement of persons in real-world uncooperative scenarios is very challenging in great part due to limited and unreliably annotated data. For instance annotating a single human body pose for activity recognition requires 40-60 seconds in complex sequences, leading to long-winded and costly annotation processes. Therefore increasing the sizes of annotated datasets through crowdsourcing or automated annotation is often used at a great financial costs, without reliable validation processes and inadequate annotation tools greatly impacting the annotation quality. In this work we combine multiple techniques into a single web-based general-purpose annotation application. Pretrained machine learning models enable annotators to interactively detect pedestrians, re-identify them throughout the sequence, estimate their poses, and correct annotation suggestions in the same interface. Annotations are then inter- and extrapolated between frames. The application is evaluated through several user studies and the results are extensively analyzed. Experiments demonstrate a 55% reduction in annotation time for less complex scenarios while simultaneously decreasing perceived annotator workload.

1. Introduction

Motion and behavior analysis of individuals in crowds or groups of people from different camera perspectives present several opportunities for very different actors. Applications include smart surveillance, robotics, online and offline video search, and other monitoring systems. This kind of scenario is typically uncooperative and raises several challenges. Cameras are strategically placed in elevated position in order to monitor traffic, assure the traveler's safety and security or monitor players and fans in a station. Therefore, these cameras cover wide field of views considerably increasing the difficulty of tasks such as person detection and human pose estimation. Those are complicated by (self-)occlusions, truncations at image borders, crowdedness, quality and resolution of the frames as illustrated in Figure 1. As a result, annotating a single pose may require up to 60 seconds. Scaling this time with the number of pedestrians per frame, frame rate, length of the sequence, number of scenarios, and camera angles, it rapidly becomes obvious that annotation processes are extremely laborious and thus cost-intensive and long-winded. For comparison, the classification dataset ImageNet was annotated with image-level annotation by tens of thousands of annotators over the course of multiple years [12, 28]. However, the skeleton representation of pedestrians is advantageous in regards to ethics and data privacy, since it decouples activities from visual information, such as gender or skin complexion and also reduces the computational power required to infer activities. Due to the high cost of such high-level annotation only few datasets provide bounding box annotations [11, 29], video instance segmentation [38] or Pose Tracking [1] with action recognition for surveillance videos [25] and remain comparatively small. In this paper we propose to enhance annotation capabilities for human pose estimation in surveillance videos. Our contribution is

Figure 1: Section of a surveillance video including pose annotations cropped and zoomed-in. Human pose estimation faces several challenges due to (self-)occlusions (occluded keypoints are here annotated in pink), truncations at image borders, crowdedness and low resolution.

threefold: (1) we propose a collection of interactive tools for annotating and tracking persons and their poses without requiring pre- or post-processing; (2) we extensively evaluate five workflows based on these tools with 27 human annotators on a sequence from the publicly available PoseTrack18 dataset [1] and discuss the limits of such datasets; (3) we evaluate the use of assisted annotation against manual annotation with 40 experiments with a total of 10 human annotators on surveillance footage used for activity recognition. We provide extensive experimental results for annotation time, quality and perceived workload of the annotators. Advantages and limitations of the proposed workflows are shown and discussed. Finally our results show a 55% reduction in annotation time for less complex scenarios while simultaneously decreasing perceived annotator workload and visibly stabilize annotation quality over time.

2. Related Work

Human Pose Estimation. Recent Human Pose Estimation methods are mainly divided into bottom-up and top-down approaches. In general bottom approaches first detects joints and then groups them into individuals, which is challenging [6, 17, 30]. While the main advantage of bottom up methods is that their run-time remains almost constant for increasing size of crowds, there is still an important performance gap, due to the difficulty to estimate human poses at different scale in a single path [7, 50]. However top-down methods [5, 35, 44, 47, 45, 49] depend on reliable person detection, which is often challenging to obtain in crowded surveillance scenes [43, 11]. As shown in [33], provided ground truth bounding boxes improve pose estimation greatly, which proves to be practical for the interactive annotation of human poses.

The collection and creation of video surveillance datasets with human pose annotation is often difficult due to privacy and ethical concerns. Furthermore the annotation of several dozens pedestrians per frame is hugely time-consuming and cost-intensive. In recent years several Human pose annotation datasets has been proposed [2, 23, 24, 51] however mostly for static images. Currently only two benchmarks contain poses and tracks for videos [1, 25]. Only [25] is annotated for surveillance videos and action recognition. There are scarcely any video datasets with crowds in surveillance videos [11, 25, 29, 36] and hence only few works focus on challenges specific to crowd pose estimation in the real world such as (self-)occlusion [21, 10], and therefore mostly use simulated data [13, 16], domain adaptation [15] or data augmentation [31]. Therefore large datasets for surveillance are demanded and efficient video annotations are required.

