The Benefits of Dense Stereo for Pedestrian Detection

Christoph G. Keller, Markus Enzweiler, Marcus Rohrbach, David Fernández Llorca, Christoph Schnörr, and Dariu M. Gavrila

***Abstract*—This paper presents a novel pedestrian detection system for intelligent vehicles. We propose the use of dense stereo for both the generation of regions of interest and pedestrian classification. Dense stereo allows the dynamic estimation of camera parameters and the road profile, which, in turn, provides strong scene constraints on possible pedestrian locations. For classification, we extract spatial features (gradient orientation histograms) directly from dense depth and intensity images. Both modalities are represented in terms of individual feature spaces, in which discriminative classifiers (linear support vector machines) are learned. We refrain from the construction of a joint feature space but instead employ a fusion of depth and intensity on the classifier level. Our experiments involve challenging image data captured in complex urban environments (i.e., undulating roads and speed bumps). Our results show a performance improvement by up to a factor of 7.5 at the classification level and up to a factor of 5 at the tracking level (reduction in false alarms at constant detection rates) over a system with static scene constraints and intensity-only classification.**

***Index Terms*—Active safety, computer vision, intelligent vehicles, pedestrian detection.**

I. Introduction

VISION-BASED pedestrian detection is a key problem in the domain of intelligent vehicles (IVs). Large variations in human pose and clothing, as well as varying backgrounds and environmental conditions, make this problem particularly challenging. The first stage in most systems consists of identifying generic obstacles as regions of interest (ROIs) using some computationally efficient method. Subsequently, a more expensive pattern classification step utilizing features from intensity images (gray scale or color) is applied.

Manuscript received September 30, 2010; revised February 22, 2011; accepted April 5, 2011. Date of publication May 12, 2011; date of current version December 5, 2011. The Associate Editor for this paper was D. J. J. Dailey.

C. G. Keller and C. Schnörr are with the Image and Pattern Analysis Group, Department of Mathematics and Computer Science, University of Heidelberg, 69120 Heidelberg, Germany (e-mail: uni-heidelberg.keller@daimler.com; schnoerr@math.uni-heidelberg.de).

M. Enzweiler is with the Environment Perception, Group Research, Daimler AG, 89081 Ulm, Germany (e-mail: markus.enzweiler@daimler.com).

M. Rohrbach is with the Computer Vision and Multimodal Computing Department, Max Planck Institute for Informatics, 66123 Saarbrücken, Germany (e-mail: mrohrbach@mpi-inf.mpg.de).

D. F. Llorca is with the Computer Engineering Department, University of Alcalá, 28871 Alcalá de Henares, Spain (e-mail: llorca@aut.uah.es).

D. M. Gavrila is with the Intelligent Autonomous Systems Group, University of Amsterdam, 1098 SJ Amsterdam, The Netherlands, and also with Daimler Research, 89081 Ulm, Germany (e-mail: dariu.gavrila@daimler.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2011.2143410

Previous IV applications have typically used sparse feature-based stereo approaches (e.g., [1], [15], and [30]) because of lower processing cost. However, with recent hardware advances, real-time dense stereo has become feasible [41] (here, we use a hardware implementation of the semiglobal matching (SGM) algorithm [13], [20]).

Both sparse and dense stereo approaches have proved suitable to dynamically estimate camera height and pitch angle to deal with road imperfections, speed bumps, car accelerations, etc. However, dense stereo also holds the potential to reliably estimate the vertical road profile. The more accurate estimation of ground location of pedestrians can be expected to improve system performance, particularly when considering undulating hilly roads.

Dense stereo can, furthermore, provide additional cues for pedestrian recognition. Up to now, the use of stereo information has been mainly limited to recovering 3-D scene structure [11], [25] and partial occlusion [8] and providing a focus-of-attention mechanism (e.g., [15], [17], [30], and [50]).

In this paper, we propose the use of dense stereo information in two modules of our pedestrian detection system: First, we estimate the varying road profile and camera orientation from dense stereo to refine ROIs with respect to possible pedestrian locations (see Section IV). Second, we enrich an intensity-based feature space with features operating on dense depth images to improve pedestrian classification performance (see Section V).

II. Previous Work

Many interesting approaches for pedestrian detection have been proposed. See [6], [9], [14], [16], [21], and [29] for relevant surveys and benchmark studies. Most benchmark studies have dealt with monocular pedestrian detection. Recently, Keller *et al.* [22] have introduced a large publicly available stereo-based pedestrian data set, involving a 27-min test drive through urban environment, including vehicle data. In terms of methods, previous work has mostly followed a module-based strategy comprising generation of possible pedestrian locations (ROIs), followed by pedestrian classification and tracking.

Various modalities (e.g., intensity, motion, and depth) are used in ROI generation to extend the sliding-window technique, where detector windows at various scales and locations are shifted over the image to obtain object hypotheses for classification. This preprocessing step is applied to reduce the number of hypotheses that are processed by a more powerful but computationally expensive classifier. In [15], the locations where the number of depth features exceeds a percentage of the

search window area are added to the ROI list for the subsequent shape detection module. In [50], a foreground region is obtained by clustering in disparity space. In [3] and [18], it is proposed that ROIs be selected by considering the x- and y-projections of the disparity space following the v-disparity representation [24]. In [1], object hypotheses are obtained by using a subtractive clustering in the 3-D space in world coordinates. Motion information is utilized in [10] as a preprocessing step for ROI generation.

Most approaches for ROI generation involve the assumption of a planar road, as well as constant camera height and pitch. Violations of these constraints are typically handled by relaxing the scene constraints, e.g., allowing a certain amount of deviation from the ground-plane assumption. Recently, some approaches for estimating road shape and camera parameters have been presented. To estimate camera height and pitch, linear fitting in the v-disparity space [32], in world coordinates [12], [17], and in the so-called virtual-disparity image [39] has been proposed. In [24], the road surface is modeled by fitting piecewise linear functions in the v-disparity space. Other approaches involve fits of quadratic polynomials [34] or clothoid functions [32] in the v-disparity space.

Regarding pedestrian classification, most approaches use discriminative models comprising a combination of intensity-based feature extraction and classification. Such features can be categorized into texture based and gradient based. Nonadaptive Haar wavelet features have been popularized by [35] and adapted by many others [28], [42], with manual [28], [35] and automatic feature selection [42]. Adaptive feature sets were proposed, e.g., local receptive fields (LRFs) [45], where the spatial structure is able to adapt to the data. Another class involves code-book patches that are extracted around salient points in the image, e.g., [25]. Gradient-based features have focused on discontinuities in image brightness. Normalized local histograms of oriented gradients have found wide use in both sparse (scale-invariant feature transform) [26] and dense representations [histogram of oriented gradient (HOG)] [4], [8], [33], [46], [49], [51]. Spatial variation and correlation of gradients have been encoded using covariance descriptors enhancing robustness toward brightness variations [40].

In terms of discriminative models, support vector machines (SVMs) are widely used in both linear [5], [8], [46], [49], [51] and nonlinear variants [28], [35]. Other popular classifiers include neural networks [15], [45] and AdaBoost cascades [27], [40], [42], [46], [47], [49], [51]. Some approaches additionally apply a component-based representation of pedestrians as an ensemble of body parts [8], [27], [28], [47].

