Tracking Pedestrian Heads in Dense Crowd
Ramana Sundararaman, Cédric de Ameida Braga, Eric Marchand, Julien Pettré

To cite this version:
Ramana Sundararaman, Cédric de Ameida Braga, Eric Marchand, Julien Pettré. Tracking Pedestrian Heads in Dense Crowd. CVPR 2021 - IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 2021, Virtual, United States. pp.1-11. hal-03184673

HAL Id: hal-03184673
https://inria.hal.science/hal-03184673
Submitted on 29 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Tracking Pedestrian Heads in Dense Crowd

Ramana Sundararaman  Cédric De Almeida Braga  Eric Marchand  Julien Pettré
Univ Rennes, Inria, CNRS, Irisa, Rennes, France
ramanasubramanyam.sundararaman@polytechnique.edu,
{cedric.de-almeida-braga,julien.pettre}@inria.fr,
eric.marchand@irisa.fr

Abstract

Tracking humans in crowded video sequences is an important constituent of visual scene understanding. Increasing crowd density challenges visibility of humans, limiting the scalability of existing pedestrian trackers to higher crowd densities. For that reason, we propose to revitalize head tracking with Crowd of Heads Dataset (CroHD), consisting of 9 sequences of 11,463 frames with over 2,276,838 heads and 5,230 tracks annotated in diverse scenes. For evaluation, we proposed a new metric, IDEucl, to measure an algorithm’s efficacy in preserving a unique identity for the longest stretch in image coordinate space, thus building a correspondence between pedestrian crowd motion and the performance of a tracking algorithm. Moreover, we also propose a new head detector, HeadHunter, which is designed for small head detection in crowded scenes. We extend HeadHunter with a Particle Filter and a color histogram based re-identification module for head tracking. To establish this as a strong baseline, we compare our tracker with existing state-of-the-art pedestrian trackers on CroHD and demonstrate superiority, especially in identity preserving tracking metrics. With a light-weight head detector and a tracker which is efficient at identity preservation, we believe our contributions will serve useful in advancement of pedestrian tracking in dense crowds. We make our dataset, code and models publicly available at https://project.inria.fr/crowdscience/project/dense-crowd-head-tracking/.

1. Introduction

Tracking multiple objects, especially humans, is a central problem in visual scene understanding. The intricacy of this task grows with increasing targets to be tracked and remains an open area of research. Alike other subfields in Computer Vision, with the advent of Deep Learning, the task of Multiple Object Tracking (MOT) has remarkably advanced its benchmarks [12, 24, 25, 37, 42, 61] since its inception [21]. In the recent past, the focus of MOT Challenge benchmark [13] has shifted towards tracking pedestrians in crowds of higher density. This has several applications in fields such as activity recognition, anomaly detection, robot navigation, visual surveillance, safety planning etc. Yet, the performances of trackers on these benchmark suggests a trend of saturation\(^1\). Majority of online tracking algorithms today follow the tracking-by-detection paradigm and several research works have well-established object detector’s performance to be crucial in tracker’s performance [3, 5, 11]. As the pedestrian density in a scene increases, pedestrian visibility reduces with increasing mutual occlusions, leading to reduced pedestrian detection as visualized in Figure 1. To tackle these challenges yet track humans efficiently in densely crowded environments, we rekindle the task of MOT with tracking humans by their distinctly visible part - heads. To that end, we propose a new dataset, CroHD, Crowd of Heads Dataset, comprising 9 sequences of 11,463 frames with head bounding boxes annotated for tracking. We hope that this new dataset opens

\(^1\)https://motchallenge.net/results/MOT20/
up opportunities for promising future research to better understand global pedestrian motion in dense crowds.

Supplementing this, we develop two new baseline methods on CroHD, a head detector, HeadHunter and a head tracker, HeadHunter-T. We design HeadHunter peculiar for head detection in crowded environments, distinct from standard pedestrian detectors and demonstrate state-of-the-art performance on an existing head detection dataset. HeadHunter-T extends HeadHunter with a Particle Filter framework and a light-weight re-identification module for head-tracking. To validate HeadHunter-T to be a strong baseline tracker, we compare it with three published top performing pedestrian trackers on the crowded MOTChallenge benchmark, evaluated on CroHD. We further perform comparisons between tracking by head detection and tracking by body detection to illustrate the value of our contribution.

