Integration of visual and depth information for vehicle detection

Alexandros Makris, Mathias Perrollaz, Igor E. Paromtchik, Christian Laugier

***Abstract*— In this work an object class recognition method is presented. The method uses local image features and follows the part based detection approach. It fuses intensity and depth information in a probabilistic framework. The depth of each local feature is used to weigh the probability of finding the object at a given scale. To train the system for an object class only a database of annotated with bounding boxes images is required, thus automatizing the extension of the system to different object classes. We apply our method to the problem of detecting vehicles from a moving platform. The experiments with a data-set of stereo images in an urban environment show a significant improvement in performance when using both information modalities.**

I. INTRODUCTION

The state-of-the-art visual object class recognition systems operate with local descriptors and codebook representation of the objects. Various local features (e.g. gradient maps, edges) are used to create the descriptors. Then kernel based classifiers are commonly employed to classify the detected features in one of several object classes [1][2][3][4]. The recognition of vehicles or pedestrians from sensors mounted on a moving platform is achieved by different approaches using various types of sensors, e.g. stereo camera, laser [5][6][7][8]. The approaches that perform data fusion from various sensors have proven to be the more robust in a variety of road conditions [9][10].

This work focuses on the development of an object class recognition system which follows the part based detection approach [2]. The system fuses intensity and depth information in a probabilistic framework. To train the system for a specific object class, a database of annotated with bounding boxes images of the class objects is required. Therefore, extending the system to recognize different object classes is straightforward. We apply our method to the problem of detecting vehicles by means of on-board sensors. Initially, depth information is used to find regions of interest. Additionally, the depth of each local feature is used to weigh its contribution to the posterior of the object position in the corresponding scale. In the following we provide a brief review of the methods related to our approach.

In the object recognition literature there is a large amount of works that follow the part-based approach. In [2], a codebook of object part appearance is constructed using interest point detector-descriptor pairs. The detected features are grouped into clusters and linked to the center of the object. A method that builds upon the aforementioned approach is presented in [11]. An approach to discriminatively learn a mapping between image patches and Hough votes is presented. Random trees are used to learn the above mapping in a supervised way (instead of clustering). In [12] shape and appearance information is used to perform object class recognition based on part detection and Hough transform. The codebook entries are selected using the boosting algorithm according to their significance, which is related to its discrimination capacity and the precision of the localization information for the object's centroid. In [1], a grouping of local features into pairs is proposed in order to increase their discriminative power. Selecting features connected by lines ensures finding features pairs with high repeatability.

Stereo-vision is widely used in the field of intelligent vehicles, mainly for generic obstacle detection [13][14]. A different approach for vehicles recognition is presented in [15], where the authors detect possible cars using 3D points provided by stereo-vision, and confirm the recognition of cars through a symmetry criterion. In [16], the author generates hypotheses of pedestrian as connected areas of constant disparity, and uses the aspect ratio of the corresponding regions as a clue to recognize pedestrians.

Lately, several methods that combine intensity with depth information have been proposed. In [17], vehicle and pedestrian detection is performed following the approach of [2] but also filtering the search regions by using the ground plane constraints. In [18], a method for pedestrian detection from a moving vehicle is presented. Stereo cues and a clustering algorithm are used to find candidate areas. In the following several detection windows are constructed around each area. The detection takes place in these windows using multiple features applied in manually predetermined subregions. In [9], a pedestrian classification method using depth and intensity features is developed. In this method the holistic detection approach is used extracting features from the whole region and feeding a classifier. The authors demonstrate that using both depth and intensity information outperforms any single modality method. Integration of stereo-vision with visual recognition has been proposed in [19], for estimating the road surface, reducing the hypotheses for a sliding window approach. In [20][21][22], video and laser data are fused to achieve robust vehicle and pedestrian detection.

The novelty of our approach is the fusion of depth and intensity information to form a probabilistic part-based detector. Firstly, we develop a framework to estimate the probability of finding an object at a position given all the available information. The depth of the detected local features is used to weigh (w.r.t. the corresponding distance) their contribution for the scale of the object. Using the depth information in this way takes into account the context in which we expect to

INRIA Grenoble Rhône-Alpes, 38334 St. Ismier, France (alexandros.makris, mathias.perrollaz, igor.paromtchik, christian.laugier)@inria.fr

find the objects (e.g. distant view, close-up). This is beneficial for the robustness of the approach, by avoiding for example many noisy detections resulting from false matches between features of different scales. Additionally, the computational gain from filtering out regions is very important for the on-line operation of the system which is required in the intelligent vehicles application. The method is tested with stereo video sequences captured in an urban environment.

