Efficient Stixel-Based Object Recognition

Markus Enzweiler, Matthias Hummel, David Pfeiffer and Uwe Franke

Environment Perception Group
Daimler AG Group Research & Advanced Engineering
71059 Sindelfingen, Germany

E-Mail: firstname.lastname@daimler.com

***Abstract*— This paper presents a novel attention mechanism to improve stereo-vision based object recognition systems in terms of recognition performance and computational efficiency at the same time. We utilize the Stixel World, a compact medium-level 3D representation of the local environment, as an early focus-of-attention stage for subsequent system modules. In particular, the search space of computationally expensive pattern classifiers is significantly narrowed down. We explicitly couple the 3D Stixel representation with prior knowledge about the object class of interest, i.e. 3D geometry and symmetry, to precisely focus processing on well-defined local regions that are consistent with the environment model.**

Experiments are conducted on large real-world datasets captured from a moving vehicle in urban and rural traffic. In case of vehicle recognition as an experimental testbed, we demonstrate that the proposed Stixel-based attention mechanism significantly reduces false positive rates at constant sensitivity levels by up to a factor of 8 over state-of-the-art. At the same time, computational costs are reduced by more than an order of magnitude.

I. Introduction

Many advanced driver assistance systems (ADAS) rely on visual cues derived from camera sensors to interpret and understand their environment. A key ability is to recognize and discriminate between different object classes, such as pedestrians, bicyclists or vehicles. At the core, this problem is usually tackled using statistical pattern recognition techniques which provide powerful classification capabilities, however at a large computational burden. A typical approach starts by identifying regions-of-interest (ROIs) in the image and thereafter moves on to a more expensive pattern classification and tracking step, e.g. [6], [12].

In this paper, we aim to both improve the recognition performance and reduce the computational costs of a state-of-the-art stereo-based object recognition system. For this, we employ the Stixel World [29] as an attention stage.

The general idea of Stixel-based focus-of-attention is independent of the actual object recognition system. In this work, we use stereo-based vehicle recognition, e.g. [30], as an experimental testbed: ROIs are generated from the Stixel World and subsequently classified by a neural network using local receptive field features (NN/LRF) [6], [33]. Temporal integration is provided by an α-β tracker, see Fig. 1. Our results are expected to generalize to other object classes and pattern classifiers that are sufficiently complex to represent the large training sets, e.g. [5], [6], [17], [28], [32].

The authors thank Prof. Dr. Johannes Maucher, Stuttgart Media University, for valuable feedback and discussions.

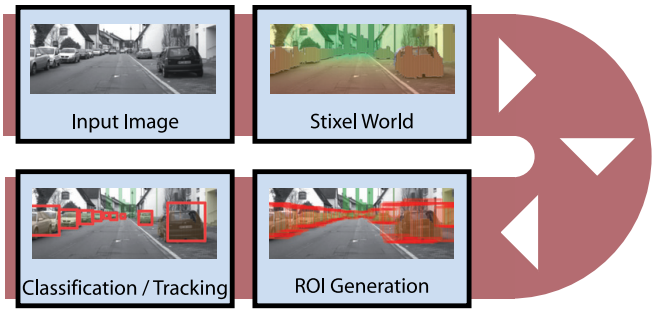

Fig. 1. Vehicle recognition system overview. The Stixel World is utilized to precisely generate regions-of-interest for subsequent classification and tracking modules.

II. Previous Work

The architecture of most vision-based object recognition systems involves three parts: a pre-processing step to select initial object hypotheses (ROI), object classification and a post-processing step to integrate classification results over time (tracking). A variety of features, classifiers and trackers to solve the classification and tracking problems have been proposed in the literature. At this point, we refer to corresponding surveys, e.g. [6], [14], [17], [30], and focus on the hypotheses generation step.

A straightforward step to obtain object hypotheses is the sliding window approach, where detector windows at various scales are shifted over the image constrained by camera geometry [6], [12], [15], [23]. Additionally, a significant speed-up results from applying an increasingly complex cascade of classifiers [9], [32], [35], [36].

Besides geometric constraints, some approaches derive cues for potential object locations directly from the image data using depth (stereo vision) [3], [8], [12], [15], [19], [20], [23], [27] or motion information [7], [21]. A more specialized attention-focusing strategy involves interest point detectors to recover regions with high information content based on local discontinuities of the image brightness function which often occur at object boundaries [4], [23], [24]. Prior knowledge about the target object class at hand, e.g.

