Information Fusion on Oversegmented Images: An Application for Urban Scene Understanding

Philippe Xu[1,2] Franck Davoine[2] Jean-Baptiste Bordes[1] Huijing Zhao[2] Thierry Denœux[1]

[1]UMR CNRS 7253, Heudiasyc
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex, France

[2]LIAMA, CNRS
Key Lab of Machine Perception (MOE)
Peking University, Beijing, P.R. China

philippe.xu@hds.utc.fr

Abstract

The large number of tasks one may expect from a driver assistance system leads to consider many object classes in the neighborhood of the so-called intelligent vehicle. In order to get a correct understanding of the driving scene, one has to fuse all sources of information that can be made available. In this paper, an original fusion framework working on segments of over-segmented images and based on the theory of belief functions is presented. The problem is posed as an image labeling one. It will first be applied to ground detection using three kinds of sensors. We will show how the fusion framework is flexible enough to include new sensors as well as new classes of objects, which will be shown by adding sky and vegetation classes afterward. The work was validated on real and publicly available urban driving scene data.

1 Introduction

Scene understanding is a very important task for advanced driver assistance systems and, more generally, for modern robotics. Within it, subtasks such as road detection, pedestrian detection or traffic signs 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 the existing works, one has to find a way to properly fuse all relevant sources of information. Some typical questions then arise. Is it possible to take advantage of a vegetation detector to help a pedestrian detector, or the other way round? How can the output of a LIDAR (laser-based) sensor, which only perceives a set of discrete impacts from obstacles, be fused with a sky detector module from a camera? How to include a new sensor or a new class of object in the system?

More generally, two critical goals have to be achieved. The first one is to fuse modules that deal with different classes of objects and be flexible enough to include new ones. The second goal is to represent, in a common space, the outputs of sensors that perceive the world differently.

Related Work In the field of intelligent vehicles, cameras and LIDAR are the most common sensors. LIDAR sensors have often been considered to detect static structures, but also moving objects [1], whereas cameras have been used for a much wider range of applications. Pedestrian detection is one of the most studied cases [2], but more general traffic scene understanding have also been considered [3]. Depth information from stereo camera systems has proven to be useful to detect obstacles and navigable space [4]. Regarding the fusion aspect, many methods based on multiple sensor systems use a region of interest approach, whereby a first sensor, for example a LIDAR [5], is used to select a set of interesting regions that are further analyzed. Another typical kind of fusion is the combination of several features [2], which can include depth [3]. Such fusion approaches are specialized to achieve a single task and are often implemented sequentially. They only partially achieve our goals. In contrast, the method presented in this paper makes it possible to directly fuse the outputs of different modules regardless of their order or their specific task.

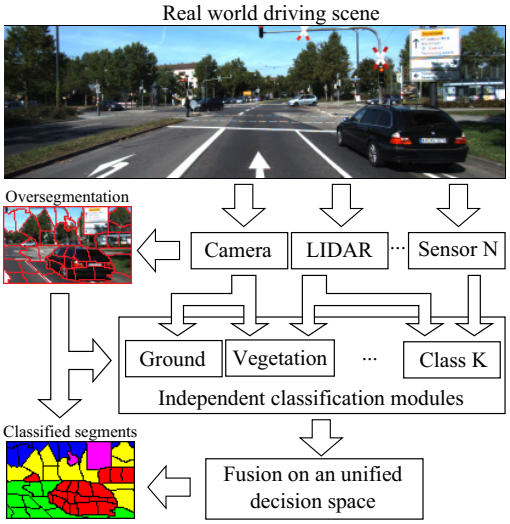


Figure 1. Overview of the fusion framework. N sensors, including a camera, observe the scene and provide data to K independent modules. The classification outputs are then fused in an unified decision space built from an oversegmented image.

Contributions To achieve the first goal, the belief functions or Dempster-Shafer theory [6] is used. We will show how each module can reason independently in its own decision space, before being combined in a common space. To handle the second goal (i.e. common space), we will formulate the problem as an image segment labeling one. Given an oversegmented image,

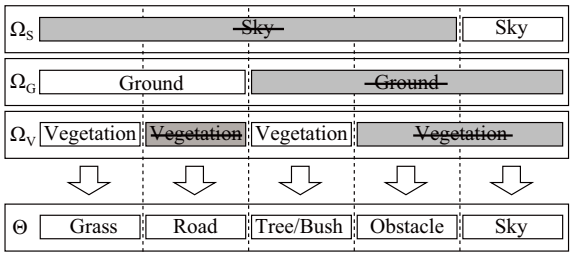


Figure 2. Illustration of a multiclass fusion. Three different class decompositions are combined to give a finer one. The "obstacle" class actually refers to anything that is not the sky, nor the ground, nor vegetation.

each module, regardless of how it perceives the environment, will have to classify each image segment.

