

A New Benchmark for Stereo-Based Pedestrian Detection

C. G. Keller¹, M. Enzweiler² and D. M. Gavrilă^{2,3}

¹Image & Pattern Analysis Group, Department of Math.
and Computer Science, Univ. of Heidelberg, Germany

²Environment Perception, Group Research, Daimler AG, Ulm, Germany

³Intelligent Autonomous Systems Group, Univ. of Amsterdam, The Netherlands

{uni-heidelberg.keller, markus.enzweiler, dariu.gavrila}@daimler.com

Abstract—Pedestrian detection is a rapidly evolving area in the intelligent vehicles domain. Stereo vision is an attractive sensor for this purpose. But unlike for monocular vision, there are no realistic, large scale benchmarks available for stereo-based pedestrian detection, to provide a common point of reference for evaluation. This paper introduces the Daimler Stereo-Vision Pedestrian Detection benchmark, which consists of several thousands of pedestrians in the training set, and a 27-min test drive through urban environment and associated vehicle data. The data, including ground truth, is made publicly available for non-commercial purposes. The paper furthermore quantifies the benefit of stereo vision for ROI generation and localization; at equal detection rates, false positives are reduced by a factor of 4-5 with stereo over mono, using the same HOG/linSVM classification component.

I. INTRODUCTION

Vision-based pedestrian detection is a key problem in the domain of intelligent vehicles (IV). Large variations in human pose and clothing, as well as varying backgrounds and environmental conditions make this problem particularly challenging.

The main contribution of this paper is to carefully quantify the benefit of stereo-vision over an otherwise identical monocular system for pedestrian detection, see Figure 1. We do not present entirely new systems, but evaluate a variant of the well-known HOG-based pedestrian detector, e.g. [4], in both monocular and stereo vision set-ups. We assume our results to generalize to other established pedestrian detectors, e.g. [5], [7], [10], [13], [16], [17].

A second contribution involves a new large real-world stereo dataset for pedestrian detection which is used in our experiments. We make this dataset publicly available for non-commercial purposes to encourage research and benchmarking¹. The data is based on the established monocular *Daimler Pedestrian Detection Benchmark* [7], which is extended in several ways. First, the new benchmark includes the corresponding (left and right) stereo image pairs for the same 27-minute urban test sequence as used in [7], where previously only the left image was published. We further present a new stereo-vision sequence not containing pedestrians for bootstrapping. Instead of generating 3D ground-truth by

back-projecting manually acquired pedestrian labels from the image into the world using the ground-plane constraint, we now derive more exact 3D ground-truth using shape information and stereo-vision. Finally, we enrich our test sequence by releasing vehicle data (velocity, yaw rate) estimated by on-board sensors to develop and evaluate more robust tracking algorithms.

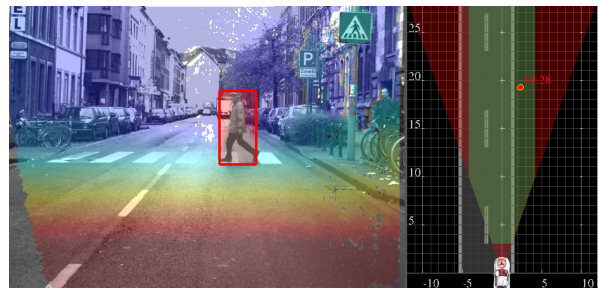


Fig. 1. Pedestrian detection using the stereo-based system.

II. PREVIOUS WORK

Many interesting approaches for vision-based pedestrian detection have been proposed. Most approaches follow a module-based strategy comprising generation of possible pedestrian location hypotheses (regions-of-interest, ROI), followed by pedestrian classification and tracking ([3], [11], [18]). A detailed review of state-of-the-art pedestrian systems is beyond the scope of this paper. We refer the reader to recent surveys and benchmarks, i.e. [5], [7], [10], [13], [16], [17].

