Where are we with Human Pose Estimation in Real-World Surveillance?

Mickael Cormier[3,1] Aris Clepe[1,2] Andreas Specker[3,1] Jürgen Beyerer[1,3]

[1]*Fraunhofer IOSB*, Karlsruhe, Germany; [2]*Fraunhofer Center for Machine Learning*;
[3]*Vision and Fusion Lab, Institute for Anthropomatics and Robotics*,
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

firstname.lastname@iosb.fraunhofer.de

Abstract

The rapidly increasing number of surveillance cameras offers a variety of opportunities for intelligent video analytics to improve public safety. Among many others, the automatic recognition of suspicious and violent behavior poses a key task. To preserve personal privacy, prevent ethnic bias, and reduce complexity, most approaches first extract the pose or skeleton of persons and subsequently perform activity recognition. However, current literature mainly focuses on research datasets and does not consider real-world challenges and requirements of human pose estimation. We close this gap by analyzing these challenges, such as inadequate data and the need for real-time processing, and proposing a framework for human pose estimation in uncontrolled crowded surveillance scenarios. Our system integrates mitigation measures as well as a tracking component to incorporate temporal information. Finally, we provide a detailed quantitative and qualitative analysis on both a scientific and a real-world dataset to highlight improvements and remaining obstacles towards robust real-world human pose estimation in uncooperative scenarios.

1. Introduction

An aim of research on intelligent video surveillance and human activity recognition in the real world is to provide near real-time detection of persons in need of assistance due to an accident or fall with subsequent injury, as well as detection of violence in public places. In all instances, the surveillance footage requires immediate processing in order to provide human assistance on-site in a matter of minutes. However, the rapidly increasing number of surveillance cameras rightly raises important concerns regarding privacy as well as proven bias towards distinct groups of persons based on their appearance [32]. An attempt to circumvent such problems relies on skeleton-based activity recognition, which promises an interesting degree of anonymization [13]. Furthermore, since a skeleton representation is a heavy abstraction, it further reduces computing considerably which is an essential aspect against traditional two-streams RGB and Optical Flow-based methods [57]. Public services such as hospitals or police departments often lack essential funding and infrastructure to support large-scale models. Therefore, video processing of a live feed is usually required to fit on a single consumer-grade GPU or even on smaller embedded systems. Hence, processing pipelines are constrained to be lightweight and thus models must have a low memory footprint, which often correlates with reducing the number of model parameters.

Figure 1: In a real-world setting person detection and pose prediction encounter various obstacles: train tracks, power cables and moving trams block different parts of the image; camera angle creates variance in person size.

While Human Pose Estimation (HPE) arguably provides promising results for scientific datasets, there is no large-scale dataset for outdoor HPE in surveillance scenarios. Publicly available datasets lack large fields of view, different elevated views with multiple steep angles, etc. As illustrated in Figure 1, real-world data distribution highly differs from conventional datasets. Public transportation infrastructure such as rails, pavement, and power cables may dominate a surveillance area. Due to the elevated view

persons are partially or almost totally occluded by diverse structures, each other or even themselves. Furthermore, depending on the time of the day and the year, different challenges such as brightness and contrast arise. Consequently, detections from a person vary over time or may be absent shortly, and pose estimation is accordingly difficult. Even small perturbations in the size of the bounding boxes often lead to different poses over time, which strongly impairs the quality of action recognition. In this paper, we propose a framework for crowd pose estimation in real-world surveillance. In contrast to similar approaches for detection and tracking of human poses which use memory costly 3D CNNs [40, 14], our framework fits in a single consumer GPU and can be replicated and scaled for multiple cameras with only little effort. Our framework contains three main components: a person detection module for real-world person detection, a robust and fast visual tracking module delivering temporal consistent tracks, and a crowd pose estimation module for fast pose estimation with multiple tracks.

In summary, we address the challenges of human pose estimation for real-world surveillance and review both quantitatively and qualitatively the state-of-the-art HPE methods against a real-world dataset designed to highlight those challenges. Our contribution is threefold: (1) we propose a framework for real-world HPE in surveillance and crowd, (2) we show that CNNs in HPE don't generalize well to the surveillance domain and highlight remaining obstacles towards a fast and robust real-world HPE, (3) and we propose strategies to stabilize poses over time in order to facilitate skeleton-based activity recognition.

