Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking

Ricardo Omar Chavez-Garcia and Olivier Aycard

***Abstract*—The accurate detection and classification of moving objects is a critical aspect of advanced driver assistance systems. We believe that by including the object classification from multiple sensor detections as a key component of the object's representation and the perception process, we can improve the perceived model of the environment. First, we define a composite object representation to include class information in the core object's description. Second, we propose a complete perception fusion architecture based on the evidential framework to solve the detection and tracking of moving objects problem by integrating the composite representation and uncertainty management. Finally, we integrate our fusion approach in a real-time application inside a vehicle demonstrator from the *interactIVe* IP European project, which includes three main sensors: radar, lidar, and camera. We test our fusion approach using real data from different driving scenarios and focusing on four objects of interest: pedestrian, bike, car, and truck.**

***Index Terms*—Intelligent vehicles, sensor fusion, classification algorithms, vehicle detection, vehicle safety.**

I. Introduction

INTELLIGENT vehicles have moved from being a robotic application of tomorrow to a current area of extensive research and development. The most striking characteristic of an intelligent vehicle system is that it has to operate in increasingly unstructured environments, which are inherently uncertain and dynamic.

ADAS help drivers to perform complex driving tasks to avoid dangerous situations. Assistance tasks include: warning messages in dangerous driving situations (e.g., possible collisions), activation of safety devices to mitigate imminent collisions, autonomous maneuvers to avoid obstacles, and attention-less driver warnings.

Perceiving the environment involves the selection of different sensors to obtain a detailed description of the environment and an accurate identification of the objects of interest. Vehicle perception is composed of two main tasks: simultaneous localization and mapping (SLAM) which generates a map of the environment while simultaneously localizing the vehicle within the map given all the measurements from sensors; and DATMO which detects and tracks the moving objects surrounding the vehicle and estimates their future behavior. Fig. 1 shows the main components of the perception task.

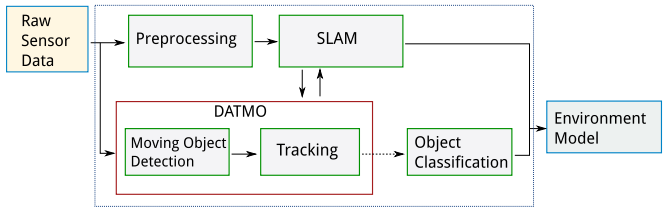

Fig. 1. General architecture of the perception task and its two main components: SLAM and DATMO. Perception provides a model of the environment usually composed by the vehicle's location, map of static objects, and a list of moving objects.

Management of incomplete information is an important requirement for perception systems. Incomplete information can be originated from sensor-related reasons, such as calibration issues, hardware malfunctions, uncertain detections and asynchronous scans; or from scene perturbations, like occlusions, weather issues and object shifting. The tracking process assumes that its inputs correspond uniquely to moving objects, and then focus on data association and tracking problems. However, in most of the real outdoor scenarios, these inputs include non-moving detections, such as noisy detections or static objects. Correctly detecting moving objects is a critical aspect of a moving object tracking system. Usually, many sensors are part of such systems.

Knowing the class of objects surrounding the ego-vehicle provides a better understanding of driving situations. Classification is seen as a separate task within the DATMO task or as an aggregate information for the final perception output [1], [2]. However, classification can help to enrich the detection stage by including information from different sensor views of the environment, e.g., impact points provided by lidar and image patches provided by camera. Evidence about the class of objects can provide hints to discriminate, confirm and question data associations. Moreover, knowing the class of a moving object benefits the motion model learning and tracking. We believe that classification information about objects of interest gathered from different sensors at early stages can improve their detection and tracking, by reducing false positive detections and mis-classifications [1], [3].

Regarding the state of the art approaches, we assume the SLAM stage as a solved task, and focus on the detection,

Manuscript received January 20, 2015; revised July 29, 2015; accepted September 4, 2015. Date of publication September 29, 2015; date of current version January 29, 2016. This work was supported by the European Commission under interactIVe, a large-scale integrating project part of the FP7-ICT for Safety and Energy Efficiency in Mobility. The Associate Editor for this paper was S. Sun.

The authors are with the Informatics Laboratory of Grenoble, Université Joseph Fourie, 38000 Grenoble, France (e-mail: ricardo.chavez-garcia@imag.fr; oliver.aycard@imag.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.2015.2479925

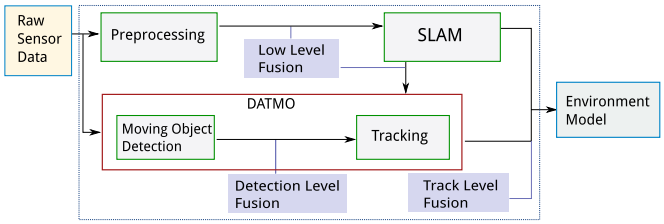

Fig. 2. Fusion levels within the SLAM and DATMO components interaction.

classification and tracking of moving objects. Precisely, we include object's class as the key component of an Evidential fusion approach that includes uncertainty management from sensor detections. The goal is to improve the results of the perception task, i.e., a more reliable list of moving objects of interest represented by their kinematic state and appearance information. Therefore, we address the problems of sensor data association, sensor fusion for object detection, and tracking. We assume that a richer list of tracked objects can improve future stages of an ADAS.

The rest of this paper is organized as follows. Section II reviews the related works. Section III introduces the concepts behind the Evidential framework. Sections IV and V describe the vehicle demonstrator and the software architecture of our vehicle application inside the *interactIVe* project. Sections VI and VII detail our proposed strategies to detect objects and extract their classification information using different sensors. In Section VIII, we present our fusion approach at detection level, and the tracking of moving objects. Experimental results are presented in Section IX. Finally, the conclusion and perspectives are stated in Section X.

II. Related Work

Fig. 2 shows the different fusion levels inside a perception system. Whilst *low level* fusion is performed within SLAM component, *detection* and *track level* fusions are performed within DATMO component. At *detection level*, fusion is performed between lists of moving object detections provided by individual sensors. At *track level*, lists of tracks from individual sensor modules are fused to produce the final list of tracks.

Promising SLAM results obtained in [1]–[3] motivated our focus on the DATMO component. Whilst Vu [1] and Wang [3] use an almost deterministic approach to perform the association in tracking, we use an evidential approach based on mass distributions over the set of different class hypotheses. Our review focuses on the fusion methods inside DATMO that use lidar, camera and radar sensors. This decision comes from our sensor set-up described in Section IV.

Multi-sensor fusion at *track level* requires a list of updated tracks from each sensor to fuse them into a combined list of tracks. The works in [2], [4], [5] solve this problem focusing on the association problem between lists of tracks, and implementing stochastic mechanisms to combine the related objects. By using an effective fusion strategy at this level, false tracks can be reduced. This level is characterized by including classification information as complementary to the final output.