Evaluation, comparison and ranking of pedestrian detection systems requires publicly available datasets which can be used as a common reference ground to benchmark many different systems. As a result of various systems having different requirements in terms of data used (e.g. gray-level appearance, optical flow, stereo, color or vehicle data), a multitude of datasets are available. Data acquisition further varies with the actual application area of the system, e.g. surveillance, IV or action recognition. Roughly, pedestrian datasets can be categorized into classification and detection datasets.

¹See <http://www.science.uva.nl/research/isla/downloads/pedestrians/index.html> or contact the last author.

	Training				Testing										Year
	# pedestrians	# pos. image	#neg. samples	# neg. images	# labels	# images	Traj. Labels	Focal Length (mm)	Stereo	Baseline (cm)	Vehicle Data	Platform	City / Setup		
ETH [9]	1578	490	-	-	10k	2293	-	8	✓	40	-	stroller	Zurich / city	2007	
CALTECH [5]	192k	67k	-	61k	155k*	65k	$\approx 1k$	7.5	-	-	-	vehicle	Los Angeles / urb.	2009	
TUD-Brussels [22]	1776	1092	-	192+26	1326	508	-	8	-	-	-	vehicle	Brussels / city	2009	
Daimler Mono [7]	3915	-	-	6744	56k	22k	259	12	-	-	-	vehicle	Aachen / urb.	2009	
CVC-02 [12]	1016	-	7650	153	7983	4634	-	6	✓	12	-	vehicle	Barcelona / urb.	2010	
<i>Daimler Stereo (this paper)</i>	3915	-	-	7129	56k	22k	259	12	✓	30	✓	vehicle	Aachen / urb.	2011	

TABLE I

SUMMARY OF THE AVAILABLE PEDESTRIAN DATASETS RECORDED FROM A MOVING PLATFORM IN AN URBAN ENVIRONMENT. * TEST DATA IS NOT PUBLICLY AVAILABLE.

Classification datasets, e.g. [4], [6], [8], [14], [17], [19], [20], are mainly used to evaluate a combination of a feature set and a pattern classifier using a given set of pedestrian (positive) and non-pedestrian (negative) cut-out samples. For pedestrians, such samples are typically extracted from manually labeled image data resulting in accurately aligned pedestrian cut-outs. Non-pedestrian cut-outs can be extracted randomly or by some pre-processing method from images not containing pedestrians. In this context, pre-processing is used to focus on application-relevant “difficult” samples. A fixed set of positive and negative training and test samples is supplied for benchmarking. To allow for classifier bootstrapping, additional negative images are often provided.

Detection datasets, e.g. [1], [5], [7], [9], [12], [22], [23], [24], containing cut-outs for training and full images for test data are used to benchmark integrated pedestrian detection systems. Although the pedestrian classifier is the most important module of most systems, differences in relative performance can also arise from varying hypotheses generation or tracking modules. Further, the extended scanning of an image skews the relation of pedestrian and non-pedestrian windows used for testing - typically, the test images only contain a few pedestrians, whereas many thousands of regions not corresponding to pedestrians may be scanned per image.

Although a classification dataset allows the isolated performance analysis of a classification module, results do not necessarily generalize to the performance of a fully integrated pedestrian detection system, as noted above. On the other hand, evaluating the classification module of an integrated system in an isolated brute-force (monocular) sliding window detection setting, e.g. [5], does not necessarily correspond to the actual application context either. Both evaluation methodologies have their justification and the choice strongly depends on the application and evaluation context.

