ISSN 1213-6972 Journal of WSCG http://www.wscg.eu Vol.28, No.1-2, 2020

OPEDD: Off-Road Pedestrian Detection Dataset

Peter Neigel[1,2], Mina Ameli[1], Jigyasa Katrolia[1], Hartmut Feld[1], Oliver Wasenmüller[1], Didier Stricker[1,2]

[1]German Research Center for Artificial Intelligence
Trippstadter Str. 122
67663 Kaiserslautern
Germany

[2]Technische Universität Kaiserslautern
Gottlieb-Daimler-Str., Gebäude 42
67663 Kaiserslautern
Germany

{firstname.lastname}@dfki.de

ABSTRACT

The detection of pedestrians plays an essential part in the development of automated driver assistance systems. Many of the currently available datasets for pedestrian detection focus on urban environments. State-of-the-art neural networks trained on these datasets struggle in generalizing their predictions from one environment to a visually dissimilar one, limiting the use case to urban scenes. Commercial working machines like tractors or excavators make up a substantial share of the total number of motorized vehicles and are often situated in fundamentally different surroundings, e.g. forests, meadows, construction sites or farmland. In this paper, we present a dataset for pedestrian detection which consists of 1018 stereo-images showing varying numbers of persons in differing non-urban environments and comes with manually annotated pixel-level segmentation masks and bounding boxes.

Keywords

Pedestrian Detection, Instance Segmentation, Non-Urban Environment, Off-Road, ADAS, Commercial Vehicles

1 INTRODUCTION

The detection of pedestrians is a major problem for Advanced Driver Assistance Systems (ADAS) and substantial effort has been made in the past decade to advance the performance of detection methods. Current state-of-the-art approaches to pedestrian detection in monocular RGB-images rely on convolutional neural networks (CNN) that output either bounding boxes or pixel-level segmentation masks for every depicted person [1]–[6]. To train these networks for pedestrian detection in a supervised manner large image datasets are needed that come with ground truth annotations for person bounding boxes or segmentation masks. The most commonly used datasets for this task portray scenes in *urban environments*, motivated by the need for and the recent progress in autonomous driving systems for private passenger cars and commercial cargo trucks.

In contrast, most industrial vehicles operate in completely different environments. Although used in dozens of industries, from coarse earthwork operations to tactful harvesters, and making up a substantial share of the total number of motorized vehicles, they are currently neglected by published datasets available for pedestrian detection. In many cases, neural networks trained on urban images fail to generalize from the context of the city to a visually different one, resulting in reduced detective capabilities for off-road environments and posing a problem for ADAS in the context of mobile working vehicles, e.g. automated emergency brakes for tractors, excavators or harvesters. Additionally, urban environments constrain the variety of poses that pedestrians are portrayed in: Most are seen walking or standing upright on the pavement. Industrial or agricultural vehicles in off-road environments can find people in unusual poses, e.g. crouching or lying down while picking crops or doing construction work. These points pose an obstacle to the safety of humans around autonomously operating vehicles in non-urban contexts.

Since there is evidence suggesting that data may be more important than algorithms for performance [7], [8], in this paper we aim to contribute a stereo image dataset of pedestrians in 5 different off-road environments: *Meadows, woods, construction sites, farmland and paddocks*. The persons shown in the

Figure 1: Example images from different datasets. Left: Cityscapes dataset [9]. Center: KITTI Stereo 2015 Dataset [10]. Right: OPEDD. From colour spectrum, over gradient orientations to pedestrian's poses: The visual makeup of urban scenes is substantially different from off-road environments, impeding generalization in detections by neural networks.

dataset are portrayed in varying poses, with some being highly unusual in the ADAS context, e.g. extended limbs, handstands, crouching or lying down. Additionally, our dataset shows significant occlusion of persons from vegetation, crops, objects or other pedestrians. The dataset itself consists of 1018 stereo images, where the left image comes with manually created ground truth pixel-level segmentation masks and individual IDs for every portrayed pedestrian, allowing the data to be used for tasks like object detection (bounding boxes), semantic segmentation (pixel masks) or instance segmentation (pixel masks and IDs). In addition to the dataset itself, we provide depth maps generated from the stereo images and the stereo video sequences from which the images of the dataset were selected.