Video and Interactive Annotation. Existing approaches use linear [42, 9] or geometric [14] interpolation for bounding boxes in videos. [22] relies on visual interpolation and frame selection guidance to reduce annotation time for bounding boxes. In [27] the problem of annotating people in crowded videos is tackled, by first annotating the center of the subjects manually through the video and then integrating detection from a person detector. In a similar fashion, it is argued in [39] that bounding box annotations are sufficient to semi-automatically segment objects with pseudo-labels. In [3] a single unified ML-aided interface is proposed to perform full image segmentation in a single pass, using initial ML segmentation. [26] propose an interactive approach to address annotation of human pose with an active learning framework, which requires annotating the most uncertain images among a large set of unlabelled images, re-training the pose estimator and using the predictions. [8] use three tools for video object segmentation: First, the Interaction-to-Mask tool predicts segmentation masks based on clicks. These are then extrapolated to other frames using their propagation tool. Finally their Difference-Aware-Fusion Tool allows annotators to interactively subtract and add regions to the annotation. Several works use similar approaches and focus on refining their models [34, 20, 48, 37, 4].

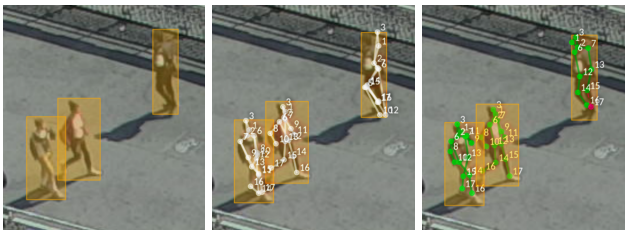
3. Overview of our Approach

Our aim is to determine whether tool assistance is able to expedite annotation processes for human pose estimation in surveillance scenarios, especially in the context of activity recognition. Nowadays, such video footage is typically in full HD resolution, with a frame rate of 30 FPS, acceptable brightness and contrast during the day and potentially blurry at night. Activity recognition concentrates on 2-5 main subjects and potentially several dozens of pedestrians due to the large field of view of elevated cameras. The human poses are various, including cyclists, persons sitting as well as poses including occlusions, truncation, and unusual activity such as lying or kicking. Therefore, based on these properties, we can argue about different degrees of complexity for scenes for which annotation is required and adapt interactive tools to those. Thus, a complex sequence contains less common pose articulations, occlusions, truncation, and multiple subjects in a low quality video, whereas a rather simple sequence features a small number of subjects in common poses in a high FPS, high quality video. In order to design an interactive tool, these properties are required to be taken advantage of.

According to [41], interpolation of any kind reduces annotation effort by an order of magnitude. Furthermore, visual annotation [22] reduces the annotation time for less complex annotations such as bounding boxes considerably. Moreover, as stated previously, interactive machine learning-aided tool assistance is greatly beneficial to instance segmentation.

3.1. Web based Annotation Tool

As the time of writing, only custom forks of Sloth [32] and VATIC [40] are known to provide open-source software for human pose estimation, which have limited capabilities and require installation. For flexibility and distributed work, a web based annotation tool is therefore developed. With feedback from annotators using Sloth, several missing and simple functionalities are included for instance, in order to toggle the visibility or change the colors and opacity of annotations, topologies, keypoints, and labels. It comes with an editing history with the ability to undo and redo actions as well as the functionality to reuse annotations between frames. Load times and memory usage for long sequences are reduced through chunking of annotations and frames. Shortcuts are implemented for frequently used actions. Further, several tools are implemented to assist the annotator in his work and generate annotation suggestions, which share the same properties as annotations. The difference is visualized through colors. Annotation suggestions are grey until they are accepted as illustrated in Figure 2. Suggestions can be accepted through editing or for the entire frame or sequence. Multiple suggestions may exist for the same entity in the same frame. In that case, the viewer displays the annotation suggestion that was generated by the currently selected tool.



(a) Box annotations (b) Pose suggestions (c) Final annotations

Figure 2: Provided bounding boxes, such as in (a), the pose estimation processing tool suggests pose annotations (b), that annotators accept or correct (c).

3.2. Interpolation and Extrapolation

First, interpolation is implemented as annotation generator. In this case, the coordinates of a keypoint are treated as vectors; each component is extrapolated/interpolated separately using a variety of interpolation methods, such as linear interpolation, cubic spline interpolation, or Lanczos resampling, whereas other annotation tools such as CVAT [9] or VATIC [40] are restricted to linear interpolation. Extrapolation is implemented analogously. Here, the user is allowed to flexibly choose at any moment the number of frames to use and may adapt this number along the annotation process.

