Pedestrian Recognition Using Combined Low-Resolution Depth and Intensity Images

Martin Rapus, Stefan Munder, Gregory Baratoff
VDO Automotive AG
Peter-Dornier-Str.10, 88131 Lindau, Germany
{martin.rapus, stefan.munder, gregory.baratoff}@continental-corporation.com

Joachim Denzler
Lehrstuhl für Digitale Bildverarbeitung
Fakultät für Mathematik und Informatik
Friedrich Schiller Universität
Ernst-Abbe-Platz 2, 07743 Jena, Germany
joachim.denzler@inf.uni-jena.de

***Abstract*— We present a novel system for pedestrian recognition through depth and intensity measurements. A 3D-Camera is used as main sensor, which provides depth and intensity measurements with a resolution of 64x8 pixels and a depth range of 0-20 meters.**

The first step consists of extracting the ground plane from the depth image by an adaptive flat world assumption. An AdaBoost head-shoulder detector is then used to generate hypotheses about possible pedestrian positions. In the last step every hypothesis is classified with AdaBoost or a SVM as pedestrian or non-pedestrian. We evaluated a number of different features known from the literature. The best result was achieved by Fourier descriptors in combination with the edges of the intensity image and an AdaBoost classifier, which resulted in a recognition rate of 83.75 percent.

I. Introduction

Pedestrian recognition is a key for driver assistant systems. The task of such a system would be to protect pedestrians in time by warning the driver, applying the breaks or raising the vehicles hood if an accident is not avoidable.

In this paper a time-of-flight (TOF) range camera [1] delivers 3D measurements. The sensor uses the so called "multiple double short time integration" (MDSI) approach. A low resolution intensity (Fig. 1) and depth image (Fig. 2) of 64 pixels width and 8 pixels height is produced based on the time of flight and amplitude of infrared impulses. The camera working range is 0-20 meters and the detectable pedestrian height lies between 4 and 8 pixels. Advantages of the used camera: The sensor operates independent of the external illumination and delivers furthermore for every pixel depth information. A challenge of recognizing pedestrians is to handle the low camera resolution and the large appearance variety induced by clothing and posture.

The used classification procedure is shown in Figure 3. In a first step the ground plane as well as points which exceed the working depth range are removed using depth information (section III). A head-shoulder detector delivers possible pedestrian positions in the remaining image (section IV). As a next step the pedestrian verification is done by extracting important features (section V-A) and a classification through SVM or AdaBoost (section V-B). Section VI shows the experimental results and the paper ends with the conclusion in section VII.

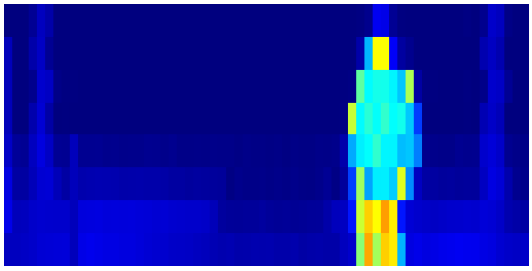

Fig. 1. intensity image

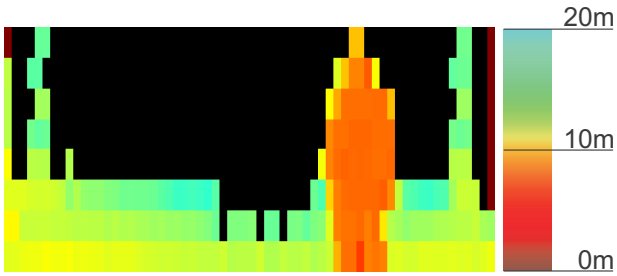

Fig. 2. depth image

II. Previous Work

The first step of the classification procedure limits the search region for pedestrians. Therefore depth information is used to detect the ground plane and points which exceed the used depth range in the image. A common approach like in [2], [3] is to assume a flat world (FWA). By that the ground plane can be marked with the help of the known camera parameters and the depth image. Another way is presented in [4]. Hough transformation in conjunction with a 2D-histogram of depth values and vertical position (v-disparity) is used to estimate the ground plane parameters. In [5], [6] a thresholding technique is used to limit the search region.