Cascaded architectures for pedestrian detection, involving modules using different cues to narrow down the image search space, have been prevalent (e.g., [15], [17], [30], and [31]). A recent trend involves the integration of multiple features (Haar wavelets, HOG, LRF, etc.) or/and modalities (intensity, depth, motion, etc.) into a single pattern classification module [8], [33], [37], [38], [43], [46], [48]. One fusion approach involves integration of all cues into a single joint feature space [38], [43], [46]. Here, the enlarged dimensionality of the joint space can cause overfitting problems or is practically intractable, cf., [38]. Boosting approaches have also been proposed to automatically

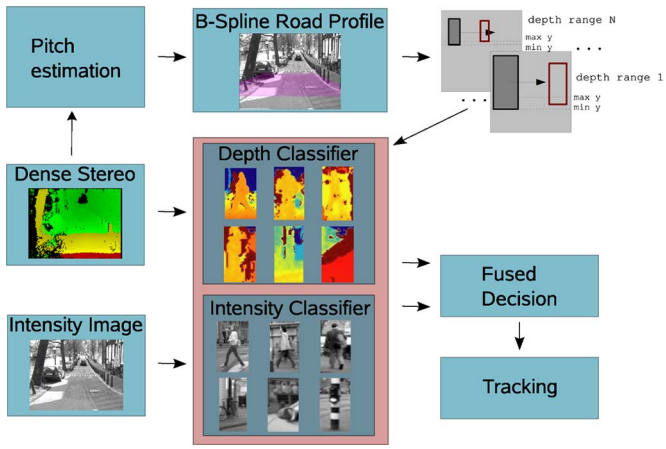


Fig. 1. Overview of the dense-stereo-based ROI generation and high-level fusion of intensity and depth classifiers. For depth images, warmer colors represent closer distances to the camera. Dense stereo is used for pitch estimation, B-Spline road profile modeling, obstacle detection, and depth-based classification.

select the "best" features from a pool of different features and modalities [46], [48]. In contrast, [8], [33], and [37] utilize fusion on the classifier level by training a specialized classifier for each feature or modality. Classifier fusion is done using fuzzy integration [33], simple classifier combination rules [37], or a mixture-of-experts framework [8].

There has been extensive work on the tracking of pedestrians to infer trajectory-level information. Most approaches apply recursive filtering of frame-level detections with additional information from different cues. For a detailed overview, see [9].

We consider the main contribution of this paper to be the use of dense stereo information in two modules of our pedestrian detection system: ROI generation and pedestrian classification. For ROI generation, we recover scene geometry in terms of camera height, camera pitch, and road profile from dense stereo information on a frame-by-frame basis. Constraints on possible pedestrian locations are dynamically derived from the recovered models of camera and road geometry. With regard to pedestrian classification, we extract spatial features from dense depth images at medium resolution (pedestrian heights up to 80 pixels) and fuse them with an intensity-based feature set on the classifier level. This paper builds upon our earlier work [23], [37] and presents an integrated pedestrian system that significantly outperforms the state of the art.

See Fig. 1 for a system overview. First, the camera pitch angle is estimated by determining the slope with the highest probability in the v-disparity map, for a reduced distance range. Second, a corridor of predefined width is computed using the vehicle velocity and the yaw rate. Only points that belong to that corridor will be used for subsequent road surface modeling. The ground surface is represented as a parametric B-Spline surface and tracked using a Kalman filter [44]. Reliability on the road profile estimation is an important issue that has to be considered for real implementations. ROIs are finally obtained by analyzing the multiplexed depth maps as in [15]. The remaining ROIs are classified using linear SVM classifiers operating on HOG features, extracted from both intensity and dense depth

data. We follow a classifier-level fusion strategy that bases the final decision on a combined vote of the individual classifiers. As opposed to fusion approaches using a joint feature space, e.g., [38], [43], and [46], this strategy does not suffer from the increased dimensionality of the joint space; see [37] and [38]. We assume our approach to generalize to other state-of-the-art features and classifiers, which are complex enough to capture the appearance of the pedestrian class; see [9]. Finally, the detected pedestrians are tracked over time.

III. Dense Stereo

With two (or more) cameras, 3-D information of the environment can be derived by finding the corresponding points across multiple cameras. A known stereo camera configuration constraints the location of corresponding image points to be on a single epipolar line. To simplify the matching process, camera images are often rectified, resulting in epipolar lines that are parallel to image lines. For a point $l(u, v)$ in the left image and the corresponding point $r(u, v)$ in the right image, the disparity $d(u, v)$ can be computed using

$$d(u, v) = l(u, v) - r(u, v). \tag{1}$$

Feature-based stereo vision systems typically provide depth measurements at points with sufficient image structure, whereas dense stereo algorithms estimate disparities at every pixel, including untextured regions. Only for regions that are visible in only one image can no disparity values be computed causing a "stereo shadow." Here, we use a hardware implementation of the SGM [20] algorithm that provides dense disparity maps in real time; see Fig. 7(b).

Given the camera geometry with focal length f and the distance between the two cameras B, dense depth maps containing distance information can be computed using

$$Z(u, v) = \frac{fB}{d(u, v)} \quad \text{at pixel } (u, v). \tag{2}$$

These dense disparity/depth maps are used for the following ROI generation, road profile estimation, obstacle detection, and pedestrian classification.

IV. Dense Stereo-Based Region-of-Interest Generation

A. Modeling of Nonplanar Road Surface

Before computing the road profile, the camera pitch angle α is estimated using the v-disparity space. We assume that the camera is installed in a way that the roll angle is insignificant. A planar road surface in the camera coordinate system can be described using

$$Y(Z) = e \cdot Z - H \tag{3}$$

with $e = \tan\alpha$ and camera height H. In v-disparity space, this road is described using

$$v(d) = ad + c \tag{4}$$

where v is the image row, and a and c are the slope and the offset that depend on the camera height and tilt angle,

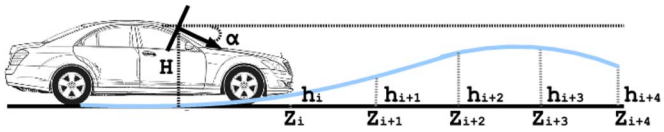


Fig. 2. Road surface modeling. Distance grid and their corresponding height values along with camera height and tilt angle.

respectively. With the assumption of a fixed camera height H, only the offset c of the line needs to be estimated in v-disparity space. Integrating the camera projection formula allows the computation of the slope

$$e(u, v) = \frac{v_0 - v}{f} + \frac{H}{Bf} d(u, v) \tag{5}$$

with the camera principal point v_0. Results are accumulated into a slope histogram, and the slope with the highest probability is selected for obtaining a first estimation of the camera pitch angle. Outliers are suppressed by computing a maximum disparity deviation for each image row, depending on the tolerance of the camera height and tilt angle.

The next step consists of computing the predicted driving corridor in front of the vehicle. This is particularly important when the vehicle is taking a curve, since most of the points in front of the vehicle do not correspond to the road. Using a single-track model with yaw-rate measurements $\dot{\psi}$ and velocity v from onboard sensors, the vehicle path can be predicted. Moving on the curve radius $r = v \cdot \dot{\psi}$, the lateral (X) and longitudinal (Z) positions in the future t are calculated as

$$X(t) = v(\dot{\psi})^{-1} \left[1 - \cos(\dot{\psi}t)\right] \tag{6}$$

$$Z(t) = v(\dot{\psi})^{-1} \sin(\dot{\psi}t). \tag{7}$$

The ROI for selecting disparity values is computed by projecting the corridor into image space using the estimated camera pitch. Here, we use a corridor of width ± 1.5 m and distance range 3–40 m in the camera coordinate system.