To establish correspondence between a tracking algorithm and pedestrian motion, it is necessary to understand the adequacy of various trackers in successfully representing ground truth pedestrian trajectories. We thus propose a new metric, IDEucl to evaluate tracking algorithms based on their consistency in maintaining the same identity for the longest length of a ground truth trajectory in the image coordinate space. IDEucl is compatible with our dataset and can be extended to any tracking benchmark, recorded with a static camera.

In summary, this paper makes the following contributions (i) We present a new dataset, CroHD, with annotated pedestrian heads for tracking in dense crowd, (ii) We propose a baseline head detector for CroHD, HeadHunter, (iii) We develop HeadHunter-T, by extending HeadHunter as the baseline head tracker for CroHD, (iv) We propose a new metric, IDEucl, to evaluate the efficiency of trackers in representing a ground truth trajectory and finally, (v) We demonstrate HeadHunter-T to be a strong baseline by comparing with three existing state-of-the-art trackers on CroHD.

2. Related Work

Head Detection Benchmarks: The earliest benchmarks in head detection are [29, 46, 62, 64], which provide ground truth head annotations of subjects in Hollywood movies. In the recent past, SCUT-Head [47] and CrowdHuman dataset [52] provide head annotations of humans in crowded scenes. Head detection is also of significant interest in the crowd counting and analysis literature [32]. Rodriguez et al. [50] introduced the idea of tracking by head detection with their dataset consisting of roughly 2200 head annotations. In the recent years, there has been a surge in research works attempting to narrow the gap between detection and crowd counting [39, 51, 41, 69] which attempts to hallucinate pseudo head ground truth bounding boxes in crowded scenes.

Head Detection Methods: Fundamentally, the task of head detection is a combination of multi-scale and contextual object detection problem. Objects at multiple scales are detected based on image pyramids [30, 47, 55, 56, 66] or feature pyramids [26, 38, 70]. The former is computationally intensive task requiring multiple forward passes of images while the latter generates multiple pyramids in a single forward pass. Contextual object detection has been widely addressed in the literature of face detection, such as [14, 43, 60] who show improved detection accuracy by using convolutional filters of larger receptive size to model context. Sun et al. [58] employ such a contextual and scale-invariant applied to head detection.

Tracking Benchmarks and Metrics: The task of Multi-Object Tracking (MOT) is to track an initially unknown number of targets in a video sequence. The first MOT dataset for tracking humans were the PETS dataset [21], soon followed by [1, 16, 24, 25]. Standardization of MOT benchmarks were later proposed in [37] and since then, it has been updated with yearly challenges with increasingly crowded scenes [13, 42]. Recently, the TAO dataset [12] was introduced for Multi-object tracking, which focuses on tracking 833 object categories across 2907 short sequences. Our dataset pushes the challenge of tracking in crowded environments with pedestrian density reaching 346 humans per frame. Other relevant pedestrian tracking dataset include [8, 9, 61].

To evaluate algorithms on MOTChallenge dataset, classical MOT metrics [63] and CLEAR MOT metrics [4] have been de facto established as standardised way of quantifying performances. The CLEAR Metric proposes two important scores MOTA and MOTP which concisely summarise the classical metrics based on cumulative per frame accuracy and precision of bounding boxes respectively. Recently, Ristani et al. [49] propose the ID metric, which rewards a tracker based on its efficiency in preserving an identity for the longest duration of the Ground Truth trajectory.

Tracking Algorithms: Online Multi-object tracking algorithms can be summarised into: (i) Detection, (ii) Motion Prediction, (iii) Affinity Computation and, (iv) Association steps. R-CNN based networks have been common choice for the detection stage due to the innate advantage of proposal based detectors over Single-Stage detection methods [31]. Amongst online Multiple Object Tracking algorithm, Chen et al. [10] use Particle Filter framework and weigh the importance of each particle by their appearance classification score, computed by a separate network, trained independently. Earlier works such as [7, 33] use Sequential Importance Sampling (SIS) with Constant Velocity Assumption for assigning importance weight to particles. Henschel et al. [28] demonstrated the the limitation of single object detector for tracking and used a head detector [57] in tandem with the pedestrian detector [48]. However, in the recent past, research works in MOT have attempted to bridge
Scene 1 | Scene 2 | Scene 3 | Scene 4 | Scene 5
--- | --- | --- | --- | ---

Figure 2. Depiction of a frame per each scene from our Crowd of Heads Dataset, CroHD. The top row shows frames from the training set while the bottom row illustrates frames from the test set. Scene 5 is kept exclusive for the test set.

the gap between tracking and detection through a unified framework [3, 18, 19, 35, 40, 61]. Most notable amongst them is Tracktor [3], who demonstrated that an object detector alone is sufficient to predict locations of targets in subsequent frames, benefitting from the high-frame rates in video.