Overview The system considered here consists of several sensors observing an urban scene, including a camera which produces an oversegmented image as pictured on Fig.(1). Each sensor may provide data to one or more modules which will run totally or partially in parallel to classify each image segment by providing a belief function [6], which will be crucial to perform the final fusion step. 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 method to process any number of classes will be then illustrated by adding a vegetation and a sky detection module. The experimental validation of this approach will be performed on the KITTI Vision Benchmark Suite [7].

2 Theory of belief functions

To show the importance of the theory of belief functions in practical situations, let us take the example of the fusion of a ground detector with a sky and a vegetation detector. As illustrated on Fig. 2, one can create new classes by intersecting those three initial classes. The ground class is, for example, divided into grass and road by intersecting it with vegetation.

As a ground detector cannot discriminate the grass from the road, the belief one has about the ground class should be put on the union of grass and road. The natural way of uniformly distributing the belief onto the two new refined classes is fundamentally incorrect, as it creates artificial knowledge about the grass and road classes. It is thus fundamental to be able to reason on subsets of classes. The theory of belief functions offers a well-founded and elegant framework to do so. It is also very well suited for information fusion.

Let Ω be a set of mutually exclusive classes called the *frame of discernment*, which corresponds to the set of all classes. A *basic belief assignment* (BBA), or *mass function*, is a function $m : 2^\Omega \to [0, 1]$ verifying:

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

Given a subset A of Ω, $m(A)$ represents the belief committed exactly to A and to none of its subsets. If $m(A) > 0$, A is said to be a *focal element* of m. In the case of total ignorance one simply has $m(\Omega) = 1$, which is called the *vacuous* mass. It should be noted that usual *Bayesian* probability distributions are just a particular type of belief function, which only has singletons as focal elements.

From a frame of discernment Ω, one can define a refinement Θ by splitting some or all its elements into new classes. The mass initially assigned to each subset of Ω is just transferred to the union of its elements in the refined frame. On the example given in Fig. 2, the mass committed on the ground class of Ω is simply transfered to the union of grass and road of Θ. Additionally, given two BBAs m_1 and m_2 built from two independent sources, they can be fused to give a new mass function $m_{1,2} = m_1 \oplus m_2$ by using the Dempster's rule of combination:

$$\begin{aligned} m_{1,2}(\emptyset) &= 0 \ , \\ m_{1,2}(A) &= \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \ , \end{aligned} \tag{2}$$

where $\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ measures the conflict between the two BBAs. This combination rule is commutative and associative, which means the order in which the fusion is done has no influence. To fuse two mass functions defined on two different frames of discernment, one just has to find a common refinement and use the Dempster's rule of combination.

When a reliability measure is available, it can be useful to weaken the belief. This can be done by using a discounting factor $\alpha \in [0, 1]$ so that:

$$\begin{aligned} {}^\alpha m(A) &= (1-\alpha) m(A), \quad \forall A \subsetneq \Omega, \\ {}^\alpha m(\Omega) &= (1-\alpha) m(\Omega) + \alpha. \end{aligned} \tag{3}$$

It means that the mass assigned to each focal element is decreased by a factor $1 - \alpha$ and the remaining is put on the ignorance.

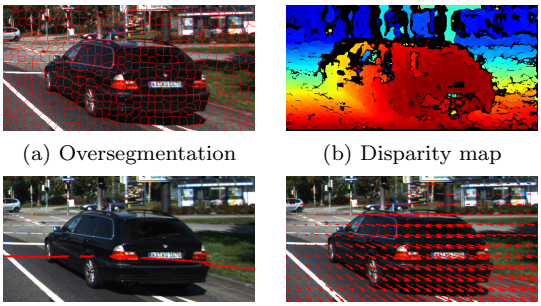
Finally, for classification, a BBA can be first transformed into a plausibility measure:

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \tag{4}$$

then the singleton with maximum plausibility is chosen.