3.3. Machine-Learning aided Suggestions

Second we long for using machine-learning aided tools. Therefore, we require modular components performing computationally expensive tasks server-side. To this aim we implement tools using deep learning models on a remote server over REST-API. Such models are then integrated in different tools with their own parameters such as a person detector, a re-identification tool, and a pose estimator. In order to conduct experiments and user studies with those tools, the means to collect, visualize, and export statistics are integrated directly into the annotation tools. Statistics are made available for a whole dataset, annotation process, and job, and can be filtered by annotation type, author, generator, sequence, and frame. Here, annotation generators may be human annotators or tools, such as interpolation, extrapolation or the different model-based tools.

Particularly, a human pose estimation tool is implemented and adds the ability to generate pose suggestions by providing bounding box annotations, generated manually, by the person detection processing tool or other means. This is achieved using a pre-trained top-down pose estimator [46], as shown in Figure 2; the processing tool receives multiple bounding boxes as its input and outputs the corresponding pose suggestions. Annotators are able to adjust the confidence threshold to balance quality and quantity of detected keypoints. Alternatively, annotators are able to draw bounding boxes and immediately receive pose suggestions from the pose estimation processing tool. This workflow is advantageous in sequences with few subjects or in other workflows without the detection processing tool, since annotators remain in control of entity IDs.

4. Experiments with Human Annotators

The different experiments are conducted in our lab with the same hardware in the same conditions, over a period of two weeks. The annotators come from different backgrounds: about a third regularly annotate video data, another third are involved in the tools development but rarely annotate themselves and finally the last third never annotated nor has any background in computer vision. The study is scheduled for a duration of a least two hours.

First, we evaluate the performance of the proposed tools on a challenging sequence of PoseTrack18 dataset (Section 4.1). Second, we evaluate the most promising workflow against manual annotation for simple and complex scenes in surveillance footage (Section 4.2). For this purpose, annotation time and perceived workload are measured among other metrics and annotator feedback is collected. Finally, the results are discussed and limitations are shown (Section 4.3).

4.1. Experiments on PoseTrack18

PoseTrack18 is one of the larger dataset which is publicly available for multi-person pose estimation and tracking in videos. It contains 514 videos including 66,374 frames in total. However only a part of these frames are annotated: 30 frames from the center are annotated for the training videos. Additionally every fourth frame is also annotated for evaluating pose tracking in validation and test videos. The annotations include 15 body keypoints location, a unique person id and a head bounding box for each person instance.

For our experiments 27 human annotators are tasked with annotating a selected sequence manually and with different tool combinations, providing a total of more than 3,800 bounding box and pose annotations. The main intent of this experiment is to compare individual annotation workflows in regard to annotation speed and perceived workload.

4.1.1 Design of the User Study

Due to the size of the dataset and resource constraints, we select a portion of a validation sequence for a total of 48 frames. As shown in Figure 3 it depicts multiple football players; the three players with the jersey numbers 2, 17, and 33 are annotated. The camera angle from which these sequences were recorded is similar to those usually found in surveillance videos. However due to camera motion and zoom, the scene presents additional challenges. The frame rate of the sequence is estimated at 2-7 FPS, hence much lower than in our surveillance sequences in Section 4.2.

The participants are split into five groups. One group consists of three participants annotating the sequence manually, omitting occluded keypoints. The other four groups annotate occluded keypoints and annotators are split among them evenly and randomly (Table 1). The participants are told that the focus of the user study lies on the annotation speed. A sixth workflow involving automated detection, re-identification and pose-estimation was aborted after after the first and the second subject required around three hours each. Those results are discarded.

In order to reduce the confounding effects of annotators being faster than others based on their experience, participants are directed to familiarize themselves with our web-interface and their individual workflow for 15 minutes. After this familiarization period, a timer is started to measure annotation time and the three players in the sequence are each annotated with a bounding box and pose. Following the annotation process, participants rate their perceived workload in five different subscales according to a modified NASA Task Load Index (NASA-TLX) which is detailed in the supplemental material.

4.1.2 Annotation Speed

The main focus of this experiment is measuring annotation speed, for which the statistics are collected in our application. The mean manual annotation time per bounding box and pose is 44.1 seconds. However, there is a rather large range and recorded values of individual participants span from 32.2 to 57.7 seconds. With a mean of 42.8 seconds, the group using the Human Pose Estimation (HPE) + Copy method accomplishes the task marginally faster than manually, as illustrated in Figure 4a. The usage of inter- or extrapolation on the other hand is detrimental to annotation speed in this sequence. Extrapolation increases mean annotation time to 54.7 seconds and the group that uses pose estimation suggestions on every fourth frame and then interpolates between them required 50.3 seconds. Since several datasets do not provide occluded keypoints, and in order to facilitate comparisons, the experiment included a group that annotated manually, though without occluded keypoints. This group show a mean time of 34.4 seconds, which corresponds to a speedup of 21.9% relative to manual annotation including occluded keypoints.