3. CroHD Dataset

**Description:** The objective of CroHD is to provide tracking annotation of pedestrian heads in densely populated video sequences. To the best of our knowledge, no such benchmark exists in the community and hence we annotated 2,276,838 human heads in 11,463 frames across 9 sequences of Full-HD resolution. We built CroHD upon 5 sequences from the publicly available MOTChallenge CVPR19 benchmark [13] to enable performance comparison of trackers in the same scene between two paradigms - head tracking and pedestrian tracking. We maintain the training set and test set classification of the aforementioned sequences to be the same in CroHD as the MOTChallenge CVPR19 benchmark. We further annotated 4 new sequences of higher crowd densities in two new scenarios. The new scenario centers on the Shibuya Train station and Shibuya Crossing, one of the busiest pedestrian crossings in the world. All sequences in CroHD have a frame-rate of 25 fps and are captured from an elevated viewpoint. The sequences involve crowded indoor and outdoor scenes, recorded across different lighting and environmental conditions. This ensures sufficient diversity in the dataset in order to make it viable for training and evaluating the comprehensiveness of modern Deep Learning based techniques. The maximum pedestrian density reaches approximately 346 persons per frame while the average pedestrian density across the dataset is 178. A detailed sequence-wise summary of CroHD is given in Table 1. We split CroHD into 4 sequences of 5740 frames for training and 5 sequences of 5723 frames for testing. They share three scenes in common, while the fourth scene is disparate to ensure generalization of trackers on this dataset. A representative frame from each sequence of CroHD and their respective training, testing splits are depicted in Figure 2. We will make our sequences and training set annotations publicly available. To preserve the fairness of the MOTChallenge CVPR19 benchmark, we will not release the annotations corresponding to the test set.

**Annotation:** The annotation and data format of CroHD follows the standard guidelines outlined by MOTChallenge benchmark [13, 42]. We annotated all visible heads of humans in a scene with the visibility left to the best of discretion of annotators. Heads of all humans, whose shoulder is visible were annotated, including the heads occluded by head coverings such as hood, caps etc. For sequences inherited from MOTChallenge CVPR19 benchmark, the annotations were performed independent of pedestrian tracking ground truth in order to have no dependencies between the two modalities. Due to the high frame rate in our video sequences, we interpolate annotations in between keyframes and adjust a track only when necessary.

CroHD constitutes four classes - Pedestrian, Person on Vehicle, Static and Ignore. Heads of statues or human faces on clothing have been annotated with an ignore label. Heads of pedestrians on vehicles, wheelchairs or baby transport have been annotated as Person on Vehicle. Pedestrians who do not move throughout the sequence are classified as static persons. Unlike the case of standard MOTChallenge benchmarks, we observe that overlap between bounding boxes are minimal since head bounding boxes from an elevated viewpoint are almost distinct. Hence, we limit our visibility flag to be binary - either visible (1.0) or occluded (0.0). We consider a proposal to be a match if the Intersection Over Union (IoU) with the ground truth is larger than 0.4.
Tracker B and COCO respectively. mAP for better understanding of global crowd motion in dense crowds, for a tracker to quantitatively distinguish between trackers which are more effective in tracking a larger portion of ground truth. It is important to know if a tracker is able to preserve an identity. A change of color implies an identity switch with both trackers registering 3 switches. Following existing metrics, Tracker A and Tracker B are considered equally efficient. They neither highlight the ineffectiveness of Tracker A nor the ability of Tracker B in covering an adequate portion of ground truth trajectory with consistent identity. Therefore, IDEucl is more appropriate for judging the quality of the estimated pedestrian motion.

