Probabilistic Integration of Intensity and Depth Information for Part-Based Vehicle Detection

Alexandros Makris, Mathias Perrollaz, Christian Laugier

Probabilistic Integration of Intensity and Depth Information for Part-Based Vehicle Detection

Alexandros Makris, Mathias Perrollaz, and Christian Laugier

***Abstract*—In this paper, 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 distance. To train the system for an object class, only a database of images annotated with bounding boxes 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.**

***Index Terms*—Bayes methods, object recognition, sensor fusion, vehicle detection.**

I. Introduction

RELIABLE environment perception is a very crucial component of intelligent vehicle systems. Driver assistance and autonomous or semiautonomous driving systems require a detailed modeling of the vehicle's surroundings to detect potentially dangerous situations. First, the detection of possible obstacles is required, but the characterization of the type of each obstacle also is very important in order to determine the appropriate behavior with respect to the obstacle. Generic object recognition techniques using visual sensors provide an accurate and feasible solution to this characterization problem due to low implementation costs. They can be used to recognize a variety of possible obstacles and other important features, such as pedestrians, vehicles, and traffic signs. The detections can then be used in order to warn the driver or automatically initiate appropriate protective measures. However, visual recognition is very challenging due to multiple difficulties of the on-road application, i.e., partial occlusions, moving sensor, large illumination variances, different possible object appearances, cluttered backgrounds, and real-time constraints.

In this paper, we develop an object class recognition system for intelligent vehicles, which follows the local part-based detection approach. The system fuses intensity and depth information in a probabilistic framework. The use of local features and depth information allows us to efficiently handle occlusions by determining the visible parts of each object and considering the features from that part to classify it. Additionally, depth information and planar constraints are used to conservatively filter out regions where no processing is necessary (e.g., sky and road surface). It has to be noted, however, that this is not a region-of-interest (ROI) generation step in the classical sense as it does not provide specific candidate areas. Creating ROIs using only one information modality, as is usually the case in many recent methods [1], [2], makes the system sensitive to that modality. Instead, we probabilistically fuse both information modalities so detections that have a high score on one modality but are missed by the other will be retained. We apply our method to the problem of detecting vehicles by means of on-board sensors. 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.

Manuscript received August 3, 2012; revised February 16, 2013 and May 1, 2013; accepted June 14, 2013. Date of publication August 6, 2013; date of current version November 26, 2013. The Associate Editor for this paper was D. Fernandez-Llorca.

The authors are with the INRIA Grenoble Rhône-Alpes, 38334 Saint Ismier, France (e-mail: alexandros.makris@inria.fr; mathias.perrollaz@inria.fr; christian.laugier@inria.fr).

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.2013.2271113

A. Related Work

The recognition of obstacles from sensors mounted on a moving platform has been addressed using many different approaches. Older methods focused on a specific obstacle class, and they used features and techniques suited for this class that were not easily generalized to other classes. Recently, the advances on object class recognition techniques have made possible the creation of more generic systems that can be used to recognize different object classes after appropriate training. The recognition is achieved using various types of sensors, e.g., monocular or stereo camera, laser, and radar. Several recent papers have reviewed the state-of-the-art approaches in vehicle and pedestrian detection [3]–[6]. The approaches that perform data fusion from various sensors have proven to be the more robust in a variety of road conditions [2], [7].

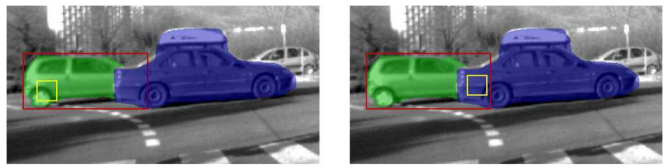
The state-of-the-art visual object class recognition systems can be split into two broad categories: 1) the methods that operate with local descriptors and codebook representation of the objects and 2) the methods that perform holistic detection, usually using a sliding-window approach [8]. The methods of the first category use various local features (e.g., gradient maps and edges) to create the descriptors. Then, kernel-based classifiers are commonly employed to classify the detected features in one of several object classes [9]–[12].

In the object recognition literature, there is a large number of works that follow the part-based approach. The basic idea of the part-based approach is that a set of detectors is independently used for each part. Subsequently, the detected parts are used to estimate the position of the whole object. In [10], 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 [13]. An approach to discriminatively learn 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 [14], 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 [9], a grouping of local features into pairs is proposed in order to increase their discriminative power. Selecting features connected by lines ensures finding feature pairs with high repeatability.

In addition to the intensity information from cameras, many recent works incorporate depth information for object recognition. Particularly in the field of intelligent vehicles, stereo vision is widely used to provide depth information. Several approaches exist that use stereo for generic obstacle detection [15], [16]. A different approach for vehicle recognition is presented in [17], where the authors detect cars using 3-D points provided by stereo vision and confirm the recognition of cars through a symmetry criterion. In [18], they generate hypotheses of pedestrians as connected areas of constant disparity and use 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 [19], vehicle and pedestrian detection is performed following the part-based approach in [10] but also filtering the search regions by using the ground plane constraints. In [20], a method for pedestrian detection from a moving vehicle is presented. Stereo cues and a clustering algorithm are used to find candidate areas. Several detection windows are constructed around each area. The detection takes place in these windows using multiple features applied to manually predetermined subregions. In [21] and [22], stereo information is used to detect ROIs for a HOG/SVM detector [8], [23]. A similar approach to generate ROIs is used in [24]. To this end, a preprocessing step is performed, where candidate obstacle regions are described as vertical rectangles with the same depth. In [7], 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 [25], for estimating the road surface, reducing the hypotheses for a sliding-window approach. In the approach in [2], a sparse disparity map is computed to establish the ROIs. Shape matching based on chamfer distance is performed in the ROIs. A set of exemplars covering the possible pedestrian shapes is used for this matching. A texture-based classification follows using a neural network with local receptive fields. Then, a dense stereo-based verification step is performed in the candidate locations. In [26], we presented an approach for improving the part-based



(a) (b)

Fig. 1. Benefits of using depth information on the local level instead of using it globally. Here, we consider a scenario in which the object of interest is partially occluded, and we detect features in two regions with different depth values, which are represented by shaded green and shaded blue color. In (a), a local feature is detected within the object, whereas in (b), it lies in the occluded area. If depth information is used on a local feature level in (a), the contribution of the feature for the detection will be strong since its depth corresponds to that detection. In (b), however, its contribution will be lower since its depth is different. However, if we use depth information globally, both features will have an equal contribution to the detection despite the fact that in (b) the feature is an outlier for the detection and should have been disregarded. Consequently, these two cases will falsely lead to a detection with equal confidence.

(a) (b)

Fig. 2. Possible detection scales from a local feature. (a) Only intensity information is used. The distance of the detection cannot be accurately determined, resulting in false-positive detections at inconsistent distances. (b) Fusion of depth and intensity information. Depth information is used to weigh the possible detections. Detections with inconsistent distances receive a lower weight.

methodology by integrating depth information. In this paper, we refine the mathematical model and extend the approach with an explicit occlusion-handling strategy. We also include an extended qualitative and quantitative evaluation, including comparison with state-of-the-art approaches.

B. Contribution

The main contribution of this work is the development of a probabilistic local part-based object recognition framework fusing intensity and depth information. Although methods that fuse intensity with depth information already exist, here, the fusion is performed on the local feature level. This approach has several advantages over the existing methods. First, it allows for an efficient way to treat partial occlusions. In general, methods based on local features are more robust to partial occlusions. Additionally, our approach integrates the depth information on the local feature level so the disambiguation of the possible occlusion scenarios is facilitated (see Fig. 1). Compared with the existing part-based generic object recognition methods, the use of depth information significantly increases robustness since it narrows the search over the possible detection scales (see Fig. 2). In this way, the context in which we expect to find the objects is taken into account (e.g., distant view and close-up).