2 RELATED WORK

Since the popularization of neural networks for object detection, many datasets with annotated pedestrians have been published, either explicitly for the task of pedestrian detection or as part of complete scene segmentation.

Cityscapes [9] is a widely used dataset for urban street scenes. Captured with a stereo camera setup on a car in 50 different cities, it consists of 25000 images, out of which 5000 are provided with fine-grained semantic labels and 20000 are coarsely labelled. The depicted classes include persons, cyclists, cars and other motorized vehicles, while pixel-level semantic labels and instance IDs enable the evaluation of object detection, semantic- , instance- and panoptic-segmentation tasks.

KITTI [10] is a popular driving dataset similar to Cityscapes in terms of portrayed environments, offering benchmarks for different object detection and segmentation tasks. The capture setup consists of a stereo camera in addition to a 360° laser scanner and GPS, providing video sequences and ground truth for the evaluation of tasks like detection, segmentation, sceneflow, depth estimation, odometry, tracking and drivable road detection.

The Caltech Pedestrian Detection Benchmark [11] consists of about 250,000 images frames of regular traffic in an urban environment. A total of 350,000 bounding boxes label circa 2300 unique pedestrians, but no annotations in terms of pixel-level segmentation masks are included.

These datasets are a subsample of publications that focus solely on urban environments and roads, making them unsuitable for the magnitude of industrial and agricultural commercial vehicles.

In contrast, the NREC Agricultural Person-Detection dataset [12] provides a large number of images for off-road pedestrian detection in apple and orange orchard rows, taken from a tractor and a pickup truck. Pedestrian poses include non-standard stances found typically in the orchard environment, only pedestrian bounding boxes are included however, making them incompatible for semantic- and instance segmentation tasks.

In total, the currently published datasets do not allow for comprehensive benchmarking on different pedestrian recognition tasks in off-road-environments. Our presented work offers manually generated ground truth segmentation masks besides bounding boxes, displays a larger and more varying number of environments and includes poses typical for the corresponding off-road environment as well as stances that are completely arbitrary and unusual, filling a gap in published datasets not yet covered.

Dataset	Number of Images	Depth	Segm. Masks	Environment	Poses
KITTI	14,999	LIDAR	✓	Urban	Std. Urban
Cityscapes	25,000	Stereo	✓	Urban	Std. Urban
Caltech	39,702	-	-	Urban	Std. Urban
NREC	23,950	Stereo	-	Agricultural	Std. Agricultural
OPEDD	**1,018**	**Stereo**	✓	**Multiple Off-Road**	**Wide Range, Unusual**

Table 1: Comparison of contents of datasets for pedestrian detection.

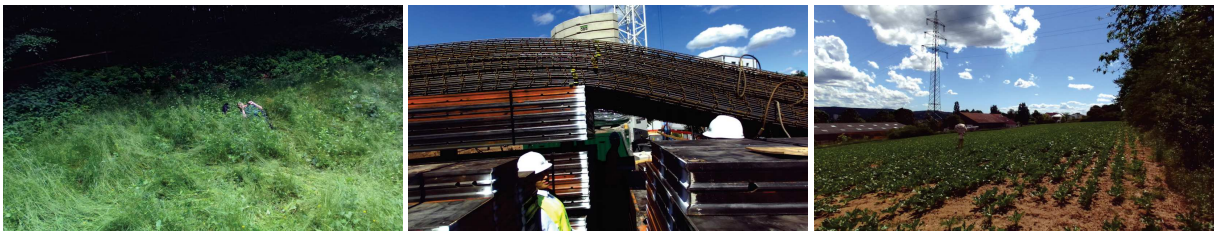

Figure 2: Our dataset shows different types of occlusion in varying environments, including naturally occurring obstacles (left: vegetation, center: construction materials) and unusual objects (right: umbrella).