The following second step generates hypotheses about pedestrian positions in the remaining image. In [2] the fact

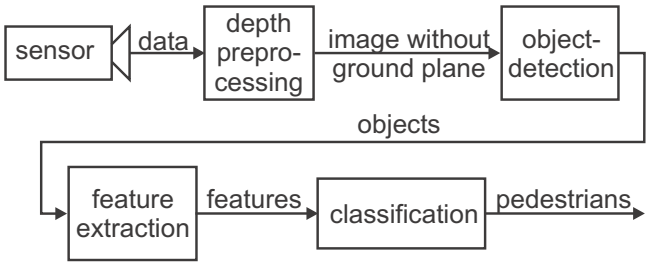

Fig. 3. classification system

that pedestrians generally have vertical symmetry is used to detect them. By shifting 3D boxes across the ground plane pedestrians are located in [3]. If the number of depth features exceeds a user defined threshold the 2D projection is considered as pedestrian object. In a subsequent step it is determined with the help of the object silhouette by hierarchic chamfer matching if it is a pedestrian object. To generate hypotheses the depth image is segmented by region growing in [6]. The v-disparity method is used in [4] to detect pedestrians. Due to graph-cut objects are generated in [7].

The last step is to verify the created hypotheses as pedestrians or non-pedestrians. First the important features for the classification are extracted. The features known from the literature are edge images [3], gradient magnitudes [5], intensity values [3] and the Fourier coefficients of the object contour [6]. Popular approaches to classify the hypotheses with the extracted features are SVM [4], [7] or AdaBoost [3]. In [5] the classification is done by an artificial neural network and in [6] a thresholding technique is used. Through filters based on the distribution of edges within the bounding box the hypotheses are classified as pedestrians or non-pedestrians in [2].

A similar low resolution 3D-camera like in this paper is used in [6]. The main differences between [6] and our approach are: We limit the search region by eliminating the ground plane while they use thresholding. Hypotheses about pedestrian positions are generated through a head-shoulder detector instead of region growing. For the final classification we use AdaBoost or a SVM and they use a thresholding technique.

III. Depth Preprocessing

Depth preprocessing is an essential step to divide the obtained depth image from the TOF camera into foreground and background. The background consists of those points that exceed the depth range or lie on the ground plane. All other points are regarded as foreground, which is searched for interesting objects (ROIs).

Through the moving vehicle the camera is not positioned at a constant height with constant orientation above the ground plane. To handle that we use an adaptation of v-disparity [8] on low resolution to estimate the ground plane. The first step of v-disparity is to calculate a 2D-histogram D of depth and vertical position:

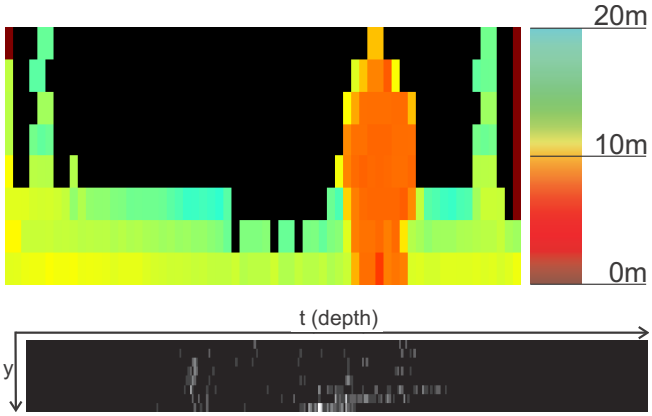

Fig. 4. depth image with the corresponding 2D-histogram, warm colors in the depth image represent near points (black area are pixels with depth $>$ 20m), intensities in the histogram represent the frequency of the respective depth in the image line

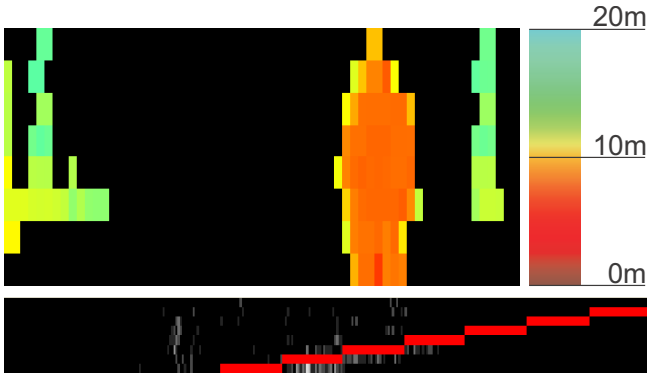

Fig. 5. depth image without ground plane (black area are pixels with depth $>$ 20m or ground plane) and the 2D-histogram with determined ground plane as red line)

$$D(y,t) = \sum_{x=0}^{M-1} \begin{cases} 1 & if\ I(x,y) = t \\ 0 & otherwise \end{cases} \quad \left(\begin{array}{c} y \in \{0,...,N-1\} \\ t \in [t_{min}, t_{max}] \end{array} \right) \tag{1}$$