Fusion at *detection level* aims at gathering and combining early data from sensor detections. Labayrade *et al.* propose to work at this level to reduce the number of mis-detections that can lead to false tracks [6]. Other works focus on data redundancy from active and passive sensors, and follow physical or learning constrains to increase the certainty of object detection [7], [8]. These works do not include all the available kinetic and appearance information. Moreover, at this level, appearance information from sensor measurements is not considered as important as the kinetic data to discriminate moving and static objects.

When classification is considered as an independent module inside the perception solution, this is often implemented as a single-class (e.g., only classifies pedestrians) or single-sensor based classification process [2], [5]. This approach excludes discriminative data from multiple sensor views that can generate multi-class modules. Research perspectives point-out the improvement of the data association and tracking tasks as a direct enhancement when classification information is managed at early levels of perception [2], [5], [9].

The most common approaches for multi-sensor fusion are based on probabilistic methods [1], [2]. However, methods based on the Evidential framework proposed an alternative not only to multi-sensor fusion but to many modules of vehicle perception [5], [6], [9]. These methods highlight the importance of incomplete and imprecise information which is not usually present in the probabilistic approaches.

An advantage of our fusion approach at the detection level is that the description of the objects can be enhanced by adding knowledge from different sensor sources. For example, lidar data can give a good estimation of the distance to the object and its visible size. In addition, classification information, usually obtained from camera, allows to make assumptions about the detected objects. An early enrichment of objects' description could allow the reduction of the number of false detections and integrate classification as a key element of the perception output rather than only an add-on.

III. Evidential Framework

The Evidential framework is a generalization of the Bayesian framework of subjective probability [10]. Evidential theory (ET) allows us to have degrees of belief for a related question according to the available evidence. ET represents the world in a set of mutually exclusive propositions known as the frame of discernment (Ω). It uses belief functions to distribute the evidence about the propositions over 2^{Ω}. The distribution of mass beliefs is done by the function $m : 2^{\Omega} \rightarrow [0,1]$, also known as Basic Belief Assignment (BBA):

$$m(\emptyset) = 0, \quad \sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

Yager's rule combines two sources of evidence while avoiding counter-intuitive results, which are present when there is a considerable degree of conflict ($m(\emptyset)$) [10]. In this rule, the conflict value is distributed among all the elements of the frame

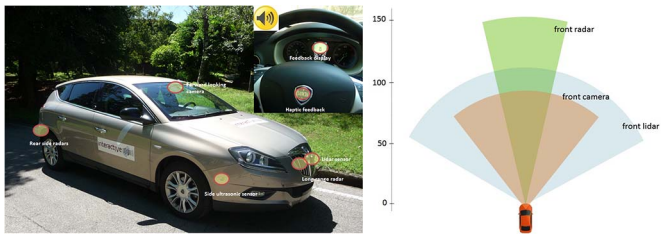

Fig. 3. Left: Images of the CRF vehicle demonstrator. Right: Field of view of the three frontal sensors used as inputs to gather datasets for our proposed fusion approach detailed in Sections VI, VII, and VIII.

of discernment rather than only the elements with intersections of the combining masses:

$$m(A) = \sum_{X_i \cap Y_j = A} m_1(X_i) m_2(Y_j), \quad A \neq \emptyset,\ A \neq \Omega$$

$$m(\Omega) = \sum_{X_i \cap Y_j = \Omega} m_1(X_i) m_2(Y_j) + \overbrace{\sum_{X_i \cap Y_j = \emptyset} m_1(X_i) m_2(Y_i)}^{\mathrm{K}}. \tag{2}$$

A. Evidential Theory for Vehicle Perception

ET has the ability to represent incomplete evidence, total ignorance and the lack of *a priori probabilities*. We can encode implicit knowledge in the definition of the structure of the frame of discernment. In addition, discounting factors are an important mechanism to integrate the reliability of the sources of evidence, such as sensor performance. Moreover, combination rules are useful tools that integrate information from different bodies of evidence. Late stages of intelligent systems, such as reasoning & decision, can integrate evidence distributions into the decision making process [11].

When the number of hypotheses is large, ET becomes less computationally tractable because the belief is distributed over the power set of all the hypotheses, 2^{Ω}. However, the application domain may allow to make assumptions to transform Ω into a reduced version of the set of possible hypotheses.

IV. Vehicle Demonstrator

We used the CRF (Fiat Research Center) demonstrator, from the *interactIVe* European project, to obtain datasets from different driving scenarios. In order to accomplish the Continuous Support functions, the Lancia Delta car (see Fig. 3) is equipped with processing units, driver interaction components, and the following front-facing set of sensors: TRW TCAM+camera gathers B&W images and has a FOV of $\pm 21^{\circ}$; TRW AC100 medium range radar provides information about moving targets. It has a detection range up to 150 m, velocity range up to 250 kph, FOV of $\pm 12^{\circ} - \pm 8^{\circ}$ (close-medium range), and angular accuracy of 0.5°; and an IBEO Lux laser scanner provides a 2D list of impact points, it has a range up to 200 m with an angular and distance resolution of 0.125° and 4 cm respectively, and a FOV of 110°.

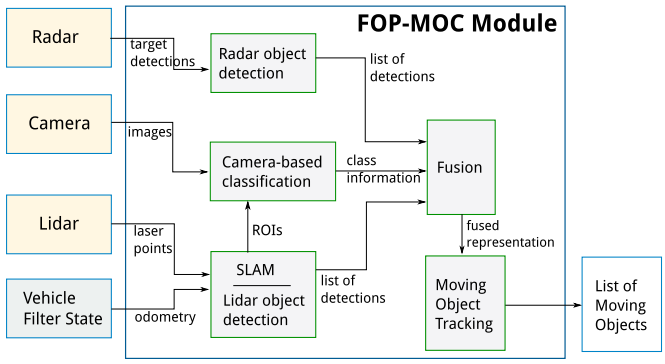

Fig. 4. Schematic of our multiple sensor perception system, also known as Frontal Object Perception (FOP)—Moving Object Detection (MOC) module. Kinetic and appearance information are extracted from lidar and radar sensors, and only appearance information from camera.

V. Software Architecture

Our contribution inside the *interactIVe* project takes place at the Perception System (PS) that aims at improving the efficiency and quality of sensor data fusion, focusing on object detection and classification. In the PS, multiple functions are developed for continuous driver support, and also for executing active interventions for collision avoidance and collision mitigation.