3 CHARACTERISTICS OF OFF-ROAD ENVIRONMENTS

Off-road, agricultural or rural environments show several characteristics that differentiate them from urban surroundings in a number of ways:

Visuals The largest differences are recognizable in the visual domain. In urban images, the background is mostly characterized by buildings and paved roads, yielding a colour spectrum dominated by greys. In contrast, off-road environments can depict a multitude of backgrounds. Agricultural and wooded surroundings usually show ample vegetation with a colour spectrum controlled by greens and browns, while construction sites display a mix of urban and non-urban components. In terms of texture, backgrounds dominated by vegetation show heavy textural repetition. Moreover, Tabor et al. [13] have shown that the gradient orientation alignment is very distinct between the different types of environments.

Composition In urban settings, pedestrians are one visually distinct object class out of many, including cars, cyclists, trucks and many more. In off-road environments, pedestrians tend to appear as much more strongly separated objects.

Occlusion In surroundings dominated by vegetation partial occlusion of persons by leaves, grass or branches is very common. Examples are people harvesting fruit in orchards or a person standing in field crops, having parts of the lower body obstructed. Additionally, the boundary of occlusions is often much fuzzier than in the case of occlusions by e.g. cars in the urban setting.

Poses Due to the nature of city scenes, datasets for pedestrian detection in urban environments show persons predominantly standing or walking upright. Additionally, because the data is usually captured from a vehicle driving on the road, most pedestrians are located on the lateral edges of the image, with persons only directly in front of the camera if the data-capturing-vehicle is positioned in front of a cross- or sidewalk. Contrary to that, many agricultural or industrial scenes show persons in unusual and more challenging poses: Often the person is seen working in a crouching or bent position and limbs extended in differing ways are common. Due to the hazardous environment on construction sites, the vehicle could encounter people lying on the ground. In general, off-road scenes display a much larger variety of poses than the average urban scene. Many of these difficulties are addressed in our dataset, described in the following chapter.

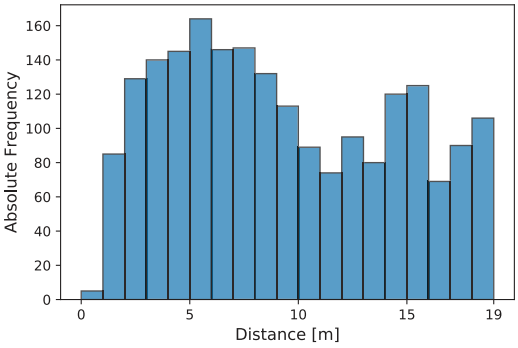

Figure 4: Histogram of distances of the portrayed pedestrians to the camera.

Figure 3: Special attention was paid to capture a wide range of poses not usually encountered in urban driving datasets. Left: Handstand. Center: Jumping with extended limbs. Right: Head covered with clothes.

Task	Implementation	Trained On	AP50	AP75	AP
Instance Segmentation	Mask R-CNN	COCO	0.6831	0.4355	0.3935
Instance Segmentation	Mask R-CNN	COCO + Ours	0.80031	0.4880	0.4500
Object Detection	YOLOv3	COCO	0.5666	0.3966	0.3437

Table 2: Test-set results on object detection and instance segmentation tasks with Mask R-CNN and YOLOv3.

4 DATASET

4.1 Data Capturing

We record all sequences of our dataset using a Stereolabs ZED Camera [14]. The stereo camera has a baseline of 120mm and is able to capture video sequences with a side-by-side output resolution of 4416x1242 pixels at 15 frames per second. In order to prevent compression artifacts, which can impair detection performance [15], the video sequences are captured with lossless compression.