An extra reweighting scheme has also been developed to ensure that the weights of the detections are comparable. We intensify by the occlusion ratio the contribution of the unoccluded features of a partially occluded object. With this technique, the

weight of an occluded object will be equal to the weight of an unoccluded one provided that the density of the detected features at the visible part is equal.

This paper is structured as follows. Section II provides the theoretical aspect of our method. Section III details the implementation, 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 conclusion is given in Section V.

II. Probabilistic Fusion of Intensity and Depth Information

A. Method Description

The proposed method probabilistically fuses intensity and depth information. As input, it uses a grayscale image and the corresponding depth map. The method proceeds as follows. We detect a set of features in the input image. For each feature, we extract the intensity and the depth descriptor. Using the intensity descriptor, we assign it to several prelearned local object parts called codebook labels. Using the depth descriptor, we estimate the feature's distance. Each assigned local feature votes for the position of the object. The output of the algorithm is a set of detected objects with their respective categories and 3-D positions in a local coordinate system.

For further geometrical developments, let us define two coordinate systems: the *Image Coordinate System* (ICS) and the *Vehicle Coordinate System* (VCS). The ICS represents the image coordinates in pixels (u, v), whereas the VCS represents 3-D Cartesian coordinates (X, Y, Z) in meters. In the VCS, the X-axis is horizontal, the Y-axis is vertical, and the Z-axis is parallel to the optical axis. Considering a calibrated camera, for objects of known size in the VCS, there is a direct correspondence between coordinates in the ICS and in the VCS [27]. Therefore, we can switch from a detection at a given position and scale of the ICS to a 3-D position of the VCS, and vice versa.

The measurements are a set of N local features. Each feature is localized by the image coordinates of its center and its size: $\mathbf{x}_j^f = [u_j^f, v_j^f, r_j^f]^T$. The size of a feature r_j^f is a parameter that defines the support area of the feature, i.e., the area around its center that is used to calculate the descriptors. Call $\{\mathbf{f}_j, \mathbf{d}_j\}_{j=1}^N$ the set of feature descriptors, where $\mathbf{f}_j$ and $\mathbf{d}_j$ are the intensity and depth descriptors of feature j, respectively.

A codebook links the detected features with objects by attributing a probability for each detected feature to be a specific local part of the object [10]. We refer to these local parts as codebook labels. The appearance of the local parts, as well as their relative position with respect to the object, are learned offline (see Section III-B) using a data set of annotated positive and negative images. This codebook representation is an intermediate level of abstraction between low-level local features and high-level detections.

To each local feature, we attribute depth variable $z_j^c \in \mathcal{X}^d$ and codebook label variable $C_j \in \mathcal{X}^c$. With z_j^c, we represent the depth of the feature, and C_j is a random variable over the possible codebook labels of the feature $\mathcal{X}^c = \{c_i\}_{i=0}^M$. The values from c_1 to c_M represent matches with one of the M codebook labels, whereas the value c_0 reflects the possibility that no label matches with the feature.

TABLE I
Definitions of the Main Variables

Variable	Description
$\mathbf{x}_j^f$	Image position/size of a local feature
$\mathbf{f}_j$, $\mathbf{d}_j$	Intensity and depth descriptor of a local feature.
z_j^c	Estimated depth of the local feature.
C_j	Estimated codebook label of a local feature.
$\mathbf{x}^o$	Image position/depth of a detection.
R^o	Size of a detection's projection on the image plane (VCS).
r^o	Size of the detection in the image (ICS).

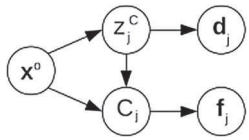


Fig. 3. Graphical model of the method. The jth feature pair, i.e., $\mathbf{f}_j$ and $\mathbf{d}_j$, is used to calculate the pdfs of the codebook labels and depths C_j and z_j^c. Using these pdfs, the position of the object $\mathbf{x}^o$ is estimated.

Each detection of an object of a specific class/viewpoint is represented by the state vector, i.e., $\mathbf{x}^o = [u^o, v^o, z^o]^T \in \mathcal{X}^o$, where (u^o, v^o) are the image coordinates of its center and z^o is its average depth. The size of the orthogonal projection of the object on plane $Z = z^o$ in the VCS, i.e., $R^o = (W^o, H^o)$, is known. Using that size, we can convert the depth, i.e., z^o, to the size of its bounding box in the ICS, i.e., $r^o = (w^o, h^o)$.

The method estimates the probability $p(\mathbf{x}^o|\mathbf{f}_j, \mathbf{d}_j)$ for each individual feature j. Subsequently, Hough voting is used to calculate the evidence $\varepsilon(\mathbf{x}^o)$ for every possible position of detection space $\mathcal{X}^o$ by accumulating the above probabilities for the set of all available features. The set of the detections is obtained by locating the local maxima of $\varepsilon(\mathbf{x}^o)$ and thresholding. The model variables are grouped in Table I.

B. Probabilistic Formulation

The graphical model depicting the conditional independence assumptions that we make is shown in Fig. 3. This model implies that the state of the object, which is conditioned on the codebook label and depth, is independent of the image features. Codebook label C_j directly depends on intensity feature $\mathbf{f}_j$. The depth of the patch z_j^c directly depends on depth feature $\mathbf{d}_j$ and indirectly on the intensity feature through the label assignment. This is justified because an intensity feature does not provide depth information; however, after assigning the feature to a codebook label of known depth, we can have an estimate of the patch's depth.

The probability of detecting an object with state $\mathbf{x}^o$ given a local feature pair is

$$p(\mathbf{x}^o|\mathbf{f}_j, \mathbf{d}_j) = \sum_{\mathcal{X}^c} \sum_{\mathcal{X}^d} p\left(\mathbf{x}^o|C_j, z_j^c\right) p\left(C_j, z_j^c|\mathbf{f}_j, \mathbf{d}_j\right). \quad (1)$$

The summations are over the codebook labels and depths. The first term of (1) is the probability of having the object at

position $\mathbf{x}^o$ given the cluster label and its depth

$$p\left(\mathbf{x}^o|C_j, z_j^c\right) = \frac{p\left(C_j, z_j^c|\mathbf{x}^o\right) p(\mathbf{x}^o)}{p\left(C_j, z_j^c\right)}. \quad (2)$$

The second term of (1) is given by the data likelihoods

$$p\left(C_j, z_j^c|\mathbf{f}_j, \mathbf{d}_j\right) \propto p\left(C_j|z_j^c\right) p(\mathbf{f}_j|C_j) p\left(\mathbf{d}_j|z_j^c\right) \quad (3)$$

where $p(\mathbf{f}_j|C_j)$ is the intensity feature likelihood. It is calculated by comparing the observed intensity feature descriptor $\mathbf{f}_j$ with the descriptor of the codebook label. $p(\mathbf{d}_j|z_j^c)$ is the depth feature likelihood. $p(C_j|z_j^c)$ is the probability distribution of the codebook label assignment given the estimated feature depth. It is calculated by comparing 1) the depth of the assigned codebook label C_j using the known size of the object part that corresponds to that label with 2) the estimated depth of feature z_j^c.

