Dense Stereo-based ROI Generation for Pedestrian Detection

C. G. Keller[1], D. F. Llorca[2] and D. M. Gavrila[3,4]

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

[2]Department of Electronics. Univ. of Alcalá. Alcalá de Henares (Madrid), Spain

[3] Environment Perception, Group Research, Daimler AG, Ulm, Germany

[4]Intelligent Systems Lab, Fac. of Science, Univ. of Amsterdam, The Netherlands

{uni-heidelberg.keller,dariu.gavrila}@daimler.com llorca@depeca.uah.es

Abstract. This paper investigates the benefit of dense stereo for the ROI generation stage of a pedestrian detection system. Dense disparity maps allow an accurate estimation of the camera height, pitch angle and vertical road profile, which in turn enables a more precise specification of the areas on the ground where pedestrians are to be expected. An experimental comparison between sparse and dense stereo approaches is carried out on image data captured in complex urban environments (i.e. undulating roads, speed bumps). The ROI generation stage, based on dense stereo and specific camera and road parameter estimation, results in a detection performance improvement of factor five over the state-of-the-art based on ROI generation by sparse stereo. Interestingly, the added processing cost of computing dense disparity maps is at least partially amortized by the fewer ROIs that need to be processed at the system level.

1 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 first stage in most systems consists of identifying generic obstacles as regions of interest (ROIs) using a computationally efficient method. Subsequently, a more expensive pattern classification step is applied.

Previous IV applications have typically used sparse, feature-based stereo approaches (e.g. [9, 1]) because of lower processing cost. However, with recent hardware advances, real-time dense stereo has become feasible [12] (here we use a hardware implementation of the semi-global matching (SGM) algorithm [7]). Both sparse and dense stereo approaches haved proved suitable to dynamically estimate camera height and pitch angle, in order to deal with road imperfections, speed bumps, car accelerations, etc. Dense stereo, furthermore, holds the potential to also reliably estimate the vertical road profile (which feature-based stereo, due to its sparseness does not). The more accurate estimation of ground location of pedestrians can be expected to improve system performance, especially when considering undulating, hilly roads.

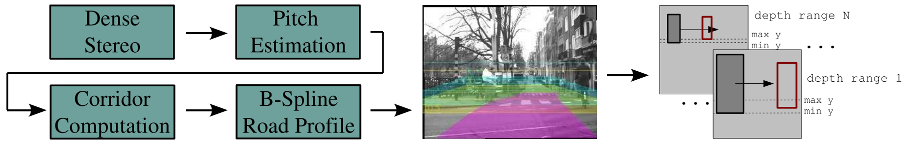


Fig. 1. Overview of the dense stereo-based ROI generation system comprising dense stereo computation, pitch estimation, corridor computation, B-Spline road profile modeling and multiplexed depth maps scanning with windows related to minimum and maximum extents of pedestrians.

The aim of this paper thus is to investigate the advantages of dense vs. sparse disparity maps when detecting generic obstacles in the early stage of a pedestrian detection system [9]. We are interested in both the ROC performance (trade-off correct vs. false detections) and in the processing cost.

2 Related Work

Many interesting approaches for pedestrian detection have been proposed. See [4] for a recent survey and a novel publicly available benchmark set. Most work has proceeded with a learning-based approach by-passing a pose recovery step and describing human appearance directly in terms of low-level features from a region of interest (ROI). In this paper, we concentrate on the stereo-based ROI generation stage.

The simplest technique to obtain object location hypotheses is the sliding window technique, where detector windows at various scales and locations are shifted over the image. This approach in combination with powerful classifiers (e.g. [13, 3, 16]) is currently computationally too expensive for real-time applications. Significant speed-ups can be obtained by including application-specific constraints such as flat-world assumption, ground-plane based objects and common geometry of pedestrians, e.g. object height or aspect ratio [9, 17].

Besides monocular techniques (e.g. [5]), which are out of scope in this work, stereo vision is an effective approach for obtaining ROIs. In [20] a foreground region is obtained by clustering in the disparity space. In [2, 10] ROIs are selected considering the x- and y-projections of the disparity space following the v-disparity representation [11]. In [1] object hypotheses are obtained by using a subtractive clustering in the 3D space in world coordinates.

Either monocular or stereo, most approaches are carried out under the assumption of a planar road and no camera height and camera pitch angle variations. In recent literature on intelligent vehicles many interesting approaches have been proposed to perform road modeling and to estimate camera pitch angle and camera height. Linear fitting in the v-disparity [14], in world coordinates [6] and in the so-called virtual-disparity image [18] has been proposed to estimate the camera pitch angle and the camera height. In [11] the road surface is modeled by the fitting of the envelope of piecewise linear functions in the v-disparity space. Other approaches are performed by fitting of a quadratic polynomial [15] or a clothoid function [14] in the v-disparity space as well.