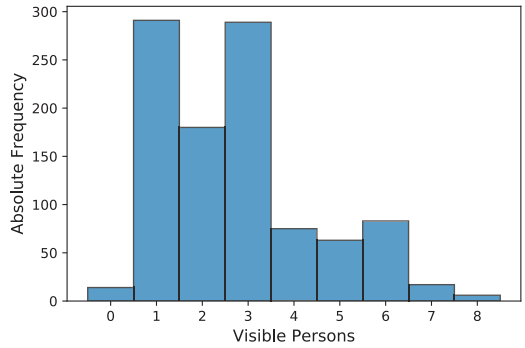

Figure 6: Histogram depicting how many pedestrians are visible in the images.

Besides rectifying the images, the ZED camera also outputs depth maps from stereo for every frame. Data was captured in short video sequences of 1 to 50 seconds with a framerate of 15 Hz.

Environments We capture data in different locations to cover a broad range of possible ADAS application scenarios: *Meadows, woods, construction sites, farmland and paddocks*.

While capturing, emphasis was laid on covering many scenarios that complicate pedestrian detection in off-road environments.

Occlusions In all of our environments, occlusion happens with locally characteristical obstacles like grass, leaves, field crops, construction materials or fences, as well as more unusual barriers like stone walls, garbage bins or objects held by persons (umbrellas, paper files). Examples can be found in Fig. 2. Moreover, we took care to include many instances of person-to-person occlusion, oftentimes by a pedestrian standing close to the camera.

Poses Our dataset shows a variety of uncommon and challenging poses including people doing handstands, lying on the ground or on objects, lying on the back or on the side, sitting, crouching or bent over, limbs extended as well as running and jumping.

Composition Special attention was paid to have multiple positions in the image covered by pedestrians, to avoid the urban situation where persons are located mainly at the sides. Additionally we vary the number of persons, see Fig. 6, and the distances they appear to the camera (Fig. 4). Most images are taken from eye-level up to 1m above, facing forward, to simulate taller vehicles like tractors or excavators, with images showing a more downward facing angle.

Image Resolution	2208 x 1242
Stereo Camera Baseline	120mm
Number of video sequences	1004
Video Framerate	15 Hz
Compression	Lossless
Number of Images	1018
Total pedestrian instances	2801

Table 3: Capturing and dataset statistics.

Lighting The light conditions vary naturally as well as intentionally, with some images being taken against direct sunlight or with people being hidden in the shadows of walls or trees.

Miscellaneous Further variations include clothing, helmets or gimmicks like clothes being thrown around or people deliberately hiding.

Image Selection From the video sequences 1018 image pairs are selected. Since the images are often

Figure 5: Samples of images with difficult lighting conditions.