The paper is structured as follows. Section II provides the theoretical background for our method. Section III gives the implementation details, providing a description of the stereoscopic sensor, the depth calculation algorithm, and the training and detection algorithms. The experimental evaluation of our method follows in Section IV and finally the conclusions in Section V.

II. METHOD DESCRIPTION

The proposed method is a probabilistic part-based object recognition method fusing intensity and depth information. The aim is to find the occurrences of a specific object category and viewpoint. Let o_n denote the object category/viewpoint with state vector $\mathbf{x} = [i_x, i_y, i_s]^T$ comprised of the image coordinates of the object's center and its scale. The method estimates the probability distribution $p(o_n, \mathbf{x}|\mathbf{I})$ where $\mathbf{I}$ denotes the image measurements.

The measurements are a set of N image features $\mathbf{I} = \{\mathbf{f}_j, \mathbf{d}_j\}_{j=1}^N$, where $\mathbf{f}_j$ and $\mathbf{d}_j$ are the intensity and depth descriptor of feature j respectively. The features are linked to the object through a codebook representation denoted by $\mathbf{C} = \{C_j, \mathbf{x}_{cj}\}_{j=1}^N$ where C_j is a random variable over the possible codebook labels of feature j and $\mathbf{x}_{cj} = [i^c_{xj}, i^c_{yj}, i^c_{sj}]^T$ its position and scale. The possible labels are the M clusters of the codebook $\{c_i\}_{i=0}^M$ where c_0 is the possibility that no cluster is observed. For each codebook cluster c_i we calculate during training the associated descriptor $\mathbf{f}_{c_i}$, and the conditional probability distribution $p(C_j = c_i, \mathbf{x}_{cj}|o_n, \mathbf{x})$. This distribution enables us to estimate the position and scale of the cluster knowing the position and scale of the object $\mathbf{x}$. If the camera parameters are known, the distance between the camera and observed cluster and thus the object can also be inferred. The graphical model depicting the conditional independence assumptions that we make is shown in Fig. 1.

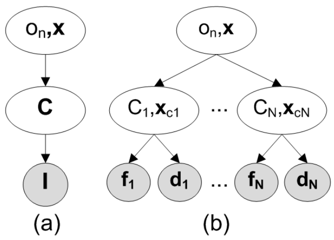

Fig. 1. Graphical Model of the method. (a) Model using $\mathbf{C}$ variable to denote the cluster labels and positions and $\mathbf{I}$ for all the available image measurements. (b) Analytic form showing the decomposition when multiple features are present. Each feature has an intensity $\mathbf{f}_j$ and a depth descriptor $\mathbf{d}_j$ and is associated with the possible clusters labels through C_j.

The probability of the object o_n at position $\mathbf{x}$ given all the available measurements is given by:

$$p(o_n, \mathbf{x}|\mathbf{I}) = \sum_{\mathbf{C}} p(o_n, \mathbf{x}|\mathbf{C}) p(\mathbf{C}|\mathbf{I}) \tag{1}$$

where the marginalization is over the values of $\mathbf{C}$.

The first term of (1) is the probability of having the object at a position given the set of observed clusters:

$$p(o_n, \mathbf{x}|\mathbf{C}) = p(o_n, \mathbf{x}) \prod_{j=1}^{N} \frac{p(C_j, \mathbf{x}_{cj}|o_n, \mathbf{x})}{p(C_j, \mathbf{x}_{cj})} \tag{2}$$

where we make the assumption that each cluster is independent from the others given the object. The second term of (1) is given by:

$$\begin{aligned} p(\mathbf{C}|\mathbf{I}) &= \prod_{j=1}^{N} p(C_j, \mathbf{x}_{cj}|\mathbf{f}_j, \mathbf{d}_j) \\ &\propto \prod_{j=1}^{N} p(\mathbf{f}_j|C_j) p(\mathbf{d}_j|C_j, \mathbf{x}_{cj}) p(C_j, \mathbf{x}_{cj}) \end{aligned} \tag{3}$$

where the probability of observing a feature given the corresponding cluster is considered independent from the rest of the features. The terms of (3) are:

- $p(\mathbf{f}_j|C_j)$ is the intensity likelihood calculated by comparing the observed feature descriptor $\mathbf{f}_j$ with the cluster's descriptor.
- $p(\mathbf{d}_j|\mathbf{x}_{cj}, C_j)$ is the depth likelihood computed by comparing the distance of the feature calculated using the depth information δ_d with the distance calculated using the scale of the cluster δ_s.
- $p(\mathbf{x}_{cj}, C_j)$ is the prior for observing the cluster C_j at a position $\mathbf{x}_{cj}$.