4.1.3 Perceived Workload

In addition to measuring the time required for annotation, the perceived workload is investigated. To this aim, the annotators are asked to rate their individual perceived workload in five different dimensions; mental demand, temporal demand, their own performance, how much effort they put in to achieve that performance, and how frustrated they were while working on the task. Similarly to annotation time, the ratings vary greatly. Nevertheless, a few trends can be observed in Figure 5. The HPE + Copy group estimate their temporal demand relatively low, even though their actual performance is almost the same as the manual annotators'. They also express little frustration. The inter- and extrapolation groups provide a high temporal demand rating, which corresponds with their results. However, the extrapolation group overestimates their performance. The perceived effort is the same for each of the six groups. Overall, the HPE + Copy group exhibits the lowest, and thus best, task load index. When not required to annotate occluded keypoints, the annotators' mental demand tends to be lower. In contrast, the group overestimates temporal demand compared to manual annotation with occlusions. Perceived performance, effort, and frustration are improved.

4.1.4 Annotation Quality

As described above only a parts of the dataset are annotated. In our case, 26 out of 48 selected frames are annotated. We compute Average Precision as in [1] for the three annotated players in each experiment and compare the annotation time

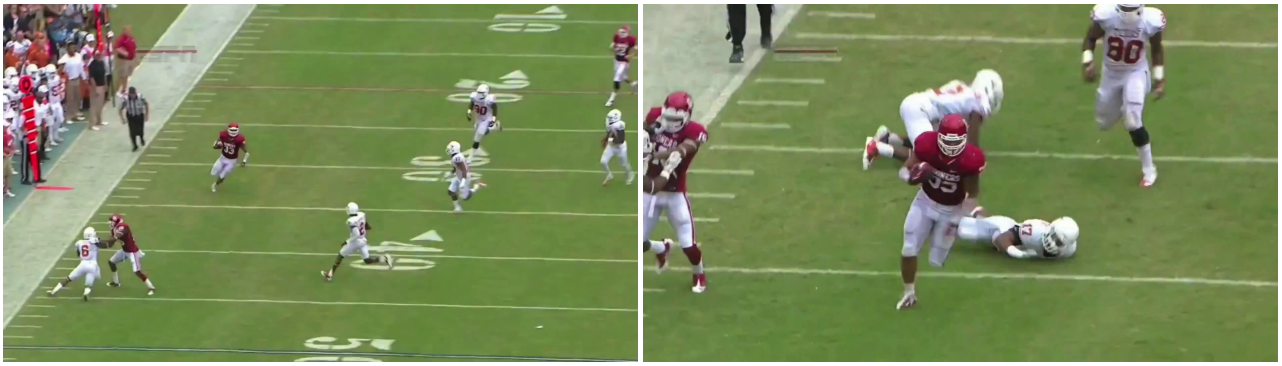
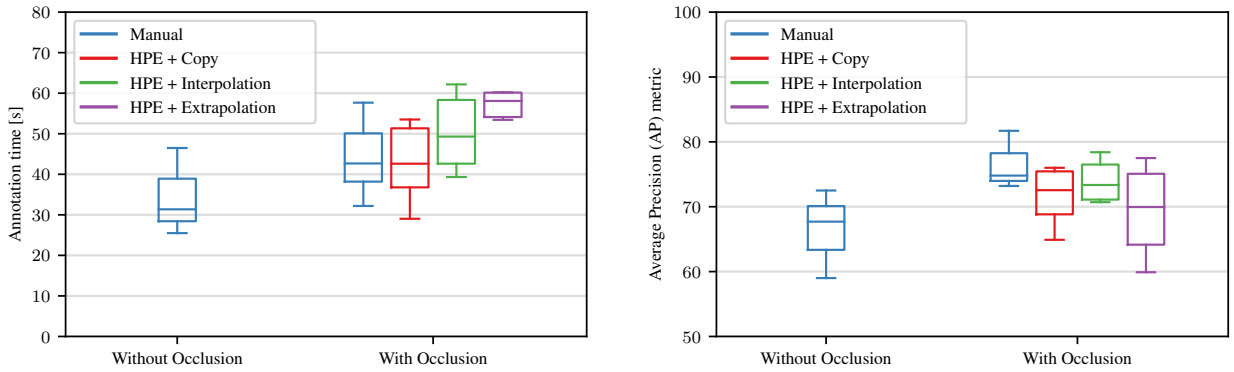


Figure 3: The first and last frame of the selected sequence from [1]. 48 consecutive frames are selected from the sequence and three players are annotated during the experiments: 2, 17 and 33.

