Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Multimodal Information Fusion for Urban Scene Understanding

Philippe Xu · Franck Davoine · Jean-Baptiste Bordes · Huijing Zhao · Thierry Denœux

Received: date / Accepted: date

Abstract This paper addresses the problem of scene understanding for driver assistance systems. In order to recognize the large number of objects that may be found on the road, several sensors and decision algorithms have to be used. The proposed approach is based on the representation of all available information in over-segmented image regions. The main novelty of the framework is its capability to incorporate new classes of objects and to include new sensors or detection methods while remaining robust to sensor failures. Several classes as ground, vegetation or sky are considered, as well as three different sensors. The approach was evaluated on real publicly available urban driving scene data.

Keywords Information fusion · Driving scene understanding · Theory of belief functions · Intelligent vehicles · Dempster-Shafer theory · Evidence theory

1 Introduction

Scene understanding is a very important task for advanced driver assistance systems and, more generally, modern robotics. Within it, subtasks such as road recognition, pedestrian detection or traffic sign understanding, among many others, are already by themselves very challenging. Many algorithms have been developed over the last decades to tackle those individual problems, each of them using different kinds of sensors. To make the most of existing techniques, one has to find a way to properly fuse all relevant sources of information.

There are two main difficulties when combining information of different nature. The first one is to fuse modules that deal with different classes of objects and be flexible enough to include new ones. The second one is to represent, in a common space, the outputs of sensors that perceive the world differently.

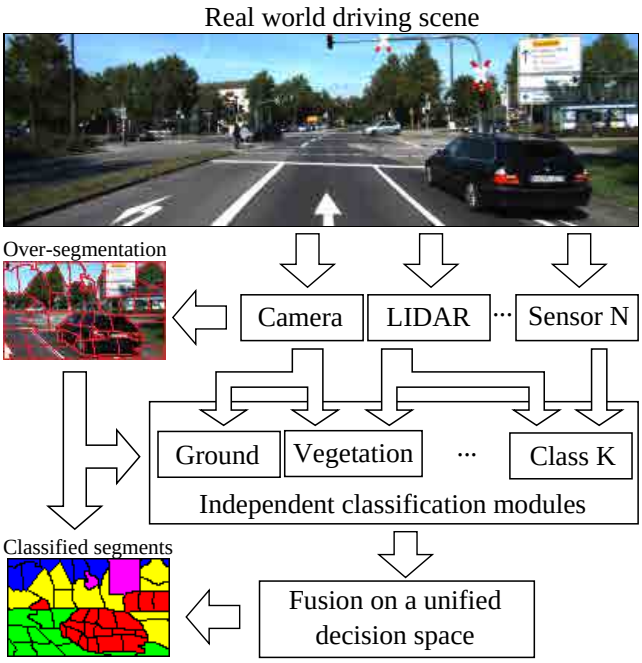


Fig. 1 Overview of the fusion framework. N sensors, including a camera, observe the scene and provide data to K independent modules. The classification outputs are then fused in a unified decision space built from an over-segmented image.

This paper is a revised and extended version of [35].

P. Xu, J.-B. Bordes, T. Denœux
UMR CNRS 7253, Heudiasyc
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex, France
E-mail: philippe.xu@hds.utc.fr

P. Xu, F. Davoine, H. Zhao
LIAMA, CNRS
Key Lab of Machine Perception (MOE)
Peking University, Beijing, P.R. China

Related Work In the field of intelligent vehicles, cameras and LiDAR (Light Detection And Ranging) are the most common sensors. LiDARs are often used to detect static structures [29] but also to detect moving objects [31]. Cameras are employed for a much wider range of applications. Pedestrian detection is one of the most studied cases [9], but more general traffic scene understanding tasks have also been considered [11,17, 20]. Depth information from stereo camera systems has proven useful to detect obstacles and mark out the navigable space [2]. Regarding the fusion aspect, many methods based on multiple sensor systems use a region of interest approach. Typically, a first sensor, for example a LiDAR [25] or a stereo camera [3], is used to select a set of interesting regions that are further analyzed. Other methods use geometric cues like the ground plane to infer constraints for object detection [21]. Another typical kind of fusion is the combination of several features [9], which can include depth [11]. Such fusion approaches are specialized to achieve a single task and are often implemented sequentially. Moreover, the outputs of classifiers such Adaboost or SVM, used in many methods [9,11,20], cannot be directly combined with other sources of information. To get probabilistic outputs, Hoiem *et al.* [17] used a logistic regression version of Adaboost while Fröhlich *et al.* [14] used a random decision forest with Gaussian process. However, if new classes of objects have to be considered, the classifiers need to be retrained completely. The existing approaches only partially achieve our goals. In contrast, the method presented in this paper makes it possible to directly fuse the outputs of different modules in any order, regardless of their specific task.

Contributions In order to combine algorithms dealing with different classes of objects, the theory of belief functions, also known as Dempster-Shafer theory [26], will be used. We will show how to model the information returned by different kinds of modules and in what ways the theory of belief functions is more adequate than probability theory. In particular, we emphasize the ability of Dempster-Shafer theory to cope with imperfectness of information such imprecision and ignorance, which are typically not well represented by probabilities. To handle different data representations from several sensors, we will formulate the problem as an image segment labeling one. Given an over-segmented image, each module, regardless of how it perceives the environment, will classify each image segment.

Overview The system considered here consists of several sensors observing an urban scene, including a camera that produces an over-segmented image as pictured in Fig. 1. Each sensor provides data to one or more modules, which are executed totally or partially in parallel to classify each image segment. The outputs from each module are expressed as belief functions [26] and combined to make a decision about the class of each region. We will show how this framework can be applied in practice by considering three sensors: a monocular camera, a stereo camera and a LiDAR. Several modules will be described for a first simplified task: ground/non-ground classification. The ability of the proposed approach to process any number of classes will be then illustrated by adding vegetation and sky detection modules. The experimental validation of this method will be performed using data from the KITTI Vision Benchmark Suite [15].

The rest of the paper is organized as follows. The task assigned to our system will first be described as an image labeling one (Section 2). The theory of belief functions will then be introduced and contrasted with the probability theory (Section 3). The construction of mass functions will then be explained and applied to scene understanding in Section 4. Finally, the whole multimodal system will be evaluated on real urban driving scene data in Section 5, and Section 6 will conclude the paper.

2 Image segment labeling formulation

As explained above, our goal is to fuse information from different sensors, which may perceive the environment in different ways. In the context of a driver assistance system, where the task is to warn drivers about potential dangers, it seems relevant to use a labeled image that reflects what the driver sees. Reasoning at the pixel level may be too local and difficult, while reasoning at the object level (e.g., inside rectangular bounding boxes) is inadequate for certain classes of objects such as the road. We chose an intermediate way by over-segmenting the image as proposed by several authors [17,20]. Many over-segmentation algorithms based on mean-shift [7], graphs [13] or the k-means algorithm [1] can be found in the literature. We chose to use the SLIC (Simple Linear Iterative Clustering) algorithm [1], as it provides a grid-like segmentation and gives the possibility to control the size of the segments. Fig. 2 shows the over-segmentation obtained by the graph-based algorithm of Felzenszwalb and Huttenlocher [13] and the SLIC algorithm. The result obtained with SLIC is more regular.

After an over-segmentation has been performed, the common task of all the modules, whatever the data representation they use (image, 3D points cloud or optical

(a) (b)

Fig. 2 (a) Over-segmentation obtained using the graph-based approach proposed by Felzenszwalb and Huttenlocher [13]. (b) Results obtained using the SLIC algorithm [1].

flow), then becomes to label each individual image segment. The labeling is done by assigning some support to each of the classes. However, as modules may consider different kinds of objects, it is necessary to use a more general representation than the one based on probabilities. The theory of belief functions will be used for this purpose.

3 Combination of imperfect information

Let $\Omega = \{\omega_1, \ldots, \omega_K\}$ be a set of mutually exclusive classes called the *frame of discernment*, which corresponds to the set of all classes. In this section, we explain how to model and combine imperfect information about the class $\omega \in \Omega$ of an observed instance. There exist many kinds of imperfect information that cannot be properly modeled by probabilities. In particular, the notions of *uncertainty*, *imprecision* and *ignorance* are crucial in our combination framework.

3.1 Probabilistic fusion

The use of probabilistic measures is the most common way to model imperfect information. The imperfect knowledge about the true class $\omega \in \Omega$ of an instance, after observing some data $\mathbf{x} \in \mathbb{X}$, is modeled by a probability distribution over Ω.

Bayesian fusion. Probabilistic fusion relies mainly on Bayes' rule. Let $P(\omega_i|\mathbf{x_1})$ and $P(\omega_i|\mathbf{x_2})$, for i=1, …, K, be the probability distributions over Ω returned by two modules after observing some data $\mathbf{x_1} \in \mathbb{X}$ and $\mathbf{x_2} \in \mathbb{X}$, respectively. By assuming conditional independence, we get:

$$p(\mathbf{x_1}, \mathbf{x_2}|\omega_i) = p(\mathbf{x_1}|\omega_i)p(\mathbf{x_2}|\omega_i), \quad \forall i \in \{1, \ldots, K\}. \tag{1}$$

Bayes' rule then yields, for all $i \in \{1, \ldots, K\}$:

$$P(\omega_i|\mathbf{x_1}, \mathbf{x_2}) = \frac{P(\omega_i)p(\mathbf{x_1}, \mathbf{x_2}|\omega_i)}{p(\mathbf{x_1}, \mathbf{x_2})} \tag{2a}$$

$$= \frac{P(\omega_i)}{p(\mathbf{x_1}, \mathbf{x_2})} p(\mathbf{x_1}|\omega_i)p(\mathbf{x_2}|\omega_i) \tag{2b}$$

$$= \frac{p(\mathbf{x_1})p(\mathbf{x_2})}{p(\mathbf{x_1}, \mathbf{x_2})} \frac{P(\omega_i|\mathbf{x_1})P(\omega_i|\mathbf{x_2})}{P(\omega_i)} \tag{2c}$$

$$\propto \frac{P(\omega_i|\mathbf{x_1})P(\omega_i|\mathbf{x_2})}{P(\omega_i)}. \tag{2d}$$

In practice, the prior class distribution $P(\omega_i)$ is difficult to estimate and is often replaced by a uniform distribution. The combination rule (2) will be referred to as the product rule. Other combination rules that replace the product by the sum, the minimum or the maximum operator can be derived from the product rule by using different approximations [19]. In the rest of the paper, the notation $P^{\Omega}_{\mathbf{x}}(\omega)$ will be used for the probability $P(\omega|\mathbf{x})$ defined over the frame of discernment Ω. The product rule will be written as $P^{\Omega}_{\mathbf{x_1},\mathbf{x_2}} = P^{\Omega}_{\mathbf{x_1}} * P^{\Omega}_{\mathbf{x_2}}$. It is important to note that, to combine two probability distributions, they have to be defined over the same frame of discernment.

