Information fusion and evidential grammars for object class segmentation

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

***Abstract*— In this paper, an original method for traffic scene images understanding based on the theory of belief functions is presented. Our approach takes place in a multi-sensors context and decomposes a scene into objects through the following steps: at first, an over-segmentation of the image is performed and a set of detection modules provides for each segment a belief function defined on the set of the classes. Then, these belief functions are combined and the segments are clustered into objects using an evidential grammar framework. The tasks of image segmentation and object identification are then formulated as the research of the best parse graph of the image, which is its hierarchical decomposition from the scene, to objects and segments while taking into account the spatial layout. A consistency criterion is defined for any parse tree, and the search of the optimal interpretation of an image formulated as an optimization problem. We show that our framework is flexible enough to include new sensors as well as new classes of object. The work is validated on real and publicly available urban driving scene data.**

I. Introduction

Automatic understanding of the scene in front of a car is an essential task for advanced driver assistance or safety systems. Automatic understanding denotes generally a segmentation of the image scene into its constituting objects, augmented eventually with spatial or functional relationships. However, there are many classes of objects which can be found in traffic scenes, and for most of them, their level of variability is very high. Indeed, detecting even a single kind of object can be very challenging since the highly cluttered environment as well as the dynamically changing backgrounds, among others, contribute to the difficulty of such a task. Many approaches have been proposed recently to tackle individual problems such as road detection or pedestrian detection, and they can use different kinds of sensors.

A. Related Work

In the last decade, the accuracy of object detection methods has increased substantially thanks to the appearance of efficient visual descriptors in images such as SIFT as well as the success of computer vision challenges such as PASCAL. In the field of intelligent vehicles, they are mainly applied to pedestrian detection which is the most studied case [7], even if more classes have also been considered [8]. However, to reach better performances, more sensors are generally used:

[1]Jean-Baptiste Bordes, Philippe Xu and Thierry Denœux are with UMR CNRS 7253, Université de Technologie de Compiègne, BP 20529, 60205 Compiègne Cedex, France. bordes@nlpr.ia.ac.cn

[2]Philippe Xu, Franck Davoine and Huijing Zhao are with LIAMA, CNRS, Key Lab of Machine Perception (MOE), Peking University, Beijing, P.R. China. philippe.xu@hds.utc.fr

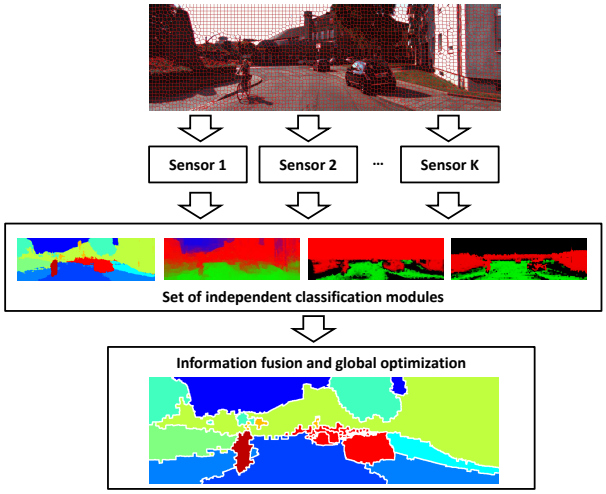

Fig. 1. Overview of the system. The scene is perceived by several sensors among which a camera provides an over-segmented image. A set of independent classification modules then gives some partial information which are finally combined through a global optimization scheme.