Fig. 2. Low-level SGM disparity image (left) and medium-level Stixel World representation (right) for an urban traffic scene. The colors encode distance from the camera from near (red) to far (green).

on-road vehicles, can provide even stronger constraints on regions-of-interest. Common cues involve the shadow below a vehicle [25], [31], vehicle symmetry [22], [37] or image entropy [18].

In light of previous research, we propose a novel attention mechanism to improve a state-of-the-art stereo-based object recognition system regarding performance and computational requirements. This attentive scheme is based on an abstract super-pixel representation of the three-dimensional scene, the Stixel World [29]. The estimate of obstacles and their location within the scene derived from the Stixel representation serves as a focus-of-attention for a hypotheses generation step. Using this prior knowledge, a small number of regions-of-interest are generated precisely in both scale and position to narrow down the search space for subsequent classification and tracking. The proposed scheme based on the medium-level Stixel representation bridges the gap between pixel-level processing and high-level object classification.

Recently, a closely related approach was presented in [2] which has been developed in parallel to our work. While focusing on a novel method to estimate Stixels without computing an intermediate disparity/depth map first, the authors of [2] evaluate a Stixel-based pruning of object detections in their experiments amongst others. In [2], the raw results of a pedestrian detector are validated in terms of their overlap with the corresponding Stixels in the 2D image space. In contrast, we explicitly couple the 3D information (location, distance and height) obtained from the Stixel world with prior knowledge about the 3D dimensions of the object class of interest to effectively constrain the search space. Additionally, we present techniques to filter out implausible hypotheses in advance based on Stixel symmetry.

In this paper, we are not concerned with establishing the best absolute vehicle recognition performance, given many state-of-the-art approaches [30]. Instead, we demonstrate the relative benefits obtained from the proposed attention mechanism using both state-of-the-art depth-based ROI generation [12], [19] and the approach of [2] as an experimental baseline.

III. System Architecture

A. The Stixel World

The Stixel World [29] is a compact medium-level representation that describes the local three-dimensional environment. Stixels are defined as vertically oriented rectangles with a fixed width (e.g. $5\,$px in the image) and a variable height. From left to right, every object within the image is approximated by a set of adjacent Stixels, see Fig. 2. Hence, Stixels allow for an enormous reduction of the raw input data, e.g. approx. 400.000 disparity measurements from a $1024 \times 440\,$px stereo image pair are reduced to a few hundred Stixels only. At the same time, Stixels give easy access to the most task-relevant information such as freespace and obstacles and thus bridge the gap between low-level (pixel-based) and high-level (object-based) vision.

As proposed in [29], Stixels are extracted from a stereo image pair in two steps: the stereo computation, e.g. using semi-global matching stereo (SGM) [13], [16], and the actual Stixel computation.

The working principle of the Stixel computation [29] is closely related to other scene labeling techniques, c.f. [10], [11]. In our approach, the three-dimensional scene is segmented into two different class types, namely *ground* and *object*. Both types are expected as planar surfaces. The difference lies within their orientation: *ground* is expected as horizontal and *object* is assumed as vertical with a constant depth.

The segmentation is regularized by a set of different physically motivated world model priors, such as gravity and ordering constraints. This way, the segmentation task leads to a typical maximum a posteriori (MAP) estimation problem. Solving for the most likely and thus optimum segmentation is achieved through the use of dynamic programming [1].

B. ROI Generation

Several methods to generate ROIs in an image have been proposed, see Sec. II. Common approaches employ a monocular dense scan with a sliding detector window and ground-plane constraints, see [6] for a review, as well as additional stereo-based filtering of this ROI grid, e.g. [12], [15], [19]. Such schemes usually discard any ROIs which do not have enough support from the corresponding pixel-based disparity (depth) estimation, e.g. a threshold on the density of depth features within each ROI [12]. We use this depth-filtering method as one experimental baseline in Sec. IV.