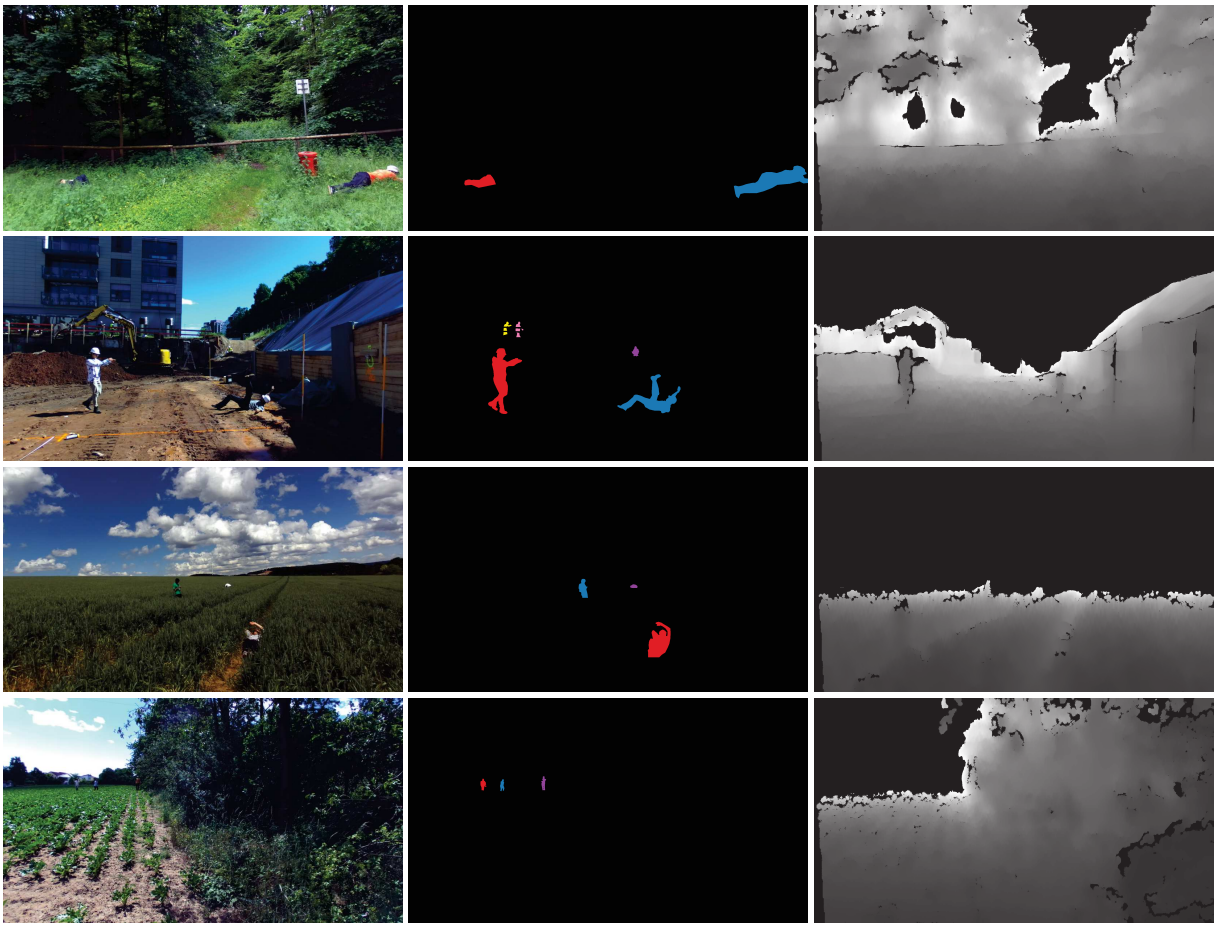

Figure 7: Selection of environments contained in the presented dataset. Left column: Left frame of stereo image. Center column: Corresponding segmentation masks. Right Column: Corresponding depth maps. Our dataset shows a variety of environments and poses.

almost identical from one frame to the next, we make sure to choose the next frame in such a way that sufficient alteration is visible, often by clear repositioning of persons or after a pan of the camera. The images that make it to the final dataset are selected by hand.

Annotation The ground truth annotations of the images were created using the VGG Image Annotation Tool (VIA) [16], [17]. All visible persons were annotated with a segmentation mask and given an (imagewise-) unique ID. Since drawn masks sometimes overlap, the IDs are assigned in increasing order with increasing depth of the person in the image. All labelling information is stored in a json file that can also be imported as a VIA project, allowing users to easily modify and expand on the annotations. The project files are available on the GitHub Repository provided at the bottom. Additionally, we supply scripts to extract segmentation masks and bounding boxes.

4.2 Related Detection Algorithms

Object Detection Object detection describes the task of extracting bounding box coordinates and dimensions of target objects in the image. We use YOLOv3 [2] trained on the COCO dataset [18] to make a first rough evaluation on our data. The CNN first predicts bounding box coordinates from anchors, regresses an objectness score and classifies the image patch.

Instance Segmentation In contrast to plain object detection, instance segmentation algorithms also output pixel-level segmentation masks for every detected bounding box. We apply Mask R-CNN [3] to our dataset, first as-supplied ([19]) trained on COCO, then fine-tuned on our training set. A region proposal network first predicts possible objects and their bounding boxes. Further branches then classify the object and output corresponding pixel masks. The results of our first evaluations can be taken from table 2. We compute the average precision (AP) similar to [20], where the number specifies the minimum intersection over union (IoU) for a predicted bounding box or segmentation mask to be assigned to a ground truth instance, e.g. AP50 meaning a minimum of 50% IoU. In addition, we average AP over multiple IoU thresholds from 0.5 to 0.95, simply denoted as AP, to avoid bias towards a specific value [9], [18].