Figure 4 shows the depth image and the generated 2D-histogram. The intensity of point (y,t) represents how often a depth value t appears in image row y. Approximate vertical lines in the resulting histogram D represent objects and a straight horizontal/diagonal line represents the ground plane. As a next step hough transformation is used to estimate and extract the ground plane line. This simple method allows the extraction of the ground plane without the need of an expensive camera calibration. Figure 5 shows the resulting histogram of the explained method applied to the input image. In Figure 4, the estimated ground plane is visualized as red line.

A common problem of the low camera resolution are weak responses of the ground plane in the 2D-histogram which causes the hough transformation to fail. To overcome that problem we use in addition a priori information about the ground plane. By the known extrinsic camera parameters a depth image is generated for the a priori ground plane.

This image is also included into the 2D-histogram from the camera depth image and serves as support. In a following step the above mentioned hough transformation is performed. Only the strongest non-vertical line represents the ground plane. For that we still assume a flat world but through v-disparity it is an adaptive FWA.

IV. Pedestrian Detection

The input image was divided into background and foreground. The task of the pedestrian detection is to generate hypotheses for pedestrian locations (ROI) in the foreground.

The head-shoulder part of a person is a characteristic feature that we use for pedestrian detection in the form of a learned head-shoulder detector. We evaluated a number of different features, based on depth, intensity or a combination of both. A detector is generated for every "possible" pixel-height of a person. The detection uses the upper third of a pedestrian. We examined popular methods for classification which are SVM and AdaBoost. In addition we compared them with a template. The used template consists of the expectation vector μ and covariance matrix S of all positive (head-shoulder part) training feature vectors. The Mahalanobis-Distance $d(\mathbf{x},\mu)$ is used to determine if an extracted feature vector $\mathbf{x}$ is a head-shoulder part.

$$d(\mathbf{x},\mu) = \sqrt{(\mathbf{x}-\mu)^T S^{-1} (\mathbf{x}-\mu)} \tag{2}$$

If the distance $d(\mathbf{x},\mu)$ lies under some predefined threshold $\theta_{template}$ the feature vector $\mathbf{x}$ represents a head-shoulder part.

To search for possible head positions in the input image a sliding window approach is used. The principle of SVM and AdaBoost can be found in the literature [9], [10].

The final ROIs are generated by expanding the head-shoulder part window onto the ground plane. For that reason the depth values in the detected head-area are used. In a first step the distance of the object to the camera is computed as the median of the depth values. With the derived object depth and the known ground plane the bottom is computed.

Experimentally we compared the proposed head-shoulder detector with two other methods from the literature. The first is an extension of region growing [6]. We evaluated different local and global homogeneity criteria in previous experiments. The best one was a global criterion where a neighboring pixel is added to the region if the distance between the pixels depth value and the regions depth as well as the regions size lie under some threshold. Two problems occur by the segmentation: On one hand is an over-segmentation and is solved like in [6] by merging neighboring regions if they have similar distance values. All original ROIs persist. On the other hand we have an under-segmentation problem. This appear if e.g. a group of people is standing nearly equidistant to the camera. For this reason a split-operator is used. This operator divides the region at certain points into smaller regions. We assume that local maxima of the regions head contour represent heads. Therefore the contour of the head area is searched for local maxima. Centered at those points a new ROI is generated with fixed ratio and the regions height.

