WEPDTOF: A Dataset and Benchmark Algorithms for In-the-Wild People Detection and Tracking from Overhead Fisheye Cameras

M. Ozan Tezcan[1], Zhihao Duan[2], Mertcan Cokbas[1], Prakash Ishwar[1], and Janusz Konrad[1]
mtezcan@bu.edu, duan90@purdue.edu, [mcokbas, pi, jkonrad]@bu.edu
[1]Boston University, Department of Electrical and Computer Engineering, Boston, MA
[2]Purdue University, Department of Electrical and Computer Engineering, West Lafayette, IN

Abstract

Owing to their large field of view, overhead fisheye cameras are becoming a surveillance modality of choice for large indoor spaces. However, traditional people detection and tracking algorithms developed for side-mounted, rectilinear-lens cameras do not work well on images from overhead fisheye cameras due to their viewpoint and unique optics. While several people-detection algorithms have been recently developed for such cameras, they have all been tested on datasets consisting of "staged" recordings with a limited variety of people, scenes and challenges. Clearly, the performance of these algorithms "in the wild", i.e., on recordings with real-world challenges, remains unknown. In this paper, we introduce a new benchmark dataset of in-the-Wild Events for People Detection and Tracking from Overhead Fisheye cameras (WEPDTOF)[1]. The dataset features 14 YouTube videos captured in a wide range of scenes, 188 distinct person identities consistently labeled across time, and real-world challenges such as extreme occlusions and camouflage. Also, we propose 3 spatio-temporal extensions[2] of a state-of-the-art people-detection algorithm to enhance the coherence of detections across time. Compared to top-performing algorithms, that are purely spatial, the new algorithms offer a significant performance improvement on the new dataset. Finally, we compare the people tracking performance of these algorithms on WEPDTOF.

1. Introduction

People detection is a challenging task that has been widely explored for side-mounted, rectilinear-lens cameras [41, 3, 5]. Adaptation of these algorithms to overhead fisheye (OF) cameras is not trivial due to the radial geometry and severe geometric distortions in the captured images. Despite significant progress in the last two years [38, 23, 16], fisheye people detection remains a challenging and relatively unexplored problem, primarily due to the limited variety of datasets for training and evaluation.

Figure 1: Sample annotated frames from current datasets and from WEPDTOF, which contains *in-the-wild* videos that are more diverse in terms of scenes, identities and densities of people, aspect ratios and real-life challenges.

Almost all state-of-the-art (SOTA) fisheye people-detection algorithms are supervised learning approaches that require annotated data for training and evaluation [38, 23, 16]. While these algorithms achieve very high evaluation scores on existing datasets, this is largely due to the limited dataset complexity and variety. The current fisheye-image datasets for people detection have been recorded in "staged" scenarios, where people move according to predefined patterns (e.g., everyone starts moving at the same time and performs similar actions). Furthermore, the variety of scenes and person identities is very limited. Therefore, the performance of SOTA algorithms in naturally-occurring scenarios, to be expected in real life, remains unknown. Clearly, a challenging dataset, recorded *in-the-wild*, with a large variety of different scenes, actions and people is essential for further advancing this area of research.

In this paper, we introduce a new benchmark dataset of in-the-Wild Events for People Detection and Tracking from Overhead Fisheye cameras (WEPDTOF). While preparing

This work was supported in part by ARPA-E under agreement DE-AR0000944 and by the donation of Titan GPUs from NVIDIA Corp.

[1]vip.bu.edu/wepdtof
[2]vip.bu.edu/rapid-t

WEPDTOF, we focused on several deficiencies of the current datasets. First, we collected *in-the-wild* (non-staged) videos from YouTube instead of recording our own. Second, we selected videos from a variety of indoor scenes recorded by OF cameras mounted at different heights, etc. Third, we focused on videos with real-life challenges (e.g., severe occlusions, camouflage of a person by adjacent background, crowded spaces) to allow performance assessment similar to an application in practice. WEPDTOF consists of 16 clips from 14 YouTube videos, each recorded in a different scene, with 1–35 people per frame and 188 person identities consistently labeled across time. Compared to the current datasets, WEPDTOF has more than 10 times the number of distinct people, ~ 3 times the maximum number of people per frame, and double the number of scenes. Fig. 1 contrasts sample frames from WEPDTOF against those from the current datasets.

The current people-detection algorithms for OF cameras apply frame-wise inference ignoring temporal dependencies. However, recent research on video-object detection has shown a significant performance boost by combining spatial *and* temporal information [36, 43, 40, 19, 6, 42, 24, 26]. We adopt this strategy for OF videos and introduce 3 spatio-temporal extensions to RAPiD [16], the top-performing algorithm on WEPDTOF. We perform a detailed performance comparison of SOTA people-detection algorithms for OF cameras with the proposed spatio-temporal extensions. First, we employ a 2-fold cross-validation to select the best set of hyper-parameters for each algorithm. Then, we evaluate each algorithm in detail by using 7 performance metrics and 2 performance trade-off plots. We show that the improved versions of RAPiD significantly outperform SOTA on WEPDTOF.

Finally, we provide a proof-of-concept illustration of people *tracking* in OF videos using a simple similarity-based association algorithm applied to the bounding boxes generated by RAPiD and its proposed extensions. To the best of our knowledge, this is the first attempt at people tracking in OF videos.