Workflow	Annotators	Description
Manual	6	Each keypoint is annotated separately
HPE + Copy	6	Annotators choose between pose suggestions for bounding boxes, copying/duplicating annotations, and manual annotation
HPE + Interpolation	6	Pose suggestions for bounding boxes on every fourth frame, interpolation in between
Extrapolation	6	Detection and pose suggestions on the first two frames, extrapolation afterwards
Manual without occlusion	3	Each keypoint is annotated separately; occluded keypoints are skipped entirely

Table 1: In the PoseTrack18 experiments, each participant is assigned one of the described workflows.



(a) Annotation time per bounding box and pose.

(b) Average Precision per bounding box and pose.

Figure 4: Annotation time (lower is better) in (a) and AP (higher is better) in (b) per bounding box and pose. A description and sample size is given for each workflow in Table 1. The results in (b) are only computed for 26 images with ground-truth from 48 images in total. While annotating manually without occlusion is much faster, this workflow also provides the worst results for this sequence. Manual annotation is both faster and more accurate compared to tool assisted workflows.

required in Figure 4b. We also report AP for each keypoint of the most and least accurate annotator as well as the average results in Table 2.

These results are paired with the baseline pose estimator reported in [33] for the whole validation set. However, with further analysis of the ground truth data, we found several mistakes greatly impairing our results, which are illustrated in Figure 6. Clearly visible keypoints are missing, left/right elbows are sometimes switched. This emphasizes the difficulty of annotating and validating such large human pose estimation datasets.

4.2. Experiments on Surveillance Footage

We conduct a second series of experiments with two sequences taken from outdoor surveillance footage for research on activity recognition in public places. In this case

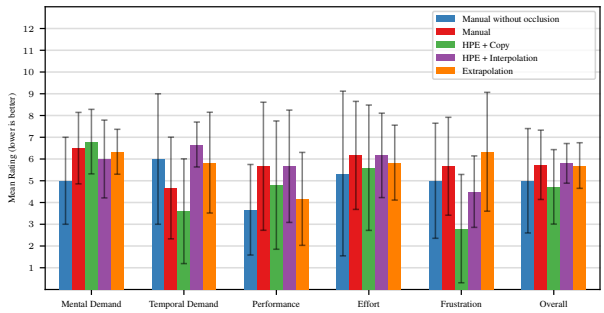
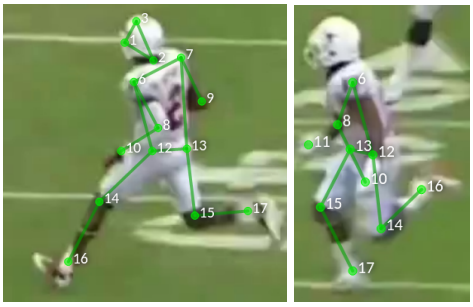


Figure 5: Perceived workload as measured by a reduced raw NASA-TLX. Scale from 1-10, lower is better. Error bars indicate standard deviation. The sample sizes are given in Table 1.

Annotation Method	AP							
	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
Mean	55.3	82.0	76.9	65.9	75.0	80.1	74.9	71.7
Most accurate (Manual w. occ.)	63.3	90.0	86.9	77.7	90.2	85.1	87.5	81.7
Least accurate (Manual w/o. occ.)	49.6	63,4	62.0	57.2	57.7	69.7	57.6	59.0

Table 2: Per-joint AP. Due to helmets and noise in the GT (see Figure 6) the AP for the three head keypoints remains quite low.



(a) Accurate GT (b) Inaccurate GT

Figure 6: The ground truth data from PoseTrack18 contain incomplete or inaccurate annotations. While the pose in (a) is reasonably accurate, the head in (b) has been skipped, and the right elbow has been annotated with keypoint 8, which actually corresponds to the left elbow.

no ground truth data is provided. The aim of this experiment series is to assess the impact that sequence complexity has on total annotation speed and advantages/usefulness of tool assistance, as well as to determine whether there is a learning curve and how steep it might be. This time, ten annotators are given a simple and a complex sequence depicting two civilians, respectively walking and fighting. The first frames of each sequence are shown in 7. The annotators are asked to annotate manually as well as with a predetermined workflow, which amounts to four tasks altogether. Each sequence consists of 20 frames, thus adding up to 800 bounding box and pose annotations in total. The two sequences are subjectively chosen as the easiest and the hardest sequence to annotate in a surveillance video. They coincide in camera angle, time of day, and general setting, however the fighting sequence includes faster, less predictable motions, unusual pose articulations, as well as (self-)occlusions. The sequences are recorded with 30 FPS.