2. Related Work

2.1. Human Pose Estimation

Since the first application of CNNs for HPE [44] multiple datasets with growing size gained in importance and were subsequently made available [28, 2, 1, 42, 12, 22, 51, 31]. The COCO dataset [28], which is one of the most commonly used, has over 200,000 images and 250,000 poses. Similar to the MPII dataset [2], COCO features non-continuous images with common poses and a frontal view. PoseTrack18 [1] bases on the MPII dataset and features continuous video frames with more complex real life scenarios in controlled environments, such as sport events. COCO defines it own topology with 17 keypoints, of which five (nose, eyes, ears) are on the head. In more realistic scenarios with steeper camera angles, reliably detecting the ears and eyes of a person is challenging. Therefore, the MPII and Posetrack18 topologies simplify the pose by reducing the head keypoints to two and three, respectively.

Such datasets primary represent the human pose in common and simple situations with favorable camera angles. OCHuman [51] and OCPose [31] are smaller datasets that address (self-)occlusion with similar frontal views on single, non-continous images with two subjects. CrowdPose [22] also belongs to the category of smaller datasets and contains crowded scenarios. This dataset also consists of non-continuous images and the crowds are in controlled environments such as group photos or sport events. Therefore, training on these datasets transfers poorly to real-world surveillance scenarios, with steep camera angles, heavy (self-)occlusion, and people in dense crowds.

Occlusion through crowd and complex poses is a challenging topic in the field of human pose estimation. A possible approach to tackle this challenge are bottom-up methods [3, 20, 6]. These methods first detect all body parts in a scenario and fuse the found keypoints to create a human pose. Since these methods detect keypoints independently from the actual person count on the image, the inference time is independent of the amount of people present. Despite these advantages, their prediction performance drops significantly in the surveillance context, as those fail to reliably create correct poses in complex human interactions, such as street fights.

An alternative approach to bottom-up are top-down methods composed of a person detector to create bounding boxes for each person, for which a pose estimator then predicts the actual pose separately. The pose estimation is done mainly by CNNs [39, 45] or transformers [46, 25]. Due to the dependency on a person detector the quality of a top-down method is linked to the person detection. Additionally, the inference time increases relatively to the person count. With recent, accurate person detectors and batch-processing, these issues are addressed and enable top-down methods as reliable approaches for HPE in surveillance context. Recently, further approaches focus on improving the detection of occluded keypoints [15] as well as directly prediction pose for crowds using GNNs [11].

2.2. Person Re-Identification

Recently, the research focus on person re-identification has shifted from developing mainly heavyweight models [5, 43, 37] to task-specific and lightweight architectures [55, 24] and generalizable methods [18, 26, 36, 56, 7, 17, 35, 38, 52]. Literature shows that optimized smaller models achieve state-of-the-art results [29, 47, 55] with increased generalization capabilities on unseen data [29, 56, 19]. In general, approaches for generalizable person re-identification are divided into two categories: Methods that learn from a single dataset [18, 26, 36, 56] and methods learning from multiple datasets [7, 17, 35, 38, 52]. Since several datasets are available that meet our requirements, namely a realistic outdoor surveillance scenario [21, 33, 53, 23, 19, 54], we rely on the second approach in this work. Most related work uses either instance [7, 17, 56] or batch [52] normalization layer to learn features that are

independent of the training domains. Other strategies in literature are the use of data augmentation techniques [38], adversarial training [49], or separate mapping networks [35]. In this work, we leverage the combination of a task-specific model with instance normalization, training on multiple datasets, and data augmentation to construct a lightweight and generalizable re-identification model.

3. Real World Challenges

(a)

(b)

(c)

Figure 2: Various challenges are present in real-world scenarios. A selection of these are: (a) Strong camera glare through reflection on mobile devices, camera distortion, train tracks interfering with human shape and sharp shadows that can lead to false positive detections, (b) people that are partially or completely obscured by environmental shadows and (c) complex, entangled and occluded poses

Human Pose Estimation in Surveillance is a challenging task. Different from laboratory conditions, a surveillance camera streams continuously days and nights through different weathers the whole year. While during summer the reflection from a cell phone may partially blend a camera as illustrated in Figure 2a, illumination within the same scene may strongly vary as shown in Figure 2b. Furthermore, due to the nature of video surveillance, the images are often heavily distorted and the cameras are mostly installed at altitude and inclined to deliver large fields of view. This camera perspective emphasizes occlusions by other objects or oneself, and sometimes even by one's own shadow, as shown in Figure 2b and 2c.