Our main contributions can be summarized as follows:

- We introduce a new dataset, WEPDTOF, for people detection and tracking from overhead fisheye cameras. This is the first *in-the-wild* fisheye video dataset with a diverse set of scenes, identities and densities of people, aspect ratios and other real-life challenges.
- We evaluate people detection performance of SOTA algorithms on WEPDTOF using various metrics and analyze each algorithm's strengths and weaknesses.
- We introduce extensions to RAPiD by leveraging spatial *and* temporal information simultaneously. The proposed extensions outperform plain RAPiD by 2-6% in terms of AP_{50} and by 3-8% in terms of *F-Score*.
- We evaluate people-tracking performance of a similarity-based object association algorithm applied to bounding boxes generated by our proposed people-detection algorithms.

2. Related Work

2.1. Datasets

There exist several datasets for people detection from OF cameras. PIROPO [10], one of the first datasets, is annotated with point-based ground-truth locations of heads of people. Although this is practical in terms of labeling, it limits its utility since it does not localize the full body of a person. BOMNI [11] and Mirror Worlds (MW) [27] address this problem by including image-axis-aligned bounding boxes around people. However, this alignment may cause a significant fraction of bounding-box pixels to capture the background for standing people who appear radially oriented in fisheye images. The most recent datasets, HABBOF [23] and CEPDOF [16], are annotated with human-aligned bounding boxes to minimize the presence of background pixels inside bounding boxes. Recently, Duan *et al.* published human-aligned bounding box annotations for a subset of MW videos, calling it MW-R [16].

Table 1 shows statistics of MW-R, HABBOF, CEPDOF and the new WEPDTOF dataset. HABBOF [23] and MW-R consist of "simple" videos that are limited in terms of the number of identities, maximum number of people labeled per frame and scene diversity. CEPDOF [21] includes more challenging scenarios with up to 13 people in a single frame, however all of its videos are recorded with the same camera hardware, in the same room and with less than 20 different identities. All these datasets have been recorded specifically for training and evaluation of people-detection algorithms under constrained settings (e.g., recorded with the same or similar camera hardware) and scenarios (e.g., people are instructed to enter a room, walk around and exit).

Here, we introduce WEPDTOF, the first dataset of *in-the-wild* events for people detection and tracking from fisheye cameras.

2.2. Algorithms

People detection from OF cameras is an emerging research area. Initial approaches extended the traditional people-detection algorithms (e.g., HOG [9], LBP [1] and ACF [14]) to the fisheye domain by considering the radial geometry and fisheye distortions [39, 7, 37, 21]. Recent CNN-based supervised learning methods have quickly superceded the traditional methods. Tamura *et al.* [38] proposed a rotation-invariant version of YOLO [30, 31, 32], a popular object-detection algorithm designed for standard images. They trained YOLO on the rotated images from the COCO dataset [25]. During inference from OF images, they

Table 1: Statistics of datasets with human-aligned bounding-box annotations ("people" refers to "annotated people").

Dataset	# of clips	# of scenes	# of distinct people	# of people per frame	# of frames	Resolution (MP)
MW-R	19	7	$\sim$15	1-6	8,752	1.1 to 2.2
HABBOF	4	2	9	2-5	5,837	$\sim$4.2
CEPDOF	8	1	17	1-13	25,504	1.1 to 4.2
WEPDTOF	16	14	188	1-35	10,544	0.6 to 5

forced all bounding boxes to be radially aligned relative to the image center. Li *et al.* [23] proposed applying YOLO repeatedly to a top-center rectangular window (where orientations of people are similar to those in standard images) extracted after $15°, 30°, 45°, ..., 345°$ rotations of the fisheye image and then post-processing the results to remove multiple detections of the same person.

Recently, Duan *et al.* introduced RAPiD, a rotation-aware version of YOLO [16]. RAPiD uses a rotated bounding-box representation with 5 parameters (spatial coordinates of the center, width, height and rotation angle of the bounding box). They also introduced an angle-aware loss function for training the network with rotated bounding boxes. To the best of our knowledge, RAPiD is the only supervised people detection algorithm for OF images which can be trained with images annotated by human-aligned bounding boxes.

Although the SOTA algorithms have been successful on existing datasets, the performance gap between the easier datasets (e.g., HABBOF) and the more challenging ones (e.g., CEPDOF) is significant. For example, AP_{50} for RAPiD on CEPDOF is $\sim 11\%$-points lower than on HABBOF [16]. Furthermore, performance of the SOTA algorithms on natural, in-the-wild videos remains unknown. We address this by introducing WEPDTOF.

Video Object Detection: With the introduction of ImageNet VID challenge [35], new algorithms leveraging both spatial and temporal information have been developed for object detection from standard videos. Several algorithms are based on post-processing the results of object detection algorithms which consider the video frames as still images [20, 2, 36]. For example, Robust and Efficient Post-Processing for video object detection (REPP) [36] links the bounding-box predictions between consecutive frames using a learning-based approach. Then, it creates and re-scores the so-called bounding-box tubelets in the temporal dimension. Another approach is to design an end-to-end video-object detection algorithm that leverages the temporal and spatial information simultaneously [43, 40, 6, 26, 24, 42]. For example, Flow-Guided Feature Aggregation (FGFA) [43] uses optical flow to warp the feature maps of past and future frames and then aggregates the warped feature maps for detecting objects in the current frame. Similarly, SELSA [40] aggregates the feature maps of all frames from a video based on their semantic similarity to the current frame. In this paper, we integrate the feature-map warping and feature aggregation introduced in FGFA as well as the post-processing introduced in REPP, into RAPiD to improve its performance.