Fig. 4 shows the schematic of our proposed PS, and the interaction between the detection and classification modules. The PS aims at detecting, classifying and tracking a set of moving objects of interest that may appear in front of the vehicle. The inputs of the fusion module are three lists of detected objects from three sensors: lidar, radar and camera. Each object is represented by its position, size and an evidence distribution of class hypotheses. Class information is obtained from the shape, relative speed and visual appearance of the detections. Lidar and radar data are used to perform moving object detection and, in cooperation with image data, they extract object classification. Three lists of composite object descriptions are taken by our fusion approach and delivered to our tracking algorithm. The final output of the fusion method comprises a fused list of object detections that will be used for the tracking module to estimate the moving object states and deliver the final output of our DATMO solution.

VI. Moving Object Detection

In this stage, we rely on the data provided by the different sensors to detect the moving objects of interest.

A. LIDAR Processing

We consider the LIDAR (LIght Detection And Ranging) scanner as the main sensor in our configuration due to its high resolution and accuracy to detect obstacles. In addition, it powers the SLAM component of our perception solution. The main goal of the lidar processing is to get precise measurements of the shape of the moving obstacles in front of the vehicle.

1) SLAM Component Solution: Although our main contributions are focused on the DATMO component, we solve the SLAM component to obtain the map and the vehicle's pose.

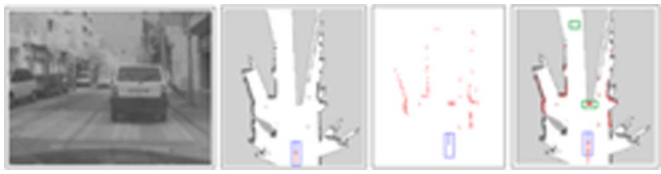

Fig. 5. Occupancy grid representation obtained by processing raw lidar data. From left to right: Reference image; static occupancy grid M_{t-1} after applying the SLAM solution; current lidar scan; detection of the moving objects (green bounding boxes).

Following the idea proposed in [1], we employ lidar data $(z_{1:t})$ for populating a two-dimensional Bayesian occupancy grid map. Each cell in the map M is associated with a measurement indicating its probability to be occupied or not by an obstacle. Vehicle's location is found by a Maximum Likelihood approach. It consists of finding the best vehicle tracks estimates (ω^*) according to a shape model $(P(\omega|z_{1:t}))$, prior model $(P(\omega))$ and likelihood model $(P(z_{1:t}|\omega))$ (Equation (3)). Afterwards, this method uses the pose estimate and the latest sensor measurements to update the grid [12].

$$\omega^* = \arg\max_{\omega\in\Omega} P(\omega|z_{1:t}) \text{ for } P(\omega|z_{1:t}) \propto P(w)P(z_{1:t}|\omega). \tag{3}$$

2) LIDAR-Based Detection: As is described in [12], we focus on identifying the inconsistencies between free and occupied cells within the grid map M while incrementally building such map. If an occupied measurement is detected on a location previously set as free, then it belongs to a moving object. If a free measurement is observed on a location previously occupied then it probably belongs to a static object.

Using a distance-based clustering process we identify clouds of cells that could belong to moving objects. This process provides information about the visible shape of the possible moving object, an estimation of its size, and the distance to the object. Fig. 5 shows an example of the evolution of the lidar-based moving object detection process. Measurements detected as parts of a moving object are not used to update the map in SLAM.

B. Camera Images

In order to obtain appearance information from images, we need to extract discriminative visual features.

1) Visual Representation: The Histograms of Oriented Gradients (HOG) descriptor has shown promising results in vehicle and pedestrian detection [13]. We took this descriptor as the core of our vehicle and pedestrian visual representation. The goal of this task is to generate visual descriptors of areas of the image to be used in future stages to determine whether these areas contain an object of interest or not.

We propose a sparse version of the HOG descriptor (S-HOG) that focuses on specific areas of an image patch. This allows us to reduce the common high-dimensional HOG descriptor [12]. Fig. 6 illustrates some of the blocks we have selected to generate the descriptors for different object classes. These blocks correspond to meaningful regions of the object (e.g., head, shoulder and legs for pedestrians). HOGs are computed over these sparse blocks and concatenated to form S-HOG descriptors. To accelerate S-HOG feature computation, we followed an *integral image* scheme [14].

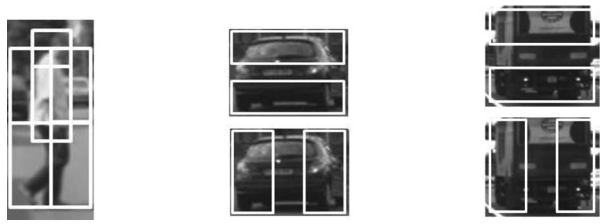

Fig. 6. Informative blocks for each object class patch, from left to right: pedestrian, car and truck. Average size of the S-HOG descriptors for pedestrians, bikes, cars and trucks are 216, 216, 288 and 288.

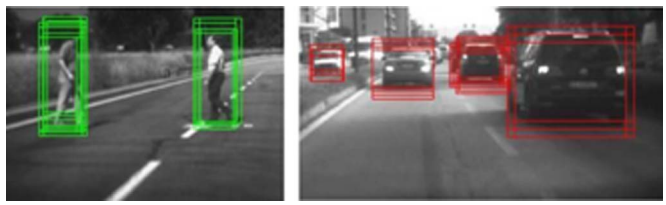

Fig. 7. Examples of successful detections of pedestrians (left) and cars (right) from camera images.

2) Object Classification: Due to performance constraints, we did not implement a visual-based moving object detection. Instead, we used the regions of interest (ROI) provided by lidar detection to focus on specific regions of the image. For each ROI, visual features are extracted, and a classifier is applied to decide if an object of interest is inside the ROI. The choice of the classifier has a substantial impact on the resulting speed and quality. We implemented a boosting-based learning algorithm called discrete Adaboost [15]. It combines many weak classifiers to form a powerful one, where weak classifiers are only required to perform better than chance.

For each class of interest (pedestrian, bike, car, truck), a binary classifier was trained off-line to identify object (positive) and non-object (negative) patches. For this training stage, positive images were collected from public (such as the Daimler dataset) and manually labeled datasets containing objects of interest from different object's viewpoints (frontal, rear, profile) [9].

Fig. 7 shows examples of the pedestrian and car detection results (green and red boxes respectively) before merging into the final objects. We estimate the confidence of object classification for each possible object. Generally, the greater the number of positive areas (containing an object of interest), the higher the confidence that the object belongs to that specific class.