3 Image segment labeling formulation

As explained in the introduction, our approach is to fuse information from different sensors, which may perceive the environment in many different ways. In the context of a driver assistance system, where the goal is to warn drivers about potential dangers, it seems relevant to use a labeled image which 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 like the road. We chose an intermediate way by oversegmenting the image. The TurboPixels algorithms proposed by Levinshtein et al. [8], which provides a grid-like segmentation, has been selected for that purpose.



(a) Oversegmentation (b) Disparity map

(c) Laser points (d) Optical flow

Figure 3. Input data of the multi-sensors system.

The common task of all the modules, whatever the data representation they use (image, 3D points cloud or map), is now to label each individual image segment. As the belief function theory makes it possible to represent all forms of partial information up to total ignorance, it does not matter if some segments are not visible from certain sensors.

4 Application to scene understanding

We applied our framework to a multi-sensor system including a stereo camera and a LIDAR, which are supposed to be calibrated. Several modules independently process the outputs of those sensors to classify each segment of the left image. The 3D information from the stereo images and the LIDAR are used to detect the ground. A monocular based approach further extends it by including a sky and a vegetation class. Finally, a temporal propagation block is used to link two consecutive images. The inputs of those different modules are shown on Fig. 3.

4.1 Belief function from models

A classification task can be seen as finding a correlation between a learned model M of a class C and an observation X of an object S (which is an image segment in our case). From an observation-to-model measure $d(X, M)$, one has to infer the class of S.

To build a BBA m over the frame of discernment $\Omega = \{C, \overline{C}\}$, where $\overline{C}$ includes everything but C, a general approach is to set two mass functions. One, m^-, that will assign belief on the class C if $d(X, M)$ is small and another one, m^+, that will, on the contrary assign belief to $\overline{C}$ if $d(X, M)$ is large.

It is important to note that, in some cases, one of those two BBAs may not be relevant. Sometimes one can only infer that S belongs to $\overline{C}$ when $d(X, M)$ is large while nothing can be said when $d(X, M)$ is small. In such situations, only one of the aforementioned mass functions should be used, otherwise they can be combined by Dempster's rule (2).

General forms for m^- and m^+ based on the one suggested in [9] are:

$$\begin{aligned} m^-(\{C\}) &= e^{-\gamma^- \left(\frac{d}{d^- - d}\right)^\beta} \text{ if } d < d^-,\ 0 \text{ otherwise,} \\ m^-(\Omega) &= 1 - m^-(\{C\}), \end{aligned} \tag{5}$$

and

$$\begin{aligned} m^+(\{\overline{C}\}) &= e^{-\gamma^+ \left(\frac{d^+}{d - d^+}\right)^\beta} \text{ if } d > d^+,\ 0 \text{ otherwise,} \\ m^+(\Omega) &= 1 - m^+(\{\overline{C}\}). \end{aligned} \tag{6}$$

The thresholds d^- and d^+ set the values under and above which, some masses can be assigned to $\{C\}$ and $\{\overline{C}\}$. The parameters $\beta \in \{1, 2, \ldots\}$, which could be arbitrarily fixed to a small value (1 or 2), and $\gamma > 0$ reflect the impact of the distance measure on the mass function. The combined mass function $m = m^- \oplus m^+$ can be further discounted by a factor α if needed.

Given a set of training data $\{(X_i, c_i)\}_{1 \leq i \leq n}$, where $c_i \in \{C, \overline{C}\}$ is the class of the observation X_i, the parameters can be chosen to minimize the following loss function:

$$L = \sum_{i=1}^{n} 1 - pl_i(\{c_i\}) + pl_i(\{\overline{c_i}\}), \tag{7}$$

where pl_i is the plausibility (4) associated to the observation X_i. The loss $L_i = 1 - pl_i(\{c_i\}) + pl_i(\{\overline{c_i}\})$ has the following properties:

$$\begin{aligned} m_i(\{c_i\}) \to 1 &\Rightarrow L_i \to 0, \\ m_i(\{\overline{c_i}\}) \to 1 &\Rightarrow L_i \to 2, \\ m_i(\Omega) \to 1 &\Rightarrow L_i \to 1. \end{aligned}$$

It assigns high cost for wrong mass assignment and an intermediate one for ignorance.