Video Multi-Object Tracking: The performance of multi-object tracking (MOT) algorithms has increased significantly over the last few years thanks to the extensive datasets such as MOT15 [22], MOT16 [28], MOT19 [12] and MOT20 [13]. A detailed analysis and comparison of these algorithms can be found in recent survey papers [8, 29]. All of these datasets and algorithms are designed specifically for the detection of image-axis-aligned bounding boxes on videos recorded by side-mounted, rectilinear-lens cameras. To the best of our knowledge, multi-object tracking has not been studied for OF cameras.

3. WEPDTOF Dataset

A dataset captured in the wild, that is composed of real-life scenes, actions, people, imaging hardware, etc., is crucial for a fair and meaningful evaluation of people-detection algorithms from OF cameras. With these constraints in mind, we introduce WEPDTOF and make it publicly-available.[3] In Fig. 2, we show a sample frame and its annotation for each video from the new dataset and discuss its salient characteristics below.

In-the-wild videos: Unlike the current fisheye people-detection datasets recorded in staged scenarios, the videos in WEPDTOF have been collected from YouTube and represent natural human behavior. This is important for assessing an algorithm's performance in real-world situations.

Variety: As shown in Table 2, WEPDTOF includes 14 different videos[4] recorded in different scenes (e.g., open office, cubicles, exhibition center, kindergarten and shopping mall). The number of people appearing in a single frame, spatial resolution and length of the videos in WEPDTOF all vary significantly. Furthermore, since the videos in WEPDTOF come from different sources, it is likely they have been captured by different camera hardware (i.e., sen-

[3] vip.bu.edu/wepdtof

[4] Two of the 14 videos in WEPDTOF are divided into two segments each, yielding a total of 16 video clips in the dataset. Each segment is a contiguous portion of the original video and corresponds to a challenging part of that video.

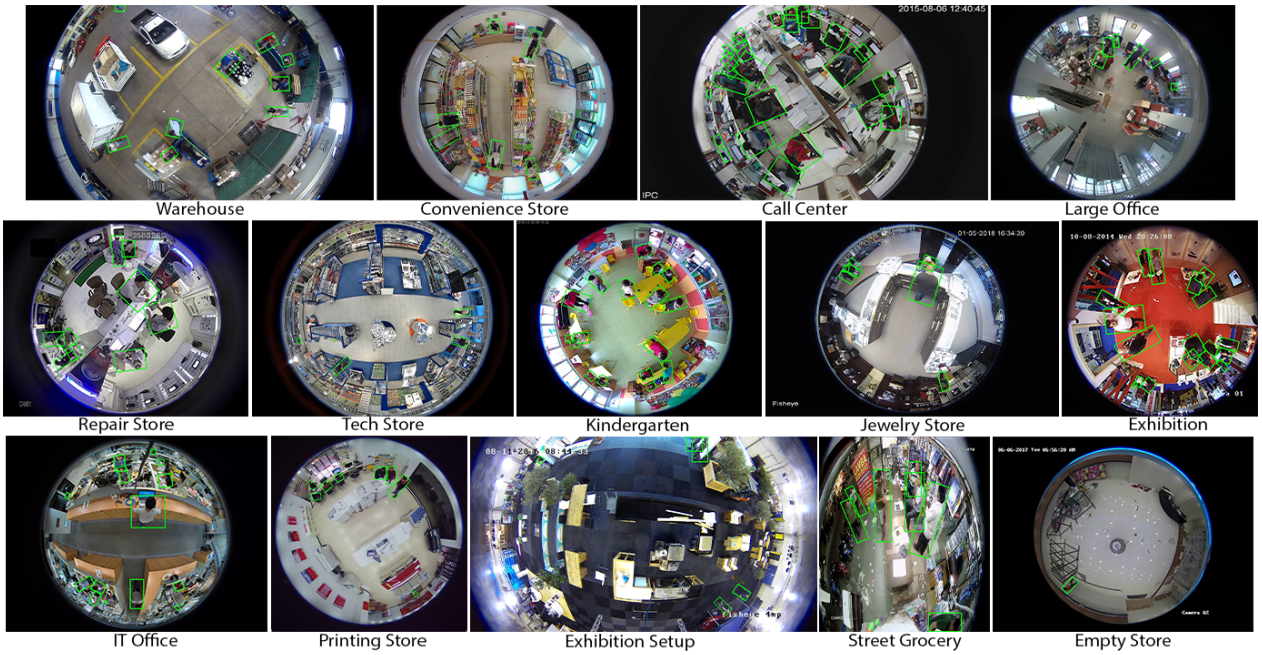


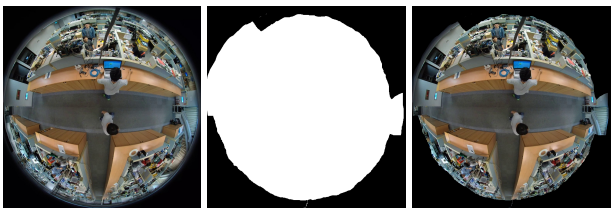
Figure 2: Sample frames with rotated bounding-box annotations from all 14 videos in WEPDTOF.

sor and lens) installed at different heights and working under different illumination conditions.