LIDAR sensors are widely used to detect static structures but also moving objects [14]. Depth information from stereo camera systems has also been used by Ess et al. [8] as well as Gavrila et al. [10] for pedestrian detection, it has also proven to be efficient to detect obstacles and navigable space [2]. Most of these methods are based on local visual clues, but some other approaches add to this local step a post-processing to take advantage of some consistency clues. Wojek et al. [15] perform joint object detection to take into account the spatial relationships between objects. Brehar [4] uses openCyc ontology to exploit inter-class relationships between classes in traffic scenes. Impressive results have also been obtained in [17] on a great variety of databases including traffic scenes using visual grammars, which is an adaptation of formal grammars for visual data. The objects and their components are first defined in the model and then, given a new image, a parse graph is computed, which is the decomposition of the scene into objects and parts of objects, down to the image primitives. Visual grammars have shown generalization capabilities and provide efficient way to face problems such as occlusion and scale.

B. Contribution

In this work, instead of presenting new efficient descriptors which are already numerous in the literature, we present a method to make the most of the existing works. For this purpose, the Dempster Shafer theory on belief functions is

used to properly fuse a set of relevant sources of information, that we call in this article "modules", even when each one of them is reasoning independently in its own decision space. This framework has several strong advantages. First of all, it provides a high level of flexibility to the system: new sensors and modules can be added easily, and their output will be fused in a common space. Reversely, the independence of the modules before fusion makes our system robust to sensor failure. Moreover, we will show that new classes can be added easily as well, since belief functions make it possible to work on sets of classes and not only on individual classes. Some expert information on the relative position of the objects in a scene is also taken into account as an other source of information by the use of an innovative framework called "evidential grammar".

C. Overview

The architecture of the system we consider, illustrated on Fig. 1, consists of a set of sensors including a camera. The image provided by the camera is over-segmented as a first step of image processing. We also consider a set of independent modules (road detection module, pedestrian detection module, etc.) receiving data from the sensors, the output of which is transformed into belief function before being fused at the segment level. Finally, the evidential grammars provide some kind of "global fusion" to this segment level information and strengthen weak detections as well as prune misdetections. We will show how this framework can be applied in practice by considering a monocular camera, stereo camera and a LIDAR. Our system is validated on the KITTI Vision Benchmark Suite [9].

II. Multi-modal and multi-class fusion

When working in a multi-modal context, several challenges arise. First of all, the sources of information may be of very different nature, they may come from several types of sensors or even from prior knowledge. Each source having its own specificity, complementary information can be fetched from them. For example, 3D information from a stereo camera or a LIDAR can be used to detect obstacles while texture and color, from a monocular camera, can be used to detect vegetation or the sky. The second challenge is now to properly combine information about different classes of objects.

We follow the framework proposed in [16] which can deal with those two issues. The information from all the sources are projected onto the image space and formulated as an image labeling problem. Meaning that each pixel of the image has to be classified. A first over-segmentation is however done so that the classification do not have to be done at the pixel level which is often too local. The combination over different sets of classes is handled using the theory of belief functions.

A. Dempster Shafer's theory of belief functions

1) Reasoning on sets with belief functions: The belief functions theory is an extension of classical probability which is especially well adapted for reasoning on sets. Given a set of classes $\Omega = \{\omega_1, \ldots, \omega_K\}$, a *mass function*, or *basic belief assignment* (BBA), is a function $m : 2^\Omega \to [0,1]$ verifying:

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

Contrary to a probability distribution which assigns a probability to every class, a mass function can assign a mass on any set of classes. Let us notice that a mass function whose non-zero values are only on singletons is equivalent to a Bayesian probability.

The plausibility is another measure often used to manipulate mass functions, it is defined as:

$$pl(A) \quad = \quad \sum_{B \cap A \neq \emptyset} m(B), \quad \forall A \subseteq \Omega. \tag{2}$$

When a decision has to be made, the singleton with maximum plausibility is usually a good choice.

Given two mass functions m_1 and m_2, they can be combined by using the Dempster's rule of combination to give a new mass $m_{1,2} = m_1 \oplus m_2$ defined as:

$$\begin{aligned} m_{1,2}(\emptyset) &= 0, \\ m_{1,2}(A) &= \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \end{aligned} \tag{3}$$

where $\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ measures the amount of conflict between the two mass functions.