Thus, to formulate this metric, we perform a global hypothesis to ground truth matching by constructing a Bipartite Graph \( G = (\mathcal{U}, \mathcal{V}, \mathcal{E}) \), similar to \([49]\). Two “regular” nodes are connected by an edge \( e \) if they overlap in time, with the overlap defined by \( \Delta_t \),

\[
\Delta_{t,t-1} = \begin{cases} 
1, & \text{if } \delta > 0.5 \\
0, & \text{otherwise}
\end{cases}
\]

Considering \( \tau_t, h_t \) to be an arbitrary ground truth and hypothesis track at time \( t \), \( \delta \) is defined as,

\[
\delta = \text{IoU}(\tau_t, h_t)
\]

The cost on each edge \( \mathcal{E} \in \mathbb{R}^N \) of this graph, \( \mathcal{M} \in \mathbb{R}^{N-1} \) is represented as the distance in image space between two successive temporal associations of “regular” node. In particular, cost of an edge is defined as

\[
\mathcal{M} = \sum_{t=1}^{N} m_t = \begin{cases} 
d(\tau_t, \tau_{t-1}), & \text{if } \Delta_{t,t-1} = 1 \\
0, & \text{otherwise}
\end{cases}
\]

where \( d \) denotes the Euclidean distance in image coordinate space. A ground truth trajectory is assigned a unique hypothesis which maintains a consistent identity for the predominant distance of ground truth in image coordinate space. We employ the Hungarian algorithm to solve

<table>
<thead>
<tr>
<th>Name</th>
<th>Frames</th>
<th>Scenario</th>
<th>Tracks</th>
<th>Boxes</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>CroHD-01</td>
<td>429</td>
<td>Indoor</td>
<td>85</td>
<td>21,456</td>
<td>50.0</td>
</tr>
<tr>
<td>CroHD-02</td>
<td>3,315</td>
<td>Outdoor, night</td>
<td>1,276</td>
<td>733,622</td>
<td>222.0</td>
</tr>
<tr>
<td>CroHD-03</td>
<td>1,000</td>
<td>Outdoor, day</td>
<td>811</td>
<td>258,012</td>
<td>258.0</td>
</tr>
<tr>
<td>CroHD-04</td>
<td>997</td>
<td>Indoor</td>
<td>580</td>
<td>175,703</td>
<td>176.2</td>
</tr>
<tr>
<td>CroHD-11</td>
<td>584</td>
<td>Indoor</td>
<td>133</td>
<td>38,492</td>
<td>65.8</td>
</tr>
<tr>
<td>CroHD-12</td>
<td>2,080</td>
<td>Outdoor, night</td>
<td>737</td>
<td>383,677</td>
<td>185.0</td>
</tr>
<tr>
<td>CroHD-13</td>
<td>1,050</td>
<td>Outdoor, day</td>
<td>725</td>
<td>257,828</td>
<td>258.0</td>
</tr>
<tr>
<td>CroHD-14</td>
<td>1,050</td>
<td>Outdoor, day</td>
<td>562</td>
<td>258,227</td>
<td>246.0</td>
</tr>
<tr>
<td>CroHD-15</td>
<td>1,008</td>
<td>Outdoor, day</td>
<td>321</td>
<td>149,821</td>
<td>149.0</td>
</tr>
<tr>
<td>Total</td>
<td>11,463</td>
<td></td>
<td>5,230</td>
<td>2,276,838</td>
<td>178</td>
</tr>
</tbody>
</table>

Table 1. Sequence-wise statistics CroHD. Sequences are named CroHD-XY, with X being either 0 or 1 depending on training set or testing set respectively. Y denotes the serial number of videos.

Figure 3. Identity prediction of two trackers - Tracker A (3a) and Tracker B (3b) for the same ground truth. A change of color implies an identity switch with both trackers registering 3 switches.
this maximum weight matching problem to obtain the best (longest) hypothesis. Once we obtain an optimal hypothesis, we formulate the metric $C$ as the ratio of length of ground truth in image coordinates covered by the best hypothesis,

$$C = \frac{\sum_{i=1}^{K} M_i}{\sum_{i=1}^{K} T_i}$$

Note that this formulation of cost function naturally weighs the significance of each ground truth track based on its distance in image coordinate space.

5. Methods: Head Detection and Tracking

In this section, we elucidate the design and working of HeadHunter and HeadHunter-T.

5.1. HeadHunter

As detection is the pivotal step in object tracking, we designed HeadHunter differently from traditional object detectors [20, 48, 65] by taking into account the nature and size of objects we detect. HeadHunter is an end-to-end two stage detector, with three functional characteristics. First, it extracts feature at multiple scales using Feature Pyramid Network (FPN) [38] using a Resnet-50 [27] backbone. Images of heads are homogeneous in appearance and often, in crowded scenes, resemble extraneous objects (typically background). For that reason, inspired by the head detection literature, we augmented on top of each individual FPNs, a Context-sensitive Prediction Module (CPM) [60]. This contextual module consists of 4 Inception-ResNet-A blocks [59] with 128 and 256 filters for $3 \times 3$ convolution and 1024 filters for $1 \times 1$ convolution. As detecting pedestrian heads in crowded scenes is a problem of detecting many small-sized adjacently placed objects, we used Transpose Convolution on features across all pyramid levels to upscale the spatial resolution of each feature map. Finally, we used a Faster-RCNN head with Region Proposal Network (RPN) generating object proposals while the regression and classification head, each providing location offsets and confidence scores respectively. The architecture of our proposed network is summarised in Figure 4.