The contribution of each feature is accumulated in a 3-D voting space $\varepsilon(\mathbf{x}^o)$. The possible detections are the local maxima of that space. In Section III-C, we describe the algorithm that we use to efficiently estimate this 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 to be identical and classically represented by a pinhole model, with f_l, u_0, and v_0 being the intrinsic parameters [27], where focal length f_l is measured in pixels. The length of the stereo baseline is b_s. For simplicity in notations, the yaw, pitch, and roll angles of the camera, relative to the VCS, are set to zero. Arbitrarily, we use the left camera of the stereo pair for the recognition task. The disparity value is denoted by Δ. The relationship between coordinates in the left image and in the VCS is given by

$$X = \frac{b_s(u - u_o)}{\Delta}, \quad Y = -\frac{b_s(v - v_o)}{\Delta}, \quad Z = \frac{f_l b_s}{\Delta}. \quad (4)$$

The stereo images are processed in order to retrieve depth information. The first stage consists of computing a disparity map. This is done by using the semiglobal matching technique proposed in [28]. This method has the advantage of providing semidense disparity maps in real time, with subpixel accuracy. The computed disparity map contains a large number of pixels that cannot belong to vehicles (e.g., road surface, buildings, sky, etc.). We propose to build a mask from the disparity image, in order to avoid processing such pixels. Two approaches are combined for this purpose, i.e., filtering based on occupancy and filtering based on geometry.

For filtering based on occupancy, we want to remove all the pixels belonging to the free space. For this purpose, an occupancy grid is computed from the disparity data. This grid is directly computed in the disparity space associated with the stereoscopic sensor. In this approach, a visibility probability is estimated for each cell as the ratio between visible pixels and possible pixels. This strategy allows for the handling of partially occluded objects. The details about this method are described in [29]. Compared with the original approach, the matching algorithm is different. The SGM algorithm does not allow automatic classification of road and obstacle pixels. This is not an issue, since road pixels do not vertically accumulate and the road and obstacle areas are clearly distinguishable on the occupancy grid, which gives a good approximation of the free space. Fig. 4 illustrates the computation of the disparity image and the associated occupancy grid. In the grid, each cell (u, Δ) is associated with a probability of being occupied $p_{u,\Delta}(Occ)$. A threshold, i.e., T_f, is applied on these occupancy values to classify the image pixels into obstacles/road: A pixel of the left image with coordinates (u, v) and disparity value Δ is filtered out if $p_{u,\Delta}(Occ) < T_f$. This filtering is more robust than just using the 3-D coordinates of each pixel individually, because the occupancy grid takes advantage of the vertical alignment of pixels along vertical objects.

Fig. 4. Occupancy grid computed from stereo vision. (Left) Left image of the stereo pair. (Right) Disparity image computed with the SGM algorithm. (Bottom) Occupancy grid computed in the u-disparity plane.

Fig. 5. Depth mask examples. The mask filters out the road surface and the objects that are over a prespecified height.

Afterward, geometrical filtering is obtained by using an arbitrarily chosen threshold for the height of the objects. This allows for the removal of pixels situated at irrelevant heights for the current application. For instance, while training the algorithm for detecting standard vehicles, there is no interest in observing above 2 m. The training data do not contain objects that have a height larger than 2 m. Similarly, pixels situated under the road surface and not filtered by the occupancy filter (generally matching errors) are removed. Examples of masks resulting from both filtering are shown in Fig. 5. After these two steps, typically about 80% of the image is discarded; thus, the computational cost of the approach is reduced by the same ratio.

B. Detector Training

The training of the visual object recognition system follows the codebook-based approach. A database of positive and

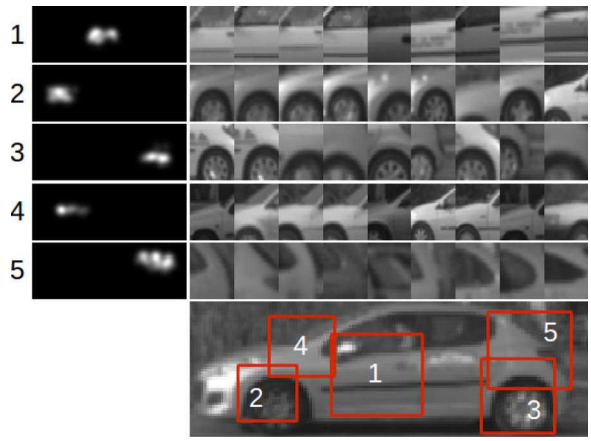


Fig. 6. Car-side codebook labels. (Top right) Several image patches belonging to five codebook labels. (Top left) Locations within the object that these patches were found, from which we compute the relative position distribution. (Bottom) Bounding boxes of the approximate position of each label within the object.

negative images is used to train the system for each object category/view we want to detect. The positive images contain the object located at distance z^{tr}. During the training phase, we calculate the SIFT descriptors [30] in a set of salient image positions. The features are grouped according to their sizes (characteristic scales) in S feature subsets. For each subset, a clustering step in the feature space, using k-means, is then performed to create the codebook of local appearances for that object class. Each cluster regroups the appearance on a specific scale range (and consequently distance range). In the literature, the most common approach for intensity-only methods is to cluster all the features directly without splitting them, in order to be able to discriminate between features at different distances. However, since during detection we use depth information to distinguish different distances, we can split the features according to their scale and then perform the clustering. This technique results to more compact clusters.

For each cluster of codebook c_i, we store 1) its appearance represented by the mean feature vector $\mathbf{f}_{c_i}$ and 2) its relative position to the center of the object $\mathbf{x}_i^c$. The latter is nonparametrically stored as the set of the relative positions of N_i features that belong to cluster $i : \{\mathbf{x}_{ik}^c\}_{k=1}^{N_i}$, where $\mathbf{x}_{ik}^c = [u_{ik}^c, v_{ik}^c, r_{ik}^c]^T$. The relative position distribution is used to approximate $p(C_j, z_j^c|\mathbf{x}^o)$. Fig. 6 shows an example of several clusters for the side view of the vehicle object class.

After the creation of the codebook, a validation step is performed in order to assess the quality of the labels. A database of positive and negative images is used to extract features. The features are matched to the codebook labels. Let N_i^{pos} and N_i^{neg} be the numbers of positive and negative features matched to label i. The probability of that label given the object $p(C_j = c_i|\mathbf{x}^o)$ is approximated by $p_r(c_i) = N_i^{pos}/(N_i^{pos} + N_i^{neg})$.

Algorithm 1 Codebook Learning Algorithm

Input:
$\{\mathbf{I}^{tr}, \mathbf{I}^{pos}, \mathbf{I}^{neg}\}$: Train, validation and negative images.
Codebook creation:
Detect N^{tr} keypoints from $\{\mathbf{I}^{tr}\}$.
for each keypoint $j = 1$ to N^{tr} **do**
  Extract intensity descriptor: $\mathbf{f}_j$.
  Store the relative feature position $\mathbf{x}_j^c$.
end for
Cluster the N^{tr} features into M clusters using k-means.
Codebook weighting:
Detect N^{pos}, N^{neg} keypoints from $\{\mathbf{I}^{pos}, \mathbf{I}^{neg}\}$.
for each cluster c_i, $i = 1$ to M **do**
  Calculate the number of matching positive and negative keypoints: N_i^{pos}, N_i^{neg}.
  Compute its weight as: $p_r(c_i) = N_i^{pos}/(N_i^{pos} + N_i^{neg})$.
endfor
Output: A set of M clusters with their associated average descriptors $\mathbf{f}_{c_i}$ and pdfs $p(C_j, z_j^c|\mathbf{x}^o)$.

C. Intensity-Depth Fusion Detector Implementation

Here, we describe the detection algorithm we use to estimate the probabilities defined in Section II. After the filtering step described in Section III-A, a salient point detector is used to locate several image patches from the rest of the image, and the descriptors are computed. For each extracted image patch j, the probability of assigning a codebook label given the intensity descriptor is computed. Subsequently, we calculate the probability density function (pdf) of its depth given the label assignment and depth descriptor. Then, the probabilistic vote of the image patch for the location of the object is cast. The overall approach is illustrated in Fig. 7. Algorithm 2 summarizes the steps of the approach.