In this work, we propose to employ the Stixel World to precisely and efficiently generate ROIs for a subsequent classification step, as follows. Each Stixel provides an estimate of 3D position and height of the underlying object (part) in the scene, see Fig. 2. We couple this estimated 3D information with 3D prior knowledge about the target object class. In particular, we generate a small set of ROIs for each Stixel,

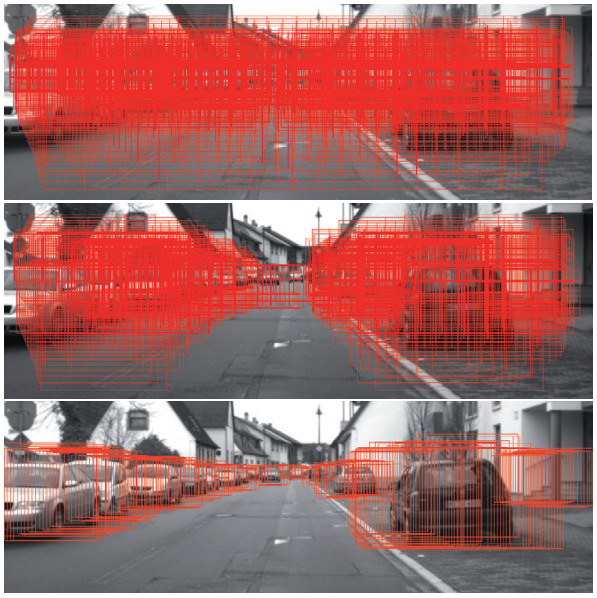

Fig. 3. ROIs generated using a monocular dense scan with ground-plane constraints [6] (top), stereo-based depth filtering [12] (center) and the proposed Stixel-based filtering scheme (bottom).

where the dimensions and locations of the ROIs in the image are constrained three-fold.

First, we assume objects to be ground-based and hence determine the vertical position of ROIs based on the planar ground model of the Stixel World. We align each ROI to the estimated Stixel bottom location in the image.

Second, we use prior knowledge about sensible 3D dimensions, e.g. vehicle width, of the object class at hand and adapt the size of ROIs accordingly. Here, we sample differently sized ROIs within a characteristic class-specific range of dimensions, in a similar fashion to regular sliding window approaches. From Fig. 2 it is observed, that the Stixel representation often (correctly) contains comparatively tall Stixels, e.g. on buildings. Hence, we additionally place a threshold on the maximum 3D height of a Stixel to assess whether that particular Stixel is to be included in the ROI generation process or not. This threshold is empirically determined on the training set, so that a sizable fraction of target class objects are covered, e.g. 99%.

Third, we combine the knowledge of 3D target geometry with the estimated distance of the Stixel to the camera to determine the scale(s) of the ROIs. An object with given 3D dimensions at a given distance has a unique back-projection to the image space.

This process is repeated for each Stixel and results in a well-defined small set of ROIs. Compared to the aforementioned monocular dense scan and additional depth-based filtering [12], the number of ROIs is reduced by approximately two orders of magnitude (vs. dense scan) and one order of magnitude (vs. depth-filtering), respectively. See Fig. 3.

The ROI generation scheme outlined above is suitable for most object classes to be detected in urban traffic, e.g. vehicles, pedestrians, etc. We propose an additional

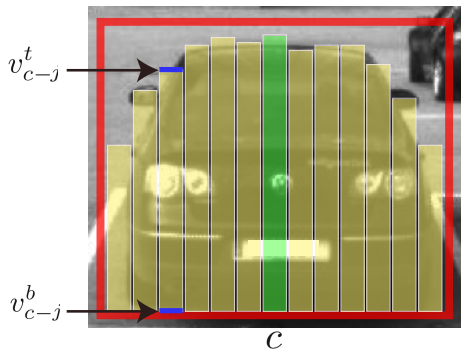

Fig. 4. Hypotheses generation using Stixels as focus-of-attention. Each Stixel serves as center of an ROI. ROI dimensions are determined using 3D constraints, see Sec. III-B. An exemplary ROI R is marked in red, generated from the Stixel marked in green. Note the symmetry of the Stixel set around the center of the object of interest.

scheme to further thin out the ROI set which is particularly tailored towards the detection of objects that exhibit a certain symmetry. In our application for example, leading vehicles are inherently symmetric. As a result, symmetry cues have been successfully applied to the vehicle detection problem [22], [37], see Fig. 4. Similar to elevating our environment model from the low-level pixel space to the medium-level Stixel World, we propose to also evaluate symmetry using the Stixel representation, as follows.