Building upon this work, we propose the use of dense stereo vision for ROI generation in the context of pedestrian detection. Dense disparity maps are provided in real-time [7]. Firstly, camera pitch angle is estimated by determining the slope with highest probability in the v-disparity map, for a reduced distance range. Secondly, a corridor of a 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. Then, the ground surface is represented as a parametric B-Spline surface and tracked by using a Kalman filter [19]. Reliability on the road profile estimation is an important issue which has to be considered for real implementations. ROIs are finally obtained by analyzing the multiplexed depth maps as in [9] (see Figure 1).

3 Dense Stereo-Based ROI Generation

3.1 Modeling of non-planar road surface

Feature-based stereo vision systems typically provide depth measurements at points with sufficient image structure, whereas dense stereo algorithms estimate disparities at all pixels, including untextured regions, by interpolation.

Before computing the road profile, the camera pitch angle is estimated by using the v-disparity space. We assume that the camera is installed such that the roll angle is insignificant. Then, the disparity of a planar road surface (this assumption can be accepted in the vehicle vicinity) can be calculated by:

$$d(v) = a \cdot v + b \tag{1}$$

where v is the image row and a, b are the slope and the offset which depend on camera height and tilt angle respectively. Both parameters can be estimating using a robust estimator. However, if we assume a fixed camera height we can compute a slopes histogram and determine the slope with the highest probability, obtaining a first estimation of the camera pitch angle. In order to put only good candidates into the histogram, a disparity range is calculated for each image row, depending on the tolerance of the camera height and tilt angle.

The next step consists in computing a corridor of a pre-defined width using the vehicle velocity, the yaw rate, the camera height and the camera tilt angle. If the vehicle is stopped, a fixed corridor is used. In this way, a considerable amount of object points are not taken into account when modeling the road surface. This is particularly important when the vehicle is taking a curve, since most of the points in front of the vehicle correspond to object points.

The road profile is represented as a parametric B-Spline surface as in [19]. 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 3D 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 and their corresponding road height values as depicted in Figure 2. The number of bins of the grid will be as accurate as the B-Spline sampling.

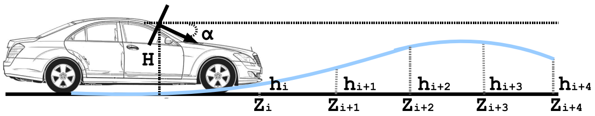


Fig. 2. Road surface modeling. Distances grid and their corresponding height values along with camera height and tilt angle.

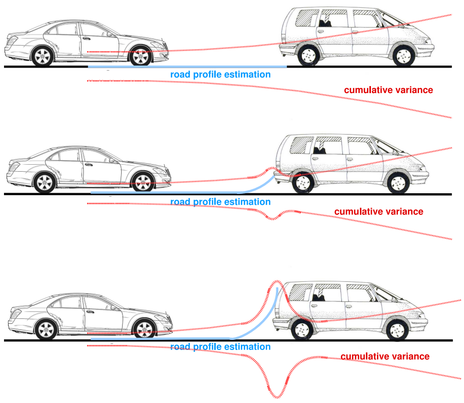


Fig. 3. Wrong road profile estimation when a vertical object appears in the corridor for a consecutive number of frames. The cumulative variance for the bin in which the vertical object is located increases and the object points are eventually passed to the Kalman filter.

3.2 Outlier Removal

In general, the method of [19] 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. 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 which *climbs* or *wraps* over the vertical objects.

In order 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 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 Figure 3.

Accordingly, a mechanism is needed in order to ensure that points corresponding to vertical objects are never passed to the filter. We compute the

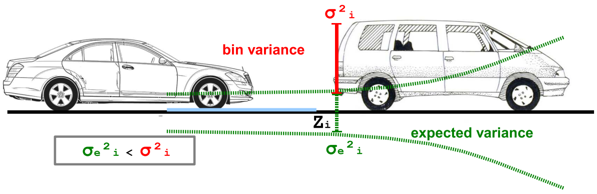


Fig. 4. Rejected measurements for bin i at distance Z_i since measurements variance σ_i^2 is greater than the expected variance σ_{ei}^2 in that bin.