4.2 Stereo-based classification

A stereo camera is used to estimate the depth of each pixel. The generated 3D point cloud, or disparity map (Fig. 3(b)), is used to detect the ground plane by using a robust plane estimator, under the assumption of a planar ground. Each segment is then classified using its distance to the ground.

The frame of discernment will be $\Omega_G = \{G, \overline{G}\}$, where G corresponds to the "ground" class. An image segment S is represented by a set of n points $X = \{p_1, p_2, \ldots, p_k, p^*_{k+1}, \ldots, p^*_n\}$, where the points p^*_i are those for which no disparity estimate is available. The distance between X and a plane Π is defined as the mean distance of the valid points p_i to Π:

$$d(X, \Pi) = \frac{1}{k} \sum_{i=1}^{k} d(p_i, \Pi), \tag{8}$$

where $d(p_i, \Pi)$ is the Euclidean distance from p_i to Π.

Here, both m^- and m^+ can be used as, by definition, a segment lying on the ground plane belongs to the ground while a segment far away from it is not. The combined mass is then discounted by a factor $\alpha = (n-k)/n$ which is the ratio of points in X for which no disparity has been estimated. In the case where a segment has no disparity estimation at all ($\alpha = 1$), the discounting will naturally lead to the vacuous mass.

4.3 LIDAR-based classification

A LIDAR sensor provides a set of 3D points that are the impacts of laser beams (Fig. 3(c)). Similarly to the

stereo camera case, a segment S is perceived as a set of k 3D points $X = \{p_1, \ldots, p_k\}$, which may be empty. If a ground plane estimation is available, the same BBA as in the stereo case can be used for S.

Additionally, all the space between a laser impact and the LIDAR's origin is known to be obstacle free. This free space can thus be associated to the ground class. For all the segments S in which some laser beams have gone through, the following BBA is used:

$$m(\{G\}) = k/n, \qquad m(\Omega) = 1 - m(\{G\}), \qquad (9)$$

where n is the maximum number of beams that could have gone through S. This BBA can be seen as the categorical mass function $m(\{G\}) = 1$ discounted by the factor $\alpha = (n-k)/n$.

4.4 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 in [10], and a *bag-of-words* [11] approach to learn models. The texture features are first vector quantized into a set of N *textons*; a model is then simply a normalized histogram $H = (h_1, \ldots, h_N)$, or discrete probability distribution, over the whole set of *textons*. A segment S is also observed as a histogram $X = (x_1, \ldots, x_N)$, any histogram distance can be used as the observation-to-model measure, such as the χ^2 distance:

$$d(X, H) \quad = \quad \frac{1}{2} \sum_{i=1}^{N} \frac{(x_i - h_i)^2}{x_i + h_i}. \qquad (10)$$

When working with appearance cues, a small value of the distance between the observation and the model is often not enough to infer the class. A typical example is a bright white sky which may have the same local appearance as a white traffic sign. Thus only m^+ can be used.

4.5 Temporal propagation

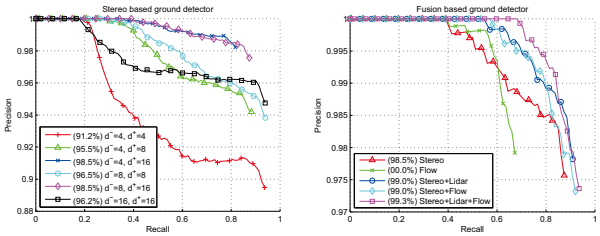
For two consecutive images at time t and $t+1$, the optical flow (Fig. 3(d)) can be used to propagate the information. To each segment S_t at time t is associated a segment S_{t+1} at t+1 defined as the one pointed by the mean flow of the pixels in S_t. The BBA m_t associated to S_t is simply propagated to S_{t+1}:

$$\forall A \subseteq \Omega, \qquad m_{t+1}(A) = m_t(A). \qquad (11)$$

It is then discounted by the ratio of pixels in S_t whose flow does not point to S_{t+1}. When several segments at time t point to the same segment at time $t+1$, the propagated BBAs are simply fused using Dempster's rule.

5 Experimental results

The KITTI dataset [7] was used to validate our approach. The stereo color camera system and Velodyne LIDAR were used as sensors. However, only one layer of the Velodyne LIDAR was used in order to simulate a single layer LIDAR, commonly employed in mobile robotics.