2) Reasoning in the product space: In the method described in this paper, it will be necessary to introduce a set of evidential variables, and thus mass functions have to be manipulated on product spaces. Some well known operations that are used for Bayesian functions have to be introduced for mass functions. In all this section, two evidential variables X and Y are defined respectively on Ω_X and Ω_Y.

a) Marginalization: In this problem, the joint mass function m_{XY} is assumed to be known, the operation of marginalization can be used to get m_X:

$$m_{XY\downarrow X}(B) \quad = \quad \sum_{A \subseteq \Omega_{XY} | A \downarrow \Omega_X = B} m_{XY}(A), \quad \forall B \subseteq \Omega_X. \tag{4}$$

b) Vacuous extension: In this problem, the belief mass m_X is assumed to be known and we wish to extend it to the product space. The belief function theory suggests to choose the least informative mass function which provides m_X after it marginalization:

$$m_{X\uparrow XY}(A) \quad = \quad \begin{cases} m_X(B) & \text{if } A = B \times \Omega_Y, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

c) Conditioning: In this problem, the joint mass function m_{XY} is assumed to be known and X is supposed to belong to B, we denote: $m_X^B(B) = 1$. The conditioning operation is defined by:

$$m_{Y|X}(.|B) \quad = \quad (m_{X\uparrow XY}^B \oplus m_{XY})_{XY\downarrow Y}. \tag{6}$$

The mass function $m_{Y|X}(.|B)$ is called conditional mass function knowing that $B \subseteq \Omega_X$.

d) Deconditioning: In this problem, the conditional mass function $m_{Y|X}(.|B)$ and we wish to evaluate m_{XY}. The belief function theory suggests to choose the least informative mass function which provides $m_{Y|X}(.|B)$ after conditioning:

$$m_{XY}(C) = \begin{cases} m_{Y|X}(A|B) & \text{if } C = (B \times A) \cup (B \times \Omega_Y), \\ 0 & \text{if different for all } C \subseteq \Omega_{XY}. \end{cases} \quad (7)$$

B. Constructing belief functions

There are different ways to construct a belief function from data. Several classifiers such as the evidential k-nearest neighbors and neural network from Denoeux [5], [6] directly give a mass function as output.

For binary classification problem ($\Omega = \{C, \overline{C}\}$), the general formulation proposed by Xu et al. [16] is used to transform the classifier output into a mass function. In this paper, our method is also enriched by taking into account the outputs of classical multiclass classifiers such as SVM or boosting which provide a set of score measures for each class which is denoted here $(s_1, s_2, \ldots, s_K)$. To extract from this output a mass function, the following steps are processed, similarly to [1]:

- The scores are transformed into a probability distribution using a softmax function:

$$p(\omega_k) = \frac{\exp(s_k)}{\sum_{j=1}^{K} \exp(s_j)}. \quad (8)$$

- The probability are then transformed into a possibility:

$$\text{poss}(\{\omega_k\}) = \sum_{\omega_j \in \{\omega_1, \ldots, \omega_K\}} \min(p(\omega_k), p(\omega_j)). \quad (9)$$

- The possibilities $\pi_k = \text{poss}(\{\omega_k\})$ are sorted so that:

$$\pi_1 \geq \pi_2 \geq \ldots \geq \pi_K. \quad (10)$$

- The possibility is finally transformed into a consonant mass function:

$$m(A) = \begin{cases} \pi_k - \pi_{k+1} & \text{if } A = \{\omega_1, \ldots, \omega_k\}, \\ \pi_K & \text{if } A = \Omega, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

III. Global fusion process using evidential grammars

The previous step is local since for every segment, the belief function describing its class is computed only with the information lying inside the segment. In this section, a global fusion process on the top of this local fusion step will be presented using evidential grammars. Thus, the mass functions of the segments will be combined, and prior information provided by experts about the possible relative positions of objects in a traffic scene will be added as well. The goals which are expected from this stage are: segmentation of the scene into objects by grouping the segments corresponding to a single instance of a class and disambiguation of the belief functions at the local level as well as reduction of false positives.