5.2. HeadHunter-T

We extended HeadHunter with two motion models and a color histogram based re-identification module for head-tracking. Our motion models consist of Particle Filter to predict motion of targets and Enhanced Correlation Coefficient Maximization [17] to compensate the Camera motion in the sequence. A Particle Filter is a Sequential Monte Carlo (SMC) process, which recursively estimates the state of dynamic systems. In our implementation, we represent the posterior density function by a set of bounding box proposals for each target, referred to as particles. The use of Particle Filter enables us to simultaneously model non-linearity in motion occurring due to rapid movements of heads and pedestrian displacement across frames.

Notation: Given a video sequence $I$, we denote the ordered set of frames in it as $\{I_0, \cdots, I_{T-1}\}$, where $T$ is the total number of frames in the sequence. Throughout the paper, we use subscript notation to represent time instance in a video sequence. In a frame $I_t$ at time $t$, the active tracks are denoted by $T_t = \{b_t^1, b_t^2, \ldots, b_t^N\}$, where $b_t^k$ refers to bounding box of the $k^{th}$ active track, denoted as $b^k_t = (x^k_t, y^k_t, w^k_t, h^k_t)$. At time $t$, the $i^{th}$ particle corresponding to $k^{th}$ track is denoted by $p_{t,i}^k$ and its respective importance weight by $w_{t,i}^k$. $L_t$ and $N_t$ denote the set of inactive tracks and newly initialized tracks respectively.

Particle Initialization: New tracks are initialized at the start of the sequence, $I_0$ from the detection provided by HeadHunter and at frame $I_t$ for detection(s) which cannot be associated with an existing track. A plausible association of new detection with existing track is resolved by Non-Maximal Suppression (NMS). The importance weights of each particle are set to be equal at the time of initialisation. Each particles represent 4 dimensional state space, with the state of each targets modelled as $(x, y, c, h, z, \hat{c}, \hat{h})$, where, $(x, y, c, h, z)$ denote the centroids, width and the height of bounding boxes.

Prediction and Update: At time $t > 0$, we perform RoI pooling on the current frame’s feature map, $F_t$, with the bounding box of particles corresponding to active tracks. Each particles’ location in the current frame is then adjusted using the regression head of HeadHunter, given their location in the previous frame. The importance weights of each particle are set to the respective foreground classification score from the classification head of HeadHunter. Our prediction step is similar to the Tracktor [3], applied to particles instead of tracks. Given the new location and importance weight of each particle, estimated position of $k^{th}$ track is computed as weighted mean of the particles,

$$S_t^k = \frac{1}{M} \sum_{i=1}^{M} w_{t,i}^k p_{t,i}^k \tag{5}$$

Resampling: Particle Filtering frameworks are known to suffer from degeneracy problems [2] and as a result we resample to replace particles of low importance weight. $M$ particles corresponding to $k^{th}$ track are re-sampled when the number of particles which meaningfully contributes to probability distribution of location of each head, $N_{eff}^k$ exceeds a threshold, where,

$$N_{eff}^k = \frac{1}{\sum_{i=1}^{M} (w_{t,i}^k)^2} \tag{6}$$
Cost Matching: Tracks are set to inactive when scores of their estimated state \( S_t \) falls below a threshold, \( \lambda_{\text{reg}}^{\text{nms}} \). Positions of such tracks are predicted following Constant Velocity Assumption (CVA) and their tracking is resumed if it has a convincing similarity with a newly detected track. The similarity, \( C \) is defined as

\[
C = \alpha \cdot IoU(L^i, N^j_t) + \beta \cdot d^1(L^i, N^j_t)
\]

(7)

where \( L^i \) and \( N^j_t \) are the \( i \)th lost track and \( j \)th new track respectively. And, \( d^1 \) denotes the Bhattacharyya distance between the respective color histograms in the HSV space [45]. Once tracks are re-identified, we re-initialize particles around its new position.