By replacing (3), (2), in (1) we get:

$$\begin{aligned} &p(o_n, \mathbf{x}|\mathbf{I}) \propto \\ &p(o_n, \mathbf{x}) \prod_{j=1}^{N} \sum_{(C_j, \mathbf{x}_{cj})} p(C_j, \mathbf{x}_{cj}|o_n, \mathbf{x}) p(\mathbf{f}_j|C_j) p(\mathbf{d}_j|C_j, \mathbf{x}_{cj}) \end{aligned} \tag{4}$$

We consider the prior $p(o_n, \mathbf{x})$ as uniform. Additionally, for each possible object position we consider only the contribution from the clusters observed within the object region. The possible detections are the local maxima of the posterior. The clusters observed outside the object region cannot affect the position of these maxima. In Section III-C, we describe the algorithm we use to estimate the posterior.

III. VEHICLE DETECTION SYSTEM IMPLEMENTATION

A. Stereo System

The vision system used in this paper is a stereoscopic sensor. It is considered as perfectly rectified. Cameras are supposed identical and classically represented by a pinhole model, $(\alpha_u, \alpha_v, u_0, v_0)$ being the intrinsic parameters. The length of the stereo baseline is b_s.

For further geometrical developments, let us define a *Vehicle Coordinate System* (VCS). For simplicity in notations, and without loss of generality, the yaw, pitch and roll angles of the camera, relative to the VCS, are set to zero. If it is not the case, homographies can be applied to the images in order to retrieve an equivalent configuration. In the VCS, X axis is parallel to the stereo baseline, Y is parallel to the optical axes and Z is oriented toward increasing height. (X_o, Y_o, Z_o) denotes the center of the stereo baseline in the VCS. Arbitrarily, we use the left camera of the stereo pair for the recognition task. Thus the coordinates $[i_x, i_y]$ will refer to the left image coordinates.

The stereo images are processed in order to retrieve depth information, following the approach described in [23]. First, a semi-dense matching algorithm is used in order to estimate a disparity value i_d for each pixel. During this stage, pixels are classified as road or obstacle by considering vertical and horizontal objects hypotheses. We use this information to discard the regions which correspond to the road surface or to objects that are not of interest (e.g. buildings, sky) using an arbitrarily chosen threshold for the height of the objects. An example of the mask resulting from this procedure is shown in Figure 2. With this step typically about 75% of the image is discarded thus the computational cost of the approach is reduced by the same ratio. For the obstacle pixels we retain the depth information. The distance of each pixel into the VCS is given by:

$$\delta_d = Y_o + \frac{\alpha_u b_s}{i_d} \tag{5}$$

Fig. 2. Depth mask example. The mask filters out the road surface and the objects that are over a prespecified height.

B. Detector Training

The training of the visual object recognition system follows the codebook based approach of [2]. For each object category/view we want to detect, a database of positive images is used to train the system. During the training phase we calculate the local SIFT [24] or SURF [25] features in a dense grid of image positions and different scales. A clustering step in the feature space using k-means is then performed to create a codebook of local appearances for each object class. For each cluster c_i we store: a) its appearance represented by the mean feature vector $\mathbf{f}_{c_i}$, b) its relative position to the center of the object. The latter is used to estimate $p(C_j = c_i, \mathbf{x}_{cj} | o_n, \mathbf{x})$. Figure 3 shows an example of several clusters for the side-view of vehicles object class.