4.2.1 Design of the User Study

According to [19] annotating the same image twice has a severe effect on performing a segmentation task. Thus, the annotators are presented with the sequences in a randomized order to avoid any confounding errors resulting from familiarity with the before-seen sequence or our user interface. The formulation of the task remains the same as in Section 4.1; drawing bounding boxes and annotating poses with the PoseTrack18 topology, including occluded keypoints. In this experiment however, annotators are more experienced. All of them previously participated in the other study and a large portion regularly annotate bounding boxes with our interface. Each sequence is annotated twice by every annotator. In the first scenario, they annotate each keypoint manually. In the other scenario, annotators follow this protocol: first, they annotate frames 1, 7, 14, and 20, choosing between pose suggestions generated from bounding boxes, copying/duplicating annotations, and manual annotation. Afterward, they interpolate the frames in between, correcting the frames with the largest discrepancy between interpolation suggestion and true pose. This way, remaining interpolation suggestions are updated in real time. They repeat this until all suggestions are precise enough. Lastly, they fill out a modified NASA-TLX form.

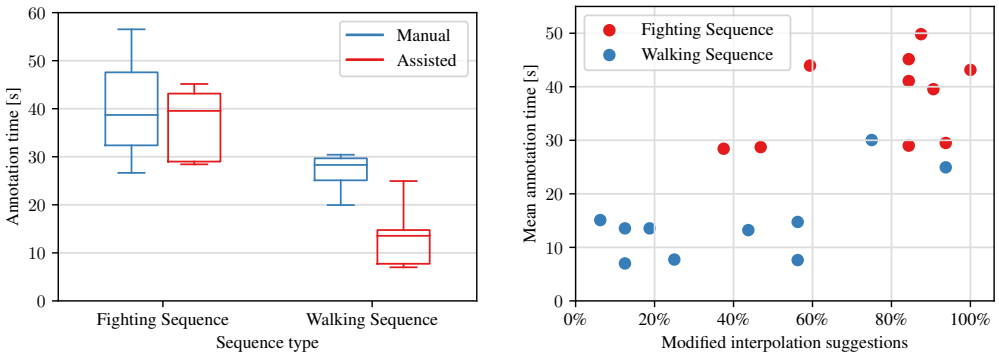
4.2.2 Annotation Speed

The results are shown in Figure 8a for both scenes and workflows. The mean manual annotation time for the fighting sequence is 41.0 seconds and 29.0 seconds for the walking sequence. This corresponds to a 29.2% difference caused by sequence complexity alone. When assistance from pose estimation and interpolation is used, mean annotation times decrease to 36.5 seconds and 13.1 seconds, which amounts to 10.9% and 55.5% speedups respectively. As observed before, the variability of the distributions for the fighting sequence and assisted annotation of the walking sequence is considerable.

The collected statistics permit to gather in-depth knowledge about the annotation process. As expected, annotators correcting a larger portion of suggestions are slower, an effect amplified by sequence complexity as illustrated in Figure 8b.

(a) Walking (b) Fighting

Figure 7: Frames from surveillance footage on a public place. In (a) the two persons are walking towards each other. The scene is simple: the movements are clear, monotonous and there are no occlusions. In contrast, (b) is a quite complex scene with several (self-)occlusions and unpredictable motions.



(a) (b)

Figure 8: (a) Annotation time per bounding box and pose for each of the four scenarios. Sample size is 10 per scenario. (b) Relationship between the percentage of modified interpolated annotation suggestions and mean annotation time per annotator for the walking and fighting sequences.

4.2.3 Perceived Workload

As might be expected the annotators consistently rate different dimensions of their perceived workload lower after the walking sequence than after the fighting sequence, as shown in Figure 9. Overall, perceived workload decreases by 27.1% for manual annotation and by 38.0% for tool assisted annotation. This is most noticeable in the mental demand dimension with 45.5% and 47.5% lower ratings. The results allow several observations. First, the mental demand is similar for manual and assisted annotation, while the temporal demand is noticeably lower in the walking scenario, which corresponds with the measured annotation time shown in 4.2.2. Finally, effort and frustration are rated lower when employing tool assistance. Overall, there is no significant decrease in perceived workload when tool assistance is employed for the fighting sequence. For the walking sequence, perceived workload is reduced to some degree. Admittedly, variability is rather high in this experiment. Standard deviations of subscales are partially greater than the mean values.