Data Acquisition and Annotation is an essential part of real-world surveillance projects. Large amounts of raw data are generated at a high rate, but due to the high cost of annotation, only a small part of this data may be used for supervised learning. Due to policies regarding personal privacy as well as logistical problems, *e.g*. persisting and storing this data, only a fraction of the available data should be authorized for storage and further processing. While active learning methods aim to select the potentially most useful data for improving deep learning models, we found no significant improvements from these methods for human pose estimation in the wild. Nevertheless, the data requires semi-automatic processing with humans in the loop to select the most promising material based on qualitative observation of the models in production at that point. For instance, pose estimation may be disturbed by strollers or specific distortions. In this work, a private dataset for Real World Surveillance Crowd Pose Estimation (RWS-CPE) is created to provide quantitative analysis from 100 frames selected out of 11 different cameras with 4,785 bounding boxes and 1,894 poses, with at least a dozen up to more than one hundred person per image. Bounding boxes and human poses are annotated interactively using model predictions as described in [9]. Poses are annotated using the Posetrack18 topology [1] for bounding boxes with more than 100px vertically, the others are annotated with bounding boxes only. The size distribution is challenging with 1,323 small ($< 32 \times 32$px), 2,834 medium (32×32px$\leq$box$\leq 96 \times 96$px) and 628 large boxes ($> 96 \times 96$px).

4. Top-Down Crowd Pose Estimation

The aim of our system is to provide HPE for a live video stream from surveillance camera in order to perform action recognition on the generated skeleton and detect violence as illustrated in Figures 2b and 2c. While skeleton-based action recognition is mostly performed with ground truth annotation in the literature, providing clean and stable skeleton tracks from surveillance footage in real-time remains extremely challenging. Large-scale video surveillance systems require flexibility and scalability in an almost plug-and-play matter when new servers and cameras are added. Therefore, our processing pipeline is compiled and delivered in a Docker container within a Kubernetes cluster to reduce administrative workload. Since authorities often have extremely limited hardware budgets, our pipeline is required to deliver acceptable results and speed on consumer hardware, *e.g*. the pipeline is required to process a camera stream on a single GPU RTX 2060 with 6GB RAM. In this section, we offer a brief overview of our system and introduce each main processing component.

4.1. Overview

Our system for live real-world HPE is designed for high flexibility. Therefore, each component is encapsulated in a separate module, which allows the visualization and configuration at each pipeline step. The video stream produced by a surveillance camera is the input of the processing pipeline.

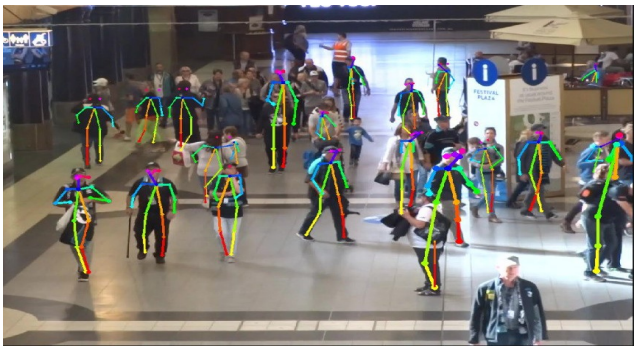

Figure 3: Examples of invalid poses return by the bottom-up approach OpenPose[3] on the MOT20 Dataset [12] which combines joints of different persons.

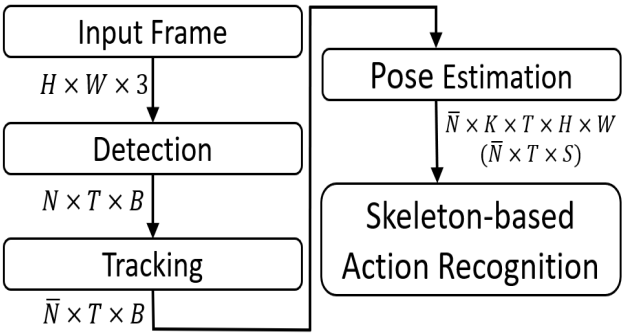

Figure 4: Overview of our real-world system for HPE. Depending on the interface of a following action recognition module, the output is either a 3D Heatmap Volume or a compressed skeleton abstraction

Unless stated otherwise, each component processes single frames. While bottom-up HPE approaches tend to scale nicely with a growing number of persons to detect, which would be beneficial to guarantee processing time, a gap in the prediction quality compared to top-down approaches exists. Moreover, they produce artifacts that strongly impair further processing for activity recognition, as illustrated in Figure 3. Therefore, we first implement a CNN person detector which generates bounding boxes for person predictions. Those are subsequently processed by a visual tracker, which creates, curates, and sorts tracks using a person re-identification module. From this step a tensor of dimensions $N \times T \times B$ is returned, where N is the number of tracks, T the temporal length, and B an abstraction of the detection parameter. This tensor serves as input to the HPE module. Here, following [13] we represent a 2D pose as a heatmap of size $K \times H \times W$, where K is the number of joints, H and W are respectively the height and width of the frame. The heatmap should be zero-padded to match the bounding box heatmap produced by the pose estimator with the size of the frame. Finally, we obtain a 3D heatmap volume of size $N \times K \times T \times H \times W$ for each track by stacking all heatmaps along the temporal dimension. Depending on the requirement of the action recognition following our system, this tensor may also be compressed with a skeleton abstraction of dimension $N \times T \times S$.