Information representation. After observing some data $\mathbf{x} \in \mathbb{X}$, the probability $P^{\Omega}_{\mathbf{x}}(\omega_i)$ can be interpreted as the confidence degree of the class $\omega_i \in \Omega$. If there exists a singleton $\omega_j \in \Omega$ such that $P^{\Omega}_{\mathbf{x}}(\omega_j) = 1$, the information is said to be certain. Otherwise, it is uncertain. The closer the probability distribution is to the uniform distribution, the less informative it is. In particular, ignorance is handled by the principle of indifference, which states that, in the absence of any evidence, a uniform distribution should be defined over all possible outcomes. Ignorance can occur when the data $\mathbf{x}$ conveys no relevant information or when is it known to be unreliable. In this case, the uniform distribution $U^{\Omega}_{\mathbf{x}}$ over Ω is used:

$$U^{\Omega}_{\mathbf{x}}(\omega_i) = \frac{1}{K}, \quad \forall \omega_i \in \Omega. \tag{3}$$

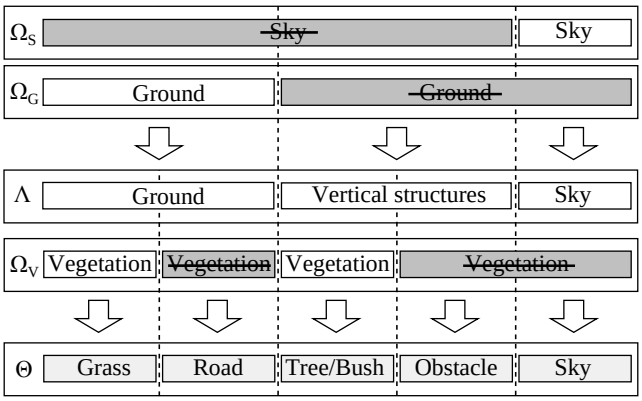


Fig. 3 Illustration of multi-class fusion. A ground detector can be combined with a sky detector by defining the "vertical" class which refers to anything that is not the ground or the sky. The combination with a vegetation detector leads to an even finer class decomposition. The "obstacle" class refers to anything that is neither the sky, the ground nor vegetation.

Reliability. If the reliability $r \in \{0,1\}$ of the source of information is known, it can be combined with the initial probability distribution $P_{\mathbf{x}}^{\Omega}$. If the source of information is reliable, *i.e.*, $r=1$, then $P_{\mathbf{x}}^{\Omega}$ is kept as it is, otherwise it is replaced by $U_{\mathbf{x}}^{\Omega}$. The combined probability $P_{\mathbf{x},r}^{\Omega}$ is derived from the law of total probability:

$$\begin{aligned} P_{\mathbf{x},r}^{\Omega}(\omega_i) =& P_{\mathbf{x}}^{\Omega}(\omega_i|r=1)P_R(r=1) && \text{(4a)}\\ &+ P_{\mathbf{x}}^{\Omega}(\omega_i|r=0)P_R(r=0) \\ =& P_{\mathbf{x}}^{\Omega}(\omega_i)P_R(r=1) + U^{\Omega}(\omega_i)P_R(r=0), && \text{(4b)} \end{aligned}$$

for all $\omega_i \in \Omega$.

Refinement. When several modules deal with different kinds of objects, it is necessary to reason with several frames of discernment with varying granularities. As stated before, two probability distributions can be combined only if they are defined over the same frame of discernment. From a frame of discernment Ω, a refinement Θ can be defined by splitting some or all its elements into new classes. A refinement from Ω to Θ can be defined [26] by an application $\rho : 2^{\Omega} \to 2^{\Theta}$ such that:

$$\bullet\ \{\rho(\{\omega\}),\ \omega \in \Omega\} \subseteq 2^{\Theta} \text{ is a partition of } \Theta; \quad \text{(5a)}$$

$$\bullet\ \forall A \subseteq \Omega,\ \rho(A) = \bigcup_{\omega \in A} \rho(\{\omega\}). \quad \text{(5b)}$$

The notation 2^{Ω} refers to the power set of Ω, which is the set of all subsets of Ω.

For instance, if a ground detector reasoning over $\Omega_G = \{\text{ground}, \cancel{\text{ground}}\}$ has to be combined with a sky detector reasoning over $\Omega_S = \{\text{sky}, \cancel{\text{sky}}\}$, a common frame of discernment $\Lambda = \{\text{ground}, \text{vertical}, \text{sky}\}$ has to be defined, as illustrated in Fig. 3. The refinement from the Ω_G to Λ is defined by:

$$\begin{cases} \rho(\{\text{ground}\}) = \{\text{ground}\}, \\ \rho(\{\cancel{\text{ground}}\}) = \{\text{vertical}, \text{sky}\}. \end{cases} \quad \text{(6)}$$

Condition (5b) will simply give $\rho(\Omega_G) = \rho(\Lambda)$. The notation $\{\cancel{\text{ground}}\}$ is used instead of $\overline{\{\text{ground}\}}$ whenever we want to specifically refer to the non-ground class as a singleton, but they both semantically refer to the same thing. The class "vertical" actually corresponds to everything that is neither the ground nor the sky, *i.e.*, $\{\text{vertical}\} = \overline{\{\text{ground}\}} \cap \overline{\{\text{sky}\}}$.

Imprecise information. An important type of imperfection that occurs when dealing with refinements is imprecise information. For example, assume that the output of a ground detector, initially defined on Ω_G, is expressed in the refined frame of discernment Λ. Let $\mathbf{x}_G \in \mathbb{X}$ be some observed data and $P_{\mathbf{x}_G}^{\Omega_G}$ be the probabilities returned by a ground detector defined as follows:

$$P_{\mathbf{x}_G}^{\Omega_G}(\text{ground}) = q,\ \ P_{\mathbf{x}_G}^{\Omega_G}(\cancel{\text{ground}}) = 1-q, \quad \text{(7)}$$

where $q \in [0,1]$. The information represented by $P_{\mathbf{x}_G}^{\Omega_G}$ can be rewritten over the refined frame Λ as:

$$P_{\mathbf{x}_G}^{\Lambda}(\text{ground}) = q,\ \ P_{\mathbf{x}_G}^{\Lambda}(\{\text{vertical}, \text{sky}\}) = 1-q. \quad \text{(8)}$$

However, expression (8) does not fully define the probability $P_{\mathbf{x}_G}^{\Lambda}$. Actually, every probability distribution P so that $P(\text{vertival})+P(\text{sky}) = 1-q$, verifies the constraints defined by (8). We say that the information represented by (8) is imprecise [30]. In such situations, the principle of indifference leads to the following probability:

$$P_{\mathbf{x}_G}^{\Lambda}(\text{ground}) = q, \quad \text{(9a)}$$

$$P_{\mathbf{x}_G}^{\Lambda}(\text{vertical}) = P_{\mathbf{x}_G}^{\Lambda}(\text{sky}) = \frac{1-q}{2}. \quad \text{(9b)}$$

As the ground detector cannot differentiate the "vertical" class from the "sky" class, the initial probability assigned the non-ground class is uniformly distributed to these two refined classes.

One major issue with such an approach is that the information represented by (9) is not exactly the same as (7). Suppose that the observation $\mathbf{x}_G$ conveys no relevant or reliable information. The initial probability $P_{\mathbf{x}_G}^{\Omega_G}$ should be uniform ($q = 1/2$), in which case the ground detector cannot make any decision. Reasoning on another frame of discernment such as Λ does not change the information at hand, which should still be modeled by a uniform distribution. However, equation (9) does not define a uniform distribution. Even worse, the ground detector would then be able to make a decision and choose the "ground" class as the most probable one. Paradoxically, if instead of $\{\cancel{\text{ground}}\}$, the

class {ground} had been refined into {grass, road}, then the same ground detector would have chosen the "non-ground" class as the most probable one. This shows that traditional probability theory cannot properly represent imprecise information.

3.2 Theory of belief functions

The issues mentioned above can be dealt by extending the notion of probability to sets of classes. The theory of belief functions, also known as Dempster-Shafer theory or evidence theory, offers a well-founded and elegant framework to do so. It is also very well suited for information fusion [18].

Information representation. A *mass function*, or *basic belief assignment*, over a frame of discernment Ω is a function $m : 2^\Omega \to [0, 1]$ verifying:

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \subseteq \Omega} m(A) = 1. \tag{10}$$

Given an object of class $\omega \in \Omega$, our belief about its membership to some subsets of Ω can be modeled by a mass function m. The quantity $m(A)$, for a given subset $A \subseteq \Omega$, represents the belief committed exactly to the hypothesis $\omega \in A$. It is important to understand that the hypothesis $\omega \in A$ does not support the membership of ω to any subset $B \subsetneq A$.

If $m(A) > 0$, then A is said to be a *focal element* of m. The state of total ignorance is then easily defined by the *vacuous mass function*, which only has Ω as focal element, *i.e.*, $m(\Omega) = 1$. The information represented by a mass function is said to be imprecise if there exists at least one non-singleton focal element. Otherwise, if a mass function has only singletons as focal elements, it actually defines a probability distribution and it is said to be *Bayesian*. Therefore, a probability distribution is a particular kind of mass function that encodes precise information. Finally, any non-vacuous mass function with only one focal element will be said to be *categorical*. These particular mass functions actually represent certain information that may precise.

Discounting. In the theory of belief functions, knowledge about the reliability of a source of information can be handled by a discounting factor [26]. It is used to weaken a mass function by transferring some mass to the ignorance state. For a factor $\alpha \in [0, 1]$, the discounted mass function ${}^\alpha m$ is defined as:

$$\begin{aligned} {}^\alpha m(A) &= (1 - \alpha)m(A), \quad \forall A \subsetneq \Omega, \\ {}^\alpha m(\Omega) &= (1 - \alpha)m(\Omega) + \alpha. \end{aligned} \tag{11}$$

If $\alpha = 0$, the information is considered reliable and is kept as is. On the other hand, if $\alpha = 1$, the information is totally unreliable and we get the vacuous mass function. Smets [27] actually showed that the discounting equation (11) can be derived by interpreting α as the probability that the source of information is not reliable. Thus, the discounting factor α plays a role equivalent to $P_R(r = 0)$ in the probabilistic case (4).

Refinement. Because mass functions are directly defined over sets of classes, refinement and imprecise information can be easily handled. Given a refinement $\rho : 2^\Omega \to 2^\Theta$, a mass function m^Ω defined over Ω can be transformed into a mass m^Θ defined over Θ, such that for all $B \subseteq \Theta$:

$$m^\Theta(B) = \begin{cases} m^\Omega(A) & \text{if } \exists A \subseteq \Omega,\ B = \rho(A), \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

On the ground detector example (Sec. 3.1, paragraph *Refinement*), a mass initially assigned to {ground} will be transferred to {grass, road}; it will not be uniformly distributed to these two subclasses as in the Bayesian case.

Evidential combination. Given two mass functions m_1 and m_2 induced, respectively, by observations $\mathbf{x_1} \in \mathbb{X}$ and $\mathbf{x_2} \in \mathbb{X}$, which are supposed to be independent, one can combine them using Dempster's rule to compute a new mass function $m_{1,2} = m_1 \oplus m_2$ defined as follows:

$$m_{1,2}(\emptyset) = 0, \tag{13a}$$

$$m_{1,2}(A) = \frac{1}{1 - \kappa} \sum_{B \cap C = A} m_1(B)m_2(C),\ \forall A \neq \emptyset, \tag{13b}$$

where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B)m_2(C). \tag{14}$$

The quantity κ measures the conflict between the two mass functions. This combination rule has the vacuous mass as unique neutral element. When Dempter's rule is used to combine two Bayesian mass functions, it is equivalent to the probabilistic product rule (2). In particular, the combination of any mass function with a Bayesian one always yields a Bayesian mass function.