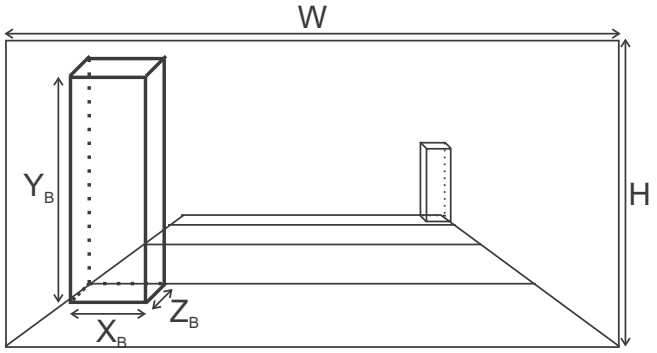

Fig. 6. generation of 3D-boxes with specific metrics (X_B-width, Y_B-height and Z_B-depth) for pedestrian detection

The second comparison method to detect pedestrians is the sliding 3D-box approach proposed in [3]. A 3D-box is positioned at all possible ground plane locations (see Figure 6) in the image. If a number of depth features exceeds a user defined threshold the box projection represents an object.

V. Pedestrian Verification

The generated ROIs have to be verified as pedestrian or non-pedestrian. This is done in two steps. First the important features have to be extracted and than classified.

A. Feature Extraction

For the classification of a detected object as pedestrian or non-pedestrian relevant features are needed, which allow a good separation of both classes.

We evaluated how the features known from the literature behave when applied to low resolution depth and intensity images. One of them are the horizontal and vertical edge images [4]. The image gradient magnitudes [5], the pure intensity or depth values and the haar-like features from Viola and Jones [11] were also tested.

The Fourier coefficients of the object contour [6] were also examined as features. For that we have to generate the contour of a ROI. The head-shoulder detector yields the object depth (see section IV). All pixel of the ROI form a region if their distance to the objects depth lie under a threshold. The contour is generated from that region. The invariance of the representation to translation is achieved by setting the zeroth Fourier coefficient α_0 to zero, because it represent the center of gravity of the object. Furthermore all coefficients are normalized by the first Fourier coefficient α_1, to achieve invariance to scale.

The evaluated features summarized:

- gradient magnitudes (intensity or depth image)
- raw values (intensity or depth image)
- Fourier descriptors of the objects contour
- haar-like rectangle features [11] (intensity or depth image)
- the (horizontal and vertical) edges (intensity or depth image)

Besides, combinations consisting of intensity and depth features of the mentioned features are examined.

B. Classification

The result of the preceding step is a feature vector $\mathbf{x}$. It still remains to associate that vector to the pedestrian or non-pedestrian class.

One popular approach is to use a Support Vector Machine (SVM) [9]. The idea behind a SVM is a linear separation of both classes through a hyperplane $H : \mathbf{x}^T\mathbf{n} + b = 0$. For more details on SVMs see [9].

We evaluated these kernel functions K:

- linear kernels:

$$K(\mathbf{x},\mathbf{y}) := \mathbf{x}^T\mathbf{y} \tag{3}$$

- polynomial kernels:

$$K(\mathbf{x},\mathbf{y}) := \left(\mathbf{x}^T\mathbf{y}+1\right)^d \tag{4}$$

- radial-basis-function (RBF) kernels:

$$K(\mathbf{x},\mathbf{y}) := \exp\left(-\gamma\cdot|\mathbf{x}-\mathbf{y}|^2\right) \tag{5}$$

The advantage of a SVM is the high generalization performance which is achieved.

Another popular classification method is AdaBoost [10]. A final decision is made by a weighted combination of multiple weak classifiers. AdaBoost uses weak classifiers which make only binary decisions. In the training the weak classifiers are added iteratively to the trained one until a wanted (minimal) error is achieved. Furthermore a weight is assigned to each weak classifier, which corresponds to its classification performance. We make use of the learning algorithm in [11].

VI. Experiments

A. Data Sets

For training a set of 1288 pedestrian and 30000 non-pedestrian samples were used. The data set for testing the classifiers consists of 57 video sequences with 7215 frames total and 2845 frames containing pedestrians.

B. Pedestrian Detection

At first different features and classifiers for the introduced head-shoulder detector are examined and then compared to the other methods out of section IV. Those other methods are region growing and the 3D-box sliding approach. Our evaluation criterion is the number of not detected pedestrians versus generated ROIs. Detections are considered as correct if they overlap at least 75% of a pedestrian. For every possible height of a pedestrian in the image a head-shoulder detector was created. Depending on the pedestrian height the used training set consists of 50-400 labeled heads and 10000-45000 random extracted non-heads. Knowledge about the average pedestrian height is used to limit the search area for the head-shoulder detection.