4.2. Person Detection

Person detection in real-world surveillance scenarios is a complex trade-off problem: on the one hand, the detector is expected to process frames at high speed to avoid becoming the main bottleneck of the pipeline. On the other hand, the detector is required to deliver accurate bounding box predictions that include all body parts to enable precise pose estimation. Furthermore, false positives cause cascading noise and thus should be avoided. Following [10] we choose a one-stage person detector, YOLOv4, for acceptable speed and its superior accuracy against other one-stage detectors for person detection in surveillance scenarios. Since real-world projects for intelligent surveillance usually start without annotated material, only a few frame examples are used to precise the project's specifications. However, such samples are insufficient for supervised training in the target domain.

4.3. Visual Tracking and Temporal Consistency

In top-down HPE the model is used to predict a pose for each bounding box provided by the person detector, regardless of time information. Only single frames are available in most popular datasets therefore evaluation of temporal aspects isn't possible. Thus, neither predictions from the past nor from future frames is involved in the prediction of a pose at a specific point in time t. Consequently, predictions on frame sequences are highly volatile over time, which impair the performance of action recognition systems. A common problem is the interruption of skeleton sequences due to missing detections caused by slight variations in illumination or pose. The whole pipeline attempts to fulfill real-time requirements and HPE accounts for a significant share of the computational costs. Therefore, multiple detections and incomplete tracks should be discarded to reduce computation time. To this aim, we develop a light, visual tracker that is required to perform at least as fast as the person detection component. Its main function is to track persons for at least a few seconds in order to enable sequence-based activity recognition. A few interruptions of the tracks are tolerable, as far as they retain a minimal length. Since the input stream from static surveillance comes in with at least 25 FPS, physical assumptions about the movement speed of persons can be met, *i.e.* a person will only move a few pixels between consecutive frames. We argue that a person re-identification model is both accurate and fast enough to allow sufficient tracking quality.

We apply the OSNet-x1.0 architecture with instance normalization [56] since it offers state-of-the-art performance at reasonable computational costs. The model is trained with the proposed parameter setting from [55]. During training, several data augmentation mechanisms are applied to increase the variance of training data. Concretely, images are randomly flipped, their brightness and contrast are adjusted, and rotated up to 30 degrees in both directions. The latter helps to overcome real-world challenges such as steep camera views, as it simulates different viewing angles of a person from an almost vertical viewpoint [30]. We leverage multiple complementary datasets with different characteristics to learn domain-invariant re-identification features. Market-1501 [53] and DukeMTMC-reID [33] consist of typical surveillance imagery captured by low-mounted static cameras, while PRAI-1581 [50] is a drone-based dataset and thus represents steep views from overview cameras. We realize the tracking by extracting l2-normalized feature vectors for all persons present in the video frame. Afterward, the Euclidean distances to feature vectors from people occurring in preceding frames (gallery) are computed. When the distance is below an adjustable threshold, the person ID of the gallery track is assigned to the matching detection. Otherwise, a new track is created.

Our module saves the feature directly in memory to provide the best speed result. Since the module runs continuously and needs to fulfill legal requirements, we implement a ring buffer and only save features from a few preceding frames. This gallery size is empirically set to five, which proved sufficient for solid re-identification performance.

For temporal consistency a tensor $\tilde{N} \times T \times B$ is constructed from the $\tilde{N}$ tracks available. If a minimum of 60% detections for the track in chosen time length are missing, we argue that completing those gaps is misleading and thus the track won't be forwarded. Given a gap size G, one or multiple gaps with size smaller than G will be filled either by copy from the last frame or linear interpolation, preventing gaps in pose estimation. Empirical studies on qualitative results indicate that pose estimation on bounding boxes with less than 100 px in vertical is highly uncertain and produces severe artifacts over time. Therefore tracks outside a user chosen size bounds are also filtered out, leading to a final tensor of size $N \times T \times B$ where $N \leq \tilde{N}$ which is sorted per track ID.