A. Evidential Grammars

A grammar is defined as a 4-tuple $\{V_N, V_T, S, \Gamma\}$ where V_N is a finite set of non-terminal nodes, V_T a finite set of terminal nodes, S a start symbol at the root, and Γ is a set of production (or derivation) rules. A production rule $\gamma \in \Gamma$ changes a string of symbols (containing at least one non-terminal symbol) into another string of symbols. The production process starts with the S symbol and stops when the string is composed only of terminal symbols. The set of all the possible strings which can be produced by a grammar is called a *language*. The strength of grammars lies in the fact the language generated by a grammar can be large even when the *vocabulary*, that is to say V_T and V_N contain few elements.

To deal with image grammars, the natural left-to-right ordering is replaced with spatial relationships such as “hinge”, “border”, or “occlude”, which are used to combine segments into complex and structured objects. Moreover, to rank alternative interpretations and take into account uncertainty (on the class of the objects, on their relationships and on the derivation process), the grammar is augmented to a 5-tuple $\{V_N, V_T, S, \Gamma, \mu\}$ by adding a fifth component μ containing a set of conditional mass functions expressing our knowledge about the decomposition of the scene and the objects. This 5-tuple is called “evidential grammar”, the global framework of which has been detailed in [3], we thus expose here briefly the main aspects of this method.

B. Model of an image interpretation

The image interpretation is represented by a parse hypergraph. A parse tree is a decomposition of a scene into its components. For this purpose, several partitions of the image into regions are considered, each one corresponding to a level of description: objects, parts-of-objects, segments etc. An evidential variable X_i is set for every region R_i to describe its class, and every region is assumed to contain one single instance of an object: let us emphasize that uncertainty on the value of X_i doesn't mean than several classes might be mixed in R_i. To group them into a single entity, the pair (R_i, X_i) is called a “node” denoted N_i. Except in the case when X_i is associated to a region at the segment level, R_i is partitioned into regions of the lower level of description the corresponding nodes of which will be called “children nodes” of N_i. To get a parse hypergraph, the parse tree is augmented with spatial and contextual relationships between the children of a given node. These relationships depend of the level of interpretation where the nodes are lying, relationships such as ”aligned” or ”borders” can be used at a part-of-object level, ”occludes” or ”supports” at an object level. These relationships are taken into account by adding an evidential variable Ξ_i in the graph taking its value in the discernment frame composed of the set of relationships for the corresponding level of description. This makes it possible for instance to model a pedestrian as a head “over a” body.

In [3], it is shown that a parse hypergraph can be set in relationship with an evidential network by assuming that the joint belief function of a node and its children can be

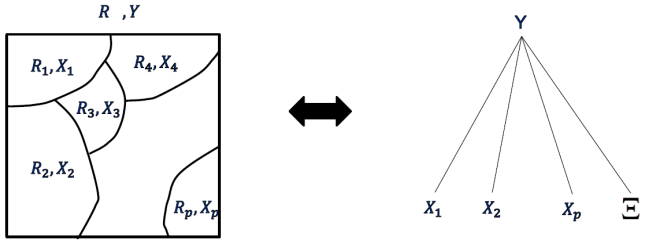

Fig. 2. Correspondence between a partition of an object into p components and the graphical dependency of the related variables. The variable Ξ describes the spatiale relationships between the regions R_1, R_2 $\ldots R_p$.