C. Radar Targets

The radar sensor uses a built-in mechanism to detect moving obstacles (targets), specially those with a cross-section similar to a car. The list of n targets is delivered as input to the perception approach. Each element of the list includes the range, azimuth and relative speed of the detected target. The sensor will produce a target for each object with a significant radar

cross section. However, targets may correspond to static objects or other moving obstacles, producing false positives. In a similar way, *weak objects* like pedestrians can not always be detected, consequently producing mis-detections. Due to different dynamics defining the objects of interest, we track every target using and Interactive Multiple Model (IMM) represented by *constant velocity, constant acceleration and turning* models. IMM provides a trade-off between a Generalized Pseudo Bayesian method of first (GPB1) and second (GPB2) degree [3], [16]. It only computes k Gaussians, as in GPB1, but it still has as output a mixture of k Gaussians as in GPB2. Data association between targets is achieved by a pruned Multi Hypothesis Tracking approach.

VII. Moving Object Classification

We enhanced the common kinetic representation by including class information within fusion at the detection level. This information can help to improve detection associations, better estimate object's motion, and reduce the number of false tracks. However, at detection level, there is not enough certainty about the object's class and keeping only one class hypothesis disables the possibility of rectifying a premature decision.

Our composite representation is formed by two parts: kinetic + appearance. The former includes position and shape information in a two dimensional space, inferred from the moving object detection process. The latter includes an evidence distribution $m(2^{\Omega})$ for all possible class hypotheses, where $\Omega = \{\mathbf{p}\text{edestrian}, \mathbf{b}\text{ike}, \mathbf{c}\text{ar}, \mathbf{t}\text{ruck}\}$ is the frame of discernment representing the classes of interest. This representation is used by the fusion approach to deliver a fused list of object detections, and to perform tracking.

A. Lidar Sensor

The first part of the object representation can be obtained by analyzing the shape of the detected moving objects. In the case of large detections this object is modeled by a box $\{x, y, w, l, c\}$, where x and y are the center of the box, w and l are the width and length according to the class of object c. For small detections (mainly pedestrians) a point model $\{x, y, c\}$ is used, where x, y and c represent the object center and class of the object, respectively. The position and size of the object is obtained by measuring the detected objects in the 2D occupancy grid. The class of the object is inferred from the visible size of the object and follows a fixed fitting-model approach. However, no precise classification decision can be made due to the temporary visibility of the moving objects. For example, if the width of a detected object is less than a threshold ω_{small}, we may think the object is a pedestrian or a bike but we are not sure of the real size of the object.

To define the typical size of the classes of interest, we used a priori knowledge from the distribution of the physical dimensions of several passenger cars, trucks and motorbikes sold in Europe [17]. However, instead of keeping only one class decision, we define a basic belief assignment $m_l(A)$ (Equation (4)) for each $A \in \Omega$, which describes an evidence distribution for the class of the moving object detected by lidar. We include class-related factors ($\alpha_p, \alpha_b, \alpha_c$ and α_t) to represent the lidar's performance to detect pedestrians, bikes, cars and trucks, respectively. Also we use discounting factors (γ_b and γ_c) to indicate the uncertainty of the lidar processing for mis-detecting a bike or car.

When a bike is detected, due to visibility issues the detected object can still be a part of a *car* or a *truck*, for that reason evidence is also put in $\{\mathbf{b}, \mathbf{c}, \mathbf{t}\}$. For the same reason, when a $\mathbf{t}$*ruck* is detected, we are almost sure it cannot be a smaller object. In all the cases, the ignorance hypothesis Ω represents the lack of knowledge and the general uncertainty about the class.

$$m_l(A) = \begin{cases} m_l(\{\mathbf{p}\}) = \alpha_p & \text{if class} = \mathbf{p} \\ m_l(\Omega) = 1 - \alpha_p & \\ m_l(\{\mathbf{b}\}) = \gamma_b \alpha_b & \text{if class} = \mathbf{b} \\ m_l(\{\mathbf{b}, \mathbf{c}, \mathbf{t}\}) = \gamma_b(1 - \alpha_b) & \\ m_l(\Omega) = 1 - \gamma_b & \\ m_l(\{\mathbf{c}\}) = \gamma_c \alpha_c & \text{if class} = \mathbf{c} \\ m_l(\{\mathbf{c}, \mathbf{t}\}) = \gamma_c(1 - \alpha_c) & \\ m_l(\Omega) = 1 - \gamma_c & \\ m_l(\{\mathbf{t}\}) = \alpha_t & \text{if class} = \mathbf{t} \\ m_l(\Omega) = 1 - \alpha_t. & \end{cases} \tag{4}$$

B. Camera Sensor

We follow the image processing described in Section VI-B2 to obtain an appearance-based evidence distribution of the object classes. Lidar detection process provides a set of ROIs which we use for hypotheses generation. For hypotheses verification, we use the built off-line classifiers to classify the different objects.

The camera-based classification generates several sub regions inside each ROI to cover many possible scale and size configurations. Sometimes a ROI can contain more than one object of interest. Once we have obtained the object classification for each ROI, we generate a basic belief assignment m_c following the Equation (5). This belief assignment represents the evidence distribution for the classes hypotheses in Ω of each object detected for camera processing, where α_p, α_c and α_t are confidence factors and c_c represents the camera sensor's accurateness, i.e., its rate of correct predictions.

$$m_c(A) = \begin{cases} m_c(\{\mathbf{p}\}) = \alpha_p c_c & \text{if class} = \mathbf{p} \\ m_c(\{\mathbf{p}, \mathbf{b}\}) = \alpha_p(1 - c_c) & \\ m_c(\Omega) = 1 - \alpha_p & \\ m_c(\{\mathbf{b}\}) = \alpha_b c_c & \text{if class} = \mathbf{b} \\ m_c(\{\mathbf{p}, \mathbf{b}\}) = \alpha_b(1 - c_c) & \\ m_c(\Omega) = 1 - \alpha_b & \\ m_c(\{\mathbf{c}\}) = \alpha_c c_c & \text{if class} = \mathbf{c} \\ m_c(\{\mathbf{c}, \mathbf{t}\}) = \alpha_c(1 - c_c) & \\ m_c(\Omega) = 1 - \alpha_c & \\ m_c(\{\mathbf{t}\}) = \alpha_t c_c & \text{if class} = \mathbf{t} \\ m_c(\{\mathbf{c}, \mathbf{t}\}) = \alpha_t(1 - c_c) & \\ m_c(\Omega) = 1 - \alpha_t. & \end{cases} \tag{5}$$

C. Radar Sensor

Radar targets are considered as preliminary moving object detections. Therefore, to obtain the object's class we use the relative target speed delivered by the sensor. Speed threshold S_p is statistically estimated using recorded data from the slowest scenario for vehicles, urban areas. We apply the basic belief assignment m_r (Equation (6)), where α and β are confidence factors for specific classes.

$$m_r(A) = \begin{cases} m_r(\Omega) = \alpha & \text{if object}_{\text{speed}} < S_p \\ m_r(\{\mathbf{p}, \mathbf{b}\}) = 1-\alpha & \\ m_r(\Omega) = 1-\beta & \text{if object}_{\text{speed}} >= S_p \\ m_r(\{\mathbf{c}, \mathbf{t}\}) = \beta. & \end{cases} \tag{6}$$

VIII. Fusion Approach

Once we have performed moving object detection for each sensor input, and defined a composite object representation, the next task is the fusion of object detections and tracking. We propose a multi-sensor fusion framework placed at the detection level. Although this approach is presented using three main sensors, it can be extended to work with more sources of evidence by defining extra detection modules that are able to deliver the object representation previously defined.