Real-life Challenges: As shown in Fig. 2 and described in Table 2, WEPDTOF captures real-world challenges such as camouflage, severe occlusions, cropped field of view and geometric distortions. For example, in "Exhibition Setup" (Fig. 2), it is very difficult to detect some people since the color of their clothing is very similar to the background, an effect known as camouflage that is frequently encountered in practice. On the other hand, severe occlusions are clearly visible in "Call Center". Finally, geometric distortions manifest themselves either as a distorted aspect ratio of the images, such as in "Street Grocery", or as a dramatically reduced bounding-box size for people at the field-of-view periphery (person far away) as seen in "IT Office". The challenge of geometric distortions was not significantly captured in any of the previous datasets.

Spatial annotations: We annotated the frames with rotated bounding boxes which are represented as 5-vectors (x, y, w, h, α), where (x, y) represents the spatial location of the center of the bounding box, w the width and and h the height of the bounding box, and α the counterclockwise rotation angle of the bounding box. In our annotations, we exclude some of the areas that are close to the field-of-view periphery since people appear very small and close to each other making it nearly impossible to annotate accurately. These excluded regions are identified by means of a manually-outlined binary region of interest (ROI) map for each video. Fig. 3 shows an example of ROI for "IT Office".

Temporally-consistent annotations: All our bounding-box annotations are temporally consistent, that is, bounding boxes of the same person carry the same ID in consecutive frames. Therefore, WEPDTOF can be used not only for people detection from OF cameras, but also for tracking.



(a) (b) (c)

Figure 3: (a) Sample frame from WEPDTOF's "IT Office", (b) its ROI map, and (c) frame shown within ROI map.

4. Leveraging Temporal Information in People Detection from Overhead Fisheye Cameras

As we discussed in Section 2.2, the current people-detection algorithms for OF cameras are all *image-based* although all of the datasets are *video-based*. Inspired by recent advances in video object detection, we introduce three new people-detection algorithms for OF cameras which use both spatial *and* temporal information. Our approach combines the best-performing fisheye people detection algorithm, RAPiD [16], with some of the best-performing video object detection algorithms, REPP [36] and FGFA [43]. RAPiD is a YOLO-based people detection algorithm that consists of three parts. The first part is the "backbone" network which takes an image as the input and produces three feature maps of the input frame at three resolutions. The second part is called "Feature Pyramid Network" which

Table 2: Details of WEPDTOF. The last column identifies two sets of video clips used in cross-validation.

Video	***Challenges***				
	# of clips	**# of people per frame**	**Total # of frames**	**Resolution** ($W \times H$)	**Set number**
Empty Store	*Occlusions*				
	1	1	358	1440×1080	1
Exhibition Setup	*Camouflage, Tiny people, Partly-visible people, Cropped view*				
	1	3-9	1800	1920×1080	1
Convenience Store	*Cropped view*				
	1	6-7	1075	960×720	1
Large Office	*Occlusions, Tiny people*				
	2	7-9	700	1350×1080	1
Warehouse	*High camera, Cropped view*				
	1	7-10	496	1920×1080	1
Exhibition	*Occlusions, Crowded space, Tiny people*				
	1	17-18	690	1080×1080	1
Call Center	*Crowded space, Tiny people, Cropped view*				
	1	33-35	610	1920×1080	1
Tech Store	*Tiny People, Camouflage, Occlusions, High camera, Cropped view*				
	1	2-4	633	2592×1944	2
Jewelry Store	*Occlusions*				
	2	3-10	900	1620×1080	2
Street Grocery	*Distorted aspect ratio, Non-circular field of view, Cropped view*				
	1	5-7	231	944×1080	2
Printing Store	*Camouflage*				
	1	7-8	1055	1440×1440	2
Repair Store	*Camouflage, Occlusions, Partly-visible people*				
	1	8-10	947	900×720	2
IT Office	*Tiny people*				
	1	14-15	500	1440×1080	2
Kindergarten	*Children*				
	1	18-21	549	900×720	2

takes the three outputs generated by the "backbone" network and transforms them into bounding-box-related feature maps. The last part, "detection head" transforms these feature maps into bounding boxes.

RAPiD+REPP: REPP [36] is a post-processing method designed for object-detection algorithms that produces image-axes-aligned bounding boxes. It uses a learning-based similarity function to link bounding boxes in consecutive frames and produce the so-called object tubelets (known earlier as object tunnels [34]). This is followed by a refinement step to smooth out the confidence score, location, and size of the bounding boxes within tubelets. This, effectively, increases the confidence scores of weaker detections (which could be missed) and decreases the scores of false detections.

RAPiD+REPP applies post-processing, similar to that of REPP, to bounding boxes detected by RAPiD. As the input to the learning-based similarity function, we use the following features for a pair of rotated bounding boxes (e.g., from consecutive frames):

- Euclidean distance between their centers,
- ratio of their widths and ratio of their heights,
- absolute difference between their angles,
- Intersection over Union (IoU) between them.

IoU is computed between the rotated bounding boxes which is computationally much more expensive than computing it for image-axes-aligned bounding boxes. In our experiments, appearance embedding used in [36] did not improve the performance, thus we did not include it as a feature. Our learning-based similarity function is trained on bounding boxes from CEPDOF using logistic regression. We form bounding-box tubelets using the greedy algorithm described in [36] with the similarity scores computed by the trained-similarity function. Then, we apply the refinement step from [36] to smooth out the confidence score, location, and size of the bounding boxes within tubelets.

RAPiD+FGFA: FGFA is an end-to-end video-object detection algorithm which aggregates feature maps computed from past, current, and future frames to aid object detection in the current frame. It consists of three parts. The first part is called "feature extraction network" and it takes a single frame as input and produces a feature map. The second part is called "flow-guided feature aggregation". It uses optical flow to warp feature maps of the past and future frames and

produces their aggregate. Finally, the "detection network" predicts bounding boxes for the current frame from the output of the aggregation step.