In the context of advanced driver assistance systems (ADAS) in the intelligent vehicles domain, video sequences acquired in a realistic urban traffic environment are crucial for an adequate evaluation of state-of-the-art systems. Depending on the design of the systems under consideration, different image cues may be required. Systems utilizing op-

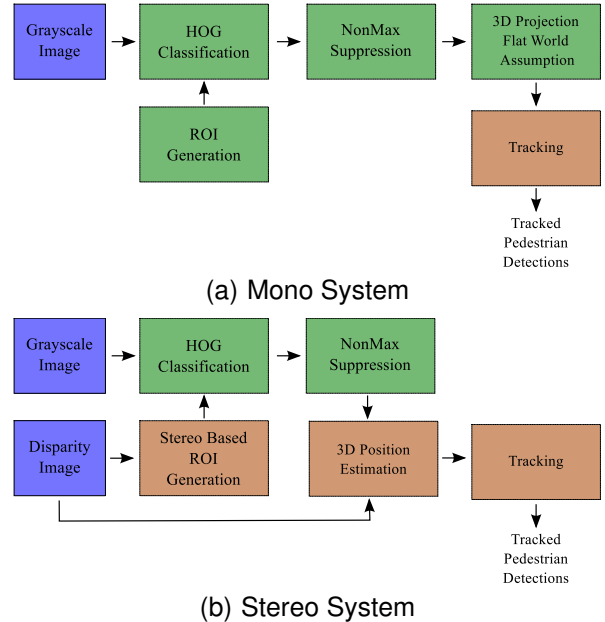


Fig. 2. Comparison of the processing steps for the stereo and mono system.

tical flow require a sufficiently large frame rate while stereo based systems need additional image data to derive depth information. Table I shows an overview of available pedestrian detection datasets recorded from a moving platform, as well as their main properties. Manually annotating video data is a time-consuming and tedious work. In [5], an interactive procedure where the system generated intermediate labels by interpolation between manually assigned labels is proposed. Especially for sequences recorded with a large frame rate this approach can reduce the costs for labeling at the expense of accuracy [12].

In the remainder of this paper, we introduce the systems used for benchmarking, present our new stereo-based benchmark dataset and present our experimental evaluation.

III. SELECTED PEDESTRIAN DETECTION SYSTEMS

In our experiments, we compare the performance of two state-of-the-art baseline systems. The first system solely depends on a monocular camera setup for detection and

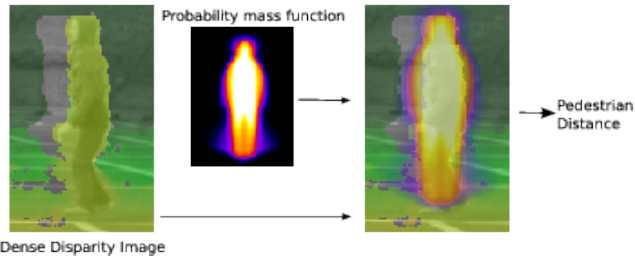


Fig. 3. Pedestrian distance estimation using weighted disparity values.

tracking, see [7]. In contrast, the second system utilizes stereo data for hypotheses generation and refined pedestrian localization, i.e. an adapted version of [11]. Stereo data is computed using the ‘‘Semi-Global Matching’’ (SGM) algorithm [15] algorithm which provides dense disparity maps. Figure 2 illustrates the processing steps of the selected systems.

Both systems utilize an initial set of ROIs generated for various detector scales and image locations using a flat-world assumption and ground-plane constraints. For the stereo-based system, ROIs at a certain distance are only generated if the number of depth features for the distance exceeds a percentage of the ROI area. ROIs are then passed to the classification module which uses histograms of oriented gradients (HOG) features [4] on gray-scale image data. Extracted features are classified by a linear support vector machine (linSVM). To speed-up the feature computation, we implemented the integral histograms of oriented gradients approach e.g. [25], which does not allow for the inclusion of tri-linear interpolation steps, as described in [4]. The resulting computational speed-up comes at the cost of a lower detection performance [25].