The road profile is represented as a parametric B-Spline surface as in [44]. B-Splines are a basis for the vector space of piecewise polynomials with degree d. The basis functions are defined on a knot vector c using equidistant knots within the observed distance interval. A simple B-Spline least square fit tries to approximate the 3-D measurements optimally. However, a more robust estimation over time is achieved by integrating the B-Spline parameter vector c, the camera pitch angle α, and the camera height H into a Kalman filter. Finally, the filter state vector is converted into a grid of distances Z_i and their corresponding road height values h_i, as depicted in Fig. 2. The number of bins of the grid will be as accurate as the B-Spline sampling.

B. Outlier Removal

In general, the method in [44] works well if the measurements provided to the Kalman filter correspond to actual road points. The computation of the corridor removes a considerable amount of object points. However, there are a few cases in which the B-Spline road modeling still leads to bad results.

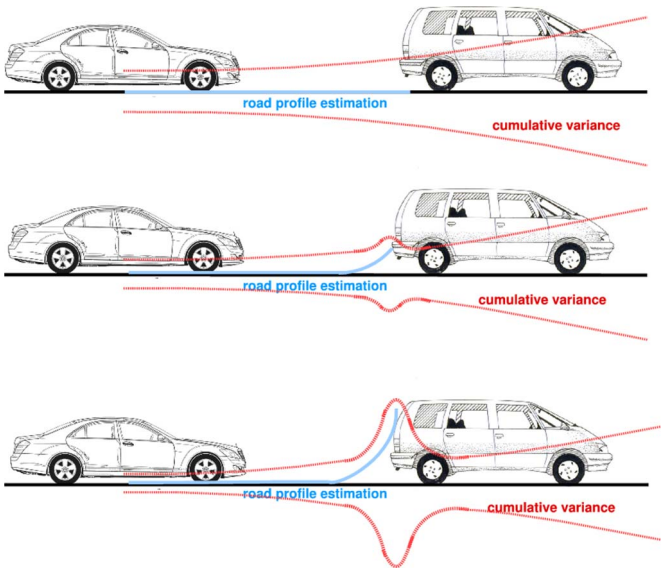


Fig. 3. Wrong road profile estimation when a vertical object appears in the corridor for a consecutive number of frames. The cumulative variance for the bin in which the vertical object is located increases, and the object points are eventually passed to the Kalman filter.

These cases are mainly caused by vertical objects (cars, motorbikes, pedestrians, cyclists, etc.) in the vicinity of the vehicle. Reflections in the windshield can cause additional correlation errors in the stereo image. If we include these points, the B-spline fitting achieves a solution that *climbs* or *wraps* over the vertical objects.

To avoid this problem, the variance of the road profile for each bin σ_i^2 is computed. Thus, if the measurements for a specific bin are out of the bounds as defined by the predicted height and the cumulative variance, they are not added to the filter. Although this alternative can deal with spurious errors, if the situation remains for a consecutive number of iterations (e.g., when there is a vehicle stopped in front of the host vehicle), the variance increases due to the inavailability of measurements, and the points pertaining to the vertical object are eventually passed to the filter as measurements. This situation is depicted in Fig. 3.

Accordingly, a mechanism is needed in to ensure that points corresponding to vertical objects are never passed to the filter. We compute the variance of all measurements for a specific bin and compare it with the expected variance in the given distance. The latter can be computed by using the associated standard deviations σ_m via error propagation from stereo triangulation [34], [44]. If the computed variance σ_i^2 is greater than the expected one σ_{ei}^2, we do not rely on the measurements but on the prediction for that bin. This is useful for cases in which there is a vertical object like the one depicted in Fig. 4.

However, in cases in which the rear part of the vertical object produces 3-D information for two consecutive bins, this approach may fail, depending on the distance to the vertical object. For example, in Fig. 5, the rear part of the vehicle yields 3-D measurements in two consecutive bins Z_i and Z_{i+1} whose variance is lower than the expected one for those bins. In this case, measurements will be added to the filter, which will yield unpredictable results.

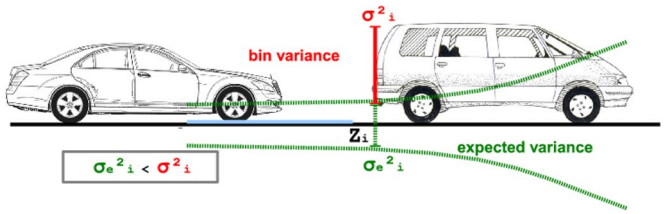


Fig. 4. Rejected measurements for bin i at distance Z_i since measurement variance σ_i^2 is greater than the expected variance σ_{ei}^2 in that bin.

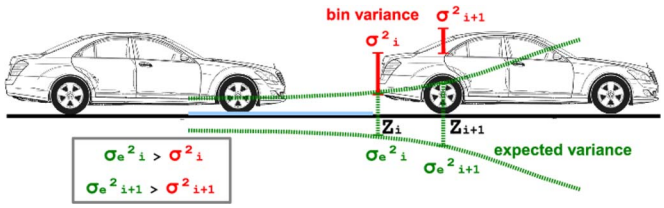


Fig. 5. Accepted measurements for bins i and $i+1$ at distances Z_i and Z_{i+1} since measurement variances σ_i^2 and σ_{i+1}^2 are lower than the expected variances σ_{ei}^2 and σ_{ei+1}^2 in these bins.

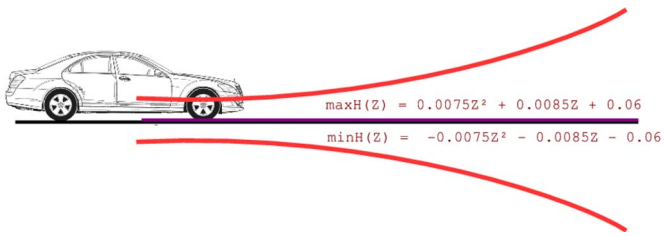


Fig. 6. Second-order polynomial function used to accept/reject measurements at all distances.

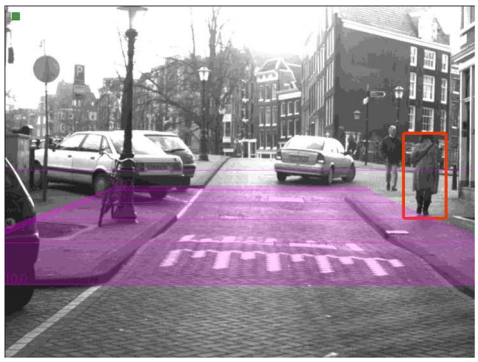
We therefore define a fixed ROI, in which we restrict measurements to lie. To that effect, we quantify the maximum road height changes at different distances and fit a second-order polynomial; see Fig. 6. The fixed region can be seen as a compromise between filter stability and response to sharp road profile changes (undulating roads). Apart from this ROI, we maintain the aforementioned test on the variance to see if measurements corresponding to a particular grid are added to the filter or not.

C. System Integration

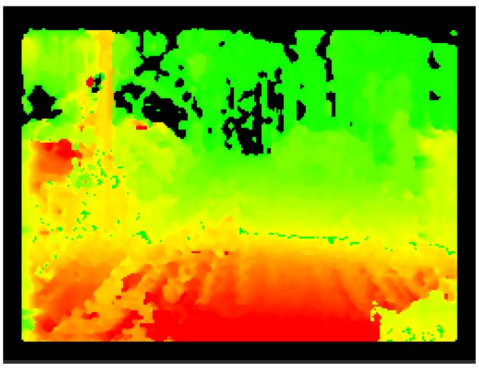
Initial ROIs R_i are generated using a sliding-window technique where detector windows at various scales and locations are shifted over the depth map. In previous work [15], the flat-world assumption along with known camera geometry restricted the search space. Pitch variations were handled by relaxing the scene constraints [15], e.g., camera pitch and camera height tolerances. In our approach, the use of dense stereo allows a reliable estimation of the vertical road profile, camera pitch, and tilt angle (see Fig. 7).