FGFA uses backward warping to warp a feature map of the j^{th} frame (f_j) to the i^{th} frame as follows, $f_{j\to i} = \mathcal{W}(f_j, \mathbf{M}_{i\to j})$, where $\mathcal{W}(\cdot)$ is bilinear warping function, $\mathbf{M}_{i\to j}$ is an estimate of optical flow field from frame i to j (computed using FlowNet [15]), and $f_{j\to i}$ represents f_j warped to frame i. The warped features are aggregated as follows, $\bar{f}_i = \sum_{j=i-K}^{i+K} w_{j\to i} f_{j\to i}$, where K represents the number of past and future frames to be aggregated and $w_{j\to i}$ are adaptive weights. For computing $w_{j\to i}$, the feature maps f_i and $f_{j\to i}$ are processed by a shallow neural network which outputs f_i^ϵ and $f_{j\to i}^\epsilon$, respectively. Then, at each spatial location p, weight $w_{j\to i}(p)$ is computed as the cosine similarity between $f_i^\epsilon(p)$ and $f_{j\to i}^\epsilon(p)$ followed by the SoftMax function to normalize the weights, i.e., $\sum_{j=i-K}^{i+K} w_{j\to i}(p) = 1$. In the last step, $\bar{f}_i$ is fed into a "detection network" to predict bounding boxes for frame i.

Following the ideas introduced in FGFA [43] and summarized above, we introduce RAPiD+FGFA which applies temporal aggregation to each of the 3 feature maps generated by the "backbone" network of RAPiD (referred to as P_1, P_2 and P_3 in [16]). We use the Farnebäck algorithm [18] to compute optical flow since it outperforms FlowNet on OF videos. We also introduce RAPiD+FA which applies feature aggregation with adaptive weights, but without feature warping, i.e., $f_{j\to i} = f_j$.

Tracking by Rotated Bounding-Box Association: As discussed in Section 2.2, all existing object-tracking algorithms are designed for standard cameras and image-axes-aligned bounding boxes. We propose a simple people-tracking algorithm for OF cameras using the following two-step approach. First, we compute bounding-box predictions on individual frames using one of the OF people-detection algorithms. Then, we associate bounding boxes between consecutive OF frames by applying the learning-based similarity function used in RAPiD+REPP.

The source code of RAPiD+REPP, RAPiD+FA and RAPiD+FGFA algorithms is available publicly.[5]

5. Experiments

5.1. Algorithms and Implementation Details

As baseline algorithms, we use AA, AB [23], Tamura *et al.* [38] and RAPiD [16][6]. We also report the results of the spatio-temporal algorithms that we introduced in Section 4: RAPiD+REPP, RAPiD+FA and RAPiD+FGFA.We do not perform additional evaluations for the earlier datasets since the performance of RAPiD on them is already very high (AP_{50} over 90% for videos recorded under normal-light conditions, as reported in [17]).

All algorithms use YOLO v3 [32] as backbone and are trained on MS COCO dataset [25] for 100,000 iterations. RAPiD and its variants are fine-tuned on the combination of MW-R, HABBOF and CEPDOF for 5,000 iterations using the angle-aware loss function introduced in [16]. Since AA, AB and Tamura *et al.* are not designed to be trained using rotated bounding boxes, fine-tuning on fisheye datasets is not possible. While training the algorithms on COCO, we use SGD with a learning rate of 0.001, momentum of 0.9 and weight decay of 0.0005. For fine-tuning on fisheye datasets, we decrease the learning rate of SGD to 0.0001. In all of our experiments, we use a minibatch size of 128 and apply data augmentations in the form of random rotations, flips, resizing and color transformations. During inference, we use input resolution of $1,024 \times 1,024$, however during training we use 608×608 due to memory limitations. In RAPiD+REPP, similarly to the original paper [16], we filter frame-wise bounding-box detections of RAPiD with a confidence score threshold of 0.1 and use a similarity score threshold of 0.7 to link bounding boxes in consecutive frames. In RAPiD+FA and RAPID+FGFA, we use the feature maps of 11 consecutive frames to predict the bounding boxes of the middle frame (K=5). In all algorithms, we perform cross-validation to first find the best-performing network weights on the validation set in terms of AP_{50}, and then to tune the confidence threshold in order to maximize the *F-Score*.

5.2. Dataset Splits

Although WEPDTOF is significantly larger than the current datasets in terms of the diversity of scenes, the number of people, etc., it is not large enough in terms of the number of videos to be split into training-validation-test sets. Also, some of the challenges introduced by WEPDTOF appear in only one video (e.g., distorted aspect ratio and children). Thus, we use *all* the videos in WEPDTOF for performance evaluation. A combination of existing datasets (MS COCO [25], MW-R [16], HABBOF [23] and CEPDOF [16]) serves as the training set for all reported algorithms. We evaluate the algorithms on WEPDTOF using 2-fold cross-validation (the last column of Table 2 identifies 2 sets of videos). In cross-validation, we use one of the sets as a validation set to find the best set of hyper-parameters and the other set for reporting the performance on unseen videos, and then we swap their roles.