A. Data Association

When working with many sources of evidence, it is important to consider the problem of finding which object detections are related among the different lists of detections provided by the sensors (sources of evidence).

The combination of information at detection level has the advantage of increasing the reliability of the detection result by reducing the influence of inaccurate, uncertain, incomplete, or conflicting information from sensor measurements or object classification modules.

Let us consider two sources of evidence S_1 and S_2. Each of these sources provides a list of detections $A = \{a_1, a_2, \ldots, a_b\}$ and $B = \{b_1, b_2, \ldots, b_n\}$, respectively. In order to combine the information of these sources, we need to find the associations between the detections in A and B. All possible associations can be expressed as a matrix of magnitude $|A \times B|$ where each cell represents the evidence m_{a_i,b_j} about the association of the elements a_i and b_j for $i \leq |A|$ and $j \leq |B|$. We can define three propositions regarding the possible association $P(a_i, b_j)$:

- 1: if a_i and b_j are the same object.
- 0: if a_i and b_j are not the same object.
- Ω : ignorance about the association (a_i, b_j).

Let us define $\Omega_d = \{1,0\}$ as the frame of discernment to represent the aforementioned propositions. Therefore, $m_{a_i,b_j}(\{1\})$ and $m_{a_i,b_j}(\{0\})$ quantify the evidence supporting the proposition $P(a_i, b_j) = 1$ and $P(a_i, b_j) = 0$ respectively, and $m_{a_i,b_j}(\{1,0\})$ stands for the ignorance, i.e., evidence that cannot support the other propositions. These propositions can be addressed by finding similarity measures between detections in A and B.

Sensors S_1 and S_2 can provide detections of a different kind. These detections can be represented by a position, shape, or appearance information, such as class. Hence, m_{a_i,b_j} has to be able to encode all these similarities. Let us define m_{a_i,b_j} in terms of its similarity value as follows:

$$\begin{aligned} m_{a_i,b_j}(\{1\}) = \alpha_{i,j}, \quad m_{a_i,b_j}(\{0\}) = \beta_{i,j} \\ m_{a_i,b_j}(\{1,0\}) = 1 - \alpha_{i,j} - \beta_{i,j} \end{aligned} \tag{7}$$

where $\alpha_{i,j}$ and $\beta_{i,j}$ quantify the evidence supporting the singletons in Ω_d for the detections a_i and b_j, i.e., the similarity measures between them.

We can define m_{a_i,b_j} as the fusion of all possible similarity measures to associate detections a_i and b_j. Therefore, we can assume that individual masses of evidence carry specific information about these two detections. Let us define m^p as the evidence measurements about the position similarity between detections in A and B provided by sources S_1 and S_2 respectively; and m^c as the evidence measurements about the appearance similarity.

Following the analysis made in Section III-A, we use Yagers's combination rule defined in Equation (2) to represent m_{a_i,b_j} in terms of $m^p_{a_i,b_j}$ and $m^c_{a_i,b_j}$ as follows:

$$\begin{aligned} m_{a_i,b_j}(A) &= \sum_{B \cap C = A} m^p_{a_i,b_j}(B) m^c_{a_i,b_j}(C) \\ K_{a_i,b_j} &= \sum_{B \cap C = \emptyset} m^p_{a_i,b_j}(B) m^c_{a_i,b_j}(C) \\ m_{a_i,b_j}(\{\Omega_d\}) &= m'_{a_i,b_j}(\{\Omega_d\}) + K_{a_i,b_j} \end{aligned} \tag{8}$$

where $m^p_{a_i,b_j}$ and $m^c_{a_i,b_j}$ represent the evidence about the similarity between detections a_i and b_j according to position and class information. In addition, the associative property of this rule allows to combine several sources of evidence (sensors) [10], [12].

Once matrix $M_{A,B}$ is built, we analyze the evidence distribution m_{a_i,b_j} for each cell to decide if there is an association $(m_{a_i,b_j}(\{1\}))$, there is not$(m_{a_i,b_j}(\{0\}))$, or we have not enough evidence to decide $(m_{a_i,b_j}(\{1,0\}))$ which can probably be due to noisy detections.

When two object detections are associated, the method combines the object representations by fusing the evidence distributions for class information. This fusion is achieved by applying the combination rule described in Equation (2). The fused object representation (kinetic+appearance) is passed as input to the tracking stage to be considered in the objects motion model estimation. Non-associated object detections are passed as well expecting to be deleted by the tracking process if they are not confirmed by new evidence.

It is important to notice that not all the sensors provide the same amount and type of information. For example, while radar data do not include information about the shape of the target, lidar data provide information about the position and the shape of the object. If two associated detections have complementary information, this is passed directly to the fused object representation; if the information is redundant, it is combined according to its type. For the position, we use the Bayesian

based fusion presented in [2], which combines the position information of two detections by integrating their covariance matrices. Shape information is usually provided only by the lidar. As stated above, class information is combined using the evidential combination rule from Equation (2).

In the next sections, we review our proposed methods to extract similarity information from the position and class of the detections. This information is included in Equation (8) to decide if two detections are associated.

1) Position Similarity: According to the position of two detections a_i and b_j, we encode their similarity evidence in $m^p_{a_i,b_j}$. Based on their positions, we can define function d_{a_i,b_j} as a distance function that satisfies the properties of a pseudo-distance metric. We choose Mahalanobis distance due to its ability to include the correlations of the set of distances. Therefore, a small value of d_{a_i,b_j} indicates that detections a_i and b_j are part of the same object; and a large value indicates the opposite. All the propositions for $m^p_{a_i,b_j}$ belong to the frame of discernment Ω_d. Hence, the BBA for $m^p_{a_i,b_j}$ is described as follows:

$$\begin{aligned} m^p_{a_i,b_j}(\{1\}) &= \alpha f(d_{a_i,b_j}), \quad m^p_{a_i,b_j}(\{0\}) \\ &= \alpha\left(1 - f(d_{a_i,b_j})\right) m^p_{a_i,b_j}, \quad (\{1,0\}) = 1 - \alpha \end{aligned} \tag{9}$$

where $\alpha \in [0,1]$ is an evidence discounting factor and $f(d_{a_i,b_j}) \to [0,1]$. The smaller the distance, the larger value given by function f. In our case we choose f as:

$$f(d_{a_i,b_j}) = \exp(-\lambda d_{a_i,b_j}) \tag{10}$$

where λ is used as a threshold factor that indicates the border between close and far distances.