To adapt the subsequent detection modules, we compute new camera heights H_i' and pitch angles α_i' for all bins of the road profile grid. After that, standard equations for projecting 3-D points into the image plane are used.

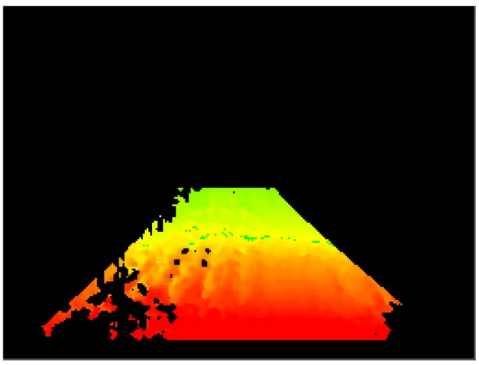
First, dense depth maps are filtered as follows: Points $P_r = (X_r, Y_r, Z_r)$ under the actual road profile, i.e., $Z_i < Z_r < Z_{i+1}$ and $Y_r < h_i$, and over the actual road profile plus the



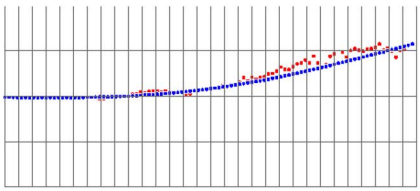
(a)



(b)



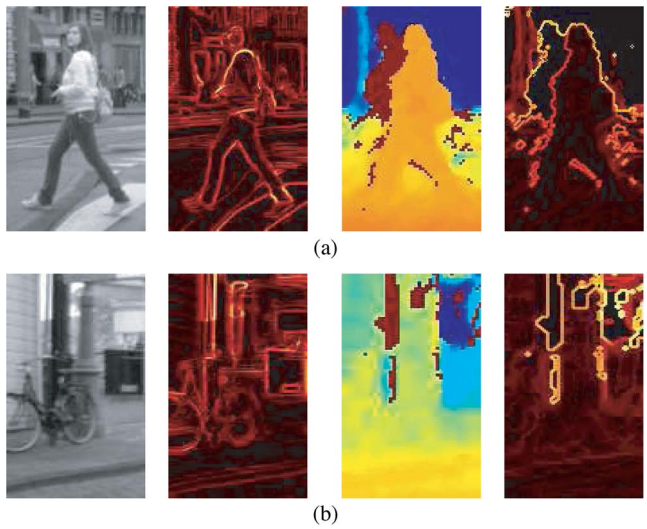
(c)



(d)

Fig. 7. System example with estimated road profile and pedestrian detection. (a) Final output with detected pedestrian marked red. The magenta area illustrates the system detection area. (b) Dense stereo image. (c) Corridor used for spline computation after outlier removal. (d) Spline (blue) fitted to the measurements (red) in system profile view.

maximum pedestrian size, i.e., $Z_i < Z_r < Z_{i+1}$ and $Y_r > h_i + H_{\max}$, are removed since they do not correspond to obstacles (possible pedestrians). The resulting filtered depth map is multiplexed into N discrete depth ranges, which are subsequently scanned with windows related to minimum and maximum extent of pedestrians. Possible window locations (ROIs) are defined according to the road profile grid (we assume that the pedestrian stands on the ground). Each pedestrian candidate



(a) (b)

Fig. 8. Intensity and depth images for (a) pedestrian and (b) nonpedestrian samples. From left to right: intensity image, gradient magnitude of intensity, depth image, gradient magnitude of depth. (a) Pedestrian. (b) Nonpedestrian.

region R_i is represented in terms of the number of depth features DF_i. A threshold θ_R governs the amount of ROIs, which are committed to the subsequent module. Only ROIs with $DF_i > \theta_R$ trigger the evaluation of the next cascade module. Others are rejected immediately.

V. Multi-Modality Classification

A. Spatial Depth and Intensity Features

Dense stereo provides disparity and depth information for most image areas, apart from regions that are visible only by one camera (stereo shadow). See the dark red areas to the left of the pedestrian torso in Fig. 8(a). Spatial features can be based on either depth Z (in meters) or disparity d (in pixels). As shown in Section III, both are inversely proportional given the camera geometry with focal length f and the distance between the two cameras B.

Objects in the scene have similar foreground/background gradients in depth space, irrespective of their location relative to the camera. In disparity space, however, such gradients are larger the closer the object is to the camera. To remove this variability, we base our spatial features on depth instead of disparity.

A visual inspection of the depth images versus the intensity images in Fig. 8 reveals distinct properties that are unique to each modality. In intensity images, lower body features (shape and appearance of legs) are the most significant features of a pedestrian (see results of part-based approaches, e.g., [28]). The texture of the pedestrian exhibits lots of gradients and characteristic structure resulting from clothing. In contrast, the upper body area has dominant foreground/background gradients and is particularly characteristic for a pedestrian in the depth image. There are no significant depth gradients on areas corresponding to the pedestrian body (we assume pedestrians to be in an upright position). Additionally, the stereo shadow is clearly visible in the upper body area (to the left of the pedestrian torso) and represents a significant local depth discontinuity. This might not

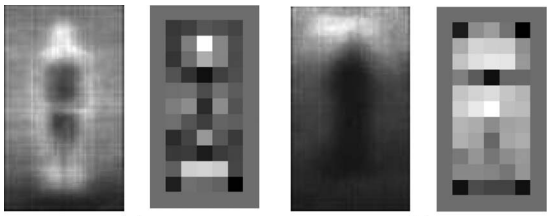


Fig. 9. Average gradient magnitude and SVM weights averaged over HOG blocks for (a) intensity and (b) depth images in the training set. (a) Intensity features. (b) Depth features.

be a disadvantage but, rather, a distinctive feature. The various salient regions in depth and intensity images motivate our use of fusion approaches between both modalities to benefit from the individual strengths; see Section V-B.

To instantiate feature spaces involving depth and intensity, we utilize well-known state-of-the-art features, which focus on local discontinuities: HOG features with a linear SVM classifier (HOG/linSVM); see [5]. We assume our approach to generalize to other state-of-the-art features and classifiers; see [9]. To get an insight into the resulting HOG features, Fig. 9 depicts the average gradient magnitude of all pedestrian training samples for both intensity and depth. We observe that the gradient magnitude is particularly high around the upper body contour for the depth image while being more evenly distributed for the intensity image. Furthermore, almost no depth gradients are present on areas corresponding to the pedestrian body. Fig. 9 further shows the weights of the linear SVM classifier after training on the corresponding feature sets. In this visualization, each "pixel" results from averaging the SVM weights over the underlying block of HOG features. In the intensity domain, HOG blocks corresponding to head/shoulder and leg regions have the highest weight. In the case of the depth features, the upper body (coarse depth contrast between foreground and background) and torso areas (uniform texture) are most indicative of a pedestrian.