Decision making. There exist several strategies for decision making [8] when reasoning within the theory of belief functions. In most cases, the mass function is first transformed into another representation. Two very important representations are the belief and plausibility functions defined, respectively, as:

$$bel(A) = \sum_{B \subseteq A} m(B), \quad \forall A \subseteq \Omega, \tag{15}$$

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \quad \forall A \subseteq \Omega. \tag{16}$$

For $A \subseteq \Omega$, $bel(A)$ measures the degree of support of A, while $pl(A) = 1 - bel(\overline{A})$ measures the lack of support to the complement of A. One has, for all $A \subseteq \Omega$, $bel(A) \leq pl(A)$. There exists a one-to-one correspondence between mass, belief and plausibility functions. For decision making, two simple strategies consist in choosing the singleton with maximum belief or plausibility [8]. They are called, respectively, the pessimistic and optimistic strategy.

Another widely used strategy is to transform the mass function into a probability measure called the pignistic probability $BetP$ [28], defined as:

$$BetP(\omega_k) = \sum_{A \subseteq \Omega, \omega_k \in A} \frac{m(A)}{|A|}, \quad \forall \omega_k \in \Omega. \tag{17}$$

The mass assigned to a set A is simply equally distributed to its elements. The singleton with maximum probability is then selected. In this paper, we adopted the optimistic strategy, which selects the singleton with maximum plausibility. This choice was motivated by the fact that this strategy is consistent with respect to refinements and is computationally efficient [4,6]. An example for which the three mentioned strategies yield different decisions is given in Appendix A.

4 Belief functions for scene understanding

We applied our framework to a multi-modal system including a stereo camera and a LiDAR sensor, which are supposed to be calibrated [15]. Several modules independently process the outputs of these sensors to classify each segment of the image in Fig 4(a). Some simple classification rules are first applied directly using pixel coordinates. The 3D information from the stereo images and the LiDAR are then used to detect the ground. Next, two monocular-based approaches allow us to infer the scene layout and further extend it by including a vegetation class. Finally, a temporal propagation module is used to link two consecutive images. The inputs of the different modules described below are shown in Fig. 4.

4.1 Classification from pixel location

Some very simple rules can be directly inferred from pixel coordinates. For example, we are certain that the "lower" part of the image cannot be the sky and the "upper" part cannot be the ground. By assuming a maximum pitch angle of $\pm 5°$, upper ($V_{\max}$) and lower ($V_{\min}$) bounds of the horizon line can be computed as illustrated by the blue and green lines in Fig. 5(a). This assumption may not hold in certain complex situations such as uphills or downhills, for which a robust horizon line estimator would be needed. A segment in the image can be described by its minimum and maximum vertical coordinate $(\underline{v}, \overline{v})$. Two distinct mass functions can then be constructed. The first one is defined over the frame of discernment $\Omega_{\mathrm{s}} = \{\text{sky}, \cancel{\text{sky}}\}$ as follows:

$$m^{\Omega_{\mathrm{s}}}_{\overline{v}}(\{\cancel{\text{sky}}\}) = \begin{cases} 1 \text{ if } \overline{v} \leq V_{\min}, \\ 0 \text{ otherwise,} \end{cases} \tag{18a}$$

$$m^{\Omega_{\mathrm{s}}}_{\overline{v}}(\{\text{sky}\}) = 0, \tag{18b}$$

$$m^{\Omega_{\mathrm{s}}}_{\overline{v}}(\Omega_{\mathrm{s}}) = 1 - m^{\Omega_{\mathrm{s}}}_{\overline{v}}(\{\cancel{\text{sky}}\}). \tag{18c}$$

This mass function states that, if the maximum vertical coordinate is lower that the lower bound $V_{\min}$, then the segment cannot be the sky. Otherwise, we do not know if the segment corresponds to the sky or not, which is represented by the vacuous mass function $m^{\Omega_{\mathrm{s}}}_{\overline{v}}(\Omega_{\mathrm{s}}) = 1$. Similarly, a second mass function is defined over $\Omega_{\mathrm{G}} = \{\text{ground}, \cancel{\text{ground}}\}$ as follows:

$$m^{\Omega_{\mathrm{G}}}_{\underline{v}}(\{\cancel{\text{ground}}\}) = \begin{cases} 1 \text{ if } \underline{v} \geq V_{\max}, \\ 0 \text{ otherwise,} \end{cases} \tag{19a}$$

$$m^{\Omega_{\mathrm{G}}}_{\underline{v}}(\{\text{ground}\}) = 0, \tag{19b}$$

$$m^{\Omega_{\mathrm{G}}}_{\underline{v}}(\Omega_{\mathrm{G}}) = 1 - m^{\Omega_{\mathrm{G}}}_{\underline{v}}(\{\cancel{\text{ground}}\}). \tag{19c}$$

These two mass functions can be combined by Dempster's rule on a common refinement $\Lambda = \{\text{ground, vertical, sky}\}$, yielding:

$$m^{\Lambda}_{\underline{v},\overline{v}}\left(\overline{\{\text{sky}\}}\right) = \begin{cases} 1 \text{ if } \overline{v} \leq V_{\min}, \\ 0 \text{ otherwise,} \end{cases} \tag{20a}$$

$$m^{\Lambda}_{\underline{v},\overline{v}}\left(\overline{\{\text{ground}\}}\right) = \begin{cases} 1 \text{ if } \underline{v} \geq V_{\max}, \\ 0 \text{ otherwise,} \end{cases} \tag{20b}$$

$$m^{\Lambda}_{\underline{v},\overline{v}}(\Lambda) = 1 - m^{\Lambda}_{\underline{v},\overline{v}}\left(\overline{\{\text{sky}\}}\right) - m^{\Lambda}_{\underline{v},\overline{v}}\left(\overline{\{\text{ground}\}}\right) \tag{20c}$$

where $\overline{\{\text{sky}\}} = \{\text{ground, vertical}\}$ and $\overline{\{\text{ground}\}} = \{\text{vertical, sky}\}$. Fig. 5(b) illustrates the combined mass functions.

Following the same reasoning, a probabilistic approach would lead to the following probabilities:

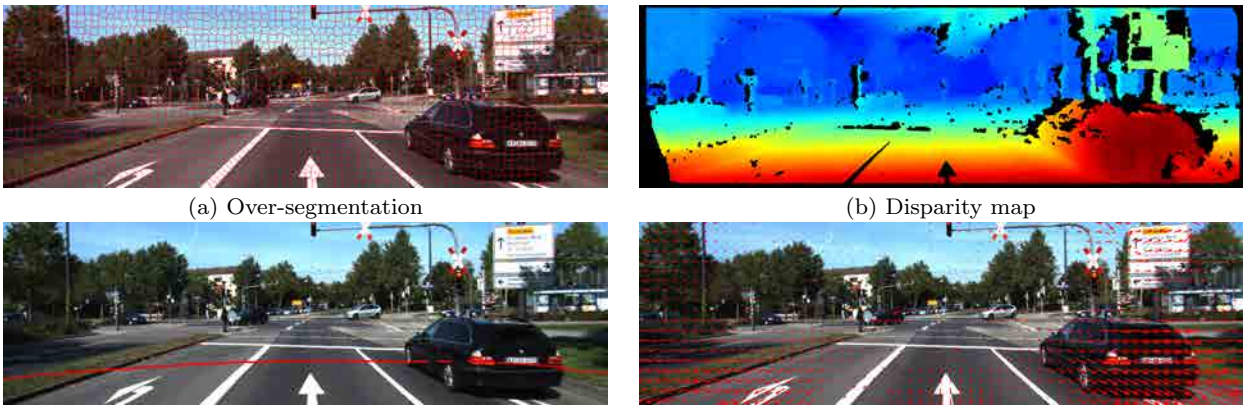
$$P^{\Omega_{\mathrm{s}}}_{\overline{v}}(\cancel{\text{sky}}) = \begin{cases} 1 \quad \text{if } \overline{v} \leq V_{\min}, \\ 1/2 \text{ otherwise,} \end{cases} \tag{21a}$$

$$P^{\Omega_{\mathrm{s}}}_{\overline{v}}(\text{sky}) = \begin{cases} 0 \quad \text{if } \overline{v} \leq V_{\min}, \\ 1/2 \text{ otherwise,} \end{cases} \tag{21b}$$

and

$$P^{\Omega_{\mathrm{G}}}_{\underline{v}}(\cancel{\text{ground}}) = \begin{cases} 1 \quad \text{if } \underline{v} \geq V_{\max}, \\ 1/2 \text{ otherwise,} \end{cases} \tag{22a}$$

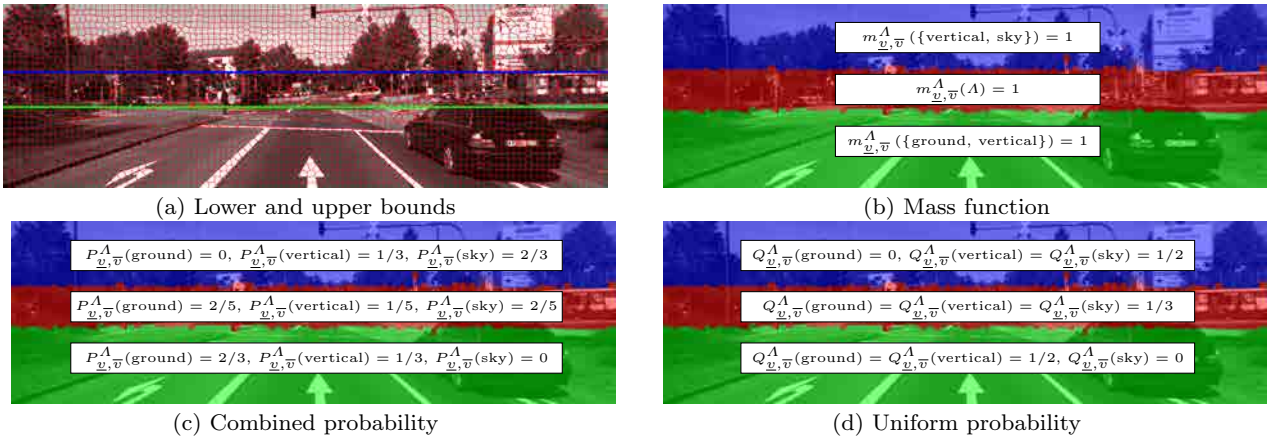
$$P^{\Omega_{\mathrm{G}}}_{\underline{v}}(\text{ground}) = \begin{cases} 0 \quad \text{if } \underline{v} \geq V_{\max}, \\ 1/2 \text{ otherwise.} \end{cases} \tag{22b}$$



(a) Over-segmentation (b) Disparity map

(c) Laser points (d) Optical flow

Fig. 4 Inputs to the multi-sensor system. (a) The over-segmentation is obtained using the SLIC algorithm [1]. (b) The disparity map is computed from the ELAS algorithm [16]. (c) A single laser layer is extracted from a Velodyne LiDAR. (d) The optical flow is computed using the TV-L1 formulation as implemented in OpenCV [36].



(a) Lower and upper bounds (b) Mass function

(c) Combined probability (d) Uniform probability

Fig. 5 Classification from pixel coordinates. (a) Lower and upper bounds of the horizon line: $V_{\min}$ (green), $V_{\max}$ (blue). (b) Mass function $m^{\Lambda}_{\underline{v},\overline{v}} = m^{\Omega_{\text{G}}}_{\underline{v}} \oplus m^{\Omega_{\text{S}}}_{\overline{v}}$. (c) Probability $P^{\Lambda}_{\underline{v},\overline{v}} = P^{\Omega_{\text{S}}}_{\overline{v}} * P^{\Omega_{\text{G}}}_{\underline{v}}$. (d) Probability $Q^{\Lambda}_{\underline{v},\overline{v}}$.