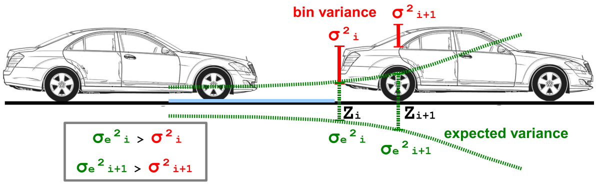


Fig. 5. Accepted measurements for bins i and $i+1$ at distances Z_i and Z_{i+1} since measurements variances σ_i^2 and σ_{i+1}^2 are lower than the expected variances σ_{ei}^2 and σ_{ei+1}^2 in these bins.

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 associate standard deviations σ_m via error propagation from stereo triangulation [15, 19]. 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 in the example depicted in Figure 4.

However, in cases in which the rear part of the vertical object produces 3D information for two consecutive bins, this approach may fail depending on the distance to the vertical object. For example, in Figure 5 the rear part of the vehicle yields 3D 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.

We therefore define a fixed region of interest, in which we restrict measurements to lie. To that effect, we quantify the maximum road height changes at different distances and we fit a second order polynomial, see Figure 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 region of interest, we maintain the beforementioned test on the variance, to see if measurements corresponding to a particular grid are added or not to the filter.

3.3 System Integration

Initial ROIs R_i are generated using a sliding windows technique where detector windows at various scales and locations are shifted over the depth map. In previous works [9] flat-world assumption along with known camera geometry were

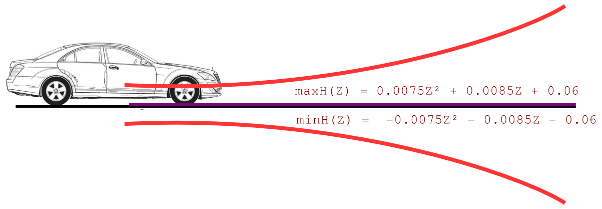

maxH(Z) = 0.0075Z² + 0.0085Z + 0.06
minH(Z) = -0.0075Z² - 0.0085Z - 0.06

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

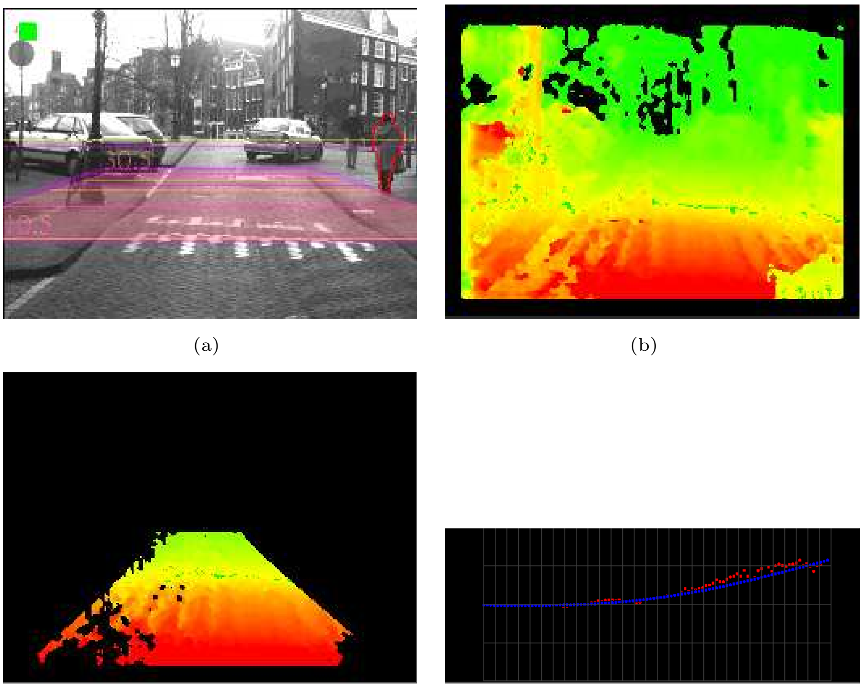
used, so that, the search space was drastically restricted. Pitch variations were handled by relaxing the scene constraints [9], e.g., including camera pitch and camera height tolerances. However, thanks to the use of dense stereo a reliable estimation of the vertical profile of the road is computed along with the camera pitch and tilt angle.

In order to easily 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 3D points into the image plane can be used.

First of all 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 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 the pedestrian stands on the ground). Each pedestrian candidate 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.

Pedestrian recognition proceeds with shape-based detection, involving coarse-to-fine matching of an exemplar-based shape hierarchy to the image data at hand [9]. Positional initialization is given by the output ROIs of the dense stereo-based ROI generation stage. The shape hierarchy is constructed off-line in an automatic fashion from manually annotated shape labels. On-line matching involves traversing the shape hierarchy with the Chamfer distance between a shape template and an image sub-window as smooth and robust similarity measure. Image locations, where the similarity between shape and image is above a user-specified threshold, are considered detections. A single distance threshold applies for each level of the hierarchy. Additional parameters govern the edge density on which the underlying distance map is based.