B. Classifier-Level Fusion Approach

A popular strategy to improve classification is to split up a classification problem into more manageable subparts on the data level, e.g., using mixture-of-experts or component-based approaches [9]. A similar strategy can be pursued on the classifier level. Here, multiple classifiers are learned on the full data set, and their outputs are combined to a single decision. Particularly, when the classifiers involve uncorrelated features, benefits can be expected. We follow a *parallel combination* strategy [7], where multiple feature sets (i.e., based on depth and intensity; see Section V-A) are extracted from the same underlying data. Each feature set is then used as input to a single classifier, and their outputs are combined. As opposed to creating a joint feature space, classifier-level fusion does not suffer from effects related to the increased dimensionality of the joint space; see [37] and [38].

For classifier fusion, we utilize a set of fusion rules that are explained below. An important prerequisite is that the individual classifier outputs are normalized so that they can homogeneously be combined. The outputs of many state-of-the-art classifiers can be converted to an estimate of posterior probabilities [36]. We use this in our experiments.

Let $\mathbf{x}_k,\ k = 1, \ldots, n$ denote a (vectorized) sample. The posterior for the kth sample with respect to the jth object class (e.g., pedestrian and nonpedestrian), which is estimated by the ith classifier $i = 1, \ldots, m$, is given by $p_{ij}(\mathbf{x}_k)$. Posterior probabilities are normalized across object classes for each sample so that

$$\sum_j \left(p_{ij}(\mathbf{x}_k)\right) = 1. \tag{8}$$

Classifier-level fusion involves the derivation of a new set of class-specific confidence values for each data point $q_j(\mathbf{x}_k)$ out of the posteriors of the individual classifiers $p_{ij}(\mathbf{x}_k)$. The final classification decision $\omega(\mathbf{x}_k)$ results from selecting the object class with the highest confidence

$$\omega(\mathbf{x}_k) = \arg\max_j \left(q_j(\mathbf{x}_k)\right). \tag{9}$$

We consider the following fusion rules to determine the confidence $q_j(\mathbf{x}_k)$ of the kth sample with respect to the jth object class:

1) Product Rule: Individual posterior probabilities are multiplied to derive the combined confidence

$$q_j(\mathbf{x}_k) = \prod_i \left(p_{ij}(\mathbf{x}_k)\right). \tag{10}$$

2) Linear SVM Rule: A linear SVM is trained as a fusion classifier to discriminate between object classes in the space of posterior probabilities of the individual classifiers.

Let $\mathbf{p}_{jk} = (p_{1j}(\mathbf{x}_k), \ldots, p_{mj}(\mathbf{x}_k))^T$ denote the m-dimensional vector of individual posteriors for sample $\mathbf{x}_k$ with respect to the jth object class. The corresponding hyperplane is defined by

$$f_j\left(\mathbf{p}_{jk}\right) = \mathbf{w}_j \cdot \mathbf{p}_{jk} + b_j. \tag{11}$$

Here, $\mathbf{w}j$ denotes the linear SVM weight vector, b_j is a bias term, and $\cdot$ is the dot product. This linear SVM fusion rule equals a weighted sum of the individual classifier outputs, with weights and an additional bias term learned from the training set. The SVM decision value $f_j(\mathbf{p}_{jk})$ (distance to the hyperplane) is used as confidence value

$$q_j(\mathbf{x}_k) = f_j(\mathbf{p}_{jk}). \tag{12}$$

VI. Experiments

We tested our integrated pedestrian detection system on a 6:40 min (5919 images) sequence recorded from a vehicle driving through the canal area of the city of Amsterdam during the daytime. Because of the many bridges and speed bumps, the sequence is quite challenging for the road profiling module. Additionally, due to the complexity of the scenery, this sequence is very demanding for a pedestrian classifier.

Our training samples comprise nonoccluded pedestrian (in an upright position) and nonpedestrian cutouts from both intensity and corresponding depth images, which are captured from a moving vehicle in an urban environment. See Table I and

TABLE I
TRAINING SET STATISTICS. THE NUMBER OF PEDESTRIAN SAMPLES IS IDENTICAL FOR DEPTH AND INTENSITY IMAGES. NONPEDESTRIAN SAMPLES FOR INTENSITY AND DEPTH SLIGHTLY VARY DUE TO THE BOOTSTRAPPING PROCESS

	Pedestrians (labelled)	Pedestrians (jittered)	Non-Pedestrians (bootstrapped)
Training Set (intensity)	16497	296946	183501
Training Set (depth)	16497	296946	188301

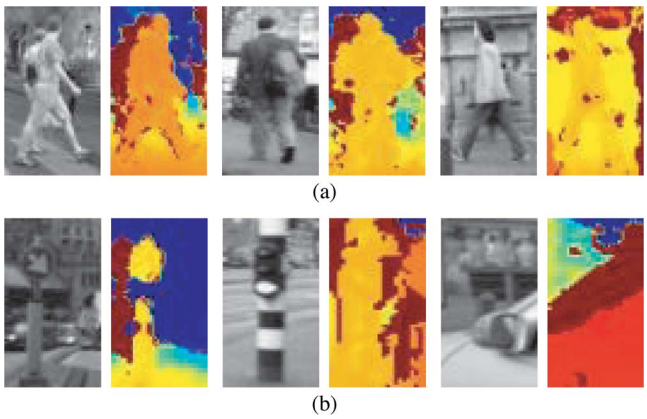


Fig. 10. Overview of (a) pedestrian and (b) nonpedestrian samples (intensity and corresponding depth images). (a) Pedestrian samples. (b) Nonpedestrian samples.

Fig. 10 for an overview. All samples are scaled to 48 × 96 pixels with an eight-pixel border to retain contour information. For each manually labeled pedestrian cutout, we randomly created 18 samples by horizontal mirroring and geometric jittering. Nonpedestrian samples were the result of a pedestrian shape detection preprocessing step with a relaxed threshold setting, i.e., containing bias toward more difficult patterns. We further applied an incremental bootstrapping technique, e.g., [10], by collecting additional false positives of the corresponding classifiers on an independent sequence and retraining the classifiers on the increased data set.

HOG features are extracted from those samples using 8 × 8 pixel cells, accumulated to 16 × 16 pixel blocks with eight gradient orientation bins. Identical feature/classifier parameters were used for intensity and depth modalities.

In our test sequence, pedestrian bounding boxes were manually labeled. Their 3-D position is obtained by triangulation in the two camera views. Only pedestrians with a distance of 12–27 m in longitudinal and ±4 m in lateral direction were considered required. Pedestrians beyond this detection area were regarded as optional, i.e., the systems are not rewarded/penalized for correct/missing detections. This results in 1684 required pedestrian single-frame instances in 66 distinct trajectories, which are required to be detected by our pedestrian detection system.

The match of a ground-truth bounding box g_i to a system alarm a_j involves bounding box coverage $\Gamma(g_i, a_j) = A(g_i \cap a_j)/A(g_i \cup a_j)$. If this ratio of intersection area and union area, is above θ_n, the ground-truth object is regarded as detected. In our experiments, we chose $\theta_n = 0.25$.

We evaluate the benefit of dense stereo on ROI generation and pedestrian classification both in isolation (see Sections VI-A and B) and in an integrated system variant (see Section VI-C). Our baseline system involves static scene geometry (flat-world assumption with fixed camera height and pitch) combined with intensity-only HOG/linSVM classification (we use the original code provided in [4]).