6. Experiments

6.1. HeadHunter

We first detail the experimental setup and analyse the performance of HeadHunter on two datasets - SCUT-HEAD [47] and CroHD respectively. For the Faster-RCNN head of HeadHunter, we used 8 anchors, whose sizes were obtained by performing K-means over ground truth bounding boxes from the training set. To avoid overlapping anchors, they were split equally across the four pyramid levels, with the stride of anchors given by \( \max(16, s/d) \) where \( s \) is square-root of the area of an anchor-box and \( d \) is the scaling factor [44]. For all experiments, we used Online Hard Example Mining [54] with 1000 proposals and a batch size of 512.

SCUT-Head is a large-scale head detection dataset consisting of 4405 images and 111,251 annotated heads split across Part A and Part B. We trained HeadHunter for 20 epochs with the input resolution to be the median image resolution of the training set (1000x600 pixels) and an initial learning rate of 0.01 halved at 5th, 10th and 15th epochs respectively. For a fair comparison, we trained HeadHunter only on the training set of this dataset and do not use any pre-trained models. We summarize the quantitative comparisons with other head detectors on this dataset in Table 2. HeadHunter outperforms other state-of-the-art head detectors based on Precision, Recall and F1 scores.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN [48]</td>
<td>0.87</td>
<td>0.80</td>
<td>0.83</td>
</tr>
<tr>
<td>R-FCN+FRN [47]</td>
<td>0.91</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>SMD [58]</td>
<td>0.93</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>HSFA2Net [53]</td>
<td>0.94</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>HeadHunter (Ours)</strong></td>
<td><strong>0.95</strong></td>
<td><strong>0.93</strong></td>
<td><strong>0.94</strong></td>
</tr>
</tbody>
</table>

Table 2. Comparison between HeadHunter’s and other state-of-the-art head detectors on the SCUT-Head dataset.

CroHD: We first trained HeadHunter on the combination of training set images from SCUT-HEAD dataset and CrowdHuman dataset [52] for 20 epochs at a learning rate of 0.001. With variations well characterized, pre-training on large-scale image dataset improves the robustness of head detection. We then fine-tuned HeadHunter on the training set of CroHD, for a total of 25 epochs with an initial learning rate of 0.0001 using the ADAM optimizer [36]. The learning rate is then decreased by a factor of 0.1 at 10th and 20th epochs respectively.

Ablation: We examined our design choices for HeadHunter, namely the use of context module and the anchor selection strategy by removing them. The head detection performance of HeadHunter and its variants on CroHD are summarised in Table 3. We threshold the minimum confidence of detection to 0.5 for evaluation. W/O Cont refers to the HeadHunter without Context Module. We further removed the median anchor sampling strategy and refer to as W/O Cont, mAn. We also provide baseline performance of Faster-RCNN with Resnet-50 backbone on CroHD, the object detector upon which we built HeadHunter. We followed the same training strategy for Faster-RCNN as HeadHunter. All variants of HeadHunter significantly outperformed Faster-RCNN. Inclusion of the context module and
the anchor initialisation strategy also has a noteworthy impact on head detection.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>MODA</th>
<th>MODP</th>
<th>mAP</th>
<th>COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN [48]</td>
<td>34.4</td>
<td>42.2</td>
<td>50.1</td>
<td>40.3</td>
<td>30.8</td>
<td>11.2</td>
<td></td>
</tr>
<tr>
<td>W/O Cont, mAn</td>
<td>40.9</td>
<td>50.8</td>
<td>57.8</td>
<td>38.1</td>
<td>37.8</td>
<td>14.4</td>
<td></td>
</tr>
<tr>
<td>W/O Cont</td>
<td>44.3</td>
<td>57.8</td>
<td>64.5</td>
<td>40.0</td>
<td>42.7</td>
<td>15.0</td>
<td></td>
</tr>
<tr>
<td>HeadHunter</td>
<td>52.8</td>
<td>63.4</td>
<td>68.3</td>
<td>50.0</td>
<td>47.0</td>
<td>19.7</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Summary of various head detector’s performances on the test set of CroHD.