For each ROI R, we obtain n (assuming n an odd number) Stixels s_i with indices $i = 0, \ldots, n-1$ that overlap with R in the image, as indicated in Fig. 4. Further, the image row coordinates corresponding to the top and bottom points of each Stixel s_i are denoted by v^t_i and v^b_i, respectively. The index of the center Stixel within R is given by $c = \lfloor \frac{n}{2} \rfloor$. We obtain a symmetry score $f_s(R)$ for R by computing:

$$f_s(R) = \frac{1}{c} \left(\sum_{j=1}^{c} |v^b_{c-j} - v^b_{c+j}| + |v^t_{c-j} - v^t_{c+j}| \right) \qquad (1)$$

Eq. (1) effectively computes a normalized symmetry score, by comparing similarity of the geometric configuration (location and height) of corresponding Stixels, mirrored at an equal distance from the center Stixel. Lower scores $f_s(R)$ indicate higher symmetry. A threshold on the symmetry score is applied to discard ROIs with low symmetry early in the processing chain. We determine this threshold in the same fashion as the height-filtering threshold outlined above.

Our application of 3D constraints effectively reduces the number of ROIs and precisely focuses subsequent search on image regions where the Stixel World is consistent with the expectation about target-class geometry. Each ROI is subject to classification by a texture-based pattern classifier, as detailed in the following section.

C. Classification, Non-Maximum Suppression and Tracking

We consider the proposed vehicle detection framework as independent of the actual pattern classifier used. Out of a

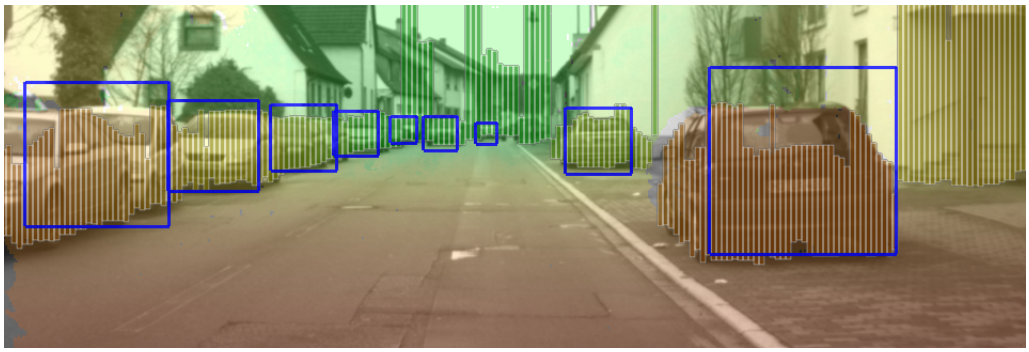

Fig. 5. Exemplary system result showing the computed SGM disparity image, the Stixel World and actual system detections after the non-maximum suppression step.

multitude of available pattern classifiers, e.g. [6], [17], [30], we chose a multi-layer neural network with 5×5 pixel local receptive field features (NN/LRF) [6], [26], [33] for experimental evaluation. Results are expected to generalize to other state-of-the-art classifiers that are sufficiently complex to solve the problem at hand. Multiple classifier responses at near-identical locations and scales are addressed by applying mean-shift-based non-maximum suppression to the recovered bounding boxes, i.e. a variant of [34]. Back-projection onto the Stixel World yields the 3D position of the detected objects.

Finally, temporal integration of detection results is employed to overcome gaps in detection and suppress spurious false positives. A 2.5D bounding box tracker is utilized, with an object state model involving bounding box position, extent and depth [6], [12]. State parameters are estimated using an α-β tracker. We acknowledge the existence of more sophisticated trackers, e.g. see [6], that are however not a focus of this paper. An exemplary result of the integrated system is shown in Fig. 5.