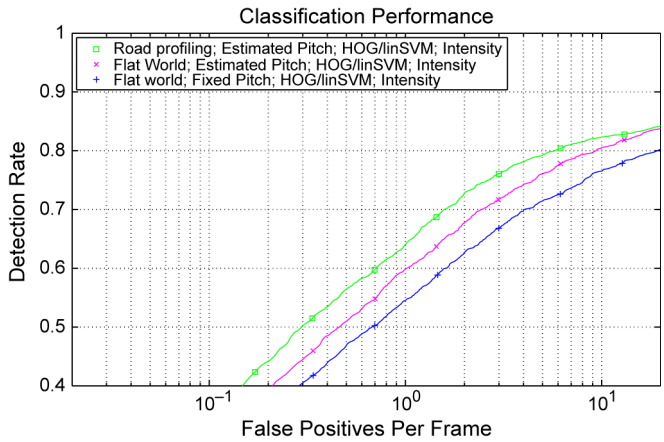


Fig. 11. Receiver operating characteristic (ROC) performance of different variants of stereo-based ROI generation combined with an intensity-only HOG/linSVM pedestrian classifier.

A. ROI Generation

The performance of the ROI generation module is evaluated in combination with the HOG/linSVM pedestrian classifier on intensity features only. Fig. 11 compares the performance of the baseline system (flat-world assumption with fixed camera height and pitch) with the proposed ROI generation technique using 1) pitch estimation with a flat-world assumption and 2) pitch estimation with road profiling. It is observed that pitch estimation (magenta ×) already improves the performance over the baseline (blue +), by distributing ROIs on a more adequate ground. An additional improvement is obtained by disregarding the flat-world assumption and estimating the actual road profile in front of the vehicle (green □). For a detection rate of, for example, 60%, the number of false positives is reduced by a factor of 2.3 using integrated pitch estimation and road profiling, compared with the baseline.

B. Multimodality Classification

Fig. 12 compares the performance of classifiers in different modalities (depth and intensity), as well as fusion strategies. All classifiers are used with the base assumption of flat world and fixed camera height and pitch, i.e., the proposed dense-stereo-based dynamic scene constraints are not (yet) in place. Our results show that a HOG/linSVM classifier on intensity features (blue +) outperforms the corresponding classifier on depth features (red ×).

The application of any proposed multimodality fusion strategies (see Section V-B) results in a significant performance boost (magenta ◇ and green □). The performance difference between both fusion strategies is only minor. At a detection rate of 60%, for example, the combined intensity–depth classifier reduces false positives by a factor of 3.3 over the intensity-only classifier. This clearly shows that the different characteristics of depth and intensity can indeed be exploited; see Section V-A.

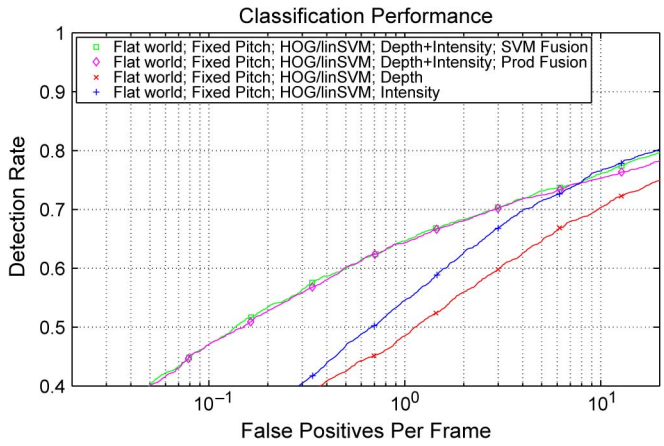


Fig. 12. ROC performance of stereo-based ROI generation combined with intensity–depth HOG/linSVM pedestrian classification.

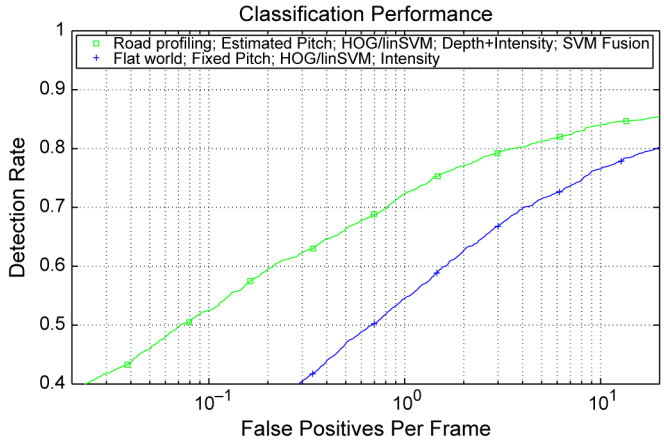


Fig. 13. ROC performance comparing the baseline system using an HOG/linSVM classifier on intensity images with the proposed system using road profiling, pitch estimation, and HOG/linSVM classifiers on depth and intensity images with SVM fusion.

C. Combined System Performance

In our next experiment, we combine the two best performing variants for ROI generation and pedestrian classification from our previous experiments: ROI generation using dense stereo based dynamic scene geometry and intensity-depth classification. Results are given in Fig. 13. The integrated system (green □) significantly boosts performance over the baseline system (blue +). At a detection rate of 60% for example, the number of false positives is reduced by a factor of 7.5, which almost equals the product (a factor of 7.6) of the individual benefits shown (factors of 2.3 for ROI generation and 3.3 for classification, respectively). This shows that the obtained performance boosts in the two different system modules are highly orthogonal to each other.

In our final experiment, we add a (rather simple) tracker to the system to obtain results on the trajectory level. We distinguish between two types of trajectories (see [15]): "class-B" and "class-A" trajectories that have at least one or at least 50% of their events matched. "class-A" trajectories include "class-B" trajectories, but the former demand stronger application performance. We compare the performance of the integrated system (dynamic scene geometry and intensity–depth classification) versus the baseline system (static scene geometry and intensity-only classification). Inputs to the tracker are pedestrian detections that were obtained from both systems by setting the classifier thresholds to correspond to a detection rate of 50% at the frame level. Nonmaximum suppression using the classifier outputs is applied to overlapping detections with a bounding box coverage of 50%. Remaining detections are tracked using a 2.5-D α–β tracker; see [15]. New tracks are started after three continuous detections and closed after two successive missed detections. Table II summarizes the performance of the two systems. The frame-level sensitivity of the system using stereo information is slightly increased compared to the baseline system. However, the main benefit lies in the reduction of false positives by a factor of approximately five. The use of dense stereo information for both road profiling and classification reduces the number of false positives per frame from 0.336 to 0.066. A comparison of the observed benefit (factor of five) to the system performance without tracking (benefit of factor 7.5) shows that tracking reduces the absolute performance differences of the systems. Similar effects have been observed in [9]. Fig. 14 illustrates system performance, including typical false positives in a cluttered image region and a missed pedestrian in a not fully upright pose.

TABLE II
SYSTEM PERFORMANCE OF THE INTEGRATED SYSTEM VERSUS THE BASELINE SYSTEM AFTER TRACKING

		F	A	B
Base System	Sensitivity	55.58%	60.53%	77.63%
	Precision	64.07%	52.36%	56.74%
	FA frame, min	0.336	40.80	37.05
Prop. System	Sensitivity	57.54%	63.16%	78.95%
	Precision	90.38%	81.71%	84.09%
	FA frame, min	0.066	9.30	8.10

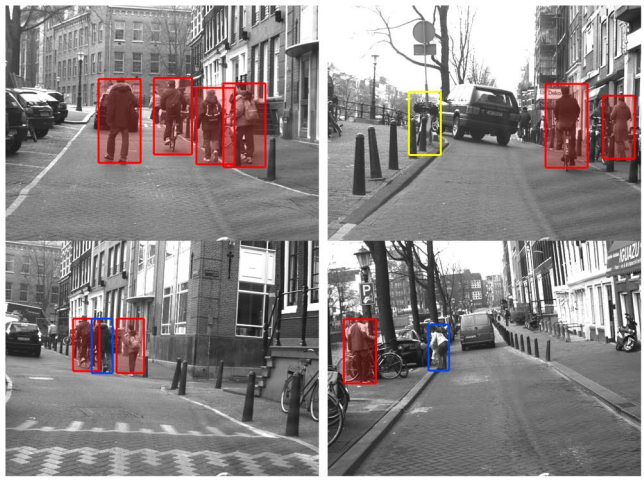


Fig. 14. Examples of system detections (red), false positives (yellow), and missed pedestrians (blue).