6.2. HeadHunter-T

For the Particle Filtering framework, we used a maximum of N=100 particles for each object. The N particles were uniformly placed around the initial bounding box. To ensure that particles were not spread immediately and were distinct enough, we sampled particles from a Uniform distribution whose lower and upper limit were \((x - 1.5w, y - 1.5h), (x + 1.5w, y + 1.5h)\) respectively. Where, \(x, y, w, h\) denote the centroid, width and height of the initial bounding box. For the color based re-identification, we used 16, 16 and 8 bins for the H, S and V channels respectively, where the brightness invariant Hue [22] was used instead of the standard Hue. \(\alpha, \beta\), which denotes the importance of IoU and color histogram matching, corresponding to Equation 7 were set to 0.8 and 0.2 respectively. We deactivated a track if it remained inactive for \(\lambda_{age} = 25\) frames or if its motion prediction falls outside the image coordinates.

We evaluated three state-of-the-art trackers on CroHD, namely, SORT [5], V-IOU [6] and Tracktor [3] to compare with HeadHunter-T. We chose methods which do not require any tracking specific training, whose implementations have been made publicly available and are top-performing on the crowded MOTChallenge CVPR19 benchmark [13]. For a fair comparison, we performed all experiments with head detection provided by HeadHunter, thresholded to a minimum confidence of 0.6. SORT is an online tracker, which uses a Kalman Filter motion model and temporally associates detection based on IoU matching and Hungarian Algorithm. V-IOU associates detection based on IoU matching and employs visual information to reduce tracking inconsistencies due to missing detection. Parameters for V-IOU and SORT were set based on fine-tuning on the training set of CroHD, as discussed in the supplementary material. We evaluated two variants of Tracktor, with and without motion model. Tracktor+MM denotes the Tracktor extended with Camera Motion Compensation [17] and CVA for inactive tracks. For the two versions of Tracktor, we set tracking parameters similar to HeadHunter. Table 6.2 summarises the performance of aforementioned methods on the test set of CroHD. HeadHunter-T outperforms all the other methods, and furthermore demonstrates superiority in identity preserved tracking. Although Tracktor [3] is similar to HeadHunter-T, there is a noticeable difference in its head tracking performance. We hypothesize the use of Particle Filter framework, which can handle arbitrary posteriors, as the reason for improvement. This claim is justified in the forthcoming section.

6.3. Ablation Experiments

HeadHunter-T: In this section, we analyse the design choices, in particular, the utility of re-identification module and Particle Filter of HeadHunter-T on the training set of CroHD. The results are summarised in Table 5. For variations in motion model, we removed the Particle Filter and used simple Camera Motion Compensation, denoted as HT w/o PF. We also experimented with a reduced number of particles initialized around the head, with n=10, denoted as HT + 10F. Further increasing the number of filters to 100 demonstrated the best performance. However, using more than 100 filters resulted in either duplicates or immoderate spreading, which are undesirable. We removed the re-identification module, to understand its influence, denoted as w/o ReID. Although color histogram is a modest image descriptor, yet it drastically reduced the number of identity switches and showed superior performance in identity preserving metrics - IDEucl, IDF1. We also experimented with \(\alpha\) and \(\beta\) values corresponding to the importance of IoU and histogram matching (Equation 7). We set \(\beta\) to 0.8 and \(\alpha\) to 0.2 and this configuration is denoted as HT + sReID. Surprisingly, we observed more identity switches and a slight decrease in performance across other tracking metrics. HeadHunter-T, our final model, outperformed all the other variants.

Choice of Filter: To further substantiate our choice of a multi-modal filter, we replaced the Particle Filter of HeadHunter-T with a Kalman Filter motion model [34]. While both Kalman Filter and Particle Filter are recursive state estimation algorithms, Kalman Filter assumes the system to be linear with Gaussian noise [2] while Particle Filter’s multimodal posterior distribution enables it to model states of nonlinear systems. We replaced the Particle Filter with a four state Kalman Filter to model the
inter-frame displacement of bounding boxes with CVA. The four states are \(x, y\) centroid coordinates, the height and aspect ratio of bounding boxes respectively, similar to the SORT [5]. The performance of this tracker, denoted as HT + KF is summarised in Table 5. HeadHunter-T with Particle Filter demonstrates superior performance than its Kalman Filter variant with respect to all the tracking metrics reported and in particular, we observe major improvement in-terms of IDEucl metric. Motion of heads along with the pedestrian displacement induces non-linearities in the position of bounding boxes. Although pedestrian motion in general is non-linear, this issue is exacerbated with the small size of head bounding boxes. Hence, using a multimodal posterior state estimation is necessary to address the perceptible impact of non-linear motion. We remark this to be the reason behind improvement in performance while using a Particle Filter in comparison to the Kalman Filter.

Table 5. Illustration of ablation studies of HeadHunter-T (denoted as HT) on the training set of CroHD. The direction of arrows indicate small or large desired metric values.