2) Class Dissimilarity: Contrary to the evidence provided by position, class information does not give direct evidence that supports the proposition $P(a_i, b_j) = 1$. This means that even if two detections are identified with the same class, one can not affirm that they are the same object. This is due to the fact that there can be multiple different objects of the same class in the current driving scenario. However, it is clear that if two detections have different classes it is more likely that they belong to different objects. Hence, we use the class to provide evidence about the dissimilarity between detections: $m^c_{a_i,b_j}$. The frame of discernment for the class evidence distribution is the set $\Omega = \{\mathbf{p}, \mathbf{b}, \mathbf{c}, \mathbf{t}\}$. The frame of discernment for detections' association is Ω_d and was described in Section VIII-A. Hence, we transfer the evidence from in Ω to Ω_d as follows:

$$\begin{aligned} m^c_{a_i,b_j}(\{1\}) &= 0 \\ m^c_{a_i,b_j}(\{0\}) &= \sum_{A \cap B = \emptyset} m^c_{a_i}(A) m^c_{b_j}(B), \quad \forall A, B \subset \Omega \\ m^c_{a_i,b_j}(\{1,0\}) &= 1 - m^c_{a_i,b_j}(\{0\}) \end{aligned} \tag{11}$$

which means that we fuse the mass evidences where no common class hypothesis is shared between detections in lists A and B. $m^c_{a_i}$ and $m^c_{b_j}$ represent the BBAs for the class hypotheses of detections in lists A and B. However, as we have no information about the possible relation of detections with the same class, we place the rest of the evidence in the ignorance hypothesis $\{1,0\}$.

B. Moving Object Tracking

Using the combined list of object detections provided by our fusion approach, we modified the model-based moving object tracking approach described in [1]. We adapted a MCMC sampling process using our composite representation to find the best trajectories of tracks (hypotheses) in a sliding window of time. Generated object hypotheses are then put into a top-down process taking into account all the object dynamics models, sensor model, and visibility constraints. However, instead of searching in all the possible neighbor hypotheses, we use the class evidence distribution of each object detection to reduce the search space by considering the hypotheses with more mass evidence. Two objects have similar classes if their classes belong to the same general set. Two sets of classes are defined as general: $\mathbf{vehicle} = \{\mathbf{c}, \mathbf{t}\}$ and $\mathbf{person} = \{\mathbf{p}, \mathbf{b}\}$.

If an object has a high evidence mass in the hypothesis $\{\mathbf{c}\}$, we only sample the possible hypotheses for $\mathbf{c}$ and $\mathbf{t}$. When the highest mass evidence is placed in a non-singleton hypothesis, such as $\mathbf{vehicle}$, the search space is expanded to include $\mathbf{c}$ and $\mathbf{t}$ samples alike.

We perform a dynamic fusion strategy, as described in [9], to associate the object's current state delivered by our fusion approach, and the object description of the current track. This allows keeping the object class information up-to-date each time new sensor data is available. Hence, the final output of our DATMO solution is composed of a list of moving objects described by their kinetic information and by a set of all the possible class hypotheses represented by masses of evidence.

IX. Experimental Results

Using the sensor set-up described in Section IV, we gathered four datasets from real scenarios: two datasets from urban areas and two data sets from highways. Both data sets were manually tagged in order to provide a ground truth reference. We analyzed the degree of improvement achieved by early inclusion of class information within the DATMO component. Moreover, we performed a comparison between the fusion approach at tracking level described in [9] and our fusion approach at detection level using the same experimental scenarios.

In our DATMO solution at detection level, we first performed SLAM with the lidar sensor measurements (see Section VI-A) to detect the possible moving entities. Among the 2D position state for each detection, we define the frame of discernment $\Omega = \{\mathbf{p}, \mathbf{b}, \mathbf{c}, \mathbf{t}\}$ for its evidence class distribution. Therefore, 2^Ω is the number of all the possible class hypotheses for each detection. Then, the object representations for lidar, radar and camera detections are extracted following the methodologies presented in Sections VI and VII. Once we obtained the object representations, we perform the fusion at detection level and the tracking of the fused list of objects as detailed in Section VIII.

Fig. 8 shows two output examples of our complete PS in highway and urban areas. Both scenarios are considered as high-traffic scenarios due to the large number of moving objects

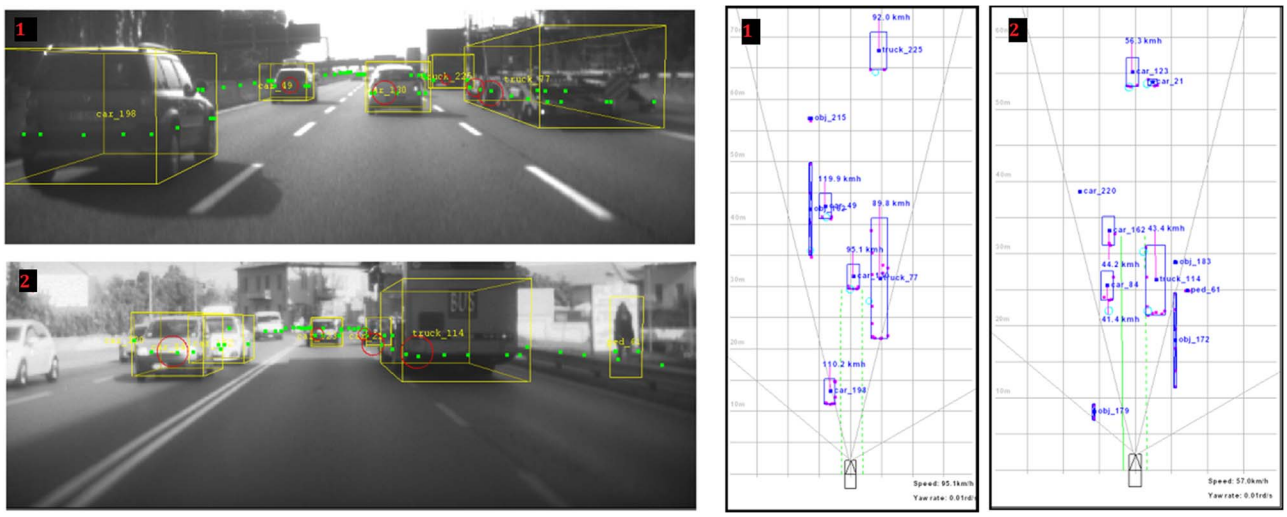

Fig. 8. Results of the PS for highway (1) and urban areas (2). Several objects of interest are detected. Left: camera image and the identified moving objects. Yellow boxes represent moving objects, red dots represent lidar hits and red circles represent radar detections. Right: top view of the same scene. Tags identify detected object's classes. Video demonstrations of our results can be found in http://goo.gl/FuMBC2.

around the vehicle. In both cases, all vehicles, including oncoming ones, are well detected, tracked and correctly classified: several cars and two trucks in the highway; and several cars, one truck and one pedestrian in the urban area. Additionally, static objects (such as barriers) are also reported and correctly identified as static obstacles using the method described in Section VI-A1. In the top view of these examples, moving objects velocity is estimated by the model-based tracking module which takes advantage of the composite object representation to deliver speed and orientation. In the early fusion stage, the radar Doppler velocity information helps to improve the target speed estimated by the lidar after its moving direction is known. Also, the class of the object is improved by the fused information from the three different sensors providing a more reliable class hypothesis in the form of a class distribution. The continuous support applications use this class distribution to decide the correct actions.