D. Processing Time

The hardware implementation of our SGM stereo requires 17 ms/frame. Other system components run in (unoptimized) C/C++ code on a single-core 2.66-GHz Intel CPU. Camera pitch estimation requires 3.5 ms/frame on the average with the additional road profiling taking 26 ms. With a static pitch and flat-world assumption, the ROI grid is generated only

once and reused in every frame. Incorporating pitch or road profile information requires an adjustment of the grid that takes 4 ms/frame. Depending on the configuration of earlier modules, the number of ROIs passed to the classifier varies. For the system using static pitch and flat world, about 700 ROIs per frame need to be classified on the average. Using pitch and road profile estimation, this number is reduced to about 600 ROIs per frame. HOG features need to be extracted and classified from the depth and intensity data, which doubles the costs for classification. On a multiprocessor architecture, feature extraction and classification for each modality could be processed in parallel. The processing time for any of the described rules to fuse the classifier decision values is minor and, hence, neglected. In our setup, feature extraction, classification, fusion, and tracking require approximately 500 ms/frame on the average. Note that processing costs do scale sublinearly with the number of ROIs, since feature computation can be shared among several overlapping ROIs (in the same modality), e.g., using integral histograms [51].

VII. Discussion

Our performance evaluation focused on demonstrating the relative improvements arising from the use of dense stereo, i.e., the reduction of false positives at constant sensitivity levels by a factor of 7.5 after the classification module and by a factor of 5 after the tracker, respectively. On absolute terms, the (class-B) trajectory-level system performance of approximately 80% sensitivity and 8 false detections per minute (cf. Table II) seems far from performance levels that would be necessary in a realistic application. However, this perceived performance gap, for the most part, stems from the exceeding difficulty of our test sequence (undulating roads, bridges, speed bumps, and very complex urban scenery), which was specifically chosen as a challenging test bed for the proposed road profiling module; see Section VI. Other studies have demonstrated differences of orders of magnitude in the performance of otherwise identical systems resulting from the use of different data sets, e.g., [6] and [42].

In this paper, we did not heavily optimize the feature sets with regard to the different modalities. Instead, we transferred general knowledge and experience from the behavior of features and classifiers from the intensity domain to the depth domain. At this point, it is not clear if (and how) additional modification and adaptation of the feature sets could further improve performance.

We did not particularly focus on processing time constraints in this paper. However, we do expect that software optimization and hardware implementation (e.g., digital signal processor and field-programmable gate array) can result in real-time applicability of the proposed algorithms, cf., [2] and [19]. Future work includes dealing with partially occluded pedestrians explicitly and integrating [8] into the current system.

VIII. Conclusion

We have investigated the benefits of dense stereo for a pedestrian detection system on challenging real-world data (i.e., undulated roads, bridges, and speed bumps). The improved ROI generation utilizes dense stereo data for pitch estimation, road profiling, and obstacle detection. Compared with our base system with flat-world assumption and fixed pitch, a reduction of false positives by a factor of 2.3 at similar detection rates has been demonstrated. By fusing classifier responses from different modalities (intensity and depth), we have additionally obtained a reduction of false positives by a factor of 3.3. Combining the proposed ROI generation and high-level fusion resulted in a reduction of false positives by a factor of 7.5 at the classification level and by a factor of 5 at the tracking level, respectively.

References

[1] I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, P. Revenga de Toro, J. Nuevo, M. Ocana, and M. A. G. Garrido, "Combination of feature extraction methods for SVM pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 292–307, Jun. 2007.

[2] S. Bauer, U. Brunsmann, and S. Schlotterbeck-Macht, "FPGA implementation of a HOG-based pedestrian recognition system," in *Proc. MPC-Workshop*, 2009, pp. 49–58.

[3] A. Broggi, A. Fascioli, I. Fedriga, A. Tibaldi, and M. Del Rose, "Stereo-based preprocessing for human shape localization in unstructured environments," in *Proc. IEEE Intell. Vehicles Symp.*, 2003, pp. 410–415.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. CVPR*, 2005, pp. 886–893.

[5] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. ECCV*, 2006, pp. 428–441.

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. CVPR*, 2009, pp. 304–311.

[7] R. P. W. Duin and D. M. J. Tax, "Experiments with classifier combining rules," in *Proc. 1st Int. Workshop Multiple Classifier Syst.*, 2000, pp. 16–29.

[8] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Conf. CVPR*, 2010, pp. 990–997.

[9] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[10] M. Enzweiler, P. Kanter, and D. M. Gavrila, "Monocular pedestrian recognition using motion parallax," in *Proc. IEEE Intell. Vehicles Symp.*, 2008, pp. 792–797.

[11] A. Ess, B. Leibe, and L. van Gool, "Depth and appearance for mobile scene analysis," in *Proc. ICCV*, 2007, pp. 1–8.

[12] D. Fernandez, I. Parra, M. A. Sotelo, P. Revenga, S. Alvarez, and M. Gavilan, "3D candidate selection method for pedestrian detection on non-planar roads," in *Proc. IEEE Intell. Vehicles Symp.*, 2007, pp. 1162–1167.

[13] U. Franke, S. K. Gehrig, H. Badino, and C. Rabe, "Towards optimal stereo analysis of image sequences," in *Proc. Robot Vis.*, 2008, pp. 43–58.

[14] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 413–430, Sep. 2007.

[15] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 41–59, Jun. 2007.

[16] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey on pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[17] D. Geronimo, A. D. Sappa, D. Ponsa, and A. M. Lopez, "2D–3D based on-board pedestrian detection system," *Comput. Vis. Image Understand.*, vol. 114, no. 5, pp. 583–595, May 2010.

[18] G. Grubb, A. Zelinsky, L. Nilsson, and M. Rilbe, "3D vision sensing for improved pedestrian safety," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2004, pp. 19–24.

[19] M. Hiromoto and R. Miyamoto, "Hardware architecture for high-accuracy real-time pedestrian detection with CoHOG features," in *Proc. 5th IEEE Workshop Embedded Comput. Vis.*, 2009, pp. 894–899.

[20] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[21] M. Hussein, F. Porikli, and L. Davis, "A comprehensive evaluation framework and a comparative study for human detectors," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 417–427, Sep. 2009.