Results for the different features and classifiers for the head-shoulder detector are shown in Figure 7 and the best detection curves of the used three methods are displayed in Figure 8. The best detection result is achieved by the head-shoulder detector. AdaBoost in combination with the pure intensity values as features give the best head-shoulder detector result of the tested features and classification methods. That detector is used in subsequent experiments.

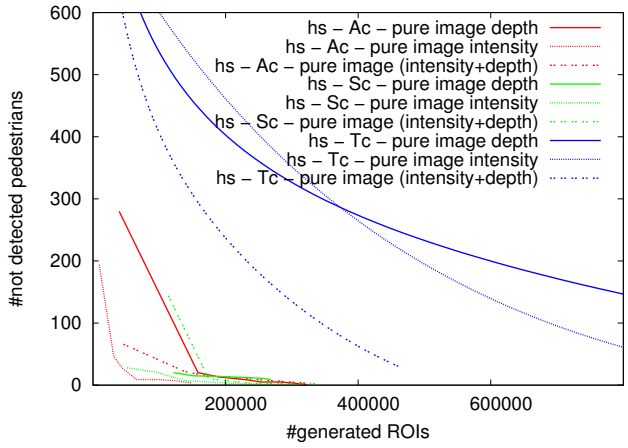

Fig. 7. head-shoulder detector results (hs - head-shoulder; Ac/Sc/Tc - AdaBoost-, SVM-, Template classifier)

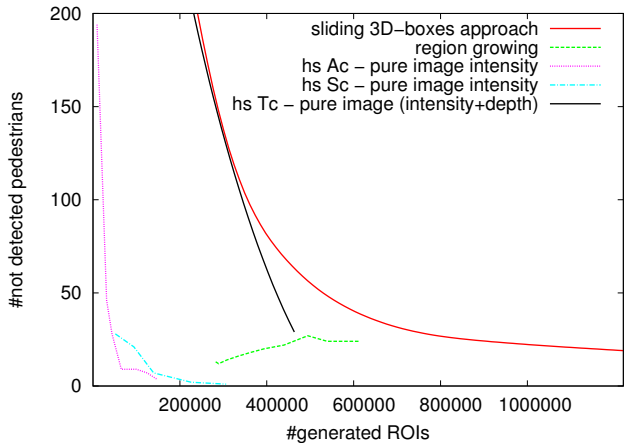

Fig. 8. best detection results (hs - head-shoulder; Ac/Sc/Tc - AdaBoost-, SVM-, Template classifier)

C. Pedestrian Verification

The introduced features in section V-A as well as combinations were tested. Used combinations consist of intensity as well as depth features.

In preliminary experiments four features are found to be promising, three of them were combinations:

1) pure intensity values
2) depth values with the gradient magnitudes and edges of the intensity image
3) Fourier descriptors of the contour with the pure intensity values
4) Fourier descriptors of the contour with the edges of the intensity image

We applied SVM and AdaBoost to these features. The SVMs were trained with the LibSVM tool [12]. The described training set in section VI-A was used for both classifiers. SVM as well as AdaBoost classifiers were tested

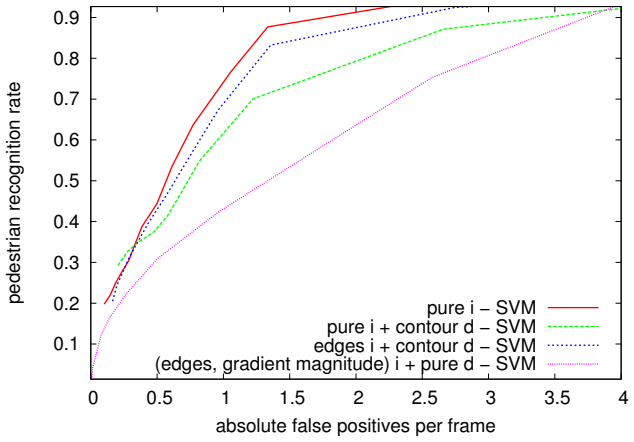

Fig. 9. ROC-curves of the SVM classifiers (i - intensity, d - depth)