4.4. Crowd Pose Estimation

While top-down HPE is more accurate for surveillance scenarios, its speed highly depends on the crowdedness of the scene. This dependency is due to the general process of top-down approaches. For each bounding box the corresponding image patch is taken as input for a backbone, which creates low-resolution features. These features serve as a bottleneck from which an upsampling process creates a heatmap for each keypoint. These heatmaps can be used directly, to take advantage of surrounding information around the keypoint, or the coordinates of each keypoint can be extracted by the position with maximum value. The resulting keypoint coordinates are then transformed from the patch to the actual input image dimension. The inference time of such a heavy model is high, and therefore the computation time increases with each detection for complex scenes. As mentioned in the last section, avoiding the processing of invalid tracks and filtering too small detections is the first step to improve runtime. A second step is to define regions of interest for which skeletons are required. Depending on a camera's field of view, the placement of the camera intends to analyze person behavior in spatially well-defined areas. Reducing the resolution of the input boxes may improve the runtime of the HPE components. However, the speed benefit is small compared to the deterioration concerning accuracy. The same applies to the use of smaller models such as MobileNet [34] or Lite-HRNet [48], but we observe that batch processing at inference and reducing precision to FP16 improves the overall speed without significant impact regarding accuracy. Since the newly introduced transformers reduce the number of parameters and computational load while improving results, we train a hybrid transformer model with truncated CNN backbone Transpose-R and Transpose-H. Furthermore, extracting the keypoints from the heatmaps or similar post-processing steps to get well-formed tensors can be done in parallel during the inference of the next frame.

5. Results

We conducted our experiments using the Person Re-identification in the Wild (PRW) dataset [54] which was captured by six surveillance cameras on a campus. PRW is based on the same footage from low-mounted cameras as the well-known Market-1501 [53] dataset. However, unlike Market-1501, it contains entire video frames and not cropped bounding boxes, which enables the evaluation of person detection methods. A total of 11,816 video frames are included, sampled every second from the original videos. These frames contain 43,110 manually annotated bounding boxes. We use the PANDA dataset [42] to improve the generalization of our detector, which is a gigapixel surveillance video dataset and contains scenes with a large field of view and large crowds. A total of 555 static giga-pixel images (390 for training, 165 for testing) are offered, *i.e.* images with at least $26,753 \times 15,052$px, however annotations for testing are not public. Each scene shows in average over 205 persons per image ranging from in average 52 up to 571 persons. Finally, we also evaluate on the RWS-CPE validation dataset introduced in Section 3, which contains 100 frames in full HD, with 4,785 Bounding boxes and 1,894 poses manually annotated.

5.1. Detection

We evaluate and compare several person detection models on the scientific PRW and the real-world RWS-CPE dataset in Table 1. Except for the YOLOv4-Panda, which is fine-tuned on the PANDA dataset, the models were trained using the COCO [28] dataset. Results indicate that YOLO architectures achieve beneficial performance on both datasets. On real-world data, the gap between the model architectures increases. Moreover, while differences regarding AP between COCO and PANDA trained models are negligible on the academic PRW dataset, the YOLOv4-Panda outperforms the other models on RWS-CPE. The PANDA dataset is more suited for transfer learning in real-world surveillance scenarios since it contains many tiny persons and crowded scenes. Unlike the other models, it is able to detect some of the small people, even if the AP^S is very small. However, the PANDA dataset mostly contains small persons and thus the AP^L is worse compared to other YOLO models. This finding highlights the necessity to select meaningful training data which show similar characteristics to the target domain and represent the huge diversity of real-world surveillance data. Nevertheless, there is still much potential for improvement, and hence training with target domain data seems inevitable at that point. With respect to inference time, the use of small YOLO models is advantageous. For instance, the YOLOv4-tiny is able to process up to 262 FPS on our hardware setup. The RetinaNet with the transformer backbone PVT v2 b0 [41] achieves processing times similar to our YOLOv4-Panda but worse AP and AR scores.

5.2. Visual Tracking and Temporal Consistency

We conducted experiments on several cameras, especially in front of a central train station where several thousand persons walk through the area covered by a static camera within a few hours. To improve speed, we fix the gallery size to five, use FP16 precision for the person re-identification model and resize detections to 64×128 px without noticing significant degradation of performance. The whole module provides visual tracking at a stable rate of 10 FPS, even when the scene is extremely crowded with around 150-200 pedestrians. In other terms, the re-identification model processes more than 100 detections per second. We attempted to further improve speed by batching inputs, however this impaired speed for less crowded scenes.

We evaluate the module qualitatively on an anonymized central station scene as illustrated in Figure 5. The results show robust tracking for the foreground of the scene. With growing distance to the camera, smaller bounding boxes, and more occlusions we see more identity switches or fragmented tracks. Furthermore, tracking allows more stable tracks and thus fewer are aborted.