IV. Experiments

A. Experimental Setup

The presented approach has been tested in experiments in the field of vehicle detection. Our training set consists of manually labeled vehicle rears in images captured from a vehicle-mounted calibrated stereo camera rig in real-world urban environments. By shifting and mirroring, 96640 vehicle training samples are created from 24160 unique vehicle labels. As non-vehicle samples, 337107 random bounding boxes (with ground-plane constraints) are sampled from image regions without any vehicles. All training samples are scaled to 36×36 pixels with an 8-pixel border on each side. The test set consists of a roughly two minute real-world sequence (2853 images) captured in urban traffic. 6297 2D ground-truth locations (bounding boxes) of vehicles have been manually labeled in the images. Corresponding 3D ground-truth positions are determined by back-projection of the rectangular labels into 3D using known camera geometry and the assumption that vehicles are constrained to the ground, similar to [6]. See Fig. 6 and Table I for details.

For system evaluation, we follow the well-established methodology of [6], [12]. The comparison of 3D system output and ground-truth involves a localization tolerance, i.e. the maximum positional deviation that allows to count the system detection as a match. Following [6], [12], we define this tolerance as percentage of distance, for longitudinal and

	Description	Labeled	Jittered/Mirrored
Train:	Vehicle Rear Samples	24,160	96,640
	Non-Vehicle Samples	84,277	337,107
Test:	Number of Images	2,853	
	Ground-Truth Labels	6,297	

TABLE I
Training and test set statistics.

(a)

(b)

Fig. 6. Training and test data. (a) 36×36 pixel vehicle and non-vehicle samples used to train the NN/LRF classifier. (b) Excerpt from our urban test sequence including manually labeled ground-truth annotations.

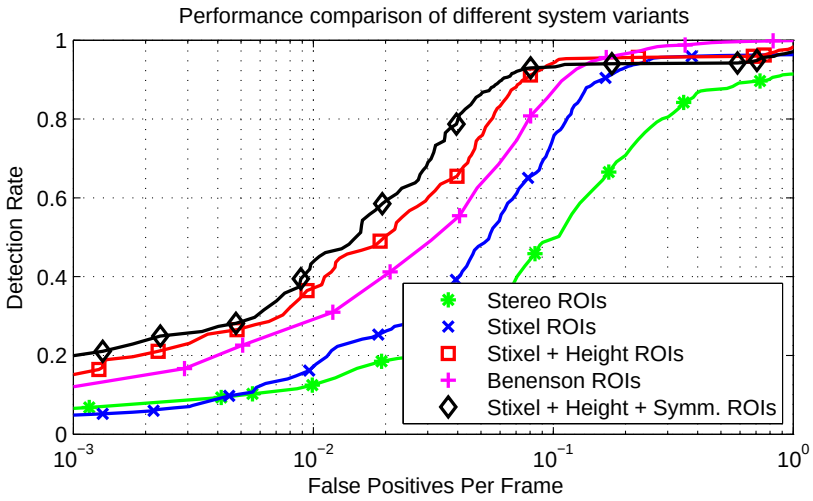

Fig. 7. ROC performance (frame-level) of the evaluated system variants.

lateral direction (Z and X), with respect to the vehicle. In our evaluation, we use $Z = 15\%$ and $X = 5\%$, which means that, for example at a distance of $10m$, we tolerate a localization error (including ground-truth back-projection error) of $\pm 1.5m$ and $\pm 0.5m$ in the position of the detected vehicle in both longitudinal and lateral direction. Ground-truth labels corresponding to vehicle frontal and side views, as well as partially occluded vehicles, are regarded as optional and are neither credited nor penalized. Additionally, we limit our evaluation to a detection range of $5m$ - $70m$ from the camera. Vehicles outside the detection area are considered optional as well. 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.

B. Frame-Level Results

In our first experiment we evaluate the performance of the proposed system on single-frame basis. As a first performance baseline, we use state-of-the-art depth-based filtering [12], as described in Sect. III-B. A system re-implementing the filtering approach presented by Benenson et al. [2] is used as a second performance baseline. Here, we combine our computation of the Stixel World (see Sect. III-A) with the filtering scheme proposed in [2]. This approach starts with a set of monocular ground-plane based ROIs, as shown in Fig. 3. A threshold is placed on the maximum allowed 2D deviation between the top (bottom) of each hypothesis in this set and the top (bottom) of the corresponding Stixel centered in this hypothesis. This margin is set to 50 pixels, the optimum value determined in [2]. All evaluated systems use the very same classifier and non-maximum suppression setup but do not incorporate tracking at this point.