In the output of our perception approach, moving objects are represented by: location, geometry, object class, speed, and moving direction. The size of the bounding box is updated using the visible lidar measurements, the fixed-size class models and the lateral information from camera classification. The height of a bounding box is set according to the class of the detected object and to the result from camera classifiers.

Tables I and II show a comparison between the results obtained by the proposed fusion approach at detection level and our previous fusion approach at track level presented in [9]. It takes into account the erroneous classifications of moving objects. We use four datasets to conduct our experiments: 2 datasets from highways and 2 datasets from urban areas. We can see that the improvement of the fusion at detection level in highways with respect to the track level fusion is not considerable. However, in high-speed situations, the certainty about the moving vehicles is quite important. Hence, this small improvement is very useful for the final applications, such as continuous support systems. Urban areas represent a modern challenge for vehicle perception. The improvement of the fusion approach at detection level was considerable compared to our previous fusion approach. Here, the richer representation of sensor detections and the data association relations allowed the early detection of real moving vehicles.

TABLE I
FUSION RESULTS. NUMBER OF **c** AND **t** MIS-CLASSIFICATIONS

Dataset	Moving objects	Number of vehicle mis-classifications	
		Tracking level	Detection level
Highway 1	110	6	4
		5.4%	3.6%
Highway 2	154	7	5
		4.5%	3.2%
Urban 1	195	20	10
		10.2%	5.1%
Urban 2	233	24	9
		10.3%	3.8%

TABLE II
FUSION RESULTS. NUMBER OF **p** AND **b** MIS-CLASSIFICATIONS

Dataset	Moving objects	Number of pedestrian mis-classifications	
		Tracking level	Detection level
Urban 1	52	11	6
		21.1%	11.53%
Urban 2	58	14	7
		24.13%	12%

Regarding the pedestrian classification results, we obtained similar improvements to those obtained for vehicle detections. The problem of small clusters detected by lidar as moving obstacles but without the certainty of being classified as pedestrians is mainly overcome by the early combination of class information from radar and camera-based classification. Furthermore, the classification of moving objects (not only pedestrians) in our proposed approach takes on average less sensor scans than the compared fusion approach described in [9]. This is due to the early integration of the knowledge about the class of detected objects placed in m_a^c and m_b^c, which is directly related to the reduced search space for the shape and motion model discovering process performed by the MCMC technique.

TABLE III
RESULTS OF OUR PS IN FOUR SCENARIOS: HIGHWAY, URBAN AREA, RURAL ROAD AND TEST TRACK. FOUR OBJECTS OF INTEREST ARE CONSIDERED: **p**EDESTRIAN, **b**IKE, **c**AR, AND **t**RUCK

Scenario	Total objects				Detections					Classifications							
					Correct				False	Correct				False			
	p	**b**	**c**	**t**	**p**	**b**	**c**	**t**	all	**p**	**b**	**c**	**t**	**p**	**b**	**c**	**t**
Highway	0	0	702	281	n/a	n/a	687	271	22	n/a	n/a	669	251	0	0	4	0
					n/a	n/a	97.8%	96.4%	2.2%	n/a	n/a	95.2%	89.3%	0%	0%	0.5%	0%
Urban	65	7	619	97	57	6	580	88	17	57	6	570	78	9	1	6	5
					87.6%	85.7%	93.6%	90.7%	2.1%	87.6%	85.7%	92.0%	80.4%	13.8%	14.2%	0.9%	5.1%
Rural	9	0	68	6	9	n/a	62	5	9	9	n/a	60	5	3	0	5	2
					100%	n/a	91.1%	83.3%	10.8%	100%	n/a	88.2%	100%	33.3%	0%	7.3%	33.3%
Test track	248	0	301	0	247	n/a	300	n/a	1	240	n/a	300	n/a	0	0	0	0
					99.6%	n/a	100%	n/a	0.1%	96.7%	n/a	100%	n/a	0%	0%	0%	0%

A. On-Line Evaluation

Based on the running-time statistics of our PS, in urban areas (the most challenging scenario), the average computing time is 40 ms which fulfills the processing time requirement of the designed real-time platform (75 ms). In rural areas and highways, the processing of the whole PS can be reduced to 30 ms.

Table III summarizes the results collected after testing our PS with on-line data in four different scenarios. Correct detections represent true moving objects. False detections represent detections wrongly recognized as moving objects. Correct classifications represent well classified moving objects. False classifications are self-explanatory. For clarity sake, the number of correct and false detections, and classifications are also represented by percentages. Four objects of interest were taken into account: pedestrian, bike, car and truck.

In test track scenarios, where only few cars and pedestrians are present, the detection and classification rate of pedestrians and cars are nearly perfect (96–100%). This scenario does not contain many common driving situations, such as several moving objects and high traffic dynamics. However, it allows us to test specific components of the PS, e.g., pedestrian and vehicle classification, and moving vehicle tracking.

In highways, the detection rate of vehicles is also improved: car (97.8%), truck (96.4%) where the missed detections are due mainly to inherently noisy and cluttered data (e.g., lidar impacts on the ground). The large size of the truck makes the truck detection not as accurate as car detection since it is sometimes confused with the barrier. The false detection rate (2.2%) is due mainly to the reflection in raw lidar data which creates ghost objects and the noisy radar target detection. However, the fusion approach allows to obtain a highly correct classification rate for both cars or trucks whilst keeping a very low false classification rate.

In urban areas, vehicle detection and classification is still high, considering the increased number of moving obstacles and the cluttered environment. However, the false detection rate is higher than in highway scenarios. This increase is due to the highly dynamic environment and to the reduced field of view in high traffic situations. Moreover, the pedestrian false classifications commonly appears when the classifiers *misclassify* traffic posts as pedestrians. These mis-classifications suggest the construction of more robust visual classifiers or the implementation of more discriminating visual descriptors.