The intensity likelihood of codebook label c_i is calculated by comparing label and feature descriptor, i.e.,

$$p(\mathbf{f}_j|C_j = c_i) \propto \exp\left\{-\frac{\|\mathbf{f}_j - \mathbf{f}_{c_i}\|^2}{2\sigma_f^2}\right\} \tag{5}$$

where σ_f is the intensity variance parameter. We consider that the feature matches with the cluster if the above likelihood is over a threshold value.

For each positive match between feature $\mathbf{f}_j$ and codebook label c_i, we calculate the N_i possible scales of the match as $s_{jk}^f = r_j^f/r_{ik}^c$, where r_j^f is the size of the image patch. The scale is converted into depth by $z_{jk}^f = z^{tr}/s_{jk}^f$. The label assignment probability given the estimated depth of the image patch is

$$p\left(C_j = c_i|z_j^c\right) = \frac{1}{N_i}\sum_{k=1}^{N_i}\mathcal{N}\left(z_{jk}^f; z_j^c, \sigma_{zf}^2\right). \tag{6}$$

For the same image patch, we calculate the distance information from stereo z_j^d. It is estimated by taking the median disparity value in the neighborhood associated to the feature and converting the value into distance using (4). To compute the depth likelihood, we take into account the fact that the disparity values are discretized. The disparity discretization results in a quantization of the distance values. The quantization step depends on the distance from the sensor, its intrinsic parameters,

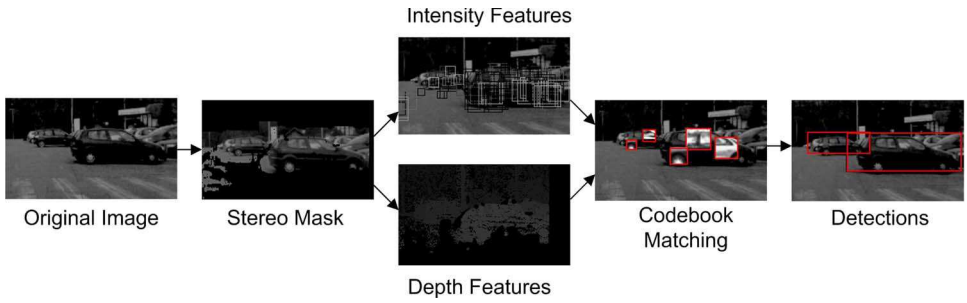


Fig. 7. Detection procedure steps. The stereo information is used to define the ROIs for the subsequent steps. Intensity and depth features are extracted from salient points within these regions. The features are matched with the codebook clusters, which are, in turn, used to estimate the posterior for the object in each position. The evidence is calculated by taking into account the contributions of all the features, and the detections are the local maxima of the evidence.

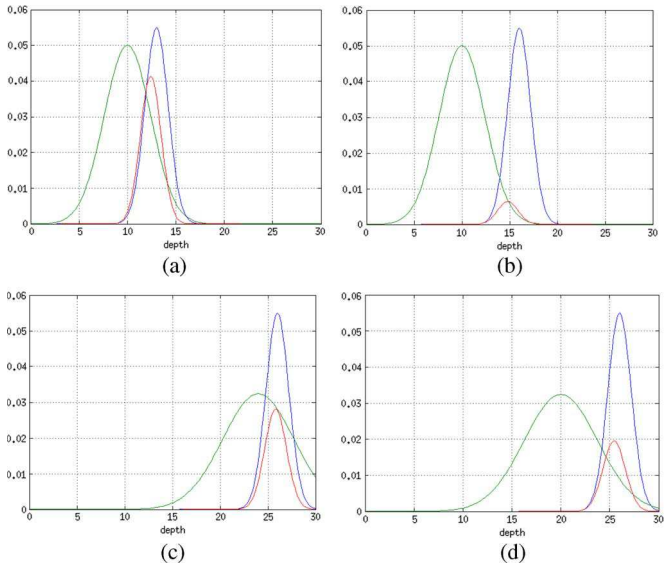


Fig. 8. Codebook label pdf over the distance for an image patch is plotted. (Green) Label distance pdf calculated using the depth likelihood, i.e., $p(\mathbf{d}_j|z_j^c)$, peaked at z_j^d. (Blue) Label distance pdf using intensity likelihood, i.e., $p(C_j = c_i|z_j^c)$, peaked at z_{jk}^f. (Red) Fusion of the two modalities. (a) For a patch that is close to the sensor, we have a small depth variance; thus, when the two distance estimations are close, the peak of the fusion is high. (b) For small distances, the peak of the fusion falls sharply when the two estimations are not close. (c) When the patch is further, the depth variance is bigger; therefore, the corresponding fusion pdf has a lower peak value. (d) The fusion peak falls slowly at bigger distances w.r.t. to the difference in depth estimations.

and the stereo baseline. In particular, the uncertainty range for distance z_j with corresponding disparity value Δ_j is given by

$$\delta z_j = \alpha_u b_s \left[\frac{1}{\Delta_j^+} - \frac{1}{\Delta_j^-} \right] \tag{7}$$

where with Δ_j^+ and Δ_j^- we denote the previous and next disparity values w.r.t. the value that corresponds to z_j. Using that uncertainty range, we calculate the depth likelihood from

$$p\left(\mathbf{d}_j|z_j^c\right) = \mathcal{N}\left(z_j^d; z_j^c, \sigma_{zd}\left(z_j^c\right)^2\right) \tag{8}$$

where $\sigma_{zd}(z_j^c) = \kappa_{zd}\delta z_j^c$ is the standard deviation and is proportional to the distance range δz_j^c.

By multiplying (5), (6), and (8), we get the probability of a codebook label at a distance given the feature pair of image patch j. Fig. 8 illustrates the codebook label pdf over the distances. Its peak value depends on the distance between z_j^d and z_{jk}^f and the variances of the distance likelihoods. The above technique allows us to use depth information to filter out the noise resulting from false-positive matches between a feature and a codebook label that correspond to detections at a different distance from the true distance of the image patch.

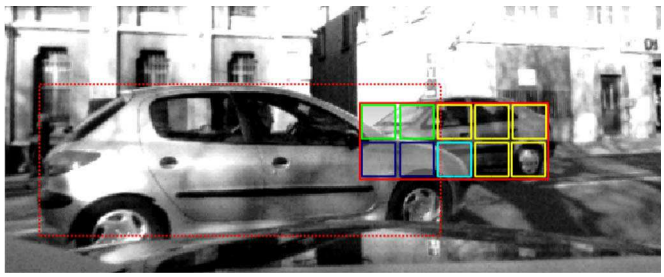


Fig. 9. Different categories of the patches of a detection. The patches can be considered as either (yellow) visible, (cyan) occluded considering stereo depth information, (green) occluded considering another detection nearer to the sensor, or (blue) occluded considering both of the above.

Having calculated the probabilities of label assignments for a feature extracted from local image patch j, the next step is to calculate the probability of the detection given codebook variable $p(\mathbf{x}^o|C_j, z_j^c)$. This is calculated using the nonparametric distribution that was learned during training, i.e., $p(C_j, z_j^c|\mathbf{x}^o)$, and considering uniform priors $p(C_j)$ and $p(\mathbf{x}^o)$, i.e.,

$$p\left(\mathbf{x}^o|c_i, z_j^c\right) = p_r(c_i)\frac{1}{N_i}\sum_{k=1}^{N_i}\delta\left(\mathbf{x}_{ik}^o\right) \tag{9}$$

where the summation is over the N_i features that where attributed to codebook label i during training, $p_r(c_i)$ is the prior weight of label i that was calculated also during training, and $\delta(\mathbf{x}_{ik}^o)$ is 1 at position $\mathbf{x}_{ik}^o$ and 0 otherwise. The position $\mathbf{x}_{ik}^o$ of the detection is given by

$$u_{ik}^o = u_j^f + u_{ik}^c s_j^c, \quad v_{ik}^o = v_j^f + \frac{v_{ik}^c}{s_j^c}, \quad z_{ik}^o = z_j^c \tag{10}$$

where s_j^c is the scale of the patch that corresponds to distance z_j^c.