Multiple detector responses at near-identical locations and scales are addressed by applying confidence-based non-maximum suppression to the detected bounding boxes using pairwise box coverage. Two system detections a_i and a_j are subject to non-maximum suppression if their coverage

$$\Gamma(a_i, a_j) = \frac{A(a_i \cap a_j)}{A(a_i \cup a_j)}, \quad (1)$$

the ratio of intersection area and union area, is above θ_n . For the following experiments $\theta_n = 0.5$ has been selected.

To allow possible collision mitigation maneuvers, the pedestrian position with respect to the vehicle is required. From the available stereo data, the pedestrian position is estimated by averaging the weighted disparity values in the detected box in the image and back-projecting the foot-point into 3D world coordinates onto the ground-plane using known camera geometry, see [11]. With manually labeled pedestrian shapes, a mask has been derived for importance weighting of disparity values depending on their location, as shown in Figure 3. Pedestrian positions for the monocular system are computed with the assumption that pedestrians are standing on the (flat) ground-plane (ground-plane constraint).

Lateral (x) and longitudinal (z) pedestrian positions are tracked using a Kalman filter [2] with measurement vector

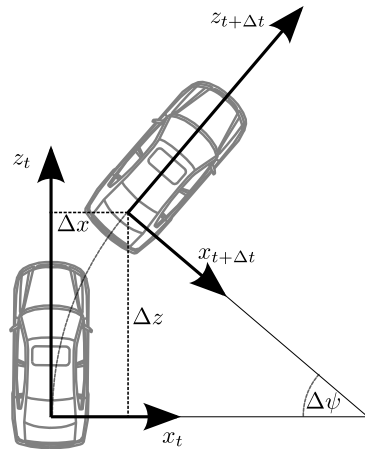


Fig. 4. Single-Track model used for ego-motion compensation.

$\mathbf{z} = (x, z)^T$ and the state vector $\mathbf{x}_k = (x, z, v_x, v_z)^T$, with v_x and v_z denoting the pedestrian velocity. We assume no abrupt velocity changes of the pedestrian and consequently use a constant velocity (CV) model. With vehicle velocity v^e and yaw-rate $\dot{\psi}^e$, estimated from on-board sensors, the vehicle ego-motion is compensated. As a possible extension, visual measurements could additionally be incorporated at this point. Figure 4 illustrates the simplified motion of the vehicle using the one-track vehicle model [21]. Ego-motion compensation is integrated into the prediction step of the Kalman filter. Between time-step t and $t + \Delta t$ the vehicle travels the distance $(\Delta x, \Delta z)$ with orientation change $\Delta\psi^e$. Moving on the curve radius $r = v^e \cdot \dot{\psi}^e$ following translation and rotation parameters apply:

$$\Delta\psi^e = \dot{\psi}^e \Delta t \quad (2)$$

$$\Delta x = v^e (\dot{\psi}^e)^{-1} [1 - \cos(\Delta\psi)] \quad (3)$$

$$\Delta z = v^e (\dot{\psi}^e)^{-1} \sin(\Delta\psi) \quad (4)$$

So the predicted pedestrian state $\hat{\mathbf{x}}_{k|k-1}$ in the vehicle coordinate system for $t + \Delta t$ is computed using

$$\hat{\mathbf{x}}_{k|k-1} = F[\hat{\mathbf{x}}_{k-1} - \mathbf{x}_{\text{cog}}] + [\mathbf{x}_{\text{cog}} - \begin{pmatrix} \Delta x \\ \Delta z \\ 0 \\ 0 \end{pmatrix}] \quad (5)$$

with \mathbf{x}_{cog} describing the translation to the vehicle center-of-gravity and F describing the state transition matrix respecting the vehicle ego-orientation change.

$$F = \begin{pmatrix} \cos(\Delta\psi) & \sin(\Delta\psi) & \cos(\Delta\psi)\Delta t & \sin(\Delta\psi)\Delta t \\ -\sin(\Delta\psi) & \cos(\Delta\psi) & -\sin(\Delta\psi)\Delta t & \cos(\Delta\psi)\Delta t \\ 0 & 0 & \cos(\Delta\psi) & \sin(\Delta\psi) \\ 0 & 0 & -\sin(\Delta\psi) & \cos(\Delta\psi) \end{pmatrix}$$