C. Detector Implementation using Depth-Vision Integration

In this section we describe the detection algorithm we use to estimate the probabilities defined in Section II. The overall

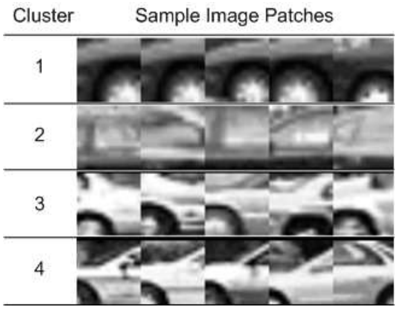

Fig. 3. Car-Side codebook clusters. Several image patches belonging to four clusters are shown. The clusters have been created with features extracted from the UIUC car database.

approach is shown in Figure 4. Algorithm 1 summarizes the steps of the approach.

In the detected regions of interest features are extracted from a dense grid and the respective descriptors are computed. The features are then matched to the clusters of the codebook. The likelihood of an intensity descriptor given a cluster $p(\mathbf{f}_j | C_j)$ is calculated by comparing the cluster's descriptor to the feature's descriptor. For the depth likelihood the scale in which the cluster is observed has to be taken into account. Let i_s^f be the scale in which a feature is detected and $i_s^{c_{in}}$ the initial scale of the matched cluster in the codebook. Then the feature will be assigned with a cluster of scale:

$$i_s^c = \frac{i_s^f}{i_s^{c_{in}}} \tag{6}$$

Knowing the scale of the cluster assigned to the feature we can determine the scale of the object. Using the predetermined size of the object class and the camera parameters we convert this scale into distance δ_s. For the same image patch we calculate the distance information we get from the stereo δ_d. As shown in equation 5, δ_d is obtained from a disparity value i_d. This value is estimated by taking the median disparity value in the neighborhood associated to the feature. Using the two distances the depth likelihood is calculated according to:

$$p(\mathbf{d}_j | C_j, \mathbf{x}_{cj}) = \exp\left\{ -\frac{(\delta_s - \delta_d)^2}{2\sigma_d^2} \right\} \tag{7}$$

where σ_d^2 is the variance parameter and is a linear function of δ_d. As the distance grows the uncertainty of the stereo distance estimation grows as well so a larger variance is required in order to have a non-negligible likelihood even with significant difference between δ_d and δ_s. The above technique allows us to group together features of the same scale, verified by the depth information. This way we filter out the noise resulting from false positive matches between different scales.

When the contribution of all features is taken into account, the mean-shift algorithm is used to find the local maxima in the $\mathbf{x}$ space. The maxima represent the positions and scales of the possible detections.

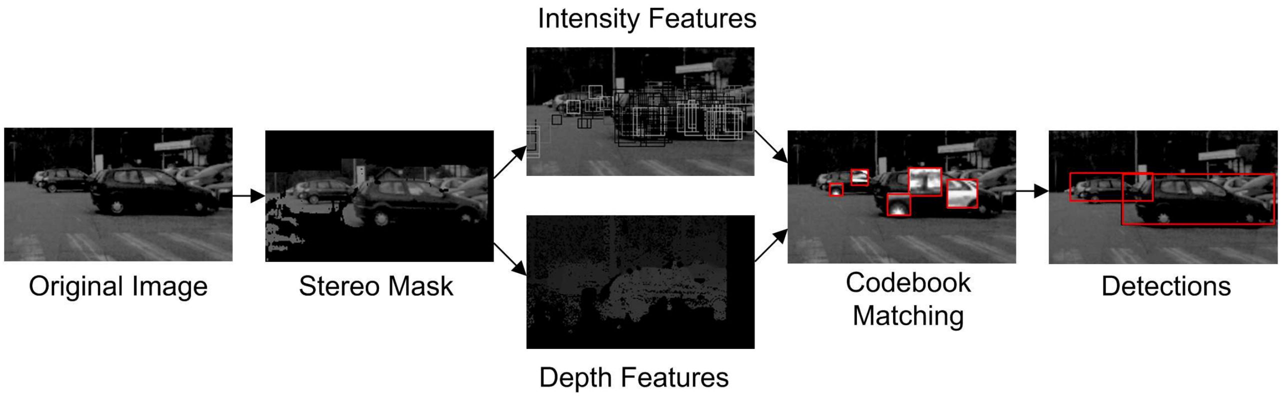

Fig. 4. The steps of the detection procedure are shown. The stereo information is used to define the regions of interest for the subsequent steps. Intensity and depth features are extracted from a dense grid within these regions. In the following the features are matched with the codebook clusters which are in turn used to estimate the posterior for the object in each position. The detections are the local maxima of the posterior.