By using the product rule (2) over Λ ={ground, vertical, sky}, with a uniform prior distribution, the combined probability is defined as follows:

$P^{\Lambda}_{\underline{v},\overline{v}}$	$\overline{v} \leq V_{\min}$	$\underline{v} \geq V_{\max}$	otherwise
ground	2/3	0	2/5
vertical	1/3	1/3	1/5
sky	0	2/3	2/5

(23)

The resulting probability $P^{\Lambda}_{\underline{v},\overline{v}} = P^{\Omega_{\text{S}}}_{\overline{v}} * P^{\Omega_{\text{G}}}_{\underline{v}}$ is counter-intuitive and does not encode the same information as $P^{\Omega_{\text{S}}}_{\overline{v}}$ and $P^{\Omega_{\text{G}}}_{\underline{v}}$. In particular, complete ignorance, which is represented differently by $P^{\Omega_{\text{S}}}_{\overline{v}} = U^{\Omega_{\text{S}}}_{\overline{v}}$ and $P^{\Omega_{\text{G}}}_{\underline{v}} = U^{\Omega_{\text{G}}}_{\underline{v}}$, is not encoded by a uniform distribution in Λ. By reasoning directly on Λ, the principle of indifference would actually lead to the following probability:

$Q^{\Lambda}_{\underline{v},\overline{v}}$	$\overline{v} \leq V_{\min}$	$\underline{v} \geq V_{\max}$	otherwise
ground	1/2	0	1/3
vertical	1/2	1/2	1/3
sky	0	1/2	1/3

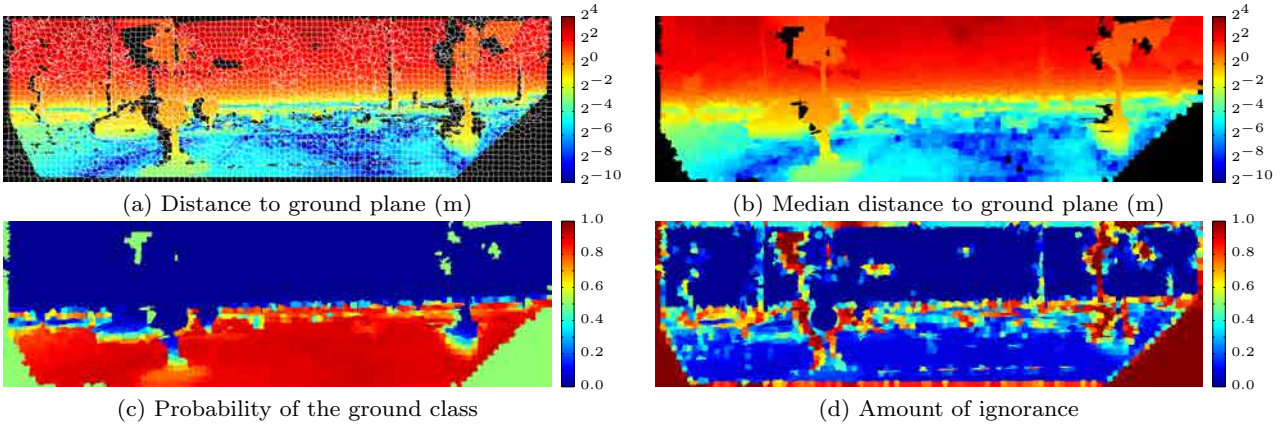
(24)

The probability distributions $Q^{\Lambda}_{\underline{v},\overline{v}}$ ans $P^{\Lambda}_{\underline{v},\overline{v}}$ are illustrated in Fig. 5(c-d). The probability $Q^{\Lambda}_{\underline{v},\overline{v}}$ seems much more reasonable than $P^{\Lambda}_{\underline{v},\overline{v}}$. In particular, the red zone, where nothing can actually be inferred, is well represented by a uniform distribution with $Q^{\Lambda}_{\underline{v},\overline{v}}$ but not with $P^{\Lambda}_{\underline{v},\overline{v}}$.

However, if a new class, such as vegetation, has to be added, none of them would actually be correct. This example clearly shows that imprecise information cannot be properly represented by probabilities. Moreover, the information on the upper and lower part of the image remains certain when using belief functions while it is encoded as uncertain with probabilities.

4.2 Stereo-based classification

3D information is very useful for scene understanding. A disparity map (Fig. 4(b)) encoding the depth of each



(a) Distance to ground plane (m) (b) Median distance to ground plane (m)

(c) Probability of the ground class (d) Amount of ignorance

Fig. 6 Stereo-based ground classification. (a) Distance to the ground plane for each pixel. (b) Median distance of segments to the ground plane. (c) Probability of the ground class $P_d^{\Omega_G}$. (d) Amount of ignorance $m_{\underline{d},\overline{d}}^{\Omega_G}(\Omega_G)$.

pixel can be estimated using a stereo camera. We used the ELAS algorithm [16] which is designed for fast high-resolution image processing.

A Euclidean 3D point cloud is first generated from the disparity map and used to estimate the ground surface. We used a robust plane estimator (RANSAC) to detect the ground plane. The assumption of a planar ground turns out to be reasonable in practice. For more robustness, the use of more complex models such as B-splines [32] could also be considered.

The estimated ground plane Π is used to build a ground detector. Each segment is seen as a set of 3D points: $\mathbf{x} = \{p_1, \ldots, p_k, p_{k+1}^*, \ldots, p_n^*\}$, where the points denoted by p_i^* are those for which no disparity has been estimated. A segment is classified as ground or non-ground depending on its distance to the ground plane. The distance d between the observation $\mathbf{x}$ and the plane Π is defined as the median distance of the valid points p_i to Π, while forgetting the invalid ones p_j^*:

$$d(\mathbf{x}, \Pi) = \underset{i=1,\ldots,k}{\text{med}}\, \delta(p_i, \Pi), \tag{25}$$

where $\delta(p_i, \Pi)$ is the Euclidean distance from p_i to Π. Fig. 6(a) illustrates the distance to the ground obtained for each pixel. Fig. 6(b) shows the median distance computed for each segment.

To get a probability measure from the distance d, a logistic regression is used by assuming that:

$$P_d^{\Omega_G}(\text{ground}) = \frac{1}{1 + \exp(ad + b)}, \tag{26}$$

where the sigmoid parameters $a, b \in \mathbb{R}$ can be optimized given some training data. Let $\{(d_i, y_i)\}_{1 \le i \le N}$ be some training data where $y_i \in \{0, 1\}$ is equal to one if the distance $d_i \in \mathbb{R}$ is associated to the ground and zero otherwise. Parameters a and b are determined by maximizing the log-likelihood function:

$$\max_{a,b \in \mathbb{R}} \sum_{i=1}^{N} y_i \log(P_i) + (1 - y_i) \log(1 - P_i), \tag{27}$$

where $P_i = P_{d_i}^{\Omega_G}(\text{ground})$. The maximization of (27) is done using Newton's method [22], which only takes $O(N)$ time per iteration. As only k out of n points are visible, the reliability of the observation is modeled by $P_R(r = 1) = k/n$. When no disparity estimates are available, *i.e.*, $P_R(r = 1) = 0$, we get the uniform distribution $P_{d,0}^{\Omega_G}(\text{ground}) = P_{d,0}^{\Omega_G}(\cancel{\text{ground}}) = 1/2$. Fig. 6(c) shows the probability obtained from the distance to the ground plane.

With belief functions, a more cautious model can be used. Instead of using the median distance, two distances $\underline{d}$ and $\overline{d}$ were considered. They correspond, respectively, to the minimum and maximum distance from the segment to the ground plane and are defined as follows:

$$\underline{d} = \min_{i=1,\ldots,k} \delta(p_i, \Pi)\ , \quad \overline{d} = \max_{i=1,\ldots,k} \delta(p_i, \Pi). \tag{28}$$

The minimum distance $\underline{d}$ is used to build a mass function $m_{\underline{d}}^{\Omega_G}$ that only supports the non-ground class. If the minimum distance is large, then we are confident about the non-ground class. However, if the minimum distance is small, nothing can actually be said. The mass function $m_{\underline{d}}^{\Omega_G}$ is defined in a way similar to (26):

$$m_{\underline{d}}^{\Omega_G}(\{\cancel{\text{ground}}\}) = \frac{1}{1 + \exp(\underline{a}\underline{d} + \underline{b})}, \tag{29a}$$

$$m_{\underline{d}}^{\Omega_G}(\{\text{ground}\}) = 0, \tag{29b}$$

$$m_{\underline{d}}^{\Omega_G}(\Omega_G) = 1 - m_{\underline{d}}^{\Omega_G}(\{\cancel{\text{ground}}\}), \tag{29c}$$

where the parameters $\underline{a}$ and $\underline{b}$ are determined by maximizing:

$$\max_{a,b \in \mathbb{R}} \sum_{i=1}^{N} y_i \log(\underline{m}_i) + (1 - y_i) \log(1 - \underline{m}_i), \tag{30}$$

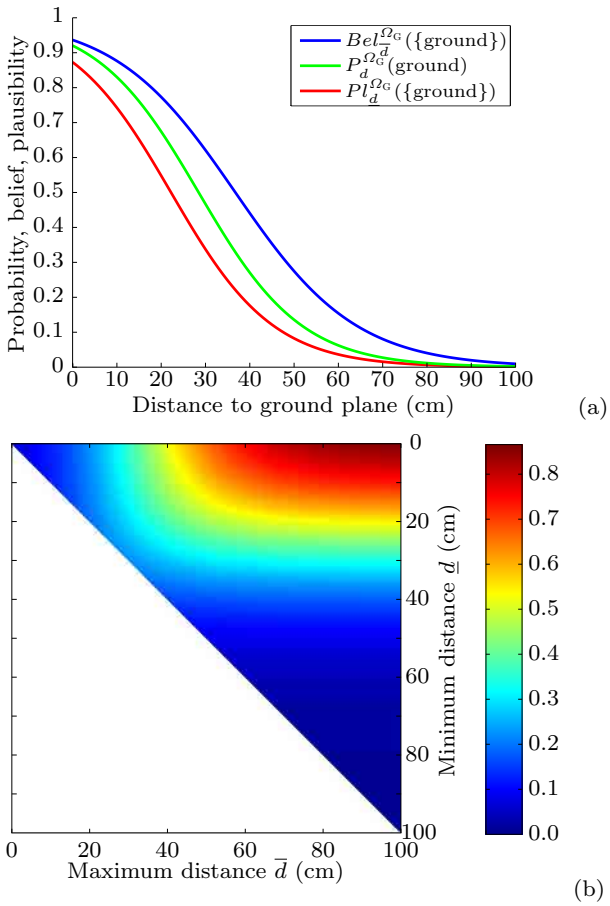


Fig. 7 (a) Probability, belief and plausibility of the ground class with respect to the distance to the ground plane. The belief and plausibility are defined as: $Bel_{\underline{d}}^{\Omega_G}(\{\text{ground}\}) = m_{\underline{d}}^{\Omega_G}(\{\text{ground}\})$ and $Pl_{\underline{d}}^{\Omega_G}(\{\text{ground}\}) = 1 - m_{\underline{d}}^{\Omega_G}(\{\not{\text{ground}}\})$. (b) Amount of ignorance $m_{\underline{d},\overline{d}}^{\Omega_G}(\Omega_G)$ given the minimum and maximum distances of a segment to the ground plane.

with $\underline{m}_i = m_{\underline{d}_i}^{\Omega_G}(\{\not{\text{ground}}\})$. In a similar way, the maximum distance $\overline{d}$ is used to build a mass function $m_{\overline{d}}^{\Omega_G}$ that only supports the ground class. A combined mass function $m_{\underline{d},\overline{d}}^{\Omega_G} = m_{\underline{d}}^{\Omega_G} \oplus m_{\overline{d}}^{\Omega_G}$ is then obtained by Dempster's rule. Finally, the mass function is discounted by a factor $\alpha = 1 - k/n$, which results in the vacuous mass function when no disparity is estimated.