To evaluate our system, we employ a three-stage evaluation procedure: We initially evaluate the Stixel-based ROI generation, as shown Sect. III-A, while disregarding height- and symmetry-filtering. Next, those filtering schemes based on 3D Stixel height and symmetry are incrementally added. Results in terms of ROC performance are given in Fig. 7. The system using depth-based filtering (*"Stereo ROIs"*) yields the worst performance in our evaluation. Stixel-based ROI generation (*"Stixel ROIs"*) considerably improves performance. Further significant performance gains are obtained by additionally incorporating height and symmetry information extracted from the Stixel World as a filter. The approach of [2] (*"Benenson ROIs"*) reaches reasonable performance, but cannot surpass our best approach. This shows the importance of our tight coupling of all available 3D information vs. the 2D ROI filtering scheme as proposed in [2]. At a constant detection rate of 80%, for example, false positives per frame are reduced by a factor of 8 between our best Stixel-based system variant and the stereo-based filtering. Compared to the approach [2], our best system variant exhibits a factor of 2 less false positives.

C. Trajectory-Level Results

For the second experiment, we select depth-filtered ROI generation (*"Stereo ROIs"*) [12] and the best performing Stixel-based system variant of the previous experiment (*"Stixel + Height + Symm. ROIs"*) and compare performance on trajectory-level after tracking. Performance is evaluated in terms of the percentage of matched ground-truth trajectories (sensitivity), the percentage of correct system trajectories (precision) and the number of false trajectories per minute, see [6]. Two types of trajectories are considered for this evaluation: *class-A* trajectories where at least 50% of the events in a trajectory have to match and *class-B* trajectories

	Stixel-Based			Stereo-Based		
	F	A	B	F	A	B
Sensitivity	68.2%	77.7%	100.0%	68.2%	77.7%	100.0%
Precision	95.7%	97.0%	97.7%	87.5%	87.6%	96.3%
FP 10^3 fr./min.	54.8	4	3	208.5	27	8

TABLE II

RESULTS OF PERFORMANCE EVALUATION ON TRAJECTORY-LEVEL.

	Number of ROIs	Avg./Image
Stereo ROIs:	16,826,788	5,897
Stixel ROIs:	3,093,035	1,084
Stixel + Height ROIs:	1,466,916	514
Stixel + Height + Symm. ROIs:	1,058,422	370
Benenson ROIs [2]:	744,407	261

TABLE III

COMPUTATIONAL COSTS OF THE EVALUATED SYSTEMS.

where at least one event has to match. Results are given in Table II. Similar to our evaluation of frame-level performance, the Stixel-based system outperforms the stereo-based system at the same sensitivity levels. The precision is considerably higher and hence the number of class-A / class-B false trajectories per minute is significantly lower (27 vs. 4 / 8 vs. 3).

D. Computational Costs

Table III examines the computational complexity of the approaches under consideration in terms of the number of ROIs generated. Regarding computational costs, the number of ROIs to classify is the dominating factor in the processing chain, given that SGM stereo and the Stixel World can be computed in real-time [29]. Compared to stereo-based ROI generation, our best Stixel-based system reduces the number of ROIs by a factor of 16 - at a better detection performance, as shown earlier. The approach of Benenson et al. [2] generates a comparatively low number of ROIs, albeit at a reduced detection performance, see Fig. 7. Note that this method has been evaluated at the optimum parameter setting, as given in [2].

V. CONCLUSION

This paper introduced a novel approach which employs the medium-level Stixel representation as a focus-of-attention stage within an integrated object recognition system. In extensive experiments in the domain of vision-based vehicle recognition, the benefit of the proposed method is quantified regarding recognition performance and computational efficiency: at equal detection rates, false positives can be reduced by up to a factor of 8 while at the same time cutting computational costs by more than an order of magnitude.

REFERENCES

[1] R. Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 6(60):503–515, 1954.

[2] R. Benenson, R. Timofte, and L. van Gool. Stixels estimation without depth map computation. *Proc. ICCV, Workshop CVVT*, 2011.

[3] M. Bertozzi and A. Broggi. GOLD: A parallel real-time stereo vision system for generic obstacle and lane detection. *IEEE Trans. on IP.*, 7:62–81, 1998.