Algorithm 1 Detection Algorithm

Input: Stereo pair: **I**, pdf: $p(C_j = c_i, \mathbf{x}_{cj}|o_n, \mathbf{x})$.
Filter image using stereo.
Extract intensity/depth feature pairs from each of the N positions of a dense scale-space grid.
for For Feature $j = 1$ to N **do**
  Calculate Intensity likelihood $p(\mathbf{f}_j|C_j)$.
  Calculate Depth likelihood $p(\mathbf{d}_j|C_j, \mathbf{x}_{cj})$.
  Posterior update with the contribution of the feature using (4).
end for
Locate local maxima of the posterior using mean-shift.
Output: A set of K detections $\left\{o_n^{(k)}, \mathbf{x}^{(k)}\right\}_{k=1}^{K}$, with associated probabilities: $p(o_n^{(k)}, \mathbf{x}^{(k)}|\mathbf{I})$.

Fig. 5. Our Experimental Platform. Lexus LS600h vehicle equipped with a TYZX stereo camera placed behind the windshield.

IV. EXPERIMENTS

In this section, we describe the experiments we conducted to evaluate the performance of our method. We apply our method to car detection and we demonstrate the improvement in robustness and computational efficiency of the complete system compared to the system using only intensity information.

For testing purposes we created a data-set using our experimental platform. The platform is a Lexus LS600h vehicle equipped with a TYZX stereo camera placed behind the windshield (Fig. 5). The stereo camera baseline is $22cm$, with a field of view of $62°$. Camera resolution is 512x320 pixels with a focal length of 410 pixels. The data-set contains 150 stereo images taken in an urban environment. We annotated the cars in these images with bounding boxes. The data-set includes challenging images, with poor illumination conditions, partial occlusions, and significant scale variations. For instance, the height of the annotated cars varies from 20 to 100 pixels.

For evaluation we compare the full method with the one using only intensity. To train both methods we used the UIUC car database. This database contains 550 images of side views of cars. Using this data-set we created a codebook of 2000 clusters. For the full method we set the variance parameter of the depth likelihood in (7) to $\sigma_d = 0.05\delta_d$. We tested the system with both SIFT and SURF descriptor. The difference in performance was negligible therefore in the experiments we used the SURF descriptor because it can be computed much faster. For the fairness of comparison we used the depth mask to find regions of interest for both methods.

In Fig. 6 we show some example detections. The proposed method detects side-views of cars in various scales, in cases with partial occlusions, and under significant background clutter. Part-based methods in general are more robust with partial occlusions. The use of depth information increases further the robustness as the features of each object are associated with a scale which in general is different from the scale of the occluding objects. An example of such situation can be seen in Fig. 7. We show a detection with and without depth information along with the features that contributed to that detection. As can be seen, in the case where no depth information is used (Fig. 7(c), (d)), many features that belong to a part of another vehicle in the background interfere with the detection resulting in inaccurate scale and position. With the use of depth information most of the features that are not on the object have been filtered out, thus resulting in a much better detection.

To perform a quantitative comparison we used a subset of our data-set, containing 60 images, where we detected the side-views of cars. For evaluation, we followed the single frame scheme which is adopted by the PASCAL object detection challenges [26]. For each frame we ran our

Fig. 6. Car-side detection examples. True and false positive detections are represented with red and yellow bounding boxes respectively. (a) Cars in different scales with significant background clutter and significant occlusions are detected. (b) Precise detection of the un-occluded vehicle, whereas a vehicle that is heavily occluded in the left is not detected. (c) Difficult detection of a vehicle which is far and partially occluded and a false detection in the region between the road surface and the trees. (d) Detection with partial occlusion. (e) Partial detection of a taller than normal vehicle(on the left). The training dataset does not contain vehicles of this type. (f) Succesfull detection of a partially occluded car and a false positive arising from a bus and a van. Trainning separate detectors for these type of vehicles as well will help to avoid these false alarms.

