EVIDENTIAL GRAMMARS FOR IMAGE INTERPRETATION. APPLICATION TO MULTIMODAL TRAFFIC SCENE UNDERSTANDING

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

[1] UMR CNRS 7253, Heudiasyc, Université de Technologie de Compiègne, BP 20529, 60205 Compiègne Cedex France
[2] LIAMA, CNRS, Peking University, Beijing, P.R. China

Abstract. In this paper, an original framework for grammar-based image understanding handling uncertainty is presented. The method takes as input an over-segmented image, every segment of which has been annotated during a first stage of image classification. Moreover, we assume that for every segment, the output class may be uncertain and represented by a belief function over all the possible classes. Production rules are also supposed to be provided by experts to define the decomposition of a scene into objects, as well as the decomposition of every object into its components. The originality of our framework is to make it possible to deal with uncertainty in the decomposition, which is particularly useful when the relative frequencies of the production rules cannot be estimated properly. As in traditional visual grammar approaches, the goal is to build the "parse graph" of a test image, which is its hierarchical decomposition from the scene, to objects and parts of objects while taking into account the spatial layout. In this paper, we show that the parse graph of an image can be modelled as an evidential network, and we detail a method to apply a bottom-up inference in this network. A consistency criterion is defined for any parse tree, and the search of the optimal interpretation of an image formulated as an optimization problem. The work was validated on real and publicly available urban driving scene data.

Keywords: Visual grammars; Belief functions; Image understanding

1 Introduction

Automatic understanding of the traffic scene in front of a car is an essential task for autonomous driving, as well as for safety systems. Generally speaking, by "understanding" we mean detecting the objects in the image and eventually describing some spatial or functional relationships between them. However, detecting even a single kind of object can be very challenging since the varying lighting conditions, the highly cluttered environment, the dynamically changing backgrounds among many others contribute to the difficulty of the task. Many algorithms have been developed which are specialized for one of these detection subtasks and each of them can use different kinds of sensors. However, even the most competitive state-of-the-art detection modules are sources of false detections, as well as misdetections (pedestrian detection for example). By modelling the consistency of the relative positions of the objects with respect to the scene, it is possible to strengthen weak detection while pruning some false detections. For better results, these post-processing methods should take into account the uncertainty of classification outputs provided by different independent modules, which is possible using the Dempster-Shafer theory [9]. Indeed, this theory provides an

elegant formalization of uncertainty and the Dempster's combination rule is an efficient way to fuse information in a multi-sensor context. The critical goal which is addressed in this paper is thus to define a precise framework to model the consistency of a scene while handling the uncertainty on the class of detected objects.

1.1 Related Work

Scene understanding is one of the main goals in computer vision and robotics. In the last decade, the accuracy of object detection methods has increased substantially. On urban scenes, Ess et al. [3] and Gavrila et al. [5] have successfully combined a set of visual cues extracted from a stereo camera for pedestrian detection in real time. More recent approaches try to annotate every pixels of the image with semantic categories: Ess et al [4] use textural features as well as some geometric information to train a set of classifiers in one versus all on small patches of the image. Some temporal smoothing is then applied as a post-processing but the annotation is purely local. Wojek et al [10] exploit the scene layout information at a further extent using joint object detection to perform the segmentation; the model includes some consistency clues about the spatial relationships between the objects (the distance between pedestrians and the ground for example).