5 CONCLUSION AND FUTURE WORK

In this paper, we have introduced a new Off-Road Pedestrian Detection Dataset (OPEDD). It consists of 1018 manually annotated images and displays persons in scenarios broadly characterized as meadows, woods, construction sites, farmland and paddocks. The people portrayed show a wide variety of poses usually not encountered in urban environments and corresponding datasets. In all settings, it supplies a variety of types of occlusions, compositions and lighting conditions. Ground truth annotations are available as pixel-level segmentation masks, with each person in an image having an individual ID, making it possible to use the dataset for object detection, semantic- and instance segmentation tasks. For future work, we aim to add additional annotations to our currently unlabelled sequences. Furthermore, instead of labelling unique images in the captured data, complete sequences could be labelled for tasks like multiple object tracking.

6 ACKNOWLEDGMENTS

We thank Maximilian Palm for aiding us in the capturing of sequences and selection of images.

GitHub https://github.com/PNeigel/OPEDD

REFERENCES

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot Multi-Box Detector," *ECCV*, pp. 21–37, 2016.

[2] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *ArXiv*, vol. abs/1804.02767, 2018.

[3] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *ICCV*, pp. 2980–2988, 2017.

[4] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting," *ECCV, Proceedings*, pp. 618–634, 2018.

[5] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-Guided Attention Network for Occluded Pedestrian Detection," *ICCV*, pp. 4967–4975, 2019.

[6] J. Zhang, L. Lin, Y.-C. Chen, Y. Hu, S. C. H. Hoi, and J. Zhu, "CSID: Center, Scale, Identity and Density-Aware Pedestrian Detection in a Crowd," *CoRR*, 2019.

[7] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *Intelligent Systems. IEEE*, 2009.

[8] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do We Need More Training Data?" *IJCV*, pp. 1–17, 2015.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," *CVPR, Proceedings*, pp. 3213–3223, 2016.

[10] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," *CVPR, Proceedings*, pp. 3354–3361, 2012.

[11] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *PAMI*, vol. 34, 2012.

[12] Z. Pezzementi, T. Tabor, P. Hu, J. K. Chang, D. Ramanan, C. Wellington, B. P. Wisely Babu, and H. Herman, "Comparing apples and oranges: Off-road pedestrian detection on the National Robotics Engineering Center agricultural person-detection dataset," *Journal of Field Robotics*, vol. 35, no. 4, pp. 545–563, 2018.

[13] T. Tabor, Z. Pezzementi, C. Vallespi, and C. Wellington, "People in the weeds: Pedestrian detection goes off-road," in *2015 IEEE SSRR*, 2015, pp. 1–7.

[14] *Https://www.stereolabs.com/zed/*, accessed May 4, 2020.

[15] M. Dejean-Servières, K. Desnos, K. Abdelouahab, W. Hamidouche, and L. Morin, "Study of the Impact of Standard Image Compression Techniques on Performance of Image Classification with a Convolutional Neural Network," *INSA Rennes; Univ Rennes;IETR; Institut Pascal*, 2017.

[16] A. Dutta, A. Gupta, and A. Zissermann, *VGG Image Annotator (VIA)*, http://www.robots.ox.ac.uk/˜vgg/software/via/, 2016, accessed May 4, 2020.

[17] A. Dutta and A. Zisserman, "The VIA Annotation Software for Images, Audio and Video," in *27th ACM Multimedia, Proceedings*, ser. MM '19, New York, NY, USA: ACM, 2019.

[18] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.

[19] W. Abdulla, *Matterport Mask R-CNN*, https://github.com/matterport/Mask_RCNN, 2017, accessed May 4, 2020.

[20] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *ECCV*, 2014.