Measurement to track associations in the track management are handled using the global nearest neighbor algorithm [2] with prior rectangular gating on the predicted pedestrian position. New tracks result from measurements that can not be assigned to an existing track. Starting in the state *hidden*, new tracks enter the state *confirmed*



Fig. 5. Overview of the detection benchmark dataset: a) pedestrian training samples. b) non-pedestrian training images. c) annotated test images.

after n measurement to track associations. After m missed associations *confirmed* tracks are terminated. Here we use $n = 2$ and $m = 2$ for the track management. Only confirmed tracks are regarded as valid system outputs.

IV. DATASET OVERVIEW

We extend the benchmarking dataset of [7] to contain stereo image pairs to allow the computation of distance data using different stereo algorithms. Stereo video data not containing pedestrians is additionally supplied to allow training and bootstrapping of different classification algorithms.

Test data has been recorded with 15 frames per second (fps) enabling the computation of optical flow data. Vehicle velocity and yaw-rate measurements from on-board sensors are provided for each frame to enable integration into a tracking and decision making system. All sequences are recorded in an urban environment representing a realistic challenge for todays pedestrian detection systems. Example images from training and testing data are given in Figure 5.

A summary of the dataset statistics is given in Table II. By shifting and mirroring, 15660 pedestrian training samples are created from 3915 unique pedestrian samples. A training sample resolution of 48×96 pixels with a border of 12 pixels around the pedestrians is used. Negative training samples (≈ 15600) are randomly cropped from the bootstrapping image

Training	
# unique pedestrians	3915
# pedestrian samples	15660
# neg. frames (stereo pairs)	7129
Testing	
# frames (stereo pairs)	21790
# labels	56484
# pedestrian traj.	259

TABLE II

DAIMLER STEREO-VISION PEDESTRIAN BENCHMARK DATASET STATISTICS.

sequence using ground-plane constraints.

In [7], 3D ground truth from camera geometry in addition to bounding box labels has been provided. The 3D ground truth data has been revised. We use 3D ground truth from stereo data because of its robustness to vehicle pitch variations and violations of the flat-world assumption. Figure 6 illustrates the ground truth generation. To increase precision of estimated 3D positions, unoccluded pedestrians in the required detection area (see Section V) have manually been shape labeled. Pedestrian distance is derived from the median of disparity values located on the pedestrian body. In combination with the pedestrian foot-point determined from the shape center-of-gravity (COG) and known camera parameters the 3D position is computed.

V. EXPERIMENTS

In the following the performance for the classifier modules and complete system configurations of the two selected baseline systems is compared. System setup and evaluation parameters are described in detail to allow reproducibility of the results.

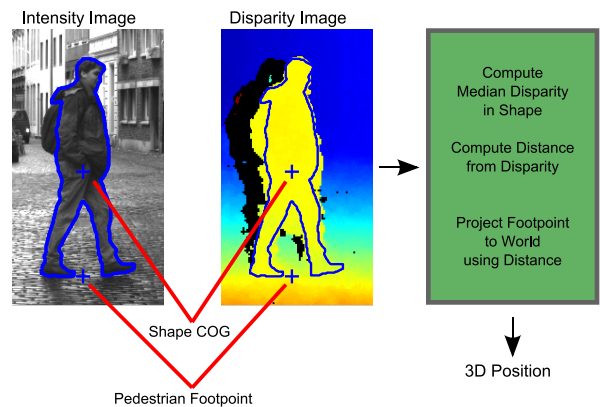


Fig. 6. Pedestrian 3D world position derived from manual labeled pedestrian shaped and dense stereo data