More generally, an efficient way to take into account spatial and functional relationships between the objects is the use of ontologies [1], which is the formal representation of concepts for a particular domain and their semantic relationships. The Literature on computer vision provides several references about ontology application on various types of databases, e.g., on the PASCAL challenge database using WordNet semantic network [8] or on traffic scenes [1] where Brehar uses the openCyc ontology. The semantic hierarchy provided by ontologies like WordNet or openCyc is useful to shed light on inter-class relations between visual elements. However, we believe it is not sufficient for modelling symbolic representation of a visual system since it does not account for some important aspects of semantic relations such as spatial and functional relationships. In [13], *and/or* graphs are used to represent the decomposition of scenes into objects, parts and low-level patterns called "primitives". This graph is set in relation with a visual grammar, which is a set of derivation rules. By augmenting these rules with probabilities, a stochastic model is defined which is trained and then used to compute the parsing graph of a test image under the Bayesian posterior. Impressive results have been achieved on various datasets containing a large number of object categories [12]. The strong advantage of visual grammars over discriminative machine learning methods is their generalization capability: parts of objects can be learnt with a relatively small amount of training data and can be used to recognize a large number of configurations, they also provide elegant and efficient ways to solve problems like occlusions and scale.

1.2 Contributions

Visual grammars take as input a set of visual primitives (texture, corners, color histograms ...) extracted from the image during a first step of image processing. Each one of these primitives provides by itself rather poor information about the class of the area it covers, but there is no doubt about its value. In this paper, we will show how it is possible to extend the visual grammar approach to the case where the input information is uncertain. The belief function theory makes it possible to deal with this situation: every input of information associated to a part of the image which will be defined as a belief function. This approach can be especially interesting in the multimodal case where modules using information incoming from several sensors can extract "higher-level" information than

traditional visual primitives: parts of objects can indeed be detected, or even objects themselves. This classification result can be taken as input to our visual grammar method, but the uncertainty of this information has to be taken into account as a module may detect an object with a given confidence degree, or may not be able to discriminate between different classes (a general obstacle detector cannot distinguish a vehicle from a pedestrian). We also propose a way to deal with another kind of uncertainty, which is uncertainty in the production rules. When the relative frequencies of occurrence of different production rules are unknown, rewriting the production rules as conditional belief function allows us not to add artificial information in the model. In brief, our work is at the intersection of two promising approaches: visual grammars and belief functions. To our knowledge, no work has yet focused on the possibility to match these two theories.

1.3 Overview

The system considered here consists of several sensors, like mono or stereo cameras and laser sensors, observing an urban driving scene. Images are over-segmented and then processed at the level of entire segments rather than single pixels. Each sensor provides data to one or more modules which will run totally or partially in parallel and assign a belief function about the class label of the image segments. The output belief functions from different modules are then fused on every segment using the Dempster's rule of combination [11] (cf Fig 1). In this article, we augment this "local" fusion process with a "global" fusion step by merging the segments into larger regions using predefined combination rules. This can achieve two main goals: disambiguate the belief one has on the segments by adding context information, and infer complex objects by combining small regions corresponding to parts of these objects.

We will show in Section 4 how to propagate the belief functions from the segment level to the region and scene levels. We will then define a criterion which makes it possible to rank all the possible parse trees for a given input image. The understanding process is thus defined as the search for the parse tree minimizing this criterion. We will then present a fast optimization algorithm implementation of this research. We call this approach a "framework" since it can be used for any set of derivation rules, spatial relationships and labels. To apply the evidential grammar, they have to be instantiated. Experimental validation is presented in Section 6, using the publicly available KITTI Vision Benchmark Suite [6].

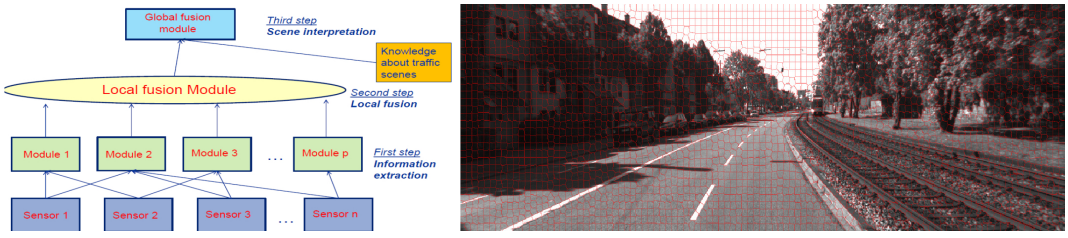


Fig. 1. (a) Global architecture of the system. The approach presented in this article corresponds to the "global fusion module" at the top of this structure. (b) Example of an oversegmented image