In Rural roads, several moving objects may appear, but high traffic dynamics are not present. Besides, there are less traffic landmarks. The high false-classification rate in this scenario is due to the increasing number of natural obstacles, such as bushes and trees. The common false classifications are due to false moving objects (mainly bushes) preliminary classified as trucks or pedestrians. One solution could be to implement a dedicated classifier to discard this type of obstacles.

X. CONCLUSION AND PERSPECTIVES

In this paper we have reviewed the problem of intelligent vehicle perception. Specifically, we have focus on the DATMO component of the perception task. We have proposed the use of classification information as a key element of a composite object representation, where not only kinetic information but appearance information plays an important role in the detection, classification and tracking of moving objects of interest. We have analyzed the impact of our composite object description by performing multi-sensor fusion at detection level. We used three main sensors to define, develop, test and evaluate our fusion approach: lidar, radar, and camera. Moreover, our complete perception solution was evaluated using on-line and off-line data from a real vehicle of the *interactIVe* European project.

Integrating class information at the detection level, allowed the fusion to improve the detection by considering an evidence distribution over the different class hypotheses of the detected objects. This improvement directly reduces the number of false detections and false classifications at early levels of the DATMO component. In addition, the tracking stage benefits from the reduction of mis-detections and from the more accurate classification information to accelerate the tracking process.

A. Perspectives

As is shown in [18], 3D-based representations (e.g., voxels segments) can provide more information about the geometry and class of the objects of interest around the ego-vehicle, and the common obstacles that generate false classifications (e.g., trees, bushes and poles).

Section IX has shown that sometimes the classification precision varies according to the current driving scenario. Promising results on the field of scene classification can power context-based learning methods to estimate parameters in the detection and classification modules, thus generating reliability factors closer to the real driving situation.

Currently, we are working on the extension of our multisensor fusion approach as an integral sensory-motor perception method for developmental systems. This method aims at learning representations and motor skills while interacting with the environment.

REFERENCES

[1] T.-D. Vu, "Vehicle perception: Localization, mapping with detection, classification and tracking of moving objects," Ph.D. thesis, Inst. Nat. Polytech. De Grenoble, Univ. Grenoble, Grenoble, France, 2009.

[2] Q. Baig, "Multisensor data fusion for detection and tracking of moving objects from a dynamic autonomous vehicle," Ph.D. dissertation, Lab. Inf. Grenoble, Univ. Grenoble, Grenoble, France, 2012.

[3] C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *Int. J. Robot. Res.*, vol. 26, no. 9, pp. 889–916, Sep. 2007.

[4] C. Mertz *et al.*, "Moving object detection with laser scanners," *J. Field Robot.*, vol. 30, no. 1, pp. 17–43, Jan. 2013.

[5] F. Fayad and V. Cherfaoui, "Detection and recognition confidences update in a multi-sensor pedestrian tracking system," in *Proc. Inf. Process. Manage. Uncertainty Knowl.-Based Syst.*, 2008, pp. 409–416.

[6] R. Labayrade, D. Gruyer, C. Royere, M. Perrollaz, and D. Aubert, "Obstacle detection based on fusion between stereovision and 2D laser scanner," in *Mobile Robots: Perception & Navigation*. Augsburg, Germany: Pro Literatur Verlag, 2007.

[7] M. Perrollaz, C. Roy, N. Hauti, and D. Aubert, "Long range obstacle detection using laser scanner and stereovision," in *Proc. IEEE Intell. Veh. Symp.*, 2006, pp. 182–187.

[8] M. Skutek, D. Linzmeier, N. Appenrodt, and G. Wanielik, "A precrash system based on sensor data fusion of laser scanner and short range radars," in *Proc. IEEE 8th Int. Conf. Inf. Fusion*, 2005, pp. 1287–1294.

[9] R. Chavez-Garcia, T.-D. Vu, O. Aycard, and F. Tango, "Fusion framework for moving-object classification," in *Proc. IEEE 16th Int. Conf. Inf. Fusion*, 2013, pp. 1159–1166.

[10] P. Smets and R. Kennes, "The transferable belief model," in *Classic Works of the Dempster–Shafer Theory of Belief Functions*, vol. 219. ser. Studies in Fuzziness and Soft Computing R. Yager and L. Liu, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 693–736.

[11] P. Smets, "Data fusion in the transferable belief model," in *Proc. IEEE 3rd Int. Conf. FUSION*, 2000, vol. 1, pp. 21–33.

[12] R. Chavez-Garcia, T. D. Vu, and O. Aycard, "Fusion at detection level for frontal object perception," in *Proc. IEEE IV*, Jun. 2014, pp. 1225–1230.

[13] P. Dollár, 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. 743–61, Apr. 2012.

[14] P. Viola and M. Jones, "Robust real-time object detection," Cambridge Res. Lab., Cambridge, MA, USA, Tech. Rep., 2001.

[15] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–655, 2000.

[16] J. Civera, A. J. Davison, and J. M. M. Montiel, "Interacting multiple model monocular SLAM," in *Proc. IEEE ICRA*, May 2008, pp. 3704–3709.

[17] K. Dietmayer, J. Sparbert, and D. Streller, "Model based object classification and tracking in traffic scenes from range images," in *Proc. IEEE Intell. Veh. Symp.*, 2001, pp. 25–30.

[18] A. Azim, "3D perception of outdoor and dynamic environment using laser scanner," Ph.D. thesis, Lab. Inf. Grenoble, Univ. Grenoble, Grenoble, France, 2013.

Ricardo Omar Chavez-Garcia received the Ph.D. degree in computer science and mathematics from the Université Joseph Fourier (formerly Université Grenoble I), Grenoble, France, in 2014. He is currently a Postdoctoral Fellow at the Institute for Intelligent Systems and Robotics (ISIR), Paris, France. He aims at proposing and developing intelligent systems for robotic platforms. His research interests are on probabilistic and evidential approaches for robotic perception. His current work focuses on multisensor approaches for detection, classification, and tracking of multiple objects, and sensory–motor representations for developmental robotics.

Olivier Aycard received the Ph.D. degree in computer science from Henri Poincaré University (Nancy 1), Nancy, France, in 1998. He is currently an Associate Professor (with Accreditation to Supervise Research) at the Department of Computer Science, Université Joseph Fourier, Grenoble, France, and a member of the Data Analysis, Modeling and Machine Learning (AMA) Team at the Informatics Laboratory of Grenoble (LIG). His research focuses on probabilistic and machine learning techniques for sensor data analysis and interpretation. In the last decade, he has been involved in several European projects (FP6 IP Prevent, FP7 STREPS Intersafe2, and FP7 IP Interactive) in the field of perception for advanced driver assistance systems and has cooperated with international car manufacturers.