(a) (b)

Figure 4. (a) Stereo based ground detection performance for different values of d^- and d^+. (b) Performance of different combination of modules. The figures in brackets correspond to the precision at a recall rate of 80%.

Texture based classification

	Grass	Road	Tree	Obst	Sky	Recall
Grass						0
Road		66		33.8	0.2	50.2
Tree						0
Obst		14.3		84.1	1.6	49.3
Sky		0		18.4	81.6	80.5

(a)

Fusion based classification

	Grass	Road	Tree	Obst	Sky	Recall
Grass	81.1	3.8	15.1	0	0	40.6
Road	7.9	89.1	0.3	2.7	0	78.8
Tree	2.5	0	94.4	3.1	0	86.7
Obst	0.8	2.3	9.1	86.8	1	52.6
Sky	0	0	0	18.4	81.6	80.5

(b)

Figure 5. Confusion matrices, values are given in percentage.

The parameters for each module were chosen by sampling them on a grid. Fig. 4(a) shows the influence of the thresholds d^- and d^+ on the stereo based ground detection. The values $d^- = 8$, $d^+ = 16$ yield a good tradeoff between precision and recall. When considered alone, the parameters β, γ^- and γ^+ have no influence on the precision and recall. β is arbitrary set to 2 while γ^- and γ^+ are chosen with respect to the loss function (7). Fig. 4(b) shows the performances on ground detection using different combinations of modules. A fusion result is illustrated in Fig. 6(b-e). It can be clearly seen that the more sources of information are used, the better is the fusion result.

Our texture-based approach was used for road, vegetation and sky detection. Fig. 5(a) shows the corresponding confusion matrix and Fig. 6(f-h) illustrates the obtained results. When kept alone, it cannot discriminate the grass from the trees. It becomes possible by combining it with the ground detector from the stereo and LIDAR modules. Fig. 5(b) shows the confusion matrix of the complete system. Again, as expected, we can see that fusion improves the performance. As our system allows us to represent ignorance, it can happen that no decision can be made on some segments, which explains why the recall rate is different from the diagonal of the confusion matrix. Fig. 6(i) shows the final result, not colored segments are the ones where no decision is made.

It can happen that the oversegmentation is erroneous, the TurboPixels algorithm has the advantage to provide segments with similar sizes, however it is not robust enough to segment fine structures such as a traffic pole. We can see on Fig. 6(a) that it is not correctly segmented thus incorrectly classified. This kind of error could be handle by combining with other segmentation algorithms such as Meanshift.