Detections of the shape matching step are verified by a texture-based pattern classifier. We employ a multi layer feed-forward neural network operating on local adaptive receptive field features [9]. Finally, temporal integration of detection results is employed to overcome gaps in detection and suppress spurious false positives. A 2D bounding box tracker is utilized, with an object state model involving bounding box position and extent [9]. State parameters are estimated using an $\alpha - \beta$ tracker.



(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 profile view.

4 Experiments

We tested our dense stereo-based ROI generation scheme on a 5 min (3942 image) sequence recorded from a vehicle driving through the canal area of the city of Amsterdam. Because of the many bridges and speed bumps, the sequence is quite challenging for the road profiling component. Pedestrians were manually labeled; their 3D position was obtained by triangulation in the two camera views. Only pedestrians located in front of the vehicle in the area 12-27m in longitudinal and ±4m in lateral direction were considered required. Pedestrians beyond this detection area were regarded as optional. Localization tolerance is selected as in [9] to be $X = 10\%$ and $Z = 30\%$ as percentage of distance for lateral (X)

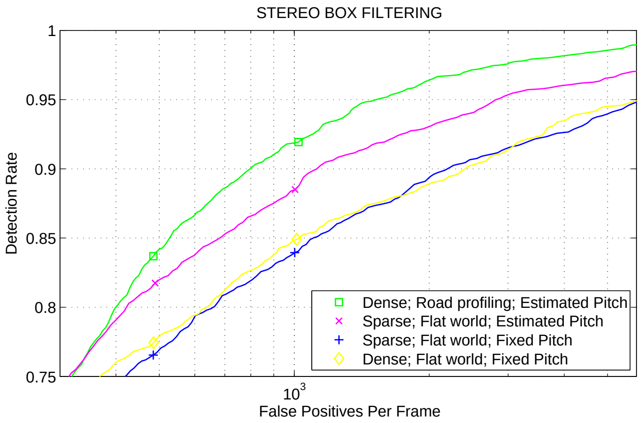


Fig. 8. ROC peformance of stereo-based ROI generation module for different variations.

	FPs/Frame	**# ROIs/Frame**
Dense - Road Profiling	1036	1549
Sparse - Pitch Estimation	1662	2345
Sparse - Fixed Pitch	3367	4388
Dense - Fixed Pitch	3395	4355

Table 1. Comparison of the number of false positives and total number of generated ROIs per frame for an exemplary threshold θ_R resulting in a detection rate of 92%

and longitudinal (Z) direction. In all, this resulted in 1684 required pedestrian single-frame instances in 66 distinct trajectories, to be detected by our pedestrian system. See Figure 7 for an illustration of the results.

We first examined the performance of the ROI generation module in isolation, see Figure 8. Shown are the ROCs (correctly vs. falsely passed ROIs) for various configurations (dense vs. sparse stereo, w/out pitch angle and road profile estimation). No significant performance difference can be observed between dense- or sparse- stereo-based ROI generation when neither pitch angle nor road profile is estimated. Estimating the pitch angle leads however to a clear performance improvement. Incorporating the estimated road profile yields an additional performance gain.

The total number of generated ROIs and false positives for an exemplary detection rate of 92% are summarized in table 1. The number of ROIs that need to be generated can be reduced by a factor of 2.8 when utilizing road profile information compared to a system with static camera position. Using camera pose information leads to an reduction of generated ROIs by a factor of 1.87. A reduced number of generated ROIs implies fewer computations in later stages of our detection system, and thus faster processing speed (approx. linear in number of ROIs).

We now turn to the evaluation on the overall system level, i.e. with the various ROI generation schemes integrated in the pedestrian classification and tracking system of [9]. Relevant module parameters (in particular density threshold θ_R for stereo-based ROI generation) were optimized for each system configuration following the ROC convex hull technique described in [9].

See Figure 9. One observes the relative ranking of the various ROI generation schemes is maintained cf. Figure 8 (the dense stereo, fixed pitch and flat world case is not plotted additionally, as is has similar performance as the equivalent sparse-stereo case). That is, there is a significant benefit of estimating pitch angle, camera height and road profile, i.e. a performance improvement of factor 5.