4.3. Discussion

Altogether, comparing our experimental results with relevant publications is challenging, since information on how large multi-person HPE and tracking datasets are annotated is almost non-existent. PoseTrack18 is annotated using VATIC [1, 40], which presents the ability to linearly interpolate annotations and is forked to incorporate other tool assistance capabilities [18]. However, it is unclear which workflow was employed during the annotation of the PoseTrack18 dataset. Furthermore, the authors chose to skip oc-

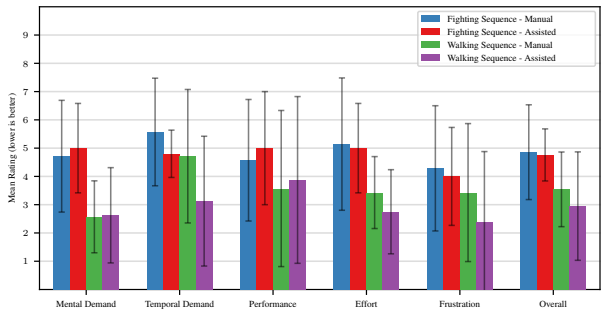


Figure 9: Perceived workload as measured by a reduced raw NASA-TLX. Scale from 1-10, lower is better. Error bars indicate standard deviation. Sample size is 10 per scenario.

cluded keypoints, which would greatly help pose tracking for action recognition.

Precise pose annotation, especially including occlusions, is time consuming. Statistics for publicly available datasets are sparse. For the selected PoseTrack18 sequence, the median manual annotation time per bounding box and pose combination is 44.1 seconds. Although tool assistance proves to be beneficial in the surveillance sequences, the annotation time for this sequence is not significantly reduced by any of the workflows employed during the PoseTrack18 experiment. This indicates that tool assistance is not always appropriate.

The pose estimator does not always accurately detect keypoints when large portions of a person's body parts are occluded, or when the image is blurry due to swift movements. Interpolation only produces precise suggestions if the person performs smooth movements; sudden motions distort the suggestions. In these cases, the annotators have to spend substantial time correcting the suggestions, sometimes more than it would take to annotate the poses manually.

These findings highlight the importance of interactivity; allowing annotators to choose the right tools for the scenario results in more precise annotation suggestions which require less time to correct. Annotator experience has a large impact on annotation speed. Experience levels were collected in the context of the survey, but are not sufficient to establish a correlation to annotation time. Inherent differences between annotators are an alternative explanation for the large variability in pose annotation speed. As illustrated in Figure 8b, two annotators modify 6.3% and 93.8% of suggestions to achieve the annotation quality they are satisfied with for the same scenario.

Furthermore, annotators perceive the same workload differently. Analogous to annotation times, distributions of the task load indices have large standard deviations as depicted in Figure 5 and Figure 9. Sequence complexity influences perceived workload to a greater extent than annotation workflow, as illustrated in Figure 9. Similarly, the absolute task load indices for both surveillance sequences are lower than the PoseTrack18 task load indices. The higher FPS availability is not only beneficial for tool assistance, but also supports human annotators in understanding the motions of the keypoints. Finally, in all cases the mental demand is quite high, although annotators consider annotation to be dull and monotonous. The task is highly repetitive, but annotators have to constantly think about the spatial positioning of keypoints over time. Several annotators had difficulties annotating the left and right keypoints from the person's perspective, or found it exhausting.

5. Conclusions

The creation of a dataset for pose based activity recognition in surveillance videos is long-winded and costly. In this work we presented and evaluated a simple framework for interactive video human pose estimation. We propose several tools from extrapolation and interpolation to deep-learning aided tools to detect and estimate the poses of persons. We extensively evaluate our approach with human annotators in over 60 experiments for simple and complex scenes. Our results show a 55% reduction of the annotation time of a simple surveillance scenario while simultaneously decreasing the perceived workload for the annotator. Further we show the diverse challenges of human pose annotation for action recognition in surveillance videos. Future works should focus on improving the annotation process for complex scenes with several dynamic occlusions such as fight scenes and multiple occlusions through scenes with larger crowds.