<table>
<thead>
<tr>
<th>Method</th>
<th>MOTA ↑</th>
<th>IDEucl ↑</th>
<th>IDF1 ↑</th>
<th>MT ↑</th>
<th>ML ↓</th>
<th>IDSw. ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT w/o PF</td>
<td>60.6</td>
<td>40.1</td>
<td>43.9</td>
<td>200</td>
<td>102</td>
<td>3652</td>
</tr>
<tr>
<td>HT + 10F</td>
<td>63.3</td>
<td>58.2</td>
<td>56.3</td>
<td>214</td>
<td>98</td>
<td>1534</td>
</tr>
<tr>
<td>HT w/o ReID</td>
<td><strong>59.5</strong></td>
<td>57.7</td>
<td>57.5</td>
<td><strong>225</strong></td>
<td><strong>91</strong></td>
<td>1411</td>
</tr>
<tr>
<td>HT + sReID</td>
<td>59.1</td>
<td>57.8</td>
<td>58.3</td>
<td>225</td>
<td>91</td>
<td>1280</td>
</tr>
<tr>
<td>HT + KF</td>
<td>63.4</td>
<td>53.8</td>
<td>55.9</td>
<td>214</td>
<td>93</td>
<td>2451</td>
</tr>
<tr>
<td><strong>HeadHunter-T</strong></td>
<td>64.0</td>
<td>61.5</td>
<td>58.5</td>
<td>225</td>
<td>91</td>
<td>1247</td>
</tr>
</tbody>
</table>

Figure 5. Comparison between HeadHunter-T and state-of-the-art trackers on common sequences of CroHD and MOTChallenge benchmark [13]. s-MOTA, s-IDF1, s-MT are scaled version of MOTA, IDF1 and Most Tracked (MT) metrics respectively.

Comparison across paradigm: We compare pedestrian and head tracking performances on the common sequences between CroHD and MOTChallenge CVPR19 dataset. The sequence being the same ensures that trackers are evaluated on full body and head bounding boxes of the same pedestrians in the scene. For this comparison, we chose published state-of-the-art methods on the aforementioned dataset, namely, Tracktor++ [3], V.IOU, DD_TMA [68] and HAM_HI [67]. We performed comparison in-terms of MOTA, IDF1, MT (Mostly Tracked in percentage) metrics. Since we used a different object detector than the rest, a straightforward comparison between performance metrics would not be fair. Hence, for each sequence, we measure the ratio of aforementioned performance metrics with their object detector’s MODA score to obtain the scaled scores - s-MOTA, s-IDF1 and s-MT. The scaled scores, averaged across five common sequences are illustrated in Figure 5. Our approach substantially outperforms other methods indicating that tracking by head detection is more suited for tracking in environments involving high pedestrian density where preserving identity is important. It is also worthy to note that HeadHunter uses a ResNet-50 backbone in contrast to a Resnet-101 backbone used by other methods. Furthermore, Tracktor++, HAM_HI and DD_TMA all use Deep Networks for extracting appearance features, while HeadHunter-T uses a color histogram based appearance feature. By compromising our tracking space (size of bounding box) to avoid mutual occlusion, we observe notable performance gain and significantly reduce the computation cost. This suggests that tracking by head detection paradigm is more desirable for real-time tracking applications focused on identity preservation.

7. Conclusion

To advance algorithms to track pedestrians in dense crowds, we introduced a new dataset, CroHD, for tracking by head detection. To further quantify the efficacy of a tracker in describing pedestrian motion, we introduced a new metric, IDEucl. We developed two new baseline methods, HeadHunter, HeadHunter-T for head detection and head tracking on CroHD respectively. We demonstrated HeadHunter-T to be consistently more reliable for identity preserving tracking applications than existing state-of-the-art trackers adapted for head tracking. Additionally, the adequacy of HeadHunter-T’s performance with a modest computational complexity, opens up opportunities for future research focused on tracking methods adapted for low computational complexity and real-time applications. We also hope that CroHD will serve useful in contiguous fields, such as Crowd Counting and Crowd Motion Analysis.

Acknowledgements: We are thankful to Dr. Vicky Kalogeiton and Prof. Dr. Bastian Leibe for their insightful feedback. We are also thankful to our annotators for their hard work. This work has received funding from the European Union’s Horizon 2020 research and innovation programme under the grant agreement No. 899739. Experiments were performed on the IGRIDA computing cluster.
References