5.3. Pose Estimation

In Table 2, we evaluate and compare several HPE models on the real-world RWS-CPE dataset, which contains 1,894 poses, with ground truth bounding boxes and with detections provided by the YOLOv4-Panda model described in Section 5.1. We select a HPE model with a MobileNetv2 [34] backbone in order to increase inference speed. We compare its inference speed with a HRNet-W48 backbone model. Furthermore, we train a Transpose hybrid-transformer model [46] with both a ResNet-S and a HrNet-S backbone and compare it with the popular ResNet-50 and HRNet-W32 backbone models trained by MMPose [8]. We pretrain the models on the COCO Human Pose Estimation task and fine-tune on the Posetrack18 dataset for 30 epochs, both with GT detections.

On our consumer machine all models perform relatively slow. MobileNetv2 and ResNet-50 achieve 5 and 7 FPS with by far the worst results regarding AP (0.32 and 0.41). In comparison, HRNet-W32 and HRNet-W48 models, which achieve state-of-the art results on the Posetrack18 dataset, perform much slower with 3 and 1 FPS, respectively, but achieve the best results with 0.52 and 0.53 AP. The transformer models Transpose-R and Transpose-H, which are competitive on Posetrack18, outperform the other models with 0.60 and 0.59 AP using ground truth boxes but achieve worse results when inaccurate bounding box predictions are used. In terms of processing speed, the transformer models achieves 1 FPS, respectively. The observations indicate that the loss of precision due to the domain gap is worse for full CNN models compared to hybrid transformers. However, the performance drops on predicted detections with 0.49 AP; these models show a greater loss of 0.10 and 0.11 AP between GT Boxes and predicted boxes. It remains unclear if this is due to greater sensibility of these models or failed detections.

5.4. Failure cases

We further evaluate the Transpose-R model qualitatively in Figure 1. The skeletons returned by the model for clearly visible persons without occlusion seem realistic, and the predictions for head and shoulder are particularly accurate. Misleading detections which include multiple person, strollers or bicycles disturb the model. Keypoints are detected on the bicycle or body parts of another person with higher confidence than on the main person represented in the bounding box. Furthermore, the model struggles to differentiate left and right. In addition, we compare Transpose-R with Transpose-H model on challenging poses in Figure 6. For a failed detection which includes two actors, both models fail to predict joints for only one subject consequently and predict keypoints on both. Depending on the angle, unusual poses such as falling or lying remain challenging. In each case, Transpose-R seems to produce more

Image Size	Model	FPS	PRW							RWS-CPE						
			AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}	AR	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}	AR
416×416	YOLOv3	53	0.40	0.73	0.40	–	0.29	0.50	0.46	0.17	0.36	0.13	0.01	0.17	0.46	0.19
	YOLOv4-tiny	**262**	0.33	0.59	0.35	–	0.18	0.47	0.38	0.08	0.15	0.07	0.00	0.04	0.36	0.09
608×608	YOLOv4	26	0.48	0.79	0.53	–	0.34	0.60	0.55	0.23	0.40	0.24	0.01	0.24	0.60	0.25
640×640	YOLOv4-csp-x-swish	16	0.45	0.73	0.50	–	0.32	0.57	0.52	0.22	0.35	0.24	0.01	0.23	0.60	0.24
896×896	YOLOv4-p5	9	0.47	0.76	0.51	–	0.34	0.58	0.53	0.21	0.34	0.24	0.02	0.22	0.57	0.23
960×544	YOLOv4-Panda	19	0.47	**0.86**	0.47	–	**0.37**	0.56	**0.57**	**0.28**	**0.56**	0.26	**0.06**	**0.32**	0.55	**0.33**
1280×1280	YOLOv4-p6	5	0.48	0.79	0.52	–	0.35	0.59	0.56	0.26	0.42	**0.28**	0.03	0.29	**0.61**	0.29
960×544	RetinaNet PVT v1 Medium	9	0.45	0.74	0.50	–	0.30	0.59	0.52	0.17	0.30	0.17	0.00	0.16	0.51	0.19
	RetinaNet PVT v2 b0	19	0.47	0.74	0.50	–	0.30	0.59	0.56	0.17	0.33	0.17	0.00	0.17	0.52	0.20
	RetinaNet PVT v2 b1	14	**0.49**	0.79	**0.54**	–	0.34	**0.61**	0.55	0.19	0.35	0.19	0.00	0.20	0.55	0.22

Table 1: Quantitative results on the public PRW Dataset and the private RWS-CPE Dataset. Models are trained on COCO Dataset [28], with exception of YOLOv4-Panda which is fine-tuned on Panda [42]. Pre-trained RetinaNet [27] models are taken from MMDetection [4]. Average Precision (AP) and Average Recall (AR) follow the COCO evaluation. The FPS is measured for the models only, not the complete detection module.