Figure 7(a) shows the measures $P_d^{\Omega_G}$, $m_{\underline{d}}^{\Omega_G}$ and $m_{\overline{d}}^{\Omega_G}$ obtained from logistic regression. Figure 7(b) illustrates the amount of ignorance $m_{\underline{d},\overline{d}}^{\Omega_G}(\Omega_G)$ for different values of $\underline{d}$ and $\overline{d}$. We can see that when $\underline{d}$ is small and $\overline{d}$ is large (top right corner), the amount of ignorance is high. In contrast, when $\underline{d}$ is large (bottom right) or $\overline{d}$ is small (top left), the information is more certain. Fig. 6(d) displays the amount of ignorance in a typical case.

4.3 LiDAR-based classification

A LiDAR sensor provides a set of 3D points that are the impacts of laser beams (Fig. 4(c)). Similarly to the stereo camera case, a segment S hit by some laser beams is perceived as a set of k 3D points. By using the ground plane estimated from the disparity map, the same form of mass function as in the stereo case can be used for S. Additionally, the space between the projections on the ground plane of the laser impacts and the LiDAR's origin is considered to be obstacle free.

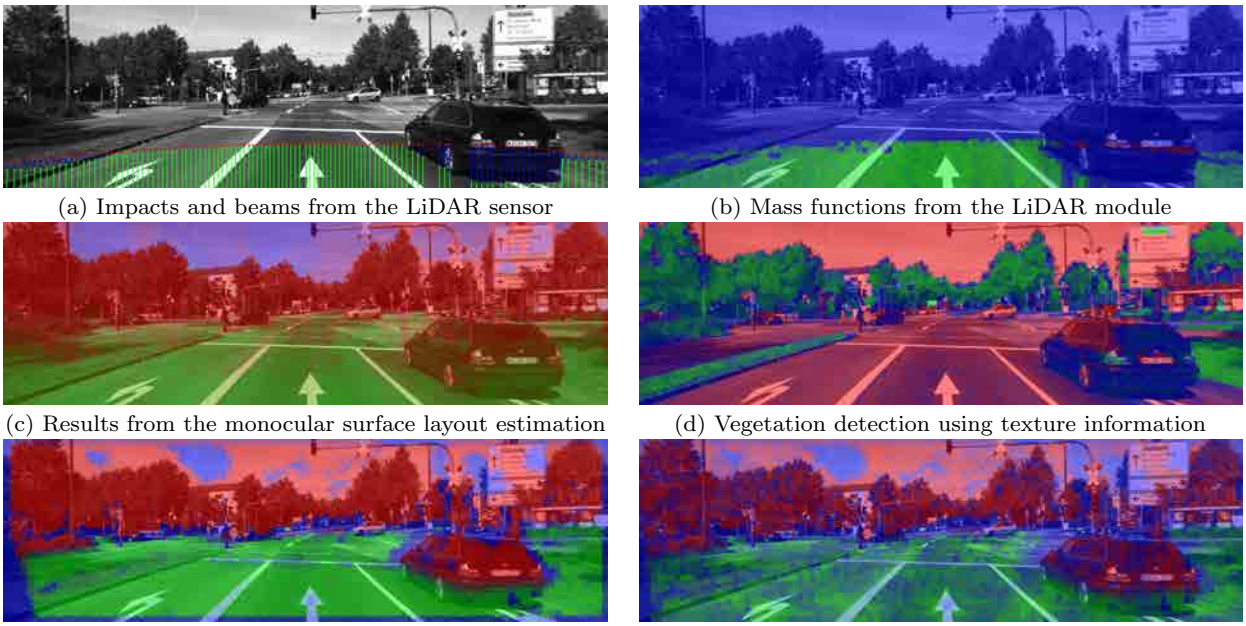
The data from the LiDAR sensor are illustrated in Fig. 8(a). The red dots represent the impacts returned by the LiDAR. The segments hit by these impacts are modeled and classified in the same way as in the stereo case. The green dots correspond to the projections of the impacts on the ground plane estimated by the stereo module. The green lines represent the laser rays from the green dots to the LiDAR's origin. The segments crossed by at least one green line are assimilated to the "ground" class. A categorical mass function $m_L^{\Omega_G}(\{\text{ground}\}) = 1$ is assigned to these segments. In the probabilistic case, the probability $P_L^{\Omega_G}(\text{ground}) = 1$ is used. Furthermore, a discounting factor or reliability measure $P_R(r = 0) = \alpha = k/n$ is considered for the segments hit or crossed by at least one laser beam. The quantity n is defined as the maximum number of beams that could have hit or crossed the segment. Finally, the segments between the red and green dots, represented by the blue lines, are ambiguous and are modeled by a vacuous mass function or uniform probability distribution. It is also the case for all the segments that are neither hit nor crossed by some laser beams. The result obtained from the LiDAR module is displayed in Fig. 8(b).

4.4 Surface layout from monocular images

Geometric structures in the scene can also be estimated directly from a single image. We used the method proposed by Hoiem et al. [17], whose code and pre-trained models are publicly available[1]. They used a set of multiple features including location, color, texture and perspective cues such as line intersections or vanishing points. Boosted decision trees were used to learn a multi-class classifier. The logistic regression version of Adaboost was used in order to get well-calibrated probabilities as output.

Hoiem et al. [17] considered three classes: "support", "vertical" and "sky". In our case, the "support" class

[1] `http://www.cs.uiuc.edu/~dhoiem`



(a) Impacts and beams from the LiDAR sensor
(b) Mass functions from the LiDAR module
(c) Results from the monocular surface layout estimation
(d) Vegetation detection using texture information
(e) Stereo-based ground detection at time $t-1$
(f) Temporal propagation of the stereo module

Fig. 8 Classification from different modules. For (b), (e) and (f), the RGB colors represent the mass assigned to {~~ground~~}, {ground} and Ω_{G} respectively. For (c), the RGB colors represent the probability of the vertical, ground and sky classes, respectively. For (d), the RGB colors represent the mass assigned to {~~vegetation~~}, {vegetation} and Ω_{V}, respectively.

corresponds to the ground. Hoiem et al. further decomposed the "vertical" class into five subclasses: "left", "center", "right", "porous" and "solid". These five subclasses are, however, of limited meaning in our case, so they were not considered. Additionally, the over-segmentation algorithm from Felzenszwald and Huttenlocher [13], which was originally used, was replaced by the SLIC over-segmentation [1].

As the output is a probability distribution, it can be directly used for probabilistic fusion. It can also be considered as a Bayesian mass function in an evidential context. However, the use of a Bayesian mass function will constrain the results of the combination to be Bayesian. To avoid such situations, the probabilistic output is considered as the pignistic probability generated by a non-Bayesian mass function.

The pignistic transformation (17) returns a probability from a mass function by distributing the mass of any subset to its singletons uniformly. This transformation is not invertible: different mass functions can lead to the same pignistic probability. A pseudo-inverse can however be defined by using the least commitment principle, which states that the least informative belief function, from a given ordering, should be selected from the set of possible candidates. Dubois et al. [10] showed that the least informative belief function with respect to the q-ordering is unique and consonant, *i.e.*, its focal elements are nested. It can be constructed as follows:

- The probability measure is first transformed into a possibility measure:
$$poss\,(\omega_i) = \sum_{\omega_j \in \Omega} \min(P(\omega_i), P(\omega_j)),\ \forall \omega_i \in \Omega. \quad (31)$$
- The possibilities $\pi_j = poss(\omega_{i_j})$ are sorted so that:
$$\pi_1 \geq \pi_2 \geq \ldots \geq \pi_C. \quad (32)$$
- The associated consonant mass function is then defined as:
$$m(A) = \begin{cases} \pi_j - \pi_{j+1} & \text{if } A = \{\omega_{i_1}, \ldots, \omega_{i_j}\}, \\ \pi_C & \text{if } A = \Omega, \\ 0 & \text{otherwise.} \end{cases} \quad (33)$$

The ignorance $m(\Omega)$ resulting from this transformation is equal to the minimum possibility of the singletons. In particular, a uniform probability distribution leads to the vacuous mass function. Finally, the accuracy of the algorithm of Hoiem *et al.* [17] on our training data is used as a discounting factor.

4.5 Texture-based classification

The textural appearance of a segment is an important cue about its class. We used the Walsh-Hadamard transform to encode the texture, as proposed by Wojek and Schiele [34]. For each segment, the Walsh-Hadamard coefficients were computed over 8×8 and 16×16 pixel patches centered at the centroid of the segment. The

three color channels were processed individually in the L*a*b* color space resulting in a feature vector of dimension 960.

This texture information was then used to build a vegetation detection module. A linear binary classifier was trained from a L1-regularized logistic regression as implemented in the Liblinear library [12]. This library was designed to efficiently learn linear classifiers from very large datasets. The probabilistic classifier output was handled as described in Sec. 4.4. No discounting or reliability estimation was used for this module as the classifier was directly trained on the KITTI dataset and can be assumed to be well-calibrated. Fig. 8(d) displays results obtained with this vegetation detector.

4.6 Temporal propagation

Given two consecutive images at times t and $t-1$, the optical flow (Fig. 4(d)) can be used to propagate the information. We used the OpenCV implementation of the TV-L1 formulation as proposed by Zach et al. [36]. To each segment S_t at time t was associated a previous segment S_{t-1} at time $t-1$, defined as the segment pointed by the mean flow of the pixels in S_t. The mass function or probability associated to S_{t-1} was simply propagated to S_t. A discounting factor corresponding to the ratio of pixels in S_t whose flow actually points to S_{t-1} was then used as reliability measure. This temporal propagation can be used with any frame of discernment. The propagation of the results from the stereo-based ground detector is illustrated in Fig. 8(e-f).

5 Experimental results

The KITTI dataset [15] was used to validate our approach, considering the stereo color camera and Velodyne 64-beam LiDAR. However, only one layer of the Velodyne LiDAR was used in order to simulate a single layer LiDAR, commonly employed in mobile robotics. A total of 110 images were manually annotated, 70 for training and 40 for testing. These images were selected to depict a high variety of scenes. The ground truth annotations are provided online[2]. Details about the annotated frames are given in Table 1.

The training data were used to learn the probabilities and mass functions for the stereo, LiDAR and texture-based modules. They were also used to get the discounting factor of the monocular surface layout estimation. No training was needed for the pixel-based and temporal propagation modules. Each classification context introduced in Section 4 was considered as an individual module. Table 2 summarizes the frames of discernment of these different modules.

[2] `https://www.hds.utc.fr/~xuphilip/dokuwiki/en/data`

Table 1 Annotated frames from the KITTI dataset. The highlighted rows correspond to the data used for testing.