expressed independently of the other nodes of the graph. The evidential variables describing the content of the segments are provided as the output of the local fusion step. Given an evidential network, the belief is propagated from the leave nodes to the other nodes of the network up to the scene level through a bottom-up inference stage. This is performed through a succession of classical operations of belief functions: deconditioning, vacuous extension, Dempster's combination in the production space, and marginalization on the variable of the father's node. More precisely, if Y is a node the children nodes of which are denoted $X_1, X_2, \ldots X_p$ and the spatial relationship between those latter nodes is denoted as an evidential variable Ξ as illustrated on Fig. 2. In a first step, the vacuous extension is applied to the functions m_{X_1}, m_{X_2}, $\ldots$, m_{X_p} and m_Ξ. The resulting functions are denoted $m_{X_1\uparrow X_1,X_2,\ldots,X_p,\Xi,Y}$, $m_{X_2\uparrow X_1,X_2,\ldots,X_p,\Xi,Y}$, $\ldots$, $m_{X_p\uparrow X_1,X_2,\ldots,X_p,\Xi,Y}$, and $m_{\Xi\uparrow X_1,X_2,\ldots,X_p,\Xi,Y}$. These belief functions characterize the contents of disjoint regions and are thus supposed to be independent pieces of evidence. These belief functions are then combined using Dempster's rule:

$$m^1_{X_1,X_2,\ldots,X_p,\Xi,Y} = \left(\bigoplus_{i=1}^{p} m_{X_i\uparrow X_1,X_2,\ldots,X_p,\Xi,Y}\right)\cdots \bigoplus m_{\Xi\uparrow X_1,X_2,\ldots,X_p,\Xi,Y}. \quad (12)$$

In a second step, all the N conditional belief functions corresponding to grammar rules involving the rewriting of a symbol into p symbols are deconditioned into a set of N functions denoted here m^k defined on the product space $\{X_1,\ldots,X_p,\Xi,Y\}$. These belief functions correspond to distinct production rules which themselves encode different semantic information about the decomposition of the objects and the scene. They are thus supposed to be independent pieces of information and Dempster's rule of combination is consequently applied. We have:

$$m^2_{X_1,X_2,\ldots,X_p,\Xi,Y} = \bigoplus_{k=1}^{N} m^k_{X_1,X_2,\ldots,X_p,\Xi,Y}. \quad (13)$$

where Ξ is the observable variable defining the spatial relation between the regions. m^2 is then combined with m^1, and a belief function taking into account all the available information is thus obtained:

$$m_{X_1,X_2,\ldots,X_p,\Xi,Y} = m^1_{X_1,X_2,\ldots,X_p,\Xi,Y} \oplus m^2_{X_1,X_2,\ldots,X_p,\Xi,Y}. \quad (14)$$

The joint mass $m_{X_1,X_2,\ldots,X_p,\Xi,Y}$ is finally marginalized to extract m_Y:

$$m_Y = m_{X_1,X_2,\ldots,X_p,\Xi,Y\downarrow Y}. \quad (15)$$

C. Search for the optimal interpretation

By using the scheme detailed in the previous section, the belief is propagated from the segments up to the root to get an interpretation of an image. However, a large number of possible parse trees can be considered and consequently as many possible interpretations of a same image. We choose here to define the optimal parse tree as the one minimizing the conflict on the root node. Since, the non-normalized Dempster combination is applied, the root node aggregates all the conflict contained in the evidential network and thus gives a measure of the quality of the hierarchy.

A greedy algorithm is finally used to search for the optimal parse hypergraph in reasonable computation time. The main idea of this algorithm is to initiate a complex configuration which is simplified step by step as long as the consistency measure of the parse tree decreases:

- A parse tree is first initialized by linking all the nodes corresponding to the segments of the image directly to the root node. This is equivalent to considering that every segment is interpreted as one object.
- As long as the consistency measure of the parse tree decreases:
  - The consistency measure is computed for a set of alternative hypergraphs, each one being obtained by applying one single elementary modification to the current parse hypergraph. The elementary modifications that we consider are the merging of every pair of nodes of the same level of the hierarchy of the parse graph. If the nodes are terminal nodes, a new node is created which is linked with this pair of nodes. If the nodes are not terminal nodes, a new node is created which is composed of all the children of this pair of nodes.
  - The parse hypergraph minimizing the consistency measure is kept for the next iteration.
- The last parse hypergraph is kept as the output of the method.