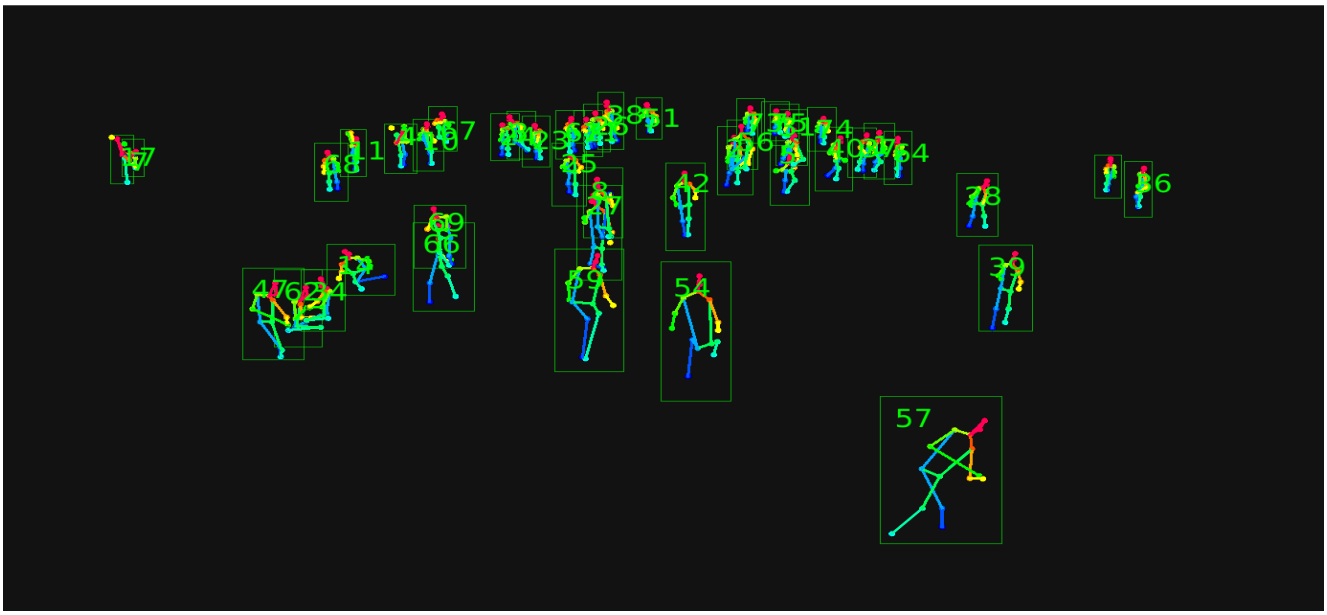

Figure 5: Anonymized results of a central station scene. Re-identification shows stable track results in the foreground. Background tracks appear more unstable due to smaller boxes and increased occlusion. Temporal consistency partially stabilized background tracks.

probable predictions and less outliers than its counterpart.

5.5. Discussion

Our results with HPE show promising results for real-world surveillance scenarios. However, since the models are trained on other domains than surveillance, several challenges remain. The YOLOv4-Panda detector, which was trained on a domain similar to the target domain, provides overall better results in a quantitative matter. Nevertheless, its performance is particularly limited for larger persons (AP^{L}) whose poses are required for action recognition systems. Furthermore, such systems are required to detect violent situations with associated movements and poses. However, poses such as punching, fighting, or lying are poorly represented in training datasets and either missed by the detector or too ambiguous for pose estimation. Consequently, in order to improve HPE in real-world surveillance, a large annotated dataset representative of the target domains and scenarios is required.

6. Conclusions

In this work, we have presented the current state-of-research on HPE in real-world surveillance. After pointing out the challenges posed by uncontrollable environments and data acquisition and annotation, we proposed a processing pipeline consisting of person detection, tracking, and subsequent pose estimation. Extensive evaluation elucidates the huge domain gap between scientific and real-

Input size	Method	FPS	Posetrack18	RWS-CPE						
			AP	gtbbox AP	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
384×288	MobileNetv2 [34]	5	0.78	0.34	0.32	0.58	0.33	0.28	0.42	0.37
	HRNet-W48 [39]	1	**0.86**	0.57	**0.53**	**0.73**	**0.60**	**0.49**	0.64	0.57
256×192	ResNet-50 [16]	**7**	0.81	0.43	0.41	0.62	0.45	0.36	0.53	0.46
	HRNet-W32 [39]	3	0.83	0.56	0.52	0.72	0.58	0.48	0.62	**0.61**
256×192	Transpose-R [46]	5	0.83	**0.60**	0.49	0.68	0.55	0.44	**0.66**	0.59
	Transpose-H [46]	1	0.82	0.59	0.44	0.69	0.54	0.44	0.65	0.58

Table 2: Comparisons on the RWS-CPE dataset provided with the same detected human boxes. Pre-training is done on the COCO Human Pose Estimation task with GT boxes. The training is done on the Posetrack18 dataset [1] with GT boxes. The transformer models TransPose [46] report competitive results with HRNet and ResNet listed for comparison for better runtimes and seems to better generalize. The versions of HRNet and ResNet are taken from MMPose [8].