A. System Configuration

Parameters for the ROI generation have been chosen to correspond to pedestrians at a longitudinal distance of $10m$ to $25m$ in front of the vehicle and $\pm 4m$ in lateral direction. Pedestrians with a height of $1.6m$ up to $2.0m$ standing on the ground are searched in the detection area. To cover the detection area, ROIs ranging from $h_{min} = 72px$ to $h_{max} = 206px$ are required. ROIs with an aspect ratio of 2:1 are generated in a multi-scale sliding window fashion on the ground-plane using a flat world assumption with a pitch tolerance of $\pm 1^\circ$. Given the pitch tolerance, ROIs are located at most $11px$ above or below the ground plane. With a scale step factor $\Delta_s = 1.1$ a total of 12 scales are generated. ROI locations are shifted at fractions $\Delta_x = 0.1$ of their height and $\Delta_y = 0.25$ of their width resulting in a total of 5920 generated ROIs, see [7].

The HOG/linSVM classifiers are trained and iteratively bootstrapped, as in [7], [17]. Gradients for the HOG features are computed with $(-1, 0, 1)$ masks. Orientation histograms with 8 bins are generated from cells with a size of 8×8 pixels. Overlapping descriptor blocks (2×2) are normalized using the L_2 -norm. An initial classifier (iter0) has been trained with the positive and negative training samples described in Section IV. For both systems, this initial classifier is iteratively applied to the set of non-pedestrian images to collect additional false positives for the next round of classifier training. This process is repeated until (test) performance saturates.

B. Evaluation

For evaluation, we follow the well-established methodology of [7], [11]. To compare system output with ground-truth, we need to specify the localization tolerance, i.e. the maximum positional deviation that still allows to count the system detection as a match. This localization tolerance is the sum of an application-specific component (how precise does the object localization have to be for the application?) and a component related to measurement error (how exact can we determine true object location?). Object localization tolerance is defined (see [7], [11]) as percentage of distance, for longitudinal and lateral direction (Z and X), with respect to the vehicle. For our evaluation of the video sensing component, we use $Z = 30\%$ and $X = 10\%$, which means that, for example at $10m$ distance, we tolerate a localization error (including ground truth measurement error) of $\pm 3m$ and $\pm 1m$ in the position of the pedestrian, longitudinal and lateral to the vehicle driving direction, respectively. Partial visible pedestrians are matched in 2D with a box coverage of $\theta_n = 0.25$. Pedestrians outside the detection area or partial visible are regarded as optional and are neither credited nor penalized. For this application we allow many-to-many correspondences, i.e. a ground truth object is considered matched if there is at least one system detection matching it.

1) *Classification Performance:* Figure 7 and 8 illustrates the performance of the two systems after each bootstrapping iteration. Both classifiers improve with additional bootstrapping iterations. For the monocular system (Figure 7) per-

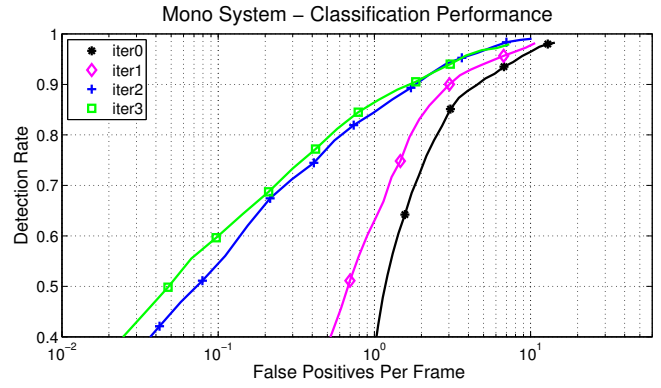


Fig. 7. Classification performance of the mono system for different bootstrapping iterations.