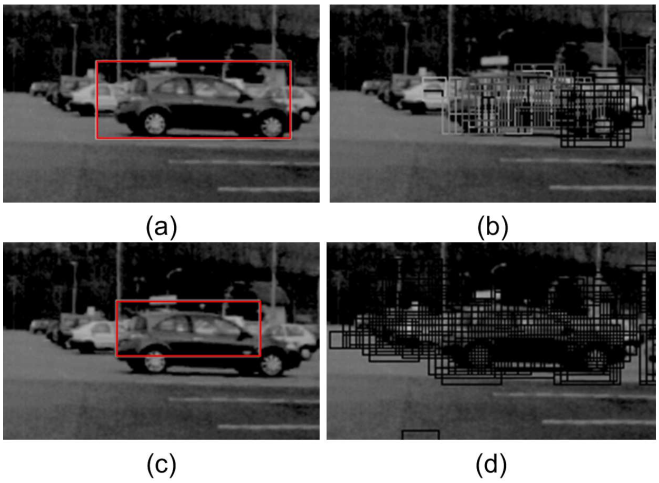

Fig. 7. Comparison of a vehicle detection. (a) Detection using depth-intensity. (b) Features that contributed to the detection. The depth information filters out the features that belong to background clutter. (c) Detection with intensity information. (d) Features that contributed to the detection.

multiscale detector resulting in a set of detected bounding boxes B_{dt} and using the ground-truth bounding boxes B_{gt} we accept a detection if:

$$\alpha = A(B_{dt} \cap B_{gt})/A(B_{dt} \cup B_{gt}) > 0.5 \tag{8}$$

where $A()$ denotes the area of the box. We associate only one detection with each ground-truth bounding box, if other detections intersect with it we count them as false positives. The output of our algorithm is a set of detection with probabilities. By adjusting the threshold to accept a detection we obtain the precision-recall curve.

In Fig. 8, the precision-recall curves are shown for our method with and without using depth information. We can see that using depth information we have a considerable increase in the performance. Additionally, this information enables us to create a mask and discard about 75% of the image thus decreasing the computational cost. As can be seen from the precision-recall curves, the challenging nature of the data-set poses difficulties for both methods. In particular, cars with poor illumination are difficult to detect with features based on image gradients. Using other type of features (e.g. based on shape) that perform better under poor illumination is expected to increase the performance. The variability in the scales of the objects is another factor that meets the limits of the used descriptors considering that they were trained using the UIUC database. This database contains cars from a single scale. Additionally, the American cars contained in the UIUC data-set have a different shape from the European cars that we have in our data-set. Nevertheless, as shown in [7], most of the state-of-the-art methods experience great difficulties in data-sets of this type (captured from a moving platform, urban environment). Under these circumstances however the increase in performance using depth information is significant. For instance the proposed method detects about one third of the vehicles, with 60% precision while the method using only intensity cannot even achieve this recall rate.

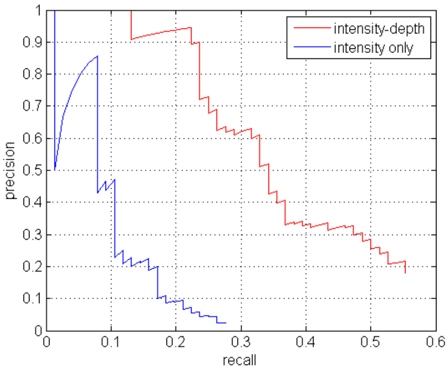

Fig. 8. Precision-Recall curves for the method using depth-intensity compared with the method using only intensity.

V. CONCLUSIONS

In this work we presented a method that fuses intensity with depth information to create a robust part-based detector. We applied the method to create a system for car detection from a moving vehicle. We tested it in a real urban environment using a data-set collected from our experimental platform. The comparison with the system using only intensity information shows a significant increase in performance.

As a first future work we consider using the stereo images dataset to train the system with intensity and depth information. This way we will be able to better estimate the parameters for the calculation of the depth likelihood. We will also be able to test the system with new types of features extracted from the depth images. As another future extension we consider to use the output probability densities of several detectors to do higher level reasoning in order to disambiguate between different object type detections for the same image region. Depth information, can be beneficial in such situations because it facilitates the reasoning in cases of occlusions.

ACKNOWLEDGMENT

This work has been done within the framework of the ArosDyn Project of INRIA. The authors thank Toyota Motor Europe for their support on our experimental work.