Figure 6: A street fight containing multiple examples of complex human interactions for pose estimation: occlusion by interacting person, compact and entangled pose, crawling, self-occlusion and lying on the ground. The first row shows the input, second row and the last row contains poses from Transpose-H and Transpose-R respectively.

world data and problems with the generalization of state-of-the-art methods. To mitigate this, annotated data from real-world surveillance systems is currently inevitable but costly. Nevertheless, we achieve promising results with our system that runs with 7-10 FPS and leverages a lightweight tracking component to enhance the temporal consistency of poses. Further research should focus on interactive methods to annotate real-world data and on semi-supervised learning methods to reduce the effort required. In addition, the temporal consistency may be improved by keypoint outlier detection. Within a track, a keypoint is expected to be detected near a certain region near which is consistent in time along the track for this keypoints. A prediction outside this region would therefore be defined as an outlier and could be removed or corrected using temporal information.

Acknowledgment

This work was supported by the Mannheim Police Headquarters. Together with the state of Baden-Württemberg and the Mannheim Police Headquarters an intelligent vision-based activity recognition will be tested and further developed in a model project in Mannheim until 2023.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt

Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018.

[2] 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)*, June 2014.

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[5] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8361, 2019.

[6] 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.

[7] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2021.

[8] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

[9] Mickael Cormier, Fabian Röpke, Thomas Golda, and Jürgen Beyerer. Interactive labeling for human pose estimation in surveillance videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1649–1658, October 2021.

[10] Mickael Cormier, Stefan Wolf, Lars Sommer, Arne Schumann, and Jürgen Beyerer. Fast pedestrian detection for real-world crowded scenarios on embedded gpu. In *IEEE EUROCON 2021-19th International Conference on Smart Technologies*, pages 40–44. IEEE, 2021.

[11] 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*, volume 35, pages 1193–1200, 2021.

[12] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003[cs]*, Mar. 2020. arXiv: 2003.09003.

[13] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. *arXiv preprint arXiv:2104.13586*, 2021.

[14] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-Track: Efficient Pose Estimation in Videos. In *CVPR*, 2018.

[15] 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.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. *arXiv preprint arXiv:1905.03422*, 2019.

[18] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2020.

[19] Philipp Kohl, Andreas Specker, Arne Schumann, and Jurgen Beyerer. The mta dataset for multi-target multi-camera pedestrian tracking by weighted distance aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1042–1043, 2020.

[20] 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.

[21] SV Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, BS Harish, and Hugo Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16:1696–1708, 2020.

[22] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019.

[23] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[24] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

[25] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. *arXiv preprint arXiv:2104.03516*, 2021.

[26] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution

and temporal lifting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 456–474. Springer, 2020.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[29] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[30] Lennart Moritz, Andreas Specker, and Arne Schumann. A study of person re-identification design characteristics for aerial data. In *Pattern Recognition and Tracking XXXII*, volume 11735, page 117350P. International Society for Optics and Photonics, 2021.

[31] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *European Conference on Computer Vision*, pages 488–504. Springer, 2020.

[32] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

[33] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016.

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[35] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 719–728, 2019.

[36] Vladislav Sovrasov and Dmitry Sidnev. Building computationally efficient and well-generalizing person re-identification models with metric learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 639–646. IEEE, 2021.

[37] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018.

[38] Masato Tamura and Tomokazu Murakami. Augmented hard example mining for generalizable person re-identification. *arXiv preprint arXiv:1910.05280*, 2019.

[39] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[40] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11088–11096, 2020.

[41] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021.

[42] 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.

[43] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.

[44] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[45] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[46] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021.

[47] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[48] 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.

[49] Qindong Zhang, Sanping Zhou, and Jinjun Wang. Learning generic feature representations with adversarial regularization for person re-identification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2358–2362. IEEE, 2021.

[50] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2020.

[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.

[52] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6277–6286, 2021.

[53] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[54] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017.

[55] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019.

[56] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[57] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. A comprehensive study of deep video action recognition, 2020.