References

[1] 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. 1, 2, 4, 5, 7

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[3] Mykhaylo Andriluka, Jasper RR Uijlings, and Vittorio Ferrari. Fluid annotation: a human-machine collaboration interface for full image annotation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1957–1966, 2018. 2

[4] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11700–11709, 2019. 2

[5] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xinyu Zhou, Erjin Zhou, Xi-

angyu Zhang, and Jian Sun. Learning delicate local representations for multi-person pose estimation, 2020. 2

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018. 2

[7] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020. 2

[8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, 2021. 2

[9] CVAT. Computer vision annotation tool (cvat). https://github.com/openvinotoolkit/cvat/, 2018. 2, 3

[10] Yan Dai, Xuanhan Wang, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Rsgnet: Relation based skeleton graph network for crowded scenes pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1193–1200, 2021. 2

[11] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1, 2

[12] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[13] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *Proceedings of the European conference on computer vision (ECCV)*, pages 430–446, 2018. 2

[14] Pedro Gil-Jiménez, Hilario Gómez-Moreno, Roberto López-Sastre, and Saturnino Maldonado-Bascón. Geometric bounding box interpolation: an alternative for efficient video annotation. *EURASIP Journal on Image and Video Processing*, 2016(1):1–13, 2016. 2

[15] Thomas Golda, Andreas Blattmann, Jürgen Metzler, and Jürgen Beyerer. Image domain adaption of simulated data for human pose estimation. In Judith Dijk, editor, *Artificial Intelligence and Machine Learning in Defense Applications II*, volume 11543, pages 112 – 127. International Society for Optics and Photonics, SPIE, 2020. 2

[16] Thomas Golda, Tobias Kalb, Arne Schumann, and Jürgen Beyerer. Human pose estimation for real-world crowded scenarios. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. 2

[17] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *European Conference on Computer Vision*, pages 718–734. Springer, 2020. 2

[18] johndoherty. Vatic - video annotation tool from irvine, california. https://github.com/johndoherty/vatic/, 2015. 7

[19] Alexander Kirillov. Joint coco and mapillary recognition challenge workshop. URL: hhttp://presentations.cocodataset.org/ECCV18/COCO18-Panoptic-Overview.pdf, 9 2018. 6

[20] Anton Konushin Konstantin Sofiiuk, Ilia Petrov. Reviving iterative training with mask guidance for interactive segmentation. *arXiv preprint arXiv:2102.06583*, 2021. 2

[21] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019. 2

[22] Alina Kuznetsova, Aakrati Talati, Yiwen Luo, Keith Simmons, and Vittorio Ferrari. Efficient video annotation with visual interpolation and frame selection guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3070–3079, 2021. 2

[23] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 2

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 2

[25] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020. 1, 2

[26] B. Liu and V. Ferrari. Active learning for human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4373–4382, 2017. 2

[27] Santiago Manen, Michael Gygli, Dengxin Dai, and Luc Van Gool. Pathtrack: Fast trajectory annotation with path supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 290–299, 2017. 2

[28] John Markoff. Seeking a better way to find web images, 2012. 1

[29] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831. 1, 2

[30] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*, 2016. 2

[31] Rafal Pytel, Osman Semih Kayhan, and Jan C van Gemert. Tilting at windmills: Data augmentation for deep pose estimation does not help with occlusions. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10568–10575. IEEE, 2021. 2

[32] Sloth. Sloth. https://github.com/cvhciKIT/sloth/, 2011. 3

[33] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 5

[34] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation, 2020. 2

[35] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[36] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking pedestrian heads in dense crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3865–3875, 2021. 2

[37] Jasper R. R. Uijlings, Mykhaylo Andriluka, and Vittorio Ferrari. Panoptic image annotation with a collaborative assistant, 2020. 2

[38] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. *arXiv:1902.03604[cs]*, 2019. arXiv: 1902.03604. 1

[39] Paul Voigtlaender, Lishu Luo, Chun Yuan, Yong Jiang, and Bastian Leibe. Reducing the annotation effort for video object segmentation datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3060–3069, 2021. 2

[40] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International journal of computer vision*, 101(1):184–204, 2013. 3, 7

[41] Carl Vondrick, Deva Ramanan, and Donald Patterson. Efficiently scaling up video annotation with crowdsourced marketplaces. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 610–623, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 2

[42] Ketaro Wada. labelme: Image Polygonal Annotation with Python. https://github.com/wkentaro/labelme, 2016. 2

[43] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3268–3278, 2020. 2

[44] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[45] Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Vipnas: Efficient video pose estimation via neural architecture search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 2

[46] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[47] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. In *CVPR*, 2021. 2

[48] Chang-Su Kim Yuk Heo, Yeong Jun Koh. Guided interactive video object segmentation using reliability-based attention maps. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[49] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2020. 2

[50] Jiabin Zhang, Zheng Zhu, Jiwen Lu, Junjie Huang, Guan Huang, and Jie Zhou. Simple: Single-network with mimicking and point learning for bottom-up human pose estimation. *arXiv preprint arXiv:2104.02486*, 2021. 2

[51] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019. 2