Up to this point, we derived the equations to calculate the probability of a detection given a single image patch, i.e., $p(\mathbf{x}^o|\mathbf{f}_j, \mathbf{d}_j)$. The contribution of each patch is summed to calculate the evidence $\varepsilon(\mathbf{x}^o)$ in each position of $\mathcal{X}^o$, i.e.,

$$\varepsilon(\mathbf{x}^o) = \sum_{j=1}^{N} p(\mathbf{x}^o|\mathbf{f}_j, \mathbf{d}_j). \tag{11}$$

The mean-shift algorithm is used to find the local maxima in the evidence space. The maxima represent the positions and scales of the possible detections.

Algorithm 2 Detection Algorithm

Input: Stereo pair: $\mathbf{I}$, pdf: $p(\mathbf{x}^o|\mathbf{C})$.
Compute depth map using the SGM algorithm.
Filter image using stereo information.
Detect N salient image points.
Extract intensity/depth feature pairs from the detected points.
for image patch $j = 1$ to N **do**
  Compute:
    Intensity likelihood: $p(\mathbf{f}_j|C_j)$.
    Intensity distance likelihood: $p(C_j|z_j^c)$.
    Depth distance likelihood: $p(\mathbf{d}_j|z_j^c)$.
    Detection pdf given the label: $p(\mathbf{x}^o|C_j, z_j^c)$.
Update the evidence $\varepsilon(\mathbf{x}^o)$ with the contribution of the jth image patch.
end for
Locate the local maxima of $\varepsilon(\mathbf{x}^o)$ using mean-shift.
Output: The set of L detections: $\{\mathbf{x}_l^o, \varepsilon(\mathbf{x}_l^o)\}_{l=1}^L$.

D. Occlusion Reweighting

Here, we further develop the proposed method by normalizing the evidence of detections to account for occluded regions. The goal of the approach is to take into account the occluded regions of a possible detection in order to reweigh the evidence that comes from the unoccluded regions. In the example in Fig. 9, the visible patches (yellow and green) account for 7/10 of the detection. If we consider that we have absolute confidence about the state of each patch, we will multiply by 10/7 the evidence of the detection to normalize its evidence with respect to a detection that is fully visible. To avoid giving high weights to almost fully occluded objects, we use a visibility threshold of 10% in order to reweigh possible detections.

In practice, for each detection patch j of a possible detection $\mathbf{x}^o$, we use depth information to calculate the probability of visibility. We then update the detection evidence as

$$\varepsilon_V(\mathbf{x}^o) = \varepsilon(\mathbf{x}^o)\frac{1}{\frac{1}{N^{rew}}\sum_{j=1}^{N^{rew}} p\left(\mathcal{V}\left(\mathbf{x}_j^o\right)|\mathbf{d}_j\right)} \tag{12}$$

where $p(\mathcal{V}(\mathbf{x}_j^o)|\mathbf{d}_j)$ is the probability of visibility of detection patch j, and N^{rew} is the total number of detection patches.[1] The visibility probability of an image patch is given by

$$p\left(\mathcal{V}\left(\mathbf{x}_j^o\right)|\mathbf{d}_j\right) = \frac{1}{2}\left[1 - erf\left(\frac{z_j^o - z_j^d}{\sigma_{zd}\left(z_j^d\right)\sqrt{2}}\right)\right] \tag{13}$$

where z_j^o is the distance of the jth patch of the detection and z_j^d is the measured distance for the same patch calculated using stereo information. This equation corresponds to 1 minus the cumulative density function of (8).

[1]We should note here that the detection patches form a nonoverlapping grid that covers the possible detection (see Fig. 9) and are different from the multiscale overlapping image patches that we used in Section III-C.

TABLE II
STEREO IMAGE DATA SET DETAILS

Imageset	Frames	Occ. Vehicles	Total Vehicles
Testing INRIA	384	283	741
Train negative	369	0	0
Train positive	880	0	880
-Vehicles front	300		
-Vehicles rear	330		
-Vehicles side	250		
Total	**1633**	**283**	**1621**

IV. EXPERIMENTS

Here, we describe the experiments we conducted to evaluate the performance of our method. We applied our method to vehicle detection, and we demonstrate the improvement in robustness and computational efficiency of the complete system, particularly for the case of occluded vehicles.

A. Experimental Setup—Data Set

For training and testing purposes, we created a data set using a TYZX stereo camera. The stereo camera baseline is 22 cm, with a field of view of 62°. Camera resolution is 512 × 320 pixels with a focal length of 410 pixels. The camera is placed behind the windshield of our vehicle. We performed several acquisitions during daytime, under varying illumination and climatic conditions in the urban area of Grenoble city in France. To avoid correlations, we used different subsets of sequences for testing and training. The details of the data set are presented in Table II. We annotated the cars in these images with bounding boxes. For training, we used about 300 positive images for each viewpoint (front, rear, and side). We tested the performance of the algorithms with different training set sizes, and we observed that there is no big performance benefit after about 200–250 samples. The data set includes challenging images, with poor illumination conditions, partial occlusions, and significant scale variations. As occluded vehicles, we consider the ones that are from 10% to 70% visible. The height of the annotated vehicles varies from 40 to 100 pixels, which corresponds to the distance range of 3–20 m. The range can be augmented by using a sensor with higher resolution or different focal length. However, as a part-based approach, our method shows its merit when detecting objects of considerable size in pixels.

We compare five methods: 1) **LPF**—the proposed Local Probabilistic Fusion method; 2) **OR-LPF**—the fusion method using the Occlusion Reweighting procedure described in Section III-D; 3) **ISM**—the intensity-only method, which is an implementation of the method proposed in [10]; 4) **GPF**—the Global Probabilistic Fusion method; and 5) **L-SVM**—the latent-SVM/HOG method [12]. The **GPF** is similar to the intensity-only method but, as a postprocessing stage, uses depth information on the object level to filter the detections. Postprocessing lowers the weights of the detections whose distance does not match with the distance calculated using stereo information. This is achieved using a likelihood function computed as the ratio of the detection pixels, which have a