IV. Experiments

The KITTI Benchmark Suite [9] was used to validate our approach. A set of 140 images has been annotated manually with a total of 14 classes as listed in Tab. I. Several modules were trained on 100 images and tested on the 40 others.

A. Sensors and modules

We used a monocular camera, a stereo camera and a LIDAR as sensors. The principal monocular classification module is the Automatic Labeling Environment (ALE) proposed by Ladický et al. [12], which can be directly learned over all the previously defined classes. The second monocular module, from the works of Hoiem et al. [11], estimates the scene geometry from one single image. The classification output

Ground	Sky		Static structures	Building	
	Road			Pole	
	Sidewalk			Fence	
	Lane marking			Other	
	Grass	Vegetation		Car/trucks	Moving obstacles
	Tree			Person/cyclist	
	Other			Other	

TABLE I

CLASSES OF OBJECTS CONSIDERED IN OUR EXPERIMENTS. THERE ARE 14 CLASSES IN TOTAL, SOME OF THEM CAN BE GROUPED INTO SETS.

is limited to the three classes: ground, obstacles and sky. In our case, the ground class is the union of road, sidewalk, lane marking and grass, while the obstacles class includes everything else except the sky. The class membership scores from those two modules are transformed into a mass function following the steps (8-11).

Then the 3D information from the stereo camera and the LIDAR are used to detect the ground as in [16], by assuming that the ground is planar. Again, the classes ground and non-ground are actually sets of other classes. We clearly see the interest of working with sets of classes.

B. Grammar model

A three levels grammar model was considered: scene, objects, and segments. No parts of objects (such as "wheel" or "head") were considered in these experiments, and the objects are supposed to be derived directly into the set of their constituting segments which can be considered as elementary pieces of the objects. Consequently, the size of the discernment space at the segment level is the same as the one at the object level. The pairwise links "occlude", "is occluded by", "bordering" and "disjoint" were used to describe spatial relationships between the objects. The root node can produce any arbitrary combination of instances of the 14 types of objects we consider under a set of 30 spatial constraints. These constraints correspond to the derivation rules which are formulated as a set of prohibited configurations between pairs of objects, for example: "The sky cannot occlude an other object", " A car cannot border a building", "The road cannot occlude an other object", etc. Each object can be decomposed in an arbitrary combination of patches of the corresponding object under a spatial constraint of neighborhood. It should be noticed that this model takes little advantage of the potential of the grammars to decompose complex objects in structured reusable components. Indeed, no database annotated with parts of objects were available for that purpose.

C. Results

The inputs and outputs of our system are illustrated on Fig. 3. The average precision of the multi-class classification is showed on Tab. II. We can see that using more information often improve the results, and it never significantly degrades them. At the local level, no notion of object is handled. After using the global approach, different cars can be separated and segmented using geometric constraints.

V. CONCLUSIONS

We have showed an information fusion based system which is flexible to include many sensors and modules defined over different sets of classes. It is based on the framework proposed in [16] and has been augmented by a global grammar-based reasoning [3]. Fusion at the segments level improves local accuracy and global fusion enables to have object level reasoning.

Future works will enhance object classification by introducing sliding windows based approaches as well as part-based object detections so as to use the full potential of visual grammars. New sources of information such maps will also be considered.

ACKNOWLEDGMENT

This work is supported and financed by the Cai Yuanpei program from the Chinese Ministry of Education, the French Ministry of Foreign and European Affairs, the French Ministry of Higher Education and Research, and by the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). It is also supported by the Blanc International ANR-NSFC Sino-French PRETIV project and the ICT-ASIA PREDiMap project.

REFERENCES

[1] A. Aregui and T. Denœux. Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. *International Journal of Approximate Reasoning*, 49(3):575-594, 2008.

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

[3] J.-B. Bordes, F. Davoine, P. Xu, and T. Denœux. Evidential grammars for image interpretation. application to multimodal traffic scene understanding. In *Third International Symposium on Integrated Uncertainty in Knowledge Modeling and Decision Making*, pp. 65-78, China, 2013.