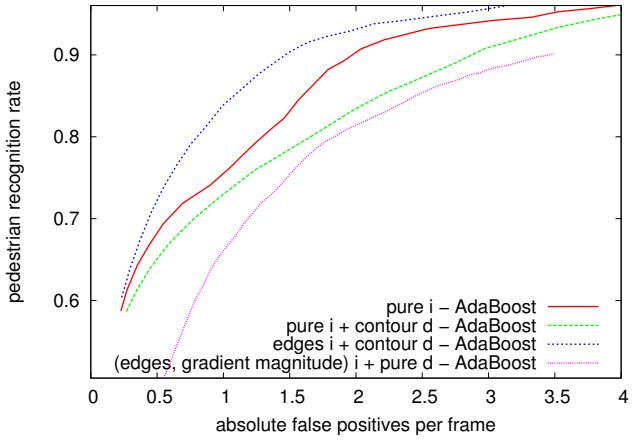

Fig. 10. ROC-curves of the AdaBoost classifiers (i - intensity, d - depth)

on the data set given in section VI-A. The resulting ROC-curves show Figure 9 and 10.

The achieved recognition rates by one false positive per frame are given in Table I. At an average error rate of one false positive per frame the feature combination Fourier descriptors of the contour with the intensity edge image and the AdaBoost classifier give the best result with a recognition rate of 83.75 percent. The average computation time for that combination on a Pentium IV 2 GHz processor is 70 ms per frame.

VII. Conclusion

This paper presents a novel pedestrian recognition system using a fusion of depth and intensity images provided by a TOF camera. Important is that the camera delivers images independent of the external illumination in which pedestrians could be recognized. The camera provides for every pixel an intensity and depth information.

In a first step (depth-preprocessing) a depth image is used to mark the ground plane pixels. Therefore an adaptation of v-disparity on low camera resolution is used.

A subsequent detection step provides pedestrian hypotheses by means of an AdaBoost head-shoulder detector with intensity features.

TABLE I

Recognition rate in percent at one false positive per frame (pi-pure intensity, ie-intensity edges, gm-gradient magnitude (intensity image), fd-Fourier descriptors

features	SVM	AdaBoost
pi	74.25	75.5
pi & fd	64.5	73.0
ie & gm & depth values	43.25	66.25
gm & fd	69.75	83.75

For the verification of the detected pedestrians a selection of the state of art features and classifiers are evaluated. The best result is obtained by the combination of Fourier descriptors of the contour with the intensity edges and an AdaBoost classifier. A pedestrian recognition rate of 83.75% was achieved by an average error rate of one false positive per frame. This shows that even with the low camera resolution pedestrian recognition is still possible.

A tracker would significantly reduce the false positive rate and therefore provides a robust pedestrian detection. The reason for that is that the tracker uses temporal integration. This remains to be explored in future work.

References

[1] P. Mengel, L. Listl, B. König, C. Toepfer, M. Pellkofer, W. Brockherde, B. Hosticka, O. Elkhalili, O. Schrey, and W. Ulfig, "Three-dimensional cmos image sensor for pedestrian protection and collision mitigation," in *10th International Forum on Advanced Microsystems for Automotive Applications*, April 2006, pp. 23–39.

[2] M. Bertozzi, A. Broggi, R. Chapuis, F. Chausse, A. Fascioli, and A. Tibaldi, "Shape-based pedestrian detection and localization," in *Proceedings of the IEEE Intelligent Transportation Systems*, vol. 1, October 2003, pp. 328–333.

[3] 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.

[4] G. Grubb, "3d vision sensing for improved pedestrian safety," Ph.D. dissertation, The Australian National University, March 2004.

[5] L. Zhao and C. Thorpe, "Stereo- and neural network-based pedestrian detection," in *IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems*, 1999, pp. 298–303.

[6] B. Fardi, F. Dousa, G. Wanielik, B. Elias, and A. Barke, "Obstacle detection and pedestrian recognition using a 3d pmd camera," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, June 2006, pp. 225–230.

[7] F. Suard, V. Guigue, A. Rakotomamonjy, and A. Benshrair, "Pedestrian detection using stereo-vision and graph kernels," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, June 2005, pp. 267–272.

[8] R. Labayrade, D. Aubert, and J.-P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "'v-disparity"' representation," in *IEEE Intelligent Vehicle Symposium*, vol. 2, June 2002, pp. 646–651.

[9] V. N. Vapnik, *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[10] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning*, 1996, pp. 148–156.

[11] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[12] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.