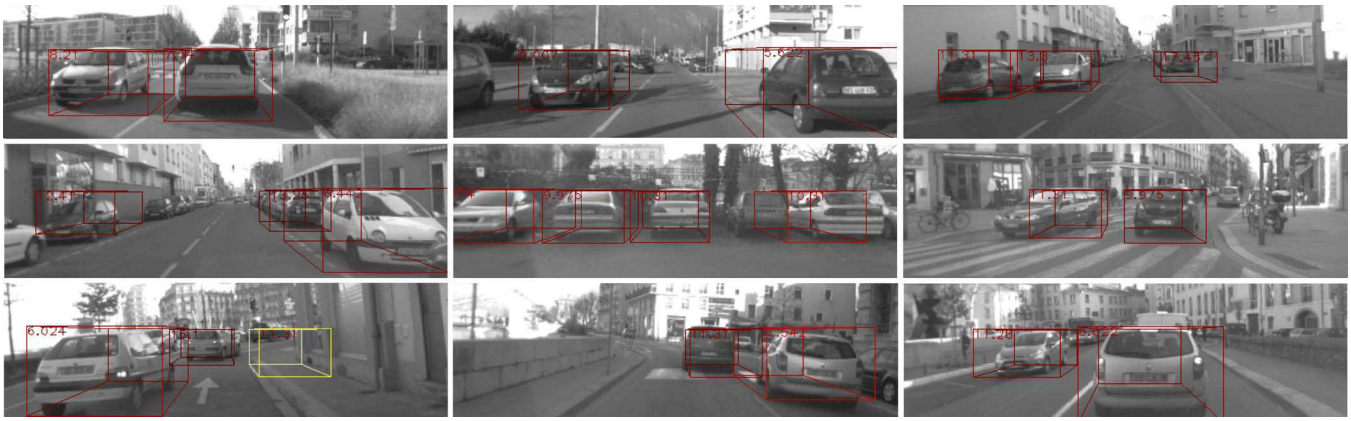


Fig. 10. Vehicle detection examples. Here, we use in parallel our detectors to detect vehicles from different viewpoints at a distance range from 3 to 20 m. Our system successfully detects most of the vehicles. Most of the missed detections are out of the detection range. (Yellow) False positive detections.

depth value that corresponds to the detection's actual distance. For the **L-SVM** method, we used the implementation provided by the authors [31].

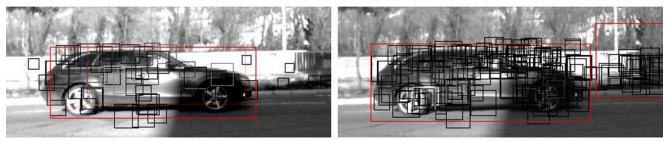
We trained all the methods using the data set in Table II. Three separate detectors were trained for frontal, rear, and side view of vehicles. The depth likelihood variance parameter, i.e., $\sigma_{zd}(z_j^c)$, is analogous to κ_{zd}, which was set to 0.7. The variance of the depth distribution given the codebook assignment, i.e., σ_{zf}, is set to a low value, i.e., 0.05, in order to ensure that the feature's estimated distance is close to the distance estimated from the label assignment. The intensity likelihood variance parameter σ_f is set as to allow a sufficient number of features to have a nonnegligible weight. For our data set, we found that 0.1 was an appropriate value. We tested the system with several intensity detectors/descriptors, and we selected the SIFT because of its superior performance. For fairness of comparison, we used the depth mask to filter out irrelevant regions for all the methods.

Using the described setup, the computational cost of the fusion method allows almost real-time operation. In a CPU implementation, we achieve a frame rate of around 3 frames/s. The most costly operations concern the extraction and matching of the features that are parallelizable. We, therefore, expect that a GPU implementation will result in a significant performance increase, and a frame rate of around 30 frames/s can be achieved.

B. Qualitative Results

Here, we illustrate the advantages of our method by showing the detection results in several images. The proposed method detects cars in various scales, even in cases with partial occlusions and under significant background clutter. In Fig. 10, we show several example detections for a series of images from our data set. To detect vehicles from different viewpoints, we run our detectors in parallel and merge the resulting detections.

The main benefit of using depth information is that each local patch contributes only to evidence of the detections that have about the same depth with the patch. In this way, many false-positive matches are avoided. An example of such a situation can be seen in Fig. 11. We show a detection with and without depth information along with the features that matched with the codebook. As can be seen, in the case where no depth



(a) (b)

Fig. 11. Comparison of a vehicle detection with and without depth information. (a) Detection using depth–intensity and features that contributed to the detection. The depth information filters out the features that belong to background clutter. (b) Detection with intensity information and features that contributed to the detection. Without depth information, background features are considered and lead to a false-positive detection.

information is used [see Fig. 11(a)], many features of the background interfere, resulting in a false-positive detection. With the use of depth information [see Fig. 11(b)], most of the features that are not on the object have been filtered out, thus resulting in a more accurate detection without any false positives.

The proposed method is well suited for situations in which frequent occlusions exist such as urban environments. Part-based methods, in general, are more robust with partial occlusions. The use of depth information in the local image patch level further increases the robustness as the features of a candidate detection have a different depth from the occluding objects so the depth descriptor can distinguish them. Fig. 12 shows a series of partially occluded detection examples. In these examples, we note that the fusion method is able to accurately detect most of the occluded vehicles, even those with high occlusion ratios on the order of 60%–80%.

In Section III-D, we use depth information to reweigh the detections that are partially occluded by multiplying the evidence from the unoccluded areas. Fig. 13 shows an example of the benefits of that technique. It allows us to have comparable evidence values for detections that are unoccluded as well as for heavily occluded detections.

C. Quantitative Results

To perform a quantitative comparison, we used several subsets of our data set in which we detected the vehicles. For evaluation, we followed the single-frame scheme, which is adopted by the PASCAL object detection challenges [32]. For each frame, we ran our multiscale detector, resulting in a set

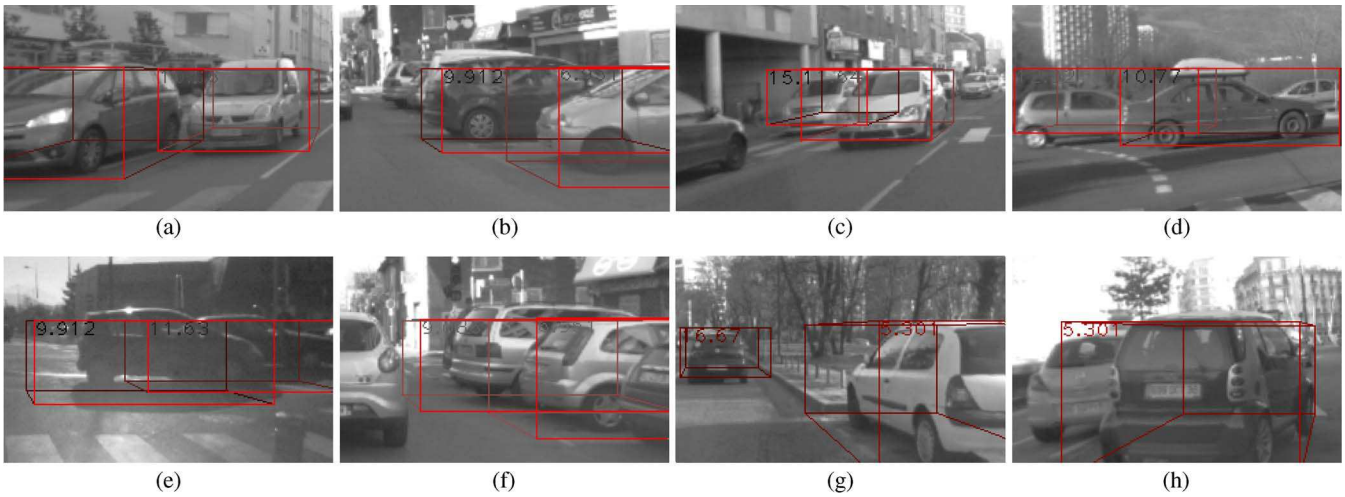


Fig. 12. Partially occluded vehicle detections. (a) Two cars in different distances are detected. (b) Precise detection of two partially occluded vehicles. (c) Detection of two frontal views of vehicles. (d) Detection of side views with partial occlusion. (e) Detection of two vehicles under difficult illumination conditions and heavy occlusion of the second one. (f) Detection of two vehicles that are less than 30% visible and very close to each other. These factors as well as the heavy background clutter from the other parked vehicles render the rear detection inaccurate. (g) Detection of two rear views of vehicles. (h) Missed detection of a partially occluded vehicle.