REFERENCES

[1] M. Awais and K. Mikolajczyk, "Feature pairs connected by lines for object recognition," in *Proc. of the Int. Conf. on Pattern Recognition (ICPR)*, 2010, pp. 3093–3096.
[2] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Int. J. of Computer Vision*, vol. 77, no. 1-3, pp. 259–289, 2008.
[3] E. Seemann, M. Fritz, and B. Schiele, "Towards robust pedestrian detection in crowded image sequences," in *CVPR*. IEEE Computer Society, 2007.
[4] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.
[5] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, 2006.
[6] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, 2010.
[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *CVPR*, 2009, pp. 304–311.
[8] M. Enzweiler and D. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.
[9] M. Rohrbach, M. Enzweiler, and D. M. Gavrila, "High-level fusion of depth and intensity for pedestrian classification," in *DAGM-Symposium*, ser. Lecture Notes in Computer Science, J. Denzler, G. Notni, and H. Süße, Eds., vol. 5748. Springer, 2009, pp. 101–110.
[10] D. 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.
[11] J. Gall and V. S. Lempitsky, "Class-specific hough forests for object detection," in *CVPR*, 2009, pp. 1022–1029.
[12] A. Opelt, A. Pinz, and A. Zisserman, "Learning an alphabet of shape and appearance for multi-class object detection," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 16–44, 2008.
[13] R. Labayrade, D. Aubert, and J. Tarel, "Real time obstacles detection on non flat road geometry through v-disparity representation," in *IEEE Intelligent Vehicles Symp.*, Versailles, France, 2002.
[14] A. Broggi, C. Caraffi, P. Porta, and P. Zani, "The single frame stereo vision system for reliable obstacle detection used during the 2005 DARPA Grand Challenge on terramax," in *Proc. of the IEEE Intelligent Transportation Systems Conf.*, Toronto, Canada, 2006.
[15] G. Toulminet, M. Bertozzi, S. Mousset, A. Bensrhair, and A. Broggi, "Vehicle detection by means of stereo vision-based obstacles features extraction and monocular pattern analysis," *IEEE Transactions on Image Processing*, vol. 15, no. 8, August 2006.
[16] T. Veit, "Connexity based fronto-parallel plane detection for stereovision obstacle segmentation," in *IEEE Int. Conf. on Robotics and Automation, Workshop on Safe Navigation in Open and Dynamic Environments: Applications to Autonomous Vehicles*, Kobe, Japan, 2009.
[17] B. Leibe, K. Schindler, N. Cornelis, and L. J. V. Gool, "Coupled object detection and tracking from static cameras and moving vehicles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1683–1698, 2008.
[18] I. P. Alonso, D. F. Llorca, M. Á. Sotelo, L. M. Bergasa, P. R. de Toro, J. Nuevo, M. Ocaña, and M. A. G. Garrido, "Combination of feature extraction methods for SVM pedestrian detection," *IEEE Trans. on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 292–307, 2007.
[19] D. Geronimi, A. D. Sappa, A. Lopez, and D. Ponsa, "Adaptive image sampling and windows classification for on-board pedestrian detection," in *Proc. of the 5th Int. Conf. on Computer Vision Systems*, Bielefeld, Germany, 2007.
[20] W. R. M. Mahlisch, R. Schweiger and K. Dietmayer, "Sensorfusion using spatio-temporal aligned video and lidar for improved vehicle detection."
[21] R. T. L. Spinello and R. Siegwart, "A trained system for multimodal perception in urban environments," in *IEEE Int. Conf. on Robotics and Automation, Workshop on Safe Navigation in Open and Dynamic Environments: Applications to Autonomous Vehicles*, 2009.
[22] P. P. M. S. L. Oliveira, U. Nunes and F. Moita, "Semantic fusion of laser and vision in pedestrian detection," in *Pattern Recognition*, vol. 43, no. 10, 2010, pp. 3648–3659.
[23] M. Perrollaz, A. Spalanzani, and D. Aubert, "A probabilistic representation of the uncertainty of stereo-vision and its application to obstacle detection," in *Proc. of the IEEE Intelligent Vehicles Symp.*, San Diego, CA, USA, 2010.
[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[25] H. Bay, T. Tuytelaars, and L. J. V. Gool, "Surf: Speeded up robust features," in *ECCV (1)*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3951. Springer, 2006, pp. 404–417.
[26] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman, "Dataset issues in object recognition," in *Toward Category-Level Object Recognition, LNCS*, vol. 4170. Springer, 2006, pp. 29–48.