[4] M. Bertozzi, A. Broggi, and S. Castelluccio. A real-time oriented system for vehicle detection. *J. Systems Arch.*, pages 317–325, 1997.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Proc. CVPR*, pages 886–893, 2005.

[6] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE PAMI*, 31(12):2179–2195, 2009.

[7] M. Enzweiler, P. Kanter, and D. M. Gavrila. Monocular pedestrian recognition using motion parallax. *IEEE IV Symp.*, pages 792–797, 2008.

[8] A. Ess, B. Leibe, and L. van Gool. Depth and appearance for mobile scene analysis. *Proc. ICCV*, 2007.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 32:1627–1645, 2010.

[10] P. F. Felzenszwalb and O. Veksler. Tiered scene labeling with dynamic programming. In *Proc. CVPR*, pages 3097–3104, 2010.

[11] D. Gallup, M. Pollefeys, and J.-M. Frahm. 3d reconstruction using an n-layer heightmap. In *Proc. DAGM*, pages 1–10, 2010.

[12] D. M. Gavrila and S. Munder. Multi-Cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73(1):41–59, 2007.

[13] S. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. In *Proc. ICVS*, 2009.

[14] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf. Survey on pedestrian detection for advanced driver assistance systems. *IEEE PAMI*, 32(7):1239–1258, 2010.

[15] D. Geronimo, A. D. Sappa, D. Ponsa, and A. M. Lopez. 2D-3D based on-board pedestrian detection system. *CVIU*, 114(5):583–595, 2010.

[16] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proc. CVPR*, pages 807–814, 2005.

[17] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE PAMI*, 22(1):4–37, 2000.

[18] T. Kalinke, C. Tzomakas, and W. v. Seelen. A texture-based object detection and an adaptive model-based classification. In *IEEE IV Symp.*, pages 341–346, 1998.

[19] C. Keller, M. Enzweiler, M. Rohrbach, D. F. Llorca, C. Schnörr, and D. M. Gavrila. The benefits of dense stereo for pedestrian recognition. *IEEE ITS*, 12(4):1096–1106, 2011.

[20] C. Knöppel, A. Schanz, and B. Michaelis. Robust vehicle detection at large distance using low resolution cameras. *IEEE IV Symp.*, pages 267–272, 2000.

[21] W. Krüger, W. Enkelmann, and S. Rössle. Real-time estimation and tracking of optical flow vectors for obstacle detection. *IEEE IV Symp.*, pages 304–309, 1995.

[22] A. Kuehnle. Symmetry-based recognition of vehicle rears. *Patt. Rec.*, 12(4):249–258, 1991.

[23] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE PAMI*, 30(10):1683–1698, 2008.

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[25] H. Mori and N. Charkari. Shadow and rhythm as sign patterns of obstacle detection. *IEEE Int. Symp. in Indust. Elec.*, 1993.

[26] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE PAMI*, 28(11):1863–1868, 2006.

[27] S. Munder, C. Schnörr, and D. M. Gavrila. Pedestrian detection and tracking using a mixture of view-based shape-texture models. *IEEE ITS*, 9(2):333–343, 2008.

[28] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38:15–33, 2000.

[29] D. Pfeiffer and U. Franke. Towards a global optimal multi-layer stixel representation of dense 3d data. *Proc. BMVC*, 2011.

[30] Z. Sun, G. Bebis, and R. Miller. On-road vehicle detection: A review. *IEEE PAMI*, 28(5):694–711, 2006.

[31] C. Tzomakas and W. v. Seelen. Vehicle detection in traffic scenes using shadows. Technical report, TR 98-06, Inst. f. Neural Computation, Ruhr University, Bochum, Germany, 1998.

[32] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005.

[33] C. Wöhler and J. K. Anlauf. A time delay neural network algorithm for estimating image-pattern shape and motion. *IVC*, 17:281–294, 1999.

[34] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. *Proc. CVPR*, 2009.

[35] B. Wu and R. Nevatia. Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. *Proc. CVPR*, 2008.

[36] X. Yong et al. Real-time vehicle detection based in haar features and pairwise geometrical histograms. *IEEE ICIA*, 2011.

[37] T. Zielke, M. Brauckmann, and W. v. Seelen. Intensity and edge-based symmetry detection with an application to car-following. *CVGIP*, 58(2):177–190, 1993.