Fig. 13. Comparison of a vehicle detection with and without occlusion reweighting. (a) The evidence is reweighted in order to account for the occluded parts of the objects. This results in accurate detections, although large parts of both objects are occluded. (b) When no reweighting takes place and keeping the same evidence threshold, only one local maximum is found, which results in one detection with inaccurate position.

of detected bounding boxes r_{dt}, and using the ground-truth bounding boxes r_{gt}, we accept a detection if

$$\alpha = A(r_{dt} \cap r_{gt})/A(r_{dt} \cup r_{gt}) > 0.5 \tag{14}$$

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 L detections, each with a corresponding evidence value $\varepsilon(\mathbf{x}_l^o)$. By adjusting the acceptance threshold for a detection, we obtain the precision–recall curve. From that curve, we calculate the average precision of our method.

We tested the methods with different values of feature subsets number S and clusters number M. Tables III–V summarize the average precision scores for the fused detector **LPF**, the intensity-only detector **ISM**, and the global fusion detector **GPF**, respectively. Using the proposed fusion method, we improve the average precision at a ratio 2–3 times compared with the intensity-only approach. In the fusion method, we observe that, on average, we get better results when we use several feature subsets during training. This is explained as being because splitting the features according to their scale during training results in more compact clusters. Since the intensity-only method cannot discriminate between features of different scales, using multiple subsets deteriorates its performance. Overall, the best results were attained using 150/5 clusters/subsets for the fusion method and 90 clusters for the intensity-only method. Comparing Tables IV and V shows that using depth on the detection level instead of doing it locally does not improve the performance. This validates our intuition about the benefits of the local fusion.

TABLE III
AVERAGE PRECISION FOR THE LPF DEPTH–INTENSITY FUSION DETECTOR. **S** IS THE NUMBER OF CODEBOOK SUBSETS

		Clusters (M)						
View	**S**	30	60	90	120	**150**	210	300
front	1	0.50	0.49	0.43	0.48	0.39	0.44	0.44
	5	0.48	0.48	0.43	0.41	**0.53**	0.44	0.41
	10	0.25	0.27	0.24	0.26	0.34	0.49	0.49
	15	0.13	0.28	0.26	0.18	0.34	0.24	0.24
rear	1	0.38	0.37	0.30	0.35	0.36	0.36	0.24
	5	0.45	0.40	0.28	0.39	**0.48**	0.44	0.36
	10	0.33	0.34	0.30	0.35	0.36	0.35	0.40
	15	0.31	0.33	0.33	0.51	0.36	0.32	0.32
side	1	**0.76**	0.56	0.70	0.60	0.64	0.63	0.17
	5	0.35	0.43	0.65	0.67	0.61	0.62	0.45
	10	0.33	0.47	0.48	0.51	0.63	0.53	0.59
	15	0.34	0.35	0.47	0.46	0.48	0.53	0.54

TABLE IV
AVERAGE PRECISION FOR THE **ISM** INTENSITY-ONLY DETECTOR

		Clusters (M)						
View	**S**	30	60	**90**	120	150	210	300
front	1	**0.16**	0.15	0.14	0.14	0.09	0.14	0.12
rear	1	0.14	0.16	0.19	**0.21**	0.16	0.09	0.07
side	1	**0.34**	0.26	0.33	0.21	0.24	0.29	0.09

TABLE V
AVERAGE PRECISION FOR THE **GPF** GLOBAL FUSION DETECTOR

		Clusters (M)						
View	**S**	30	60	**90**	120	150	210	300
front	1	**0.19**	0.17	0.15	0.16	0.11	0.15	0.17
rear	1	0.13	0.10	**0.17**	0.11	0.13	0.15	0.11
side	1	**0.34**	0.26	0.24	0.18	0.17	0.17	0.07

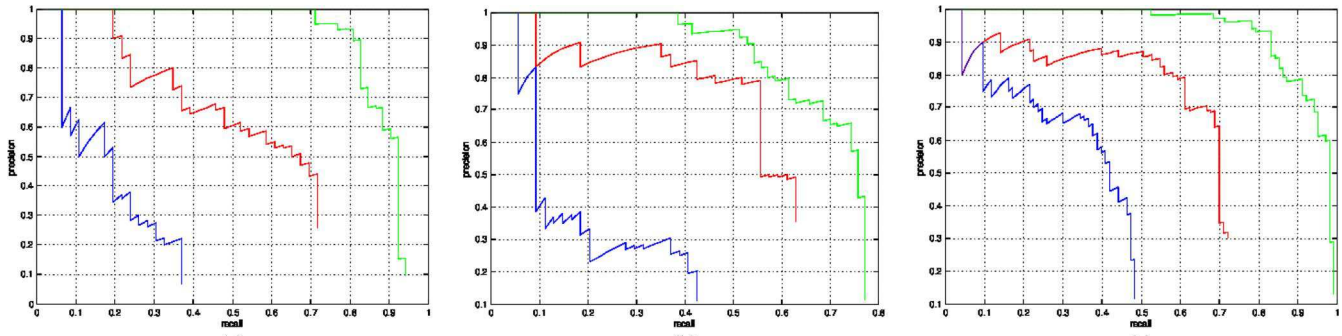


Fig. 14. Precision–recall curves for all the vehicles. (Red) **LPF**. (Blue) **ISM**. (Green) **L-SVM**. (a) Frontal view. (b) Rear view. (c) Side view.

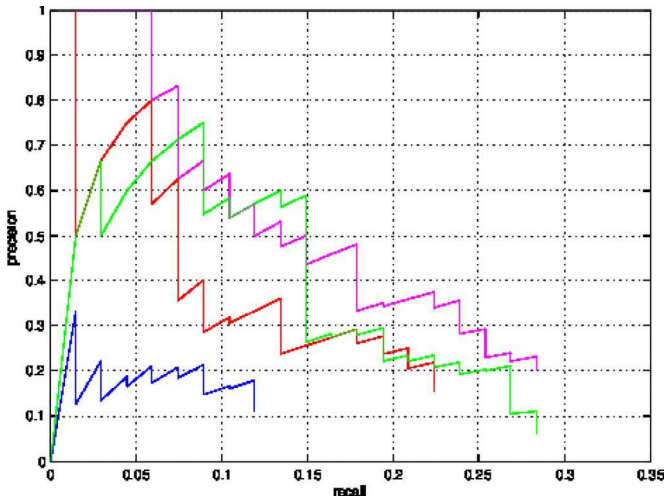


Fig. 15. Precision–recall curves for the partially occluded vehicles. Occlusion ratios: 30%–90%. (Magenta) **OR-LPF**. (Red) **LPF**. (Blue) **ISM**. (Green) **L-SVM**.

Figs. 14 and 15 show the precision–recall curves for all the vehicles and for the occluded ones, respectively. The codebooks used for these curves are the ones that gave the better overall results according to Tables III and IV (5 subsets/150 clusters for the fusion method and 90 clusters for the intensity-only method). The use of depth information results in a considerable increase in performance. In Fig. 14, we observe that the baseline method **L-SVM** outperforms the proposed **LPF**, probably due to the stronger classifier. However, compared with the **ISM** method that uses the same type of classifier, the increase in performance using depth information is significant. For instance, the proposed method detects with 80% precision, about 60% of the side views, 40% of the frontal views, and 55% of the rear views.

In Fig. 15, we test the performance of the methods for the occluded vehicles with occlusion ratios over 30%. In this setting, our **LPF** approach has comparable performance with the **L-SVM** method, whereas the **OR-LPF** approach outperforms all the other methods. It is clear that the procedure that explicitly treats occlusions significantly boosts the performance. Additionally, the difference between the fusion methods and the **ISM** is even bigger in this case; the intensity-only method has very low precision for even very small recall rates. The challenging nature of the data set with many vehicles with high occlusion ratios poses difficulties for all the methods, and therefore, the overall performance is generally low.