5.3. People Detection

Evaluation Metrics: We use average precision (AP) at 50% IoU, denoted as AP_{50}, as the main evaluation metric. Unlike regular bounding boxes, rotated bounding boxes of

[5] `vip.bu.edu/rapid-t`

[6] For AA, AB and RAPiD, we use the authors' publicly available implementations from `vip.bu.edu/habbof` and `vip.bu.edu/rapid`. We implemented Tamura *et al.* from scratch since it does not have a publicly-available source code.

people are not unique since the same person can be annotated with multiple, equally-good bounding boxes at different angles. Thus, even a very good people-detection algorithm could have a relatively low IoU with the ground truth. Therefore, we do not report AP results for IoU above 50%.

In addition to AP_{50}, similarly to the MS COCO challenge [25], we report AP_{50} for small, medium and large bounding boxes denoted as AP^S_{50}, AP^M_{50} and AP^L_{50}, respectively. Fig. 4 shows the histogram of bounding-box areas in WEPDTOF. We divide the bounding boxes into three groups: small (area $\leq 1,200$), medium ($1,200 <$ area $\leq 8,000$) and large ($8,000 <$ area) based on their areas normalized to image size of $1,024 \times 1,024$. Then, we calculate AP^S_{50} between small bounding-box annotations and small bounding-box detections, and similarly for AP^M_{50} and AP^L_{50}. We compute each of these scores as the macro-average of per-video results for the videos with at least 100 small, medium or large bounding-box annotations.

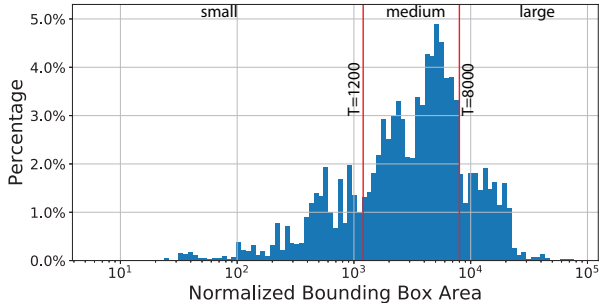


Figure 4: Histogram of bounding-box areas in the annotations of WEPDTOF. All areas are normalized to image size of $1,024 \times 1,024$.

AP_{50} can be described as the area under the *Precision-Recall* curve produced by changing the confidence threshold of the detections. Although this is very useful for summarizing the performance of an algorithm with a single number, in real-life applications the confidence threshold must be fixed. An optimal confidence threshold can be chosen as the one which results in the best *F-Score* on a validation set. We report *Precision*, *Recall* and *F-Score* metrics of the algorithms using this optimal confidence threshold found during validation.

Results: Table 3 reports the performance of baseline and proposed algorithms on WEPDTOF. The performance for individual videos is detailed in the supplementary material. Among the baseline algorithms, RAPiD outperforms its competitors on nearly all of the evaluation metrics. This was to be expected since RAPiD is the only algorithm that uses fisheye images in training with an angle-aware loss function. Surprisingly, *Precision* and *F-Score* of RAPiD are surpassed by AA and AB. To analyze these results in more detail, Fig. 5 shows plots of *Precision* versus *Recall* and *F-Score* against the confidence-score threshold. Although RAPiD produces a higher area under the PR curve, the AA and AB algorithms perform better than RAPiD for high confidence-score thresholds suggesting that RAPiD produces bounding boxes with smaller confidence scores. This is likely due to the fact that the AA and AB algorithms compute bounding-box predictions in overlapped windows extracted from the same image and combine these results in a post-processing step. Thus, they can analyze the same person at multiple orientations which may boost the confidence score of the detection. Note, however, that RAPiD is over 10 times faster than AA and AB.

The proposed extensions of RAPiD achieve 2-5%-points better AP_{50} score than the original version. This demonstrates the importance of temporal information for people detection. REPP improves the bounding boxes produced by RAPiD by changing their confidence scores, locations and sizes, but does not introduce new detections that are not produced by RAPiD. Thus, the performance gain of RAPiD+REPP is limited. By using an end-to-end integration of the temporal information, RAPiD+FA outperforms RAPiD+REPP in terms of AP_{50}, but it is still outperformed in terms of the *F-Score*. RAPiD+FGFA outperforms both of these extensions in terms of all of the metrics with the help of optical flow. The performance gains of the proposed extensions come with a trade-off in terms of efficiency. When applied to bounding boxes aligned with image axes, REPP is proven to be very efficient with just a few milliseconds of extra computing time per frame [36]. During inference, REPP computes IoU between all pairs of bounding-box predictions in consecutive frames. While this computation is efficient for bounding boxes aligned with image axes, it requires computationally-expensive geometric libraries for rotated bounding boxes thus making it inefficient.

Among all of the compared algorithms, AP^S_{50} is 4-7 times lower than AP^M_{50} and AP^L_{50}. Both MS COCO and fisheye people-detection datasets used for training are very limited in terms of small bounding boxes and this makes it challenging for learning-based algorithms to predict small bounding boxes. Clearly, this is an open research direction for people-detection algorithms from overhead fisheye cameras.

5.4. People Tracking

We use tracking by rotated bounding-box association (see Section 4) to estimate tubelets in WEPDTOF videos. We used a confidence threshold of 0.3 to filter bounding-box predictions in each frame before the association step. Folllowing the common practices in MOT, we use *MOTA* [4], *IDP*, *IDR* and *IDF1* [33] to compare the tracking performance of the algorithms. Table 4 compares the performance of RAPiD and its spatio-temporal extensions on WEPDTOF. Similarly to the people-detection results, the spatio-temporal extensions outperform RAPiD on almost all metrics. RAPiD+REPP is significantly outperformed by

Table 3: Performance of fisheye people-detection algorithms on WEPDTOF. Average run-times are computed on a node with a single NVIDIA Tesla V100 GPU.