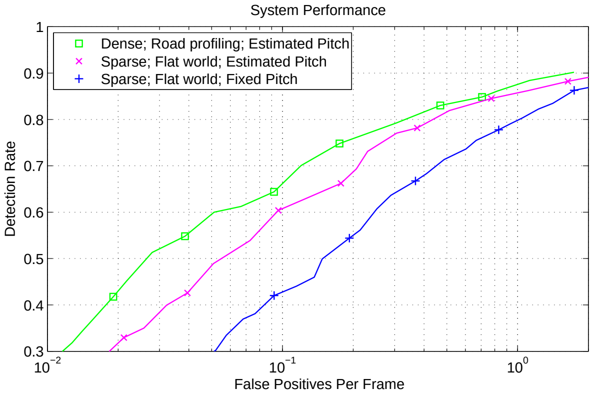


Fig. 9. Overall performance of system configurations with different ROI generation stages.

5 Conclusions

We investigated the benefit of dense stereo for the ROI generation stage of a pedestrian detection system. In challenging real-world sequences (i.e. undulated roads, bridges and speed bumps), we compared various versions of dense and sparse stereo-based ROI generation. For the case of flat world assumption and fixed camera parameters, sparse and dense stereo provided equal ROI generation performance (baseline configuration). The specific estimation of camera height and pitch angle resulted in a performance improvement of about factor three (reduction false positives at same correct detection rate). When estimating road surface as well, the benefit increased to a factor of five vs. the baseline configuration. Interestingly, the added processing cost of computing dense, rather than sparse, disparity maps is at least partially amortized by the fewer ROIs that need to be processed at the system level.

References

1. I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, P. R. de Toro, J. Nuevo, M. Ocana and M. A. Garrido. Combination of Feature Extraction Methods for SVM

Pedestrian Detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(2): 292-307, 2007.

2. 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 Intelligent Vehicle Symposium (IVS)*, 2003.
3. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In. *Proc. of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
4. M. Enzweiler and D. M. Gavrila. Monocular Pedestrian Detection: Survey and Experiments. In. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), available online: IEEE Computer Society Digital Library http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.260*, 17. Oct. 2009.
5. M. Enzweiler, P. Kanter and D. M. Gavrila. Monocular pedestrian recognition using motion parallax. In. *Proc. of the IEEE Intelligent Vehicle Symposium (IVS)*, 2008.
6. D. Fernández, I. Parra, M. A. Sotelo, P. Revenga and S. Álvarez. 3D candidate selection method for pedestrian detection on non-planar roads. In. *Proc. of the IEEE Intelligent Vehicle Symposium (IVS)*, 2007.
7. U. Franke, S. Gehrig, H. Badino and C. Rabe. Towards Optimal Stereo Analysis of Image Sequences. *Lecture Notes in Computer Science*, vol 4931, pp 43-58, 2008.
8. T. Gandhi and M. M. Trivedi. Pedestrian protection systems: Issues, survey and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 8(3): 413-430, 2007.
9. D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, vol 73, no 1, pp 41-59, 2007.
10. G. Grubb, A. Zelinsky, L. Nilsson and M. Ribbe. 3D vision sensing for improved pedestrian safety. In. *Proc. of the IEEE Intelligent Vehicle Symposium (IVS)*, 2004.
11. R. Labayrade, D. Aubert and J. P. Tarel. Real time obstacle detection on non flat road geometry through 'v-disparity' representation. In. *Proc. of the IEEE Intelligent Vehicle Symposium (IVS)*, 2002.
12. W. van der Mark and D. M. Gavrila. Real-Time Dense Stereo for Intelligent Vehicles. *IEEE Transactions on Intelligent Transportation Systems, vol. 7, nr 1, 38-50*, March 2006.
13. A. Mohan, C. Papageorgiou and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4): 349-361, 2001.
14. 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. of the IEEE Intelligent Engineering Systems (INES)*, 2004.
15. F. Oniga, S. Nedevschi, M. Meinecke and T. Binh. Road surface and obstacle detection based on elevation maps from dense stereo. In. *Proc. of the IEEE Intelligent Transportation Systems (ITSC)*, 2007.
16. P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In. *Proc. of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
17. A. Shashua, Y. Gdalyahu and G. Hayun. Pedestrian detection for driving assistance systems: single-frame classification and system level performance. In. *Proc. of the IEEE Intelligent Vehicle Symposium (IVS)*, 2004.
18. N. Suganuma and N. Fujiwara. An obstacle extraction method using virtual disparity image. In. *Proc. of the IEEE Intelligent Vehicle Symposium (IVS)*, 2007.
19. A. Wedel, U. Franke, H. Badino and D. Cremers. B-Spline modeling of road surfaces for freespace estimation. In. *Proc. of the IEEE Intelligent Vehicle Symposium (IVS)*, 2008.
20. L. Zhao and C. Thorpe. Stereo- and neural network-based pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems (ITS)*, 1(3).