Category	Date	Seq.	Annotated frames
Campus	2011-09-28	016	13, 144
Campus	2011-09-28	021	153
Campus	2011-09-28	038	29
City	2011-09-26	001	59, 107
City	2011-09-26	002	16, 56
City	2011-09-26	005	16, 56, 104, 153
City	2011-09-26	009	13, 58, 158, 265, 360, 370, 380, 390, 400, 412, 417
City	2011-09-26	011	10, 30, 50, 75, 100, 126, 150, 175, 190, 200
City	2011-09-26	013	14, 100, 143
City	2011-09-26	014	157, 200, 209
City	2011-09-26	017	32
City	2011-09-26	048	0, 21
City	2011-09-26	051	67, 86
City	2011-09-26	056	80, 158, 201
City	2011-09-26	057	41, 112
City	2011-09-26	059	26
City	2011-09-26	060	7
City	2011-09-26	084	248
City	2011-09-26	091	12, 85
City	2011-09-26	093	30, 303, 404
City	2011-09-26	095	126
City	2011-09-26	096	0, 92, 362
City	2011-09-26	104	16, 43, 239, 285
City	2011-09-26	106	1
City	2011-09-26	113	0
City	2011-09-26	117	103, 230, 384, 461, 594
City	2011-09-28	002	40, 50, 60, 70, 93, 317
City	2011-09-29	026	0
City	2011-09-29	071	11, 103, 318, 665, 906, 940
Residential	2011-09-26	019	329, 371
Residential	2011-09-26	020	0
Residential	2011-09-30	018	80, 192, 277, 329, 357, 496, 600, 650, 700, 750, 800, 850
Road	2011-09-26	015	167, 184, 220, 280
Road	2011-09-26	027	56
Road	2011-09-26	028	184, 231

Table 2 Frames of discernment of the different modules.

	Module	Frame of discernment
#1	Pixel	$\Omega_{\text{S}} = \{\text{sky}, \sout{\text{sky}}\}$
#2	Pixel	$\Omega_{\text{G}} = \{\text{ground}, \sout{\text{ground}}\}$
#3	Stereo	$\Omega_{\text{G}} = \{\text{ground}, \sout{\text{ground}}\}$
#4	LiDAR	$\Omega_{\text{G}} = \{\text{ground}, \sout{\text{ground}}\}$
#5	Surface	$\Lambda = \{\text{ground}, \text{vertical}, \text{sky}\}$
#6	Texture	$\Omega_{\text{V}} = \{\text{vegetation}, \sout{\text{vegetation}}\}$
#7	Optical flow	multiple

5.1 Ground detection

A first task was to evaluate ground detection. Modules 2, 3, 4 and 7 were first considered. Table 3 shows the results of the ground detection task. Some detection examples are shown in Fig. 10. By considering the stereo module alone, about 10% of the segments were ignored due to lack of disparity estimation. The blue regions in the images in Fig. 10(b) are the segments with high

uncertainty. Typically, the disparities could not be estimated in some textureless regions such as the sky or on white building facades.

After adding the LiDAR module, the recall rate of the ground class was increased by more than 5%. For example, we can see in Fig. 10(c-ii) that the bottom right part of the ground was detected by the LiDAR module but not by the stereo one. The LiDAR module also slightly increased the recall rate of the non-ground class ($\approx +0.2\%$). In Fig. 10(c-iii), we can see in the left part some laser impacts corresponding to some non-ground segments that were not detected by the stereo module. Finally, we can also observe a slight increase of the misclassification rate for the non-ground class ($\approx +0.2\%$). In the KITTI platform setup, the Velodyne LiDAR was installed on top of the car. This may explain that, in some particular cases, some small objects below the laser beam may be misdetected. In Fig. 10(c-i), we can see that the pole in the foreground was missed by the LiDAR sensor and thus classified as ground. Such minor issues may be dealt with by considering additional LiDAR layers.

The third considered module was the pixel-based one, for which all the segments above the horizon line upper bound were assigned to the non-ground class. As this module did not provide any information about the ground class, the results for this class remained unchanged. However, an increase of more than 5% was observed for the recall rate of the non-ground class. In particular, additional information was provided by this module in the parts of the sky and the buildings that were not classified by the stereo or LiDAR module (see Fig 10(d)).

Finally, the temporal propagation increased the recall rate or both the ground and non-ground classes by about 2%. Less than 2% of the segments were left without decision. In this ground detection module case, all the modules were correctly defined in their initial frames of discernment. The use of upper and lower bounds for the distance to the ground plane in the evidential method yielded slightly better results than the probabilistic model. But overall, the results from the probabilistic and evidential approaches were very similar.

5.2 Addition of the sky class

The sky class was added to the system with the monocular surface layout estimation module. The pixel-based module applied to the sky class was also included. The classification results are detailed in Table 4 and some examples are shown in Fig. 11.

As explained in Sec. 3, the outputs of the ground detection modules had to be transformed onto the new frame of discernment. In the probabilistic case, several effects could be noted. First, all the probabilities assigned to the non-ground class were divided by two and distributed to the vertical and sky classes. This resulted in over-confidence about the ground class. We can see from Table 4 that, after combining the monocular module with the ground detection ones, the recall rate of the ground class was increased by more than 10%. However, this came at the expense of a higher error rate ($\approx +5\%$) and a lower recall rate of the non-ground class ($\approx -5\%$). We can see in Fig. 11(c) that many non-ground regions were misclassified as ground.

A large increase of recall for the sky class ($\approx +10\%$) was also observed. This resulted from the combination of the two pixel based modules. The probability distribution resulting from their combination (23) always assigned more confidence to the sky class when a segment was not under the horizon line lower bound. We can see in Fig. 11(c-ii) that a large part of the buildings was classified as sky.

The over-confidence in both the ground and sky classes led to a large decrease of the vertical class recall rate ($\approx -10\%$). This also led to a very low error rate for the vertical class. We can see that the percentage of ground and sky segments being misclassified as vertical structures became both very low. For the ground class, this can be justified by the combination with an additional ground detector. However, for the sky, the decrease of the error rate was actually artificial. The ground detection modules did not provide any information about the sky and the pixel-based module only corrected some misclassifications occurring at the lower part of the image. In the upper part of the image, the originally misclassified sky segments were corrected only because the probability of the vertical class was artificial decreased. Overall, the accuracy was still increased ($+1.7\%$) but the error distribution became completely different.

In the evidential case, the recall rates of all three classes were increased and their error rates were decreased. Moreover, the performance of the combined system remained coherent with respect to the performances of the individual modules. We can see that the percentage of sky segments being misclassified as vertical structures only decreased slightly (-2.2%). Overall, the accuracy was increased by about 4%.

5.3 Addition of the vegetation class

Finally, the vegetation detector was added. The results are shown in Table 5. The probabilistic combination led

Table 3 Classification results of the ground detection modules 2, 3, 4 and 7. The lines correspond to the decisions made by the system and the column to the actual classes. The figures represent the recall rates in percentage.

		Stereo		Stereo+LiDAR		Stereo+LiDAR +Pixel		Stereo+LiDAR +Pixel+Flow	
		ground	~~ground~~	ground	~~ground~~	ground	~~ground~~	ground	~~ground~~
Prob	ground	87.4	4.1	93.1	4.3	93.1	4.3	95.2	4.5
	~~ground~~	4.2	85.2	4.2	85.4	4.2	91.3	3.9	93.8
	ignore	8.4	10.7	2.7	10.3	2.7	4.4	0.9	1.7
Belief	ground	87.6	3.9	93.5	4.2	93.5	4.2	95.4	4.2
	~~ground~~	4.0	85.4	3.9	85.5	3.9	91.4	3.7	94.1
	ignore	8.4	10.7	2.6	10.3	2.6	4.4	0.9	1.7

Table 4 Classification results of the combination of the surface layout estimation module with the ground detection ones. The figures represent the recall and error rates in percentage. The numbers in brackets correspond to the overall accuracy.

	Surface layout (90.5%)				Probabilistic fusion (92.2%)				Evidential fusion (94.3%)			
	ground	vert.	sky	error	ground	vert.	sky	error	ground	vert.	sky	error
ground	85.0	4.0	0.0	4.5	98.5	10.8	0.0	9.9	94.4	3.7	0.0	3.8
vert.	15.0	95.0	12.8	22.6	1.5	86.8	2.6	4.5	5.6	95.3	10.6	14.5
sky	0.0	1.0	87.2	1.1	0.0	2.5	97.4	2.5	0.0	1.0	89.4	1.1
		96.6				91.0				96.9		
		~~ground~~				~~ground~~				~~ground~~		

Table 5 Results of the combination of all the modules. The figures represent the recall and error rates in percentage. The numbers in brackets correspond to the overall accuracy.

	Probabilistic fusion (79.0%)						Evidential fusion (81.4%)					
	grass	road	tree	obst.	sky	error	grass	road	tree	obst.	sky	error
ground	86.4	3.7	4.0	5.3	0.0	13.1	73.3	1.7	1.4	1.7	0.0	6.1
road	7.0	95.0	0.4	7.6	0.0	13.6	11.2	94.5	0.4	4.2	0.0	14.3
tree	6.2	0.5	80.2	27.0	0.0	29.6	12.7	0.6	75.3	21.3	0.0	31.5
obst.	0.4	0.8	14.6	47.2	0.4	25.6	2.8	3.2	22.8	70.7	10.6	35.8
sky	0.0	0.0	0.8	12.9	99.6	12.1	0.0	0.0	0.0	2.0	89.4	2.2
	97.8		85.2				94.4		95.3			
	ground		vertical				ground		vertical			

again to over-confidence in the sky class as the probabilities on the ground and vertical classes were both distributed to two finer classes. We obtained a 99.6% recall rate of the sky class but with a very large error rate of 12.1%. Again, the lower error rate of the obstacle class in the probabilistic case compared to the evidential one ($\approx -10\%$) is artificial and was induced by this over-confidence in the sky class. Moreover, the probability originally assigned to the vegetation class is distributed among two classes while the probability of the non-vegetation class is distributed among three classes. This explains the very low recall rate of the obstacle class. Again, the evidential fusion was more robust to refinements and led to better overall accuracy. In particular, as the vegetation module did not provide any information about the ground and vertical classes, the recall rate of these two classes remained unchanged in the evidential case. On the contrary, in the probabilistic case, the recall rates changed for both the ground and vertical classes. Some examples of classification are shown in Fig. 12.

5.4 Discussion

Figure 9 shows a flowchart of the complete system. Three pre-processing blocks were first used to provide the over-segmentation, the disparity map and the optical flow. They can process the data independently and in parallel. Real time implementations of these tasks are described in the literature [24,33], but they have not been applied in this work.

The over-segmentation is needed by all the detection modules. The complexity of the SLIC over-segmentation algorithm is $O(N)$, with N the number of segments [1]. By using the C++ code provided by Achanta *et al.*[3], the computation of 3000 superpixels on 1224×370 images takes about 3 seconds. Ren and Reid [24] reported a speedup of 10x $\sim$ 20x by implementing the SLIC algorithm using GPU. In there work, the over-segmentation of a 1280×960 image is done in 86 ms.

[3] `http://ivrg.epfl.ch/supplementary_material/RK_SLICSuperpixels/index.html`

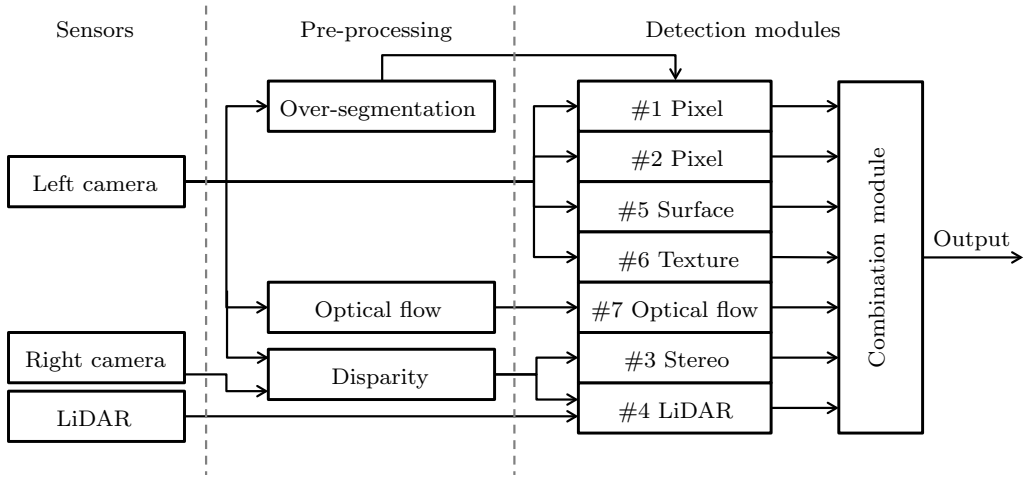


Fig. 9 Flowchart of the complete system.