Algorithm	AP_{50}	AP^S_{50}	AP^M_{50}	AP^L_{50}	*Precision*	*Recall*	*F-Score*	Avg. run-time per frame
Tamura *et al.* [38]	59.8	11.6	65.2	61.3	0.777	0.508	0.581	**98 ms**
Li *et al.* AA [23]	68.3	11.4	70.1	63.7	0.804	0.647	0.705	1477 ms
Li *et al.* AB [23]	69.8	15.8	71.3	63.1	**0.818**	0.643	0.702	1776 ms
RAPiD [16]	72.0	18.4	72.8	67.9	0.731	0.676	0.668	118 ms
RAPiD + REPP	73.7	19.8	74.2	70.2	0.794	0.679	0.703	1667 ms
RAPiD + FA	75.6	19.6	77.5	71.8	0.784	0.672	0.689	269 ms
RAPiD + FGFA	**76.6**	**20.9**	**77.9**	**72.0**	0.803	**0.691**	**0.725**	300 ms

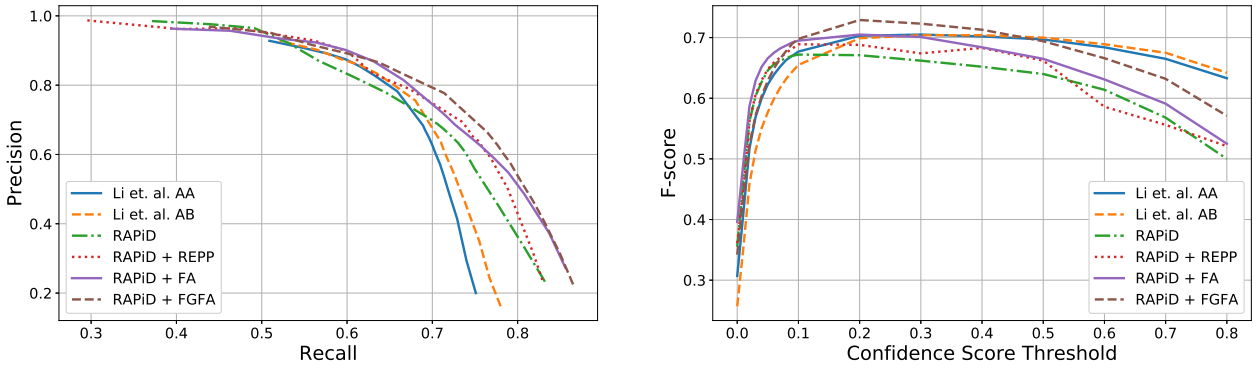


Figure 5: Performance of fisheye people-detection algorithms on WEPDTOF: *Precision-Recall* plots on left and *F-Score* versus confidence threshold plots on right. Tamura *et al.* [38] is omitted since its performance is significantly lower.

Table 4: Performance of fisheye people-tracking algorithms on WEPDTOF.

Algorithm	*MOTA*	*IDP*	*IDR*	*IDF1*
RAPiD	38.3	56.7	41.3	45.0
RAPiD + REPP	38.7	59.1	40.7	45.2
RAPiD + FA	**47.7**	**67.2**	47.9	**53.3**
RAPiD + FGFA	47.2	61.4	**49.9**	53.2

RAPiD+FA and RAPiD+FGFA. We believe this is due to the end-to-end design of RAPiD+FA and RAPiD+FGFA. Surprisingly, RAPiD+FA outperforms RAPiD+FGFA in tracking although it was outperformed in detection.

We note that people-tracking performance scores shown in Table 4 are significantly lower than respective people-detection performance scores, e.g., *IDF1* versus *F-Score*. This is likely due to the fact that we use simple post-processing to compute the tubelets. We believe these tracking results can be improved by adapting ideas proposed in MOT literature [8, 29] to OF cameras, a potentially promising future research direction.

6. Summary and Conclusions

We introduced WEPDTOF as a benchmark dataset to evaluate the real-world performance of people-detection and people-tracking algorithms for overhead fisheye cameras. WEPDTOF is the only overhead-fisheye dataset recorded *in-the-wild* and is much more diverse than the current datasets, with 14 different scenes, 188 different identities and numerous real-life challenges. It is also consistently labeled in time and thus suitable for the development of people-tracking algorithms.

We evaluated the performance of 4 state-of-the-art people-detection algorithms on WEPDTOF using multiple metrics. Then, in order to increase the coherence of detections across time, we introduced 3 spatio-temporal algorithms for people detection from overhead fisheye images. The proposed algorithms significantly outperform the current state-of-the-art on nearly all of the reported metrics.

However, even the top algorithm, RAPiD+FGFA, is far from perfect on WEPDTOF in terms of people detection (AP_{50} of 76.6%), and in terms of people tracking (*MOTA* of 47.2 and *IDF1* of 53.3%). Clearly, more research is required to produce better performing algorithms that can be used for real-world tasks. Computational efficiency is another direction for future research – the current algorithms are far from real-time execution. Further advances in these areas are needed before people detection from overhead cameras becomes a reliable tool in practice. We hope WEPDTOF will prove beneficial in this quest.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 28(12):2037–2041, 2006.

[2] H. Belhassen, H. Zhang, V. Fresse, and E.-B. Bourennane. Improving video object detection by seq-bbox matching. In *Intern. Joint Conf. on Comput. Vis. Imaging and Comput. Graph. Theory and Appl.*, pages 226–233, 2019.