[4] R. Brehar, C. Fortuna, S. Bota, D. Mladenic and S. Nedevschi. Spatio-temporal reasoning for traffic scene understanding. In *Proc. of ICCP*, pp. 377-384, USA, 2011.

[5] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(5):804-813, 1995.

[6] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 30(2):131-150, 2000.

[7] P. Dollár, C. Wojek, B. Schiele and P. Perona. Pedestrian detection: an evaluation of the state of the art. *PAMI*, 34(2):743-761, 2011.

[8] A. Ess, T. Müeller, H. Grabner and L. J. Van Gool. Segmentation-Based Urban Traffic Scene Understanding. In *Proc. of BMVC*, pp. 84-91, UK, 2009.

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

[10] D. M. Gavrila and S. Munder. Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle. *IJCV*, 73(1):41-59, 2007.

[11] D. Hoiem, A.A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):1-23, 2007.

[12] L. Ladick y, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H.S. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. In *Proc. of BMVC*, pp. 1-11, UK, 2010.

	Ground				Vegetation			Static structures				Moving obstacles			Overall
	Sky	Road	Sidewalk	Lane marking	Grass	Tree	Other	Building	Pole	Fence	Other	Car/truck	Person/ cyclist	Other	
ALE	**96.7**	82.5	**88.5**	93.5	89.9	86.7	78.5	85.1	90.5	88.4	94.7	88.8	86.0	91.1	86.7
ALE + Geo	94.8	82.6	86.4	93.4	89.9	**88.1**	87.5	**85.5**	**93.4**	**94.9**	98.7	**92.9**	94.1	**100**	87.8
ALE + Geo + 3D	**96.7**	**83.0**	87.9	**94.5**	**95.4**	87.0	**88.1**	85.1	92.8	94.8	**98.8**	92.7	**94.7**	**100**	**88.0**

TABLE II

AVERAGE PRECISION (IN PERCENTAGE) OF THE MULTI-CLASS CLASSIFICATION. ALE IS THE MONOCULAR MULTI-CLASS MODULE FROM [12]. GEO REFERS TO THE MONOCULAR GEOMETRIC CONTEXT FROM [11]. 3D IS TO THE GROUND DETECTION MODULE USING STEREO AND LIDAR AS IN [16].