[22] C. Keller, M. Enzweiler, and D. M. Gavrila, "A new benchmark for stereo-based pedestrian detection," in *Proc. IEEE Intell. Vehicles Symp.*, Baden-Baden, Germany, 2011.
[23] C. G. Keller, D. F. Llorca, and D. M. Gavrila, "Dense stereo-based ROI generation for pedestrian detection," in *Proc. DAGM Symp. Pattern Recog.*, 2009, pp. 81–90.
[24] R. Labayrade, D. Aubert, and J. P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation," in *Proc. IEEE Intell. Vehicles Symp.*, 2002, pp. 646–651.
[25] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool, "Dynamic 3D scene analysis from a moving vehicle," in *Proc. IEEE Conf. CVPR*, 2007, pp. 1–8.
[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[27] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. ECCV*, 2004, pp. 69–81.
[28] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.
[29] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.
[30] S. Munder, C. Schnörr, and D. M. Gavrila, "Pedestrian detection and tracking using a mixture of view-based shape-texture models," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 2, pp. 333–343, Jun. 2008.
[31] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 380–391, Sep. 2009.
[32] S. Nedevschi, R. Danescu, D. Frentiu, T. Marita, F. Oniga, C. Pocol, T. Graf, and R. Schmidt, "High accuracy stereovision approach for obstacle detection on non-planar roads," in *Proc. IEEE INES*, 2004, pp. 211–216.
[33] L. Oliveira, U. Nunes, and P. Peixoto, "On exploration of classifier ensemble synergism in pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 16–27, Mar. 2010.
[34] F. Oniga, S. Nedevschi, M. M. Meinecke, and T. B. To, "Road surface and obstacle detection based on elevation maps from dense stereo," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2007, pp. 859–865.
[35] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, Jun. 2000.
[36] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 1999, pp. 61–74.
[37] M. Rohrbach, M. Enzweiler, and D. M. Gavrila, "High-level fusion of depth and intensity for pedestrian classification," in *Proc. DAGM Symp. Pattern Recog.*, 2009, pp. 101–110.
[38] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proc. ICCV*, 2009, pp. 24–31.
[39] N. Suganuma and N. Fujiwara, "An obstacle extraction method using virtual disparity image," in *Proc. IEEE Intell. Vehicles Symp.*, 2007, pp. 456–461.
[40] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on Riemannian manifolds," in *Proc. IEEE Conf. CVPR*, 2007, pp. 1–8.
[41] W. Van der Mark and D. M. Gavrila, "Real-time dense stereo for intelligent vehicles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, no. 1, pp. 38–50, Mar. 2006.
[42] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, Jul. 2005.
[43] X. Wang, T. X. Han, and S. Yan, "A HOG-LBP human detector with partial occlusion handling," in *Proc. ICCV*, 2009, pp. 32–39.
[44] A. Wedel, H. Badino, C. Rabe, H. Loose, U. Franke, and D. Cremers, "B-spline modeling of road surfaces with an application to free-space estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 572–583, Dec. 2009.
[45] C. Wöhler and J. K. Anlauf, "A time delay neural network algorithm for estimating image-pattern shape and motion," *Image Vis. Comput.*, vol. 17, no. 3/4, pp. 281–294, Mar. 1999.
[46] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. IEEE Conf. CVPR*, 2009, pp. 794–801.
[47] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, Nov. 2007.
[48] B. Wu and R. Nevatia, "Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection," in *Proc. IEEE Conf. CVPR*, 2008, pp. 1–8.
[49] L. Zhang, B. Wu, and R. Nevatia, "Detection and tracking of multiple humans with extensive pose articulation," in *Proc. ICCV*, 2007, pp. 1–8.
[50] L. Zhao and C. Thorpe, "Stereo and neural network-based pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 3, pp. 148–154, Sep. 2000.
[51] Q. Zhu, M. Yeh, K. Chen, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Conf. CVPR*, 2006, pp. 1491–1498.

Christoph G. Keller received the Dipl.-Inf. degree in computer science from the University of Freiburg, Freiburg, Germany, in 2007. He is currently working toward the Ph.D. degree with the Image and Pattern Analysis Group, University of Heidelberg, Heidelberg, Germany, while on site at Daimler Research, Ulm, Germany.

In 2006 and 2007, he was a visiting Student Researcher with Siemens Corporate Research, Princeton, NJ. His current research focuses on the detection and tracking of humans with application to pedestrian recognition in the domain of intelligent vehicles.

Markus Enzweiler received the M.Sc. degree in computer science from the University of Ulm, Ulm, Germany, in 2005. He is currently working toward the Ph.D. degree with the Image and Pattern Analysis Group, University of Heidelberg, Heidelberg, Germany.

In 2002 and 2003, he was a visiting Student Researcher with the Centre for Vision Research, York University, Toronto, ON, Canada. Since 2010, he has been a Research Scientist with Environment Perception, Group Research, Daimler AG, Ulm, Germany. His current research focuses on statistical models of human appearance with application to pedestrian recognition in the domain of intelligent vehicles.

Mr. Enzweiler is a recipient of a Ph.D. scholarship from the Studienstiftung des deutschen Volkes (German National Academic Foundation).

Marcus Rohrbach received the M.Sc. degree in computer science from the University of Technology Darmstadt, Germany, in 2009. He is currently working toward the Ph.D. degree with the Computer Vision and Multimodal Computing Department, Max Planck Institute for Informatics, Saarbrücken, Germany.

From 2006 to 2007, he was with the Computer Science Department, University of British Columbia, Vancouver, BC, Canada, as a visiting student. From 2008 to 2009, he was a Student Researcher on pedestrian recognition with Daimler Research, Ulm, Germany. His current research explores the benefits of combining language and visual resources to enable scalable recognition systems.

David Fernández Llorca received the M.Sc. and Ph.D. degrees in telecommunications engineering from the University of Alcalá (UAH), Alcalá de Henares, Spain, in 2003 and 2008, respectively.

He is currently an Associate Professor with the Computer Engineering Department, UAH. His research interests are mainly focused on computer vision and intelligent transportation systems.

Dr. Llorca received the Best Ph.D. Award from UAH, the Best Research Award in the domain of Automotive and Vehicle Applications in Spain in 2008, the 3M Foundation Awards under the category of eSafety in 2009,the Master Thesis Award in eSafety from the Driver Assistance (ADA Foundation) Lectureship of the Technical University of Madrid, Madrid, Spain, in 2004, and the Best Telecommunication Engineering Student Award from IVECO Company in 2004.

Christoph Schnörr received the Dipl.-Ing. degree in electrical engineering and the Dr.rer.nat. degree in computer science from the Technical University of Karlsruhe, Karlsruhe, Germany, and the Habilitation degree in computer science from the University of Hamburg, Hamburg, Germany, in 1987, 1991, and 1998, respectively.

He has been a Full Professor with the University of Heidelberg, Heidelberg, Germany, since 2008, where he is the Head of the Image and Pattern Analysis Group, a Codirector of the Heidelberg Collaboratory for Image Processing, which is jointly funded by the German Science Foundation (DFG) and industrial partners, the Principal Investigator of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, and a speaker of the recently established Research Training Group on Probabilistic Graphical Models and Applications in Image Analysis, which is funded by the DFG. His research interests include image processing, computer vision and pattern analysis, and corresponding problems of mathematical modeling and optimization.

Dariu M. Gavrila received the M.Sc. degree in computer science from the Free University, Amsterdam, The Netherlands, in 1990 and the Ph.D. degree in computer science from the University of Maryland, College Park, in 1996.

He was a visiting Researcher with the Media Laboratory, Massachusetts Institute of Technology, Cambridge, in 1996. Since 1997, he has been a Senior Research Scientist with Daimler Research, Ulm, Germany. In 2003, he was named Professor with the Faculty of Science, University of Amsterdam, chairing the area of Intelligent Perception Systems (part time). Over the last decade, he has focused on visual systems for detecting human presence and recognizing activity, with application to intelligent vehicles and surveillance. He has published numerous papers in this area.

Dr. Gavrila received the I/O Award in 2007 from the Netherlands Organisation for Scientific Research.