V. Conclusion

In this paper, we have 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 vehicle detection from a moving platform. We tested it in a real urban environment using a data set collected from our experimental vehicle. The comparison with a standard approach using only intensity information shows a significant increase in performance. Additionally, we have demonstrated that fusion on a global detection level does not improve performance. For the occluded vehicles, we show that our approach outperforms the current state-of-the-art methods; however, for the unoccluded cases, the **L-SVM** detector gives better results.

As a first future work, we consider using the stereo images data set 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 also plan to test the system with more complex 3-D descriptors extracted from the depth images. Another future goal is to derive an algorithm that will sequentially detect over increasing distances. In this way, we will be able to use high-level information of the detections in closer distances in order to robustly identify occlusions in larger distances.

References

[1] C. Keller, M. Enzweiler, M. Rohrbach, D. Llorca, C. Schnorr, and D. Gavrila, "The benefits of dense stereo for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1096–1106, Dec. 2011.

[2] D. 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.

[3] 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, May 2006.

[4] 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, Jul. 2010.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

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

[7] M. Enzweiler and D. Gavrila, "A multilevel mixture-of-experts framework for pedestrian classification," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2967–2979, Oct. 2011.
[8] B. Dalal and N. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. CVPR*, 2005, vol. 1, pp. 886–893.
[9] M. Awais and K. Mikolajczyk, "Feature pairs connected by lines for object recognition," in *Proc. ICPR*, 2010, pp. 3093–3096.
[10] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 259–289, May 2008.
[11] E. Seemann, M. Fritz, and B. Schiele, "Towards robust pedestrian detection in crowded image sequences," in *Proc. IEEE Conf. CVPR*, 2007, pp. 1–8.
[12] 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, Sep. 2010.
[13] J. Gall and V. S. Lempitsky, "Class-specific Hough forests for object detection," in *Proc. IEEE Conf. CVPR*, 2009, pp. 1022–1029.
[14] A. Opelt, A. Pinz, and A. Zisserman, "Learning an alphabet of shape and appearance for multi-class object detection," *Int. J. Comput. Vis.*, vol. 80, no. 1, pp. 16–44, Oct. 2008.
[15] R. Labayrade, D. Aubert, and J. Tarel, "Real time obstacles detection on non flat road geometry through v-disparity representation," in *Proc. IEEE IV*, Versailles, France, 2002, pp. 646–651.
[16] 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. IEEE Intell. Transp. Syst. Conf.*, Toronto, ON, Canada, 2006, pp. 745–752.
[17] 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 Trans. Image Process.*, vol. 15, no. 8, pp. 2364–2375, Aug. 2006.
[18] T. Veit, "Connexity based fronto-parallel plane detection for stereovision obstacle segmentation," in *Proc. IEEE Int. Conf. Robot. Autom., Workshop Safe Navigat. Open Dyn. Environ., Appl. Auton. Veh.*, Kobe, Japan, 2009, pp. 1–15.
[19] 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, Oct. 2008.
[20] 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. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 292–307, Jun. 2007.
[21] R. Quintero, A. Llamazares, D. Llorca, M. Sotelo, L. Bellot, O. Marcos, I. Daza, and C. Fernandez, "Extended floating car data system—Experimental study," in *Proc. IEEE IV*, 2011, pp. 631–636.
[22] J. Vinagre Diaz, D. Fernandez Llorca, A. Rodriguez Gonzalez, R. Quintero Minguez, A. Llamazares Llamazares, and M. Sotelo, "Extended floating car data system: Experimental results and application for a hybrid route level of service," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 25–35, Mar. 2012.
[23] C. Burges (1998, Jun.). A tutorial on support vector machines for pattern recognition. *Data Mining Knowl. Discovery* [Online]. *2(2)*, pp. 121–167. Available: http://dx.doi.org/10.1023/A:1009715923555
[24] M. Enzweiler, M. Hummel, D. Pfeiffer, and U. Franke, "Efficient stixel-based object recognition," in *Proc. IEEE IV*, 2012, pp. 1066–1071.
[25] D. Geronimo, A. D. Sappa, A. Lopez, and D. Ponsa, "Adaptive image sampling and windows classification for on-board pedestrian detection," in *Proc. 5th Int. Conf. Comput. Vis. Syst.*, Bielefeld, Germany, 2007, pp. 21–24.
[26] A. Makris, M. Perrollaz, I. Paromtchik, and C. Laugier, "Integration of visual and depth information for vehicle detection," in *Proc. IEEE/RSJ IROS, Workshop Perception Navigat. Auton. Veh. Human Environ.*, 2011.
[27] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[28] H. Hirschmüller, "Stereo processing by semi-global matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, 2008.
[29] M. Perrollaz, J.-D. Yoder, A. Nègre, A. Spalanzani, and C. Laugier, "A visibility-based approach for occupancy grid computation in disparity space," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1383–1393, Sep. 2012.
[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[31] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, Discriminatively trained deformable part models, release 5. [Online]. Available: http://people.cs.uchicago.edu/rbg/latent-release5/
[32] 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*. Berlin, Germany: Springer-Verlag, 2006, pp. 29–48.

Alexandros Makris received the Diploma in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, and the Ph.D. degree in computer science from the National and Kapodistrian University of Athens, Athens, in 2010.

He is currently a Postdoctoral Researcher with the Moise team, INRIA Grenoble Rhône-Alpes, France. From 2010 to 2011, he was with the E-motion team with the same research center. His main research interests include computer vision, probabilistic models, intelligent vehicles, and perception for robotics.

Mathias Perrollaz received the M.S. degree in electrical engineering from the National Polytechnic Institute of Grenoble (INPG), Grenoble, France, in 2003, with major in signal and image processing and the Ph.D. degree from Paris-6 University (UPMC), Paris, France, in 2008, for his work on multisensor obstacle detection.

He was with the Images and Signals Laboratory in Grenoble (CNRS), working on intelligent transportation systems (ITS), and with the perception team of the LIVIC (INRETS). Since April 2009, he has been with INRIA Grenoble Rhône-Alpes, Saint Ismier, France, working on probabilistic methods for ITS. Between May and September 2011, he was with Ohio Northern University, Ada, OH, USA, working on perception for robotic manipulators. He also taught at Paris-10, Grenoble-1, 2, and INPG Universities. His main research interests include computer vision, ITS, and perception for robotics.

Christian Laugier received the Ph.D. degree in computer science from Grenoble University, Grenoble, France, in 1976.

He is a Research Director with INRIA and a Scientific Leader of the e-Motion Team. He is also responsible for scientific relations with Asia & Oceania at INRIA. In addition to his research and teaching activities, he cofounded four startup companies in the fields of robotics, computer vision, computer graphics, and Bayesian programming tools. He has served as a Scientific Consultant for ITMI, Aleph Technologies, and ProBayes companies. He has coedited several books in the field of robotics and several special issues of scientific journals, such as IJRR, Advanced Robotics, JFR, or the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. His current research interests include motion autonomy, intelligent vehicles, and probabilistic robotics.

Dr. Laugier was awarded the Nakamura Prize for his contributions to "Intelligent Robots and Systems" in 1997. He is a member of several scientific committees, including the Steering/Advisory Committees of the IEEE/RSJ IROS, FSR, and ICARCV conferences. He is also a Cochair of the IEEE RAS Technical Committee on AGV and ITS. He was the General Chair or Program Chair of such conferences as the IEEE/RSJ IROS'97, IROS'02, IROS'08, or FSR'07.