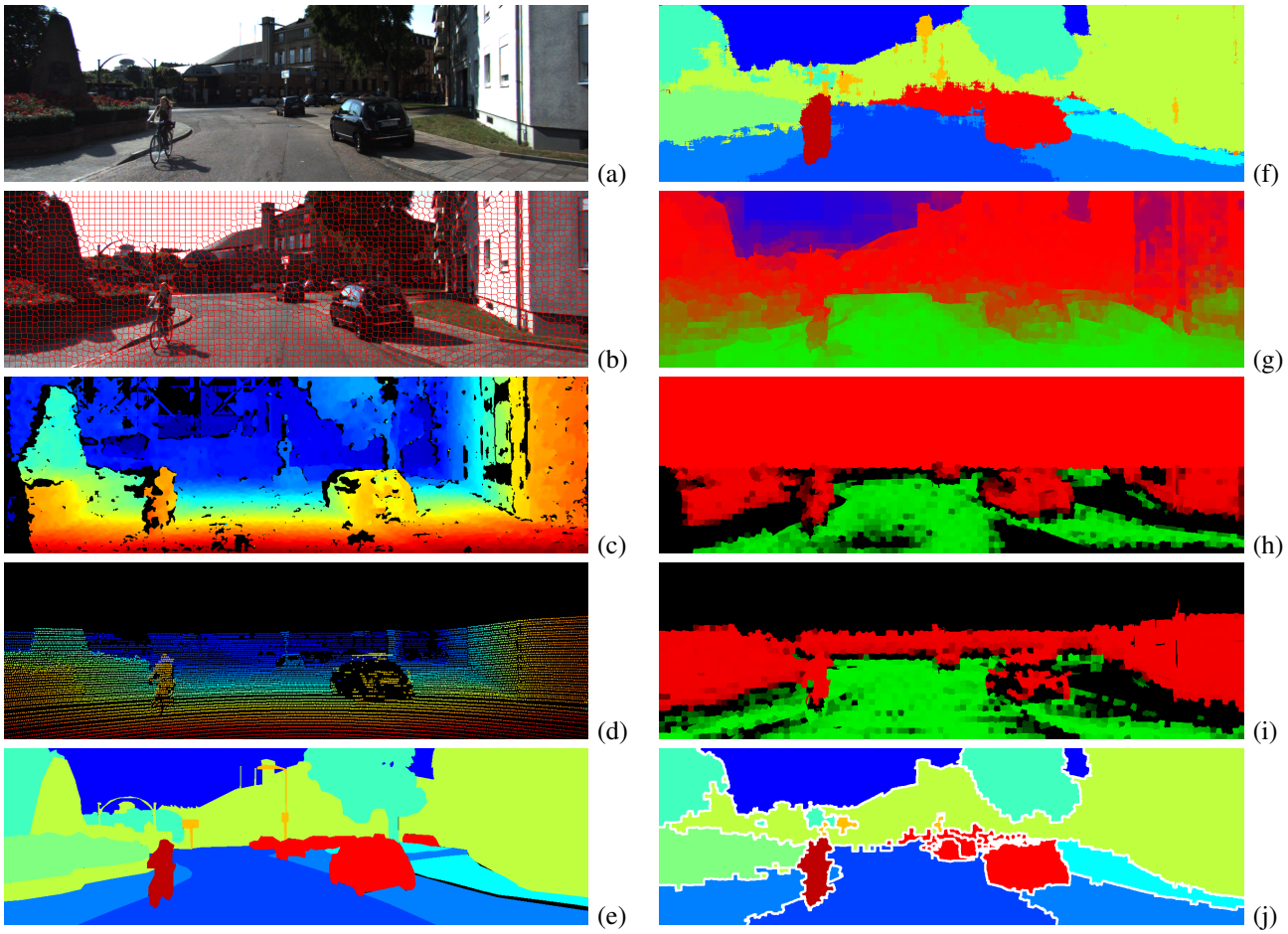

Fig. 3. Input data and results from the different modules. (a) Raw image from the left camera. (b) Over-segmented image. (c) Disparity computed from the stereo camera. (d) Lidar impact points. (e) Ground truth with 14 classes. (f) Output from ALE, the color of each pixel is the class with highest score. (g) Classification probability from the geometric context, the red, green and blue intensities represent the probability of having an obstacle, the ground and the sky respectively. (h-i) Ground/Non-ground classification using 3D information, the green color represents the mass put on the class "ground" and the red the one on "non-ground", the black color represents the ignorance. The results are from the data (c-d) respectively. (j) Final combined information and segmentation from the evidential grammar.

[13] D. Mercier, B. Quost, and T. Denœux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 9(2):246-258, 2008.

[14] S. Thrun, W. Burgard and D. Fox. Probabilistic robotics. *The MIT Press*, Cambridge, Massachusetts, 2005.

[15] C. Wojek and B. Schiele. A Dynamic Conditional Random Field Model for Joint Labeling of Object and Scene Classes. In *Proc. of ECCV*, pp. 733-747, France, 2008.

[16] P. Xu, F. Davoine, J.-B. Bordes, H. Zhao, and T. Denœux. Information fusion on oversegmented images: An application for urban scene understanding. In *Proc. of Intl. Conference on MVA*, pp. 189-193, 2013.

[17] S-C. Zhu and D. Mumford. A stochastic grammar of images. *Found. Trends. Comput. Graph. Vis.*. 2(4):259-362, 2006.