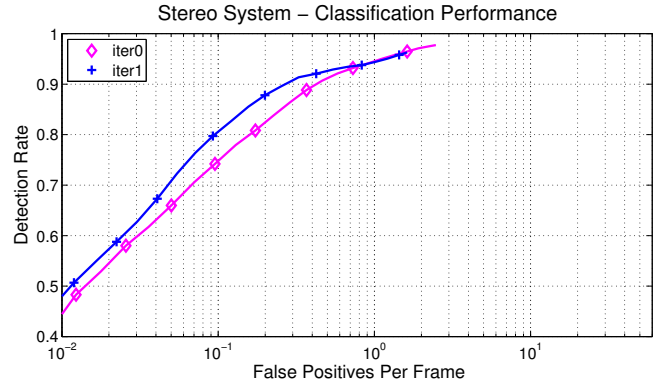


Fig. 8. Classification performance of the stereo system for different bootstrapping iterations.

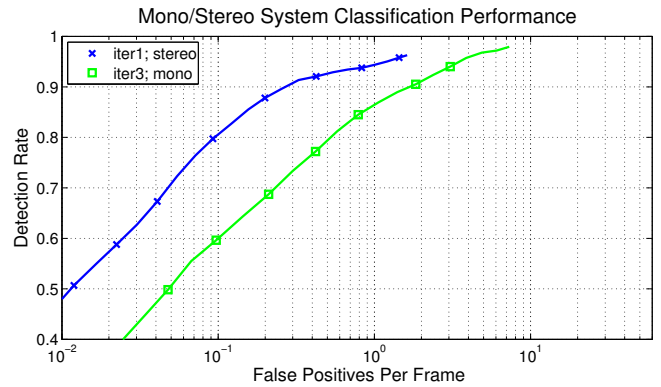


Fig. 9. Performance comparison of the mono and the stereo system.

formance saturates after three iterations. By augmenting the set of negative training samples with “difficult” examples performance is pushed by a factor of 14 at similar detection rates (60%). Because the stereo system generates ROIs only at highly structured locations the benefit of bootstrapping is less evident. After the first bootstrapping iteration performance does no longer improve.

A direct comparison of the monocular system with the stereo system (Figure 9) shows the benefit of the stereo-based ROI generation and improved localization. For a detection rate of 60% the number of false positives is reduced by a factor of 4. We attribute this to the reduced number of

		F	A	B
Mono System	Detection Rate (all)	66.58%	70.21%	78.72%
	Precision (all)	39.45%	32.50%	39.19%
	FA frame, min	0.11	13.12	11.82
Stereo System	Detection Rate (all)	58.75%	53.19%	72.34%
	Precision (all)	62.14%	50.0%	56.10%
	FA frame, min	0.02	3.05	2.68

TABLE III
SYSTEM PERFORMANCE OF THE MONO SYSTEM VS. THE STEREO
SYSTEM AFTER TRACKING.

generated ROIs containing random structures. Figures 11 and 12 illustrate some typical false positive examples of the detectors.

2) *System Performance*: Overall detection performance of the systems including the tracking module is given in Table III. Classifier thresholds are selected from Figure 9 using a common reference point of 60% detection rate. For additional insight, we consider detection rate and precision (percentage of system detections that are correct) on both the frame- and trajectory-level. For the latter, we distinguish two types of trajectories: “class-a” and “class-b” which have 50% and 1 frame entries matched. Thus, all “class-a” trajectories are also “class-b” trajectories; the different classes of trajectories represent different quality levels that might be relevant for particular applications. At comparable detection rate levels, the stereo system has a significant higher precision (approximately 20%). False alarms are reduced by a factor of 4–5 over the mono system, similar to the previous evaluation of the classification modules (see Figure 9).

VI. CONCLUSION

This paper introduced the Daimler stereo-vision pedestrian detection benchmark, and associated evaluation methodol-



Fig. 10. Examples of correct detections of the mono and stereo system.

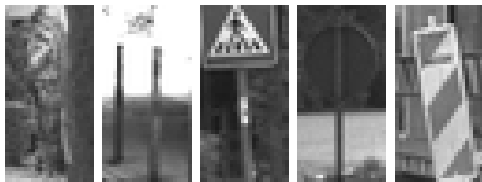


Fig. 11. Examples of false detections of the stereo system.