Table 6 Computation time per image. The computation was done using some C/C++ and MATLAB® codes on a machine running at 2.20 GHz.

Pre-processing	
Over-segmentation	3,200 ms
Disparity	220 ms
Optical flow	4,100 ms
Detection and mass functions computation	
Pixel	20 ms
Stereo	120 ms
LiDAR	17 ms
Surface	35,930 ms
Texture	220 ms
Propagation	80 ms

All the detection modules can also process data independently and in parallel, using data from one or multiple sensors. For most of the modules, simple methods were used, resulting in low computation time. Table 6 shows the computation time per image for the data processing step and the mass functions computation. The computation of the surface layout was the slowest module and required about 30 seconds per image. It was done using the MATLAB® code provided by Hoiem et al. [17]. As only low level features computation are costly, implementations on dedicated hardware can be considered to reach real time performance. Finally, the cost of the mass function combination is linear in both the number of modules and the number of considered singletons, as only the plausibilities of singletons are computed.

One drawback of our approach is that we may not reach the best performance attainable given all the information at hand. All the modules are considered independently and only use a part of the available information. A global learning, as well as an optimized combination rule [23], could yield better results. It is, however, the price to pay if we want the system to be flexible enough to allow for the inclusion of new modules and new classes without having to retrain the whole system every time. Moreover, the complexity of a global approach would grow with the increasing number of modules and classes. The modular structure of the system also makes it more robust to the failure of a sensor, such as the LiDAR.

6 Conclusion and perspectives

We have introduced an original framework for multimodal information fusion based on over-segmented images and Dempster-Shafer theory. This framework is very flexible as it makes it possible to include new classes, new sensors or new object detection algorithms without having to retrain the whole system. The information combination approach lends itself to parallel implementation and can cope with sensor failures. Future work will consider additional classes such as pedestrian. We will also adapt methods like sliding windows-based algorithms to our framework based on segments. New sources of information such as GPS or digital maps will also be considered to detect moving objects. Finally, syntactic-based approaches such the one proposed in [5] will be further studied in order to merge segments belonging to the same object and allow for a deeper understanding of the scene.

Acknowledgements This work was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). It was supported and funded by the Cai Yuanpei project 26193PE from the Chinese Ministry of Education, the French Ministry of Foreign and European Affairs and the French Ministry of Higher Education

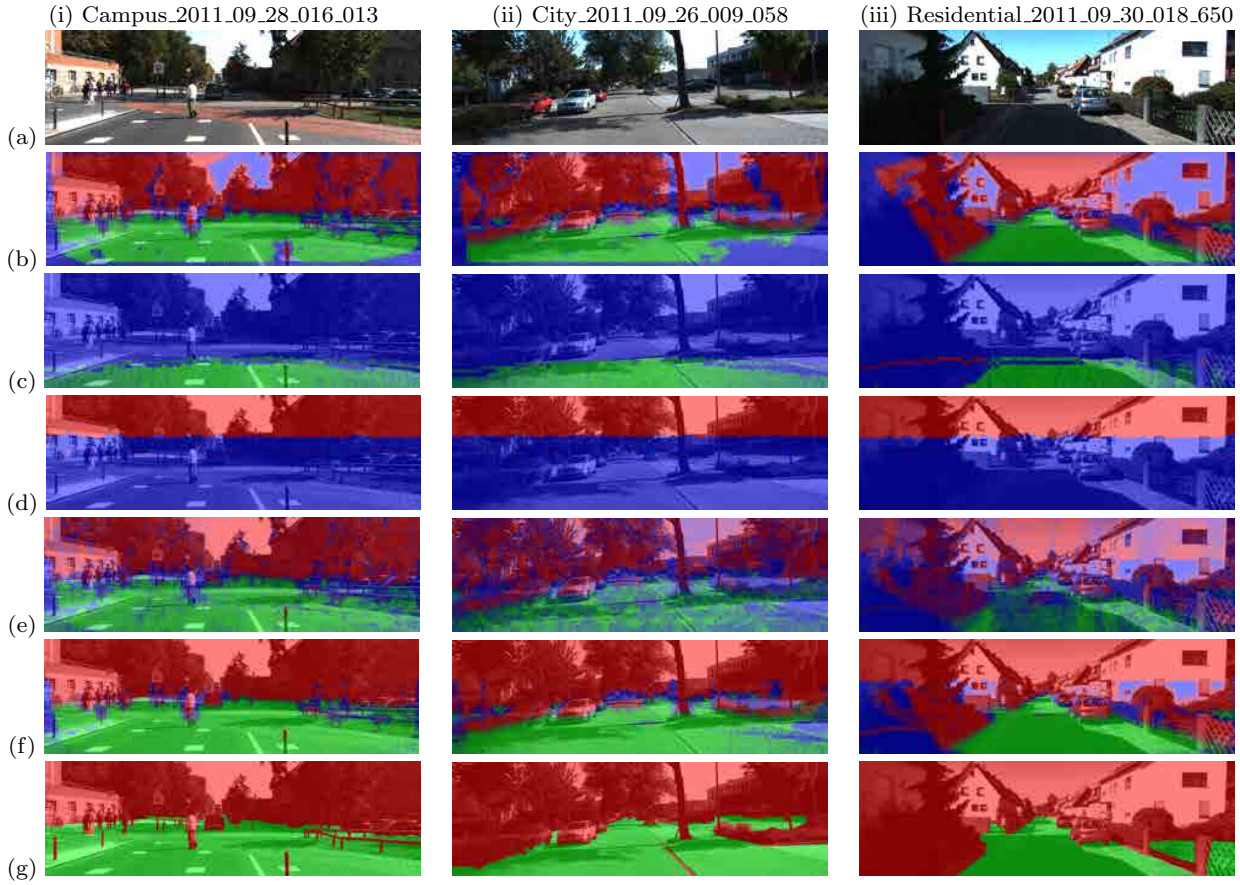


Fig. 10 Classification from the ground detection modules. The RGB colors represent the mass assigned to {~~ground~~}, {ground} and Ω_G, respectively. (a) Raw images. (b) Stereo-based module. (c) LiDAR module. (d) Pixel-based module. (e) Temporal propagation of the combined mass function from the previous frame. (f) Combined mass functions. (g) Ground truth images.

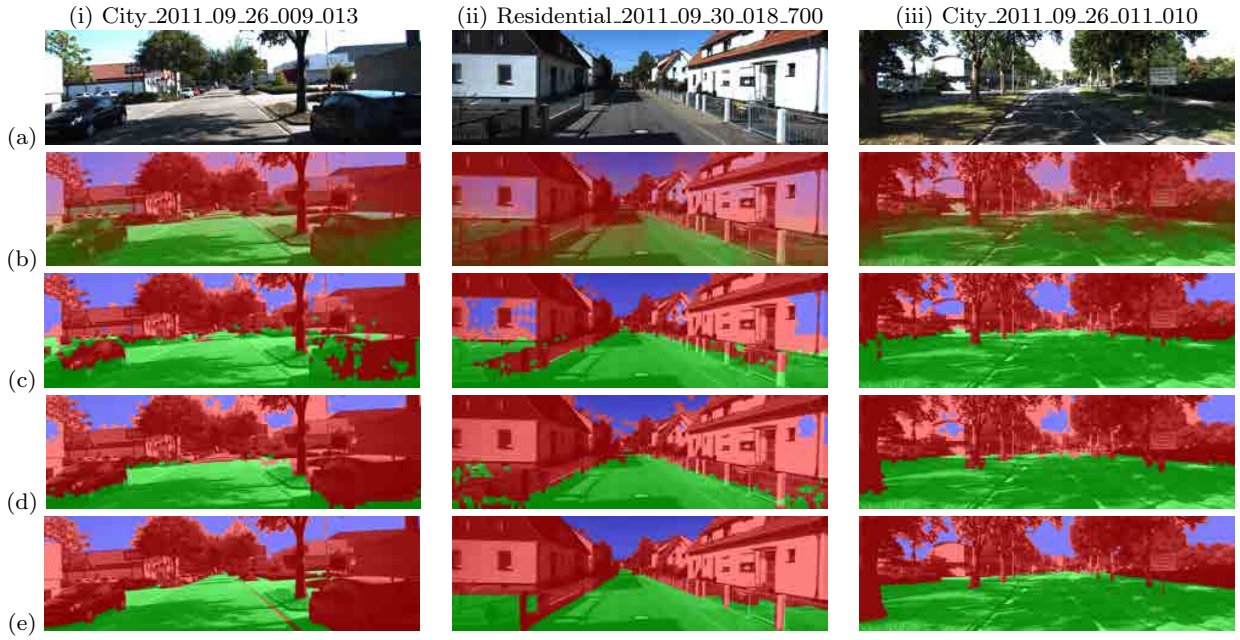


Fig. 11 Classification from different modules. The color code for (c), (d) and (e) is defined as follows: ground = green, vertical = red, sky = blue. (a) Raw image. (b) Output of the monocular surface layout estimation module, the RGB colors represent the probabilities assigned to the ground, vertical and sky classes, respectively. (c) Decisions resulting from the probabilistic combination with the ground detection modules. (d) Decision results from the evidential combination. (e) Ground truth images.