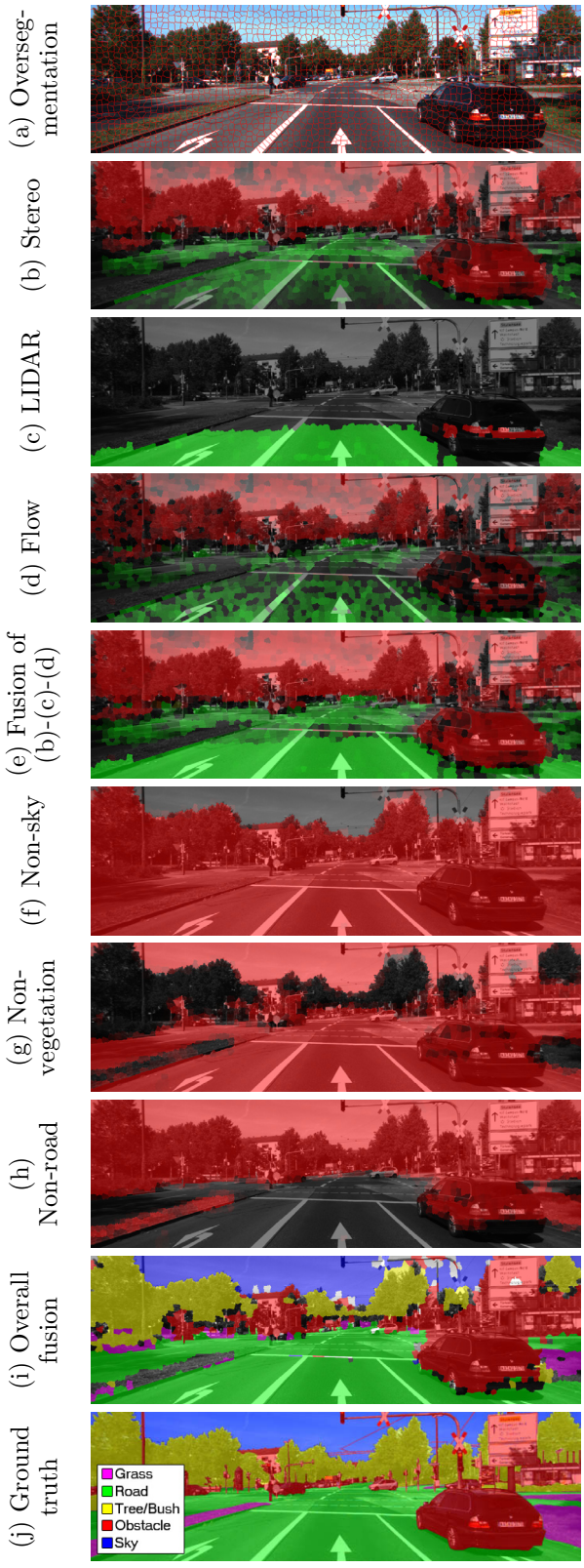


Figure 6. (a) Oversegmented raw image. (b-e) Ground/Non-Ground classification results. Green= $m(\{\text{ground}\})$, Red= $m(\{\text{non-ground}\})$, No-Color = $m(\Omega)$. (f-h) Texture based analysis for sky, vegetation and road detection, only mass on the complementary classes are assigned (represented in red). (i) Overall fusion, the color code is the same as the ground truth (j).

6 Conclusion

We have introduced an original framework for information fusion based on over-segmentation and Dempster-Shafer theory. This framework is flexible enough to allow the inclusion of new classes and new sensors or new object detection algorithms. Future work will consider additional classes such as pedestrians and adapting methods like sliding window-based algorithms to our framework based on segments. New sources of information such as GPS or maps will also be considered to detect moving objects. Finally, some global approaches will be studied in order to merge segments belonging to the same object and allow for a higher level understanding of the scene.

Acknowledgment

This work is supported and financed by the Cai Yuanpei program from the Chinese Ministry of Education (MOE), the French Ministry of Foreign and European Affairs (MAEE) and the French Ministry of Higher Education and Research (MESR). Is is also supported by the *Blanc International* ANR-NSFC Sino-French PRETIV project.

References

[1] S. Thrun, W. Burgard, D. Fox: "Probabilistic robotics". *The MIT Press*, Cambridge, Massachusetts, 2005.

[2] P. Dollár, C. Wojek, B. Schiele, P. Perona: "Pedestrian detection: an evaluation of the state of the art". *PAMI*, vol.34, no.4, pp.743-761, 2011.

[3] A. Ess, T. Müller, H. Grabner, L. Van Gool: "Segmentation based urban traffic scene understanding". *BMVC*, pp.1-11, UK, 2009.

[4] H. Badino, U. Franke, R. Mester: "Free space computation using stochastic occupancy grids and dynamic programming". *ICCV Workshop on Dynamical Vision*, Brazil, 2007.

[5] S.A. Rodríguez, V. Frémont, P. Bonnifait, V. Cherfaoui: "Multi-modal object detection and localization for high integrity driving assistance". *Machine Vision and Applications*, vol.14, pp.1-16, 2011.

[6] G. Shafer: "A mathematical theory of evidence". *Princeton University Press*, 1976.

[7] A. Geiger, P. Lenz, R. Urtasun: "Are we ready for autonomous driving? The KITTI vision benchmark suite". *CVPR*, pp.3354-3361, USA, 2012.

[8] A. Levinshtein, A. Stere, K.N. Kutulakos, D.J. Fleet, S.J. Dickinson, K. Siddiqi: "TurboPixels: fast superpixels using geometric flows". *PAMI*, vol.31, no.12, pp.2290-2297, 2009.

[9] T. Denœux: "A k-nearest neighbor classification rule based on Dempster-Shafer theory". *IEEE Transactions on Systems, Man and Cybernetics*, vol.25, no.5, pp.804-813, 1995.

[10] C. Wojek, B. Schiele: "A dynamic conditional random field model for joint labeling of object and scene classes". *ECCV*, pp. 733-747, France, 2008.

[11] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid: "Local features and kernels for classification of texture and object categories: a comprehensive study". *IJCV*, vol.73, no.2. pp.213-238, 2007.