Fig. 12. Examples of false detections of the mono system.

ogy. The paper furthermore quantified the benefit of stereo vision for ROI generation and localization; at equal detection rates, false positives are reduced by a factor of 4-5 with stereo over mono, using the same HOG/linSVM classification component.

REFERENCES

- [1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proc. of the IEEE CVPR*, 2008.
- [2] Y. Bar-Shalom, T. Kirubarajan, and X.-R. Li. *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Inc., New York, NY, USA, 2002.
- [3] A. Broggi, A. Fascioli, I. Fedriga, A. Tibaldi, and M. D. Rose. Stereo-based preprocessing for human shape localization in unstructured environments. In *Proc. of the IEEE IV*, pages 410–415, 2003.
- [4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of the IEEE CVPR*, pages 886–893, 2005.
- [5] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. *Proc. of the IEEE CVPR*, 2009.
- [6] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrilu. Multi-cue pedestrian classification with partial occlusion handling. In *Proc. of the IEEE CVPR*, 2010.
- [7] M. Enzweiler and D. M. Gavrilu. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. on PAMI*, 31(12):2179–2195, 2009.
- [8] M. Enzweiler and D. M. Gavrilu. A multi-level Mixture-of-Experts framework for pedestrian classification. *IEEE Transactions on Image Processing*, in press, 2011.
- [9] A. Ess, B. Leibe, and L. van Gool. Depth and appearance for mobile scene analysis. In *Proc. of the ICCV*, 2007.
- [10] T. Gandhi and M. M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Trans. on ITS*, 8(3):413–430, 2007.
- [11] D. M. Gavrilu and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73(1):41–59, 2007.
- [12] D. Gerónimo. *A global approach to vision-based pedestrian detection for advanced driver assistance systems*. PhD thesis, Computer Vision Center. Barcelona (Spain), February 2010.
- [13] D. Gerónimo, A. López, A. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on PAMI*, 32(7):1239–1258, 2010.
- [14] D. Gerónimo, A. Sappa, A. López, and D. Ponsa. Adaptive image sampling and windows classification for on-board pedestrian detection. *Proc. of the IEEE ICVS*, 2007.
- [15] H. Hirschmüller. Stereo processing by semi-global matching and mutual information. *IEEE Trans. on PAMI*, 30(2):328–341, 2008.
- [16] M. Hussein, F. Porikli, and L. Davis. A comprehensive evaluation framework and a comparative study for human detectors. *IEEE Trans. on ITS*, 10:417–427, 2009.
- [17] S. Munder and D. M. Gavrilu. An experimental study on pedestrian classification. *IEEE Trans. on PAMI*, 28(11):1863–1868, 2006.
- [18] S. Nedeveschi, S. Bota, and C. Tomiuc. Stereo-based pedestrian detection for collision-avoidance applications. *IEEE Trans. on ITS*, 10(3):380–391, 2009.
- [19] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Proc. of the IEEE CVPR*, pages 193–99, 1997.
- [20] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson. A new pedestrian dataset for supervised learning. In *Proc. of the IEEE IV*, pages 373–378, Eindhoven, 2008.
- [21] H. B. Pacejka. *Tyre and Vehicle Dynamics*. SAE International, Warrendale, PA, USA, 2002.
- [22] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *Proc. of the IEEE CVPR*, pages 1–8, 2009.
- [23] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proc. of the ICCV*, pages 90–97, 2005.
- [24] B. Wu and R. Nevatia. Cluster boosted tree classifier for multi-view, multi-pose object detection. In *Proc. of the ICCV*, pages 1–8, 2007.
- [25] Q. Zhu, M. Yeh, K. Chen, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *Proc. of the IEEE CVPR*, pages 1491–1498, 2006.