[3] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *Proc. European Conf. Computer Vision*, pages 613–627, 2014.

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[5] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300:17–33, 2018.

[6] Y. Chen, Y. Cao, H. Hu, and L. Wang. Memory enhanced global-local aggregation for video object detection. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 10337–10346, 2020.

[7] A.-T. Chiang and Y. Wang. Human detection in fish-eye images using hog-based detectors over rotated windows. *Proc. IEEE Intern. Conf. on Multimedia and Expo Workshops*, pages 1–6, 2014.

[8] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, volume 1, pages 886–893, 2005.

[10] C. R. del Blanco, P. Carballeira, F. Jaureguizar, and N. García. Robust people indoor localization with omnidirectional cameras using a grid of spatial-aware classifiers. *Signal Processing: Image Communication*, 93:116135, 2021.

[11] B. E. Demiröz, I. Ari, O. Eroğlu, A. A. Salah, and L. Akarun. Feature-based tracking on a multi-omnidirectional camera dataset. In *Intern. Symp. on Comm. Control and Signal Process.*, pages 1–5. IEEE, 2012.

[12] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. CVPR19 tracking and detection challenge: How crowded can it get? *arXiv:1906.04567 [cs]*, June 2019. arXiv: 1906.04567.

[13] 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.

[14] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Machine Intell.*, 36(8):1532–1545, 2014.

[15] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. IEEE Int. Conf. Computer Vision*, pages 2758–2766, 2015.

[16] Z. Duan, M. O. Tezcan, H. Nakamura, P. Ishwar, and J. Konrad. RAPiD: Rotation-aware people detection in overhead fisheye images. In *Proc. IEEE Conf. Computer Vision Pattern Recognition Workshops*, pages 636–637, 2020.

[17] Z. Duan, M. O. Tezcan, H. Nakamura, P. Ishwar, and J. Konrad. Rapid: Rotation-aware people detection in overhead fisheye images. *arXiv preprint arXiv:2005.11623*, 2020.

[18] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, pages 363–370. Springer, 2003.

[19] M. Han, Y. Wang, X. Chang, and Y. Qiao. Mining inter-video proposal relations for video object detection. In *Proc. European Conf. Computer Vision*, pages 431–446. Springer, 2020.

[20] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang. Seq-NMS for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.

[21] O. Krams and N. Kiryati. People detection in top-view fisheye imaging. In *Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance*, pages 1–6, 2017.

[22] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr. 2015. arXiv: 1504.01942.

[23] S. Li, M. O. Tezcan, P. Ishwar, and J. Konrad. Supervised people counting using an overhead fisheye camera. In *Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance*, pages 1–8, 2019.

[24] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *Proc. IEEE Int. Conf. Computer Vision*, pages 7083–7093, 2019.

[25] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common Objects in Context. *arXiv e-prints*, page arXiv:1405.0312, May 2014.

[26] M. Liu, M. Zhu, M. White, Y. Li, and D. Kalenichenko. Looking fast and slow: Memory-guided mobile video object detection. *arXiv preprint arXiv:1903.10172*, 2019.

[27] N. Ma, R. B. Knapp, N. F. Polys, J.-B. Huang, A. Ibrahim, C. Hurt, and Y. Xiao. Mirror worlds challenge. `www2.icat.vt.edu/mirrorworlds/challenge/index.html`, Mar 2018.

[28] 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.

[29] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra. Deep learning in multi-object detection and tracking: State of the art. *Applied Intelligence*, pages 1–30, 2021.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, June 2016.

[31] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, July 2017.

[32] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[33] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proc. European Conf. Computer Vision*, pages 17–35, 2016.

[34] M. Ristivojevic and J. Konrad. Space-time image sequence analysis: Object tunnels and occlusion volumes. *IEEE Transactions on Image Processing*, 15(2):364–376, 2006.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *Intern. J. Comput. Vis.*, 115(3):211–252, 2015.

[36] A. Sabater, L. Montesano, and A. C. Murillo. Robust and efficient post-processing for video object detection. *arXiv preprint arXiv:2009.11050*, 2020.

[37] M. Saito, K. Kitaguchi, G. Kimura, and M. Hashimoto. People detection and tracking from fish-eye image based on probabilistic appearance model. In *SICE Annual Conference 2011*, pages 435–440, 2011.

[38] M. Tamura, S. Horiguchi, and T. Murakami. Omnidirectional pedestrian detection by rotation invariant training. In *Proc. IEEE Winter Conf. on Appl. of Computer Vision*, pages 1989–1998, 2019.

[39] T. Wang, C. Chang, and Y. Wu. Template-based people detection using a single downward-viewing fisheye camera. In *Intern. Symp. on Intell. Signal Process. and Comm. Systems*, pages 719–723, 2017.

[40] H. Wu, Y. Chen, N. Wang, and Z. Zhang. Sequence level semantics aggregation for video object detection. In *Proc. IEEE Int. Conf. Computer Vision*, pages 9217–9225, 2019.

[41] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. Towards reaching human performance in pedestrian detection. *IEEE Trans. Pattern Anal. Machine Intell.*, 40(4):973–986, 2017.

[42] Z. Zhang, D. Cheng, X. Zhu, S. Lin, and J. Dai. Integrated object detection and tracking with tracklet-conditioned detection. *arXiv preprint arXiv:1811.11167*, 2018.

[43] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei. Flow-guided feature aggregation for video object detection. In *Proc. IEEE Int. Conf. Computer Vision*, pages 408–417, 2017.