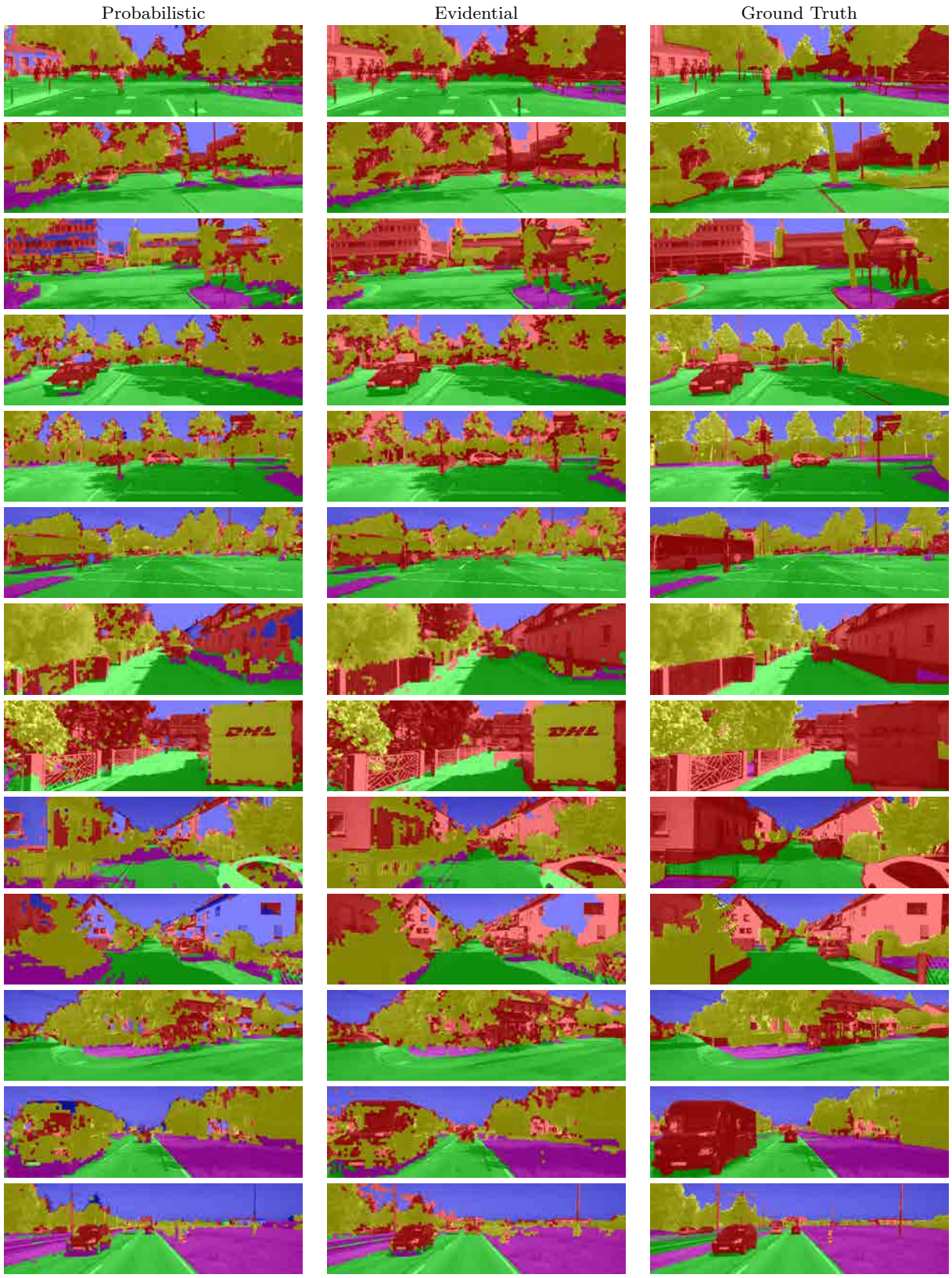


Fig. 12 Classification results considering all the modules. The color code is defined as follows: grass = magenta, road = green, tree = yellow, obstacle = red, sky = blue.

and Research. It was also supported by the ANR-NSFC Sino-French PRETIV project ANR-11-IS03-0001.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. on Pattern Analysis and Machine Intelligence **34**(11), 227–2282 (2012)
2. Badino, H., Franke, U., Mester, R.: Free space computation using stochastic occupancy grids and dynamic programming. In: Proc. International Conference on Computer Vision Workshop on Dynamical Vision. Rio de Janeiro, Brazil (2007)
3. Bansal, M., Sang-Hack, J., Bogdan, M., Jayana, E., Harpreet, S.S.: A real-time pedestrian detection system based on structure and appearance classification. In: Proc. IEEE International Conference on Robotics and Automation, pp. 903–909. Anchorage, Alaska (2010)
4. Barnett, J.A.: Calculating Dempster-Shafer plausibility. IEEE Trans. on Pattern Analysis and Machine Intelligence **13**(6), 599–602 (1991)
5. Bordes, J.B., Davoine, F., Xu, P., Denœux, T.: Evidential grammars for image interpretation - Application to multimodal traffic scene understanding. In: Z. Qin and V. N. Huyn (Eds), Integrated Uncertainty in Knowledge Modelling and Decision Making, pp. 65–78. Beijing, China (2013)
6. Cobb, B.R., Shenoy, P.P.: On the plausibility transformation method for translating belief function models to probability models. International Journal of Approximate Reasoning **41**(3), 314–330 (2006)
7. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence **24**(5), 603–619 (2002)
8. Denœux, T.: Analysis of evidence-theoretic decision rules for pattern classification. Pattern Recognition **30**(7), 1095–1107 (1997)
9. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection : an evaluation of the state of the art. IEEE Trans. on Pattern Analysis and Machine Intelligence **34**(4), 743–761 (2011)
10. Dubois, D., Prade, H., Smets, P.: A definition of subjective possibility. International Journal of Approximate Reasoning **48**(2), 352–364 (2008)
11. Ess, A., Müller, T., Grabner, H., Van Gool, L.: Segmentation based urbran traffic scene understanding. In: Proc. British Machine Vision Conference, pp. 84.1–84.11. London, UK (2009)
12. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9**, 1871–1874 (2008)
13. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59**(2), 167–181 (2004)
14. Fröhlich, B., Rodner, E., Kemmler, M., Denzler, J.: Large-scale gaussian process multi-class classification for semantic segmentation and facade recognition. Machine Vision and Applications **24**(5), 1043–1053 (2013)
15. Geiger, A., Lenz, P., Urtasun, R.: Vision meets robotics: The KITTI dataset. International Journal of Robotics Research **32**(11), 1231–1237 (2013)
16. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: Proc. Asian Conference on Computer Vision, pp. 25–38. Queenstown, New Zealand (2010)
17. Hoiem, D., Efros, A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision **75**(1), 151–172 (2007)
18. Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N.: Multisensor data fusion: A review of the state-of-the-art. Information Fusion **14**, 28–44 (2013)
19. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence **20**(3), 226–239 (1998)
20. Ladický, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., Torr, P.H.S.: Joint optimisation for object class segmentation and dense stereo reconstruction. International Journal of Approximate Reasoning **100**(2), 122–133 (2012)
21. Leibe, B., Cornelis, N., Cornelis, K., Van Gool, L.: Dynamic 3D scene analysis from a moving vehicle. In: Proc. IEEE Computer Vision and Pattern Recognition, pp. 1–8. Minneapolis, USA (2007)
22. Lin, H.T., Lin, C.J., Weng, R.C.: A note on Platts probabilistic outputs for support vector machines. Machine Learning **68**(3), 267–276 (2007)
23. Quost, B., Masson, M.H., Denœux, T.: Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. International Journal of Approximate Reasoning **52**(3), 353–374 (2011)
24. Ren, C.Y., Reid, I.: gSLIC: a real-time implementation of SLIC superpixel segmentation. Tech. rep., University of Oxford, Department of Engineering Science (2011)
25. Rodríguez, S.A., Frémont, V., Bonnifait, P., Cherfaoui, V.: Multi-modal object detection and localization for high integrity driving assistance. Machine Vision and Applications **14**, 1–16 (2011)
26. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton, New Jersey (1976)
27. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. International Journal of Approximate Reasoning **9**(1), 1–35 (1993)
28. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligence **66**, 191–243 (1994)
29. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics (Intelligent Robotics and Autonomous Agents). The MIT Press, Cambridge, Massachusetts (2005)
30. Walley, P.: Statistical reasoning with imprecise probabilities. Chapman and Hall, New York (1991)
31. Wang, C.C., Thorpe, C., Thrun, S., Hebert, M., Durrant-Whyte, H.: Simultaneous localization, mapping and moving object tracking. The International Journal of Robotics Research **26**(1), 889–916 (2007)
32. Wedel, A., Badino, H., Rabe, C., Loose, H., Franke, U., Cremers, D.: B-spline modeling of road surfaces with an application to free-space estimation. IEEE Trans. on Intelligent Transportation Systems **10**(4), 572–583 (2009)
33. Werlberger, M.: Convex approaches for high performance video processing. Ph.D. thesis, Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria (2012)
34. Wojek, C., Schiele, B.: A dynamic conditional random field model for joint labeling of object and scene classes. In: Proc. European Conference on Computer Vision, pp. 733–747 (2008)

35. Xu, Ph., Davoine, F., Bordes, J.B., Zhao, H., Denœux, T.: Information fusion on oversegmented images: An application for urban scene understanding. In: Proc. Int. Conf. on Machine Vision and Application, pp. 189–193. Kyoto, Japan (2013)
36. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. In: F. Hamprecht, C. Schnörr, B. Jähne (eds.) Pattern Recognition, *Lecture Notes in Computer Science*, vol. 4713, pp. 214–223. Springer Berlin Heidelberg (2007)

Appendix A: Decision making

In our case of study, several arguments can be stated in favor of the optimistic strategy. It is often more conclusive than the pessimistic strategy, it is coherent with frame refinement and computationally efficient. As shown by Barnett [4], given k plausibility functions, finding the singleton with maximum plausibility of the combined function only needs $O(k|\Omega|)$ operations while for the belief it is necessary to do $O(|\Omega|^k)$ operations. We indeed have the following property:

$$pl_{1,2}(\{\omega\}) = \frac{1}{1-\kappa} pl_1(\{\omega\}) pl_2(\{\omega\}) \tag{34a}$$

$$\propto pl_1(\{\omega\}) pl_2(\{\omega\}), \quad \forall \omega \in \Omega. \tag{34b}$$

To compute the pignistic probabilities, the combined mass functions need to be explicitly computed, which requires a number of operations exponential in $|\Omega|$.

To show the differences between different decision making strategies, let us consider the following mass function defined on $\Omega = \{\text{grass, road, } \sout{\text{ground}}\}$:

$$m^{\Omega}(\{\text{grass, road}\}) = 0.2, \tag{35a}$$

$$m^{\Omega}(\{\text{grass, } \sout{\text{ground}}\}) = 0.3, \tag{35b}$$

$$m^{\Omega}(\{\text{road, } \sout{\text{ground}}\}) = 0.5. \tag{35c}$$

Tab. 7 shows the beliefs, plausibilities and pignistic probabilities on the singletons. Here, the pessimistic strategy cannot lead to any decision: actually, in the worst case scenario, any decision could be wrong given the current mass function. Choosing {grass} instead of {$\sout{\text{ground}}$} would be wrong if the masses $m^{\Omega}(\{\text{grass, } \sout{\text{ground}}\})$ and $m^{\Omega}(\{\text{road, } \sout{\text{ground}}\})$ were actually entirely related to {$\sout{\text{ground}}$}. Inversely, the other decision would also be wrong if the same masses were now related respectively to {grass} and {road}. On the other hand, both pl^{Ω} and $BetP^{\Omega}$ would lead to {$\sout{\text{ground}}$}, which seems quite reasonable.

Now, if the singleton {$\sout{\text{ground}}$} is refined into {tree, obstacle, sky}, the mass function (35) will simply be rewritten as:

Table 7 bel^{Ω}, pl^{Ω} and $BetP^{\Omega}$ from mass function (35).

	{grass}	{road}	{$\sout{\text{ground}}$}
bel^{Ω}	0	0	0
pl^{Ω}	0.5	0.7	0.8
$BetP^{\Omega}$	0.25	0.35	0.4

Table 8 bel^{Θ}, pl^{Θ} and $BetP^{\Theta}$ from mass function (36).

	{grass}	{road}	{tree}	{obst.}	{sky}
bel^{Θ}	0	0	0	0	0
pl^{Θ}	0.5	0.7	0.8	0.8	0.8
$BetP^{\Theta}$	0.175	0.225	0.2	0.2	0.2

$$m^{\Theta}(\{\text{grass, road}\}) = 0.2, \tag{36a}$$

$$m^{\Theta}(\{\text{grass, tree, obstacle, sky}\}) = 0.3, \tag{36b}$$

$$m^{\Theta}(\{\text{road, tree, obstacle, sky}\}) = 0.5. \tag{36c}$$

Tab. 8 shows the measures induced by this new mass function. Following $BetP^{\Theta}$, the decision is changed and now leads to {road}. In contrast, the plausibility criterion does not discriminate between {tree}, {obstacle} and {sky}, which are still more plausible than {grass} and {road}. The optimistic strategy thus remains coherent with its previous decision.