2 Evidential grammars

2.1 Formal grammars and stochastic grammars

The modern formalization of grammars can be attributed to Chomsky [2]. 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 language 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 real data, which may include a large amount of irregular patterns, a grammar should contain a substantial amount of derivation rules to ensure that at least one sequence generates the terminal symbols which are extracted in the data. When several possible sequences of rules can generate the terminal symbols, the grammar is said to be *ambiguous*. To rank alternative interpretations of a given image, the grammar is augmented with a set of probabilities P as a fifth component to form a stochastic grammar $\{V_N, V_T, S, \Gamma, P\}$. A set of nodes can then be derived in other sets of nodes with given probabilities, taking into account the spatial relationships between the nodes. They are an essential component of the model and are estimated during the training process.

2.2 Visual grammars

The most important problem when dealing with image grammars is the loss of the natural left-to-right ordering in textual data. In language, every production rule is assumed to generate a linearly ordered sequence of nodes and, following this down to the leaves, a linearly ordered sequence of terminal symbols is obtained. In images, the implicit links between neighbouring words have to be replaced by more complex links between the nodes. Indeed, the spatial links that can combine parts of objects into an object are numerous such as "hinge", "border", "butting", "surround" as well as various alignments. 3D information can also be used to determine that an object is "occluding" another one or that an object is "supporting" another object. In image grammars, pixels seem natural candidates to be terminal symbols, but a single pixel carries very little information and usually some sets of local structures are preferred. Several works have focused on proposing *visual primitives*, sometimes called also *textons* [7] as terminal nodes for visual grammars. These terminal nodes usually correspond to corners or patches of texture; they are computed deterministically during a step of feature extraction.

In this article, a framework is proposed to deal with terminal nodes containing uncertainty. This can be used to combine outputs of a set of modules detecting various objects, parts of objects and areas but which are not perfectly reliable such as, for example in driving scenes, road extraction, wheel or pedestrian detectors, etc. The main idea is to use as terminal nodes a higher-level vocabulary taking into account uncertainty instead of low-level primitives with no uncertainty. We believe this to be especially a good choice in multimodal situation like ours: the extraction of information incoming from various sensors makes it relevant to reason from a higher level than image primitives. However, pedestrian detectors, for example, are well-known to provide many false positives and handling precisely the uncertainty is a matter of highest importance. To this end, the Dempster-Shafer

theory will be used and the set of probabilities augmenting the production rules will be replaced by a set of belief functions to constitute what we will call here *evidential grammars*.

2.3 Formalization of evidential grammars:

An evidential grammar is defined as a 5-tuple $\{V_N, V_T, S, \Gamma, \mu\}$ where the first four elements correspond to the traditional 4-tuple defining a formal grammar. The fifth component μ replaces the set of probabilities of a stochastic grammar. The set of non-terminal elements V_N and the set of terminal elements V_T contain the classes that are used to label the images. The elements of V_N correspond to the classes of objects which can be decomposed into a set of other elements belonging to $V_N \cup V_T$. The starting element S of the grammar will be interpreted here as the "Scene" annotating every image. Γ contains the production rules of the grammar. Every rule defines a decomposition of a class of object into a set of classes of objects with a defined spatial layout, and every element of V_N has at least one production rule.

More formally, let $A \in V_N$ be a class with $n_{k,\xi}(A)$ production rules into k components under spatial relation ξ (A may have production rules in more or less than k components and these components may have other spatial layouts than ξ):

- Γ_1: $A \rightarrow A_1^1 A_2^1 \dots A_k^1$
- $\dots$
- $\Gamma_{n_{k,\xi}(A)}$: $A \rightarrow A_1^{n_{k,\xi}(A)} A_2^{n_{k,\xi}(A)} \dots A_k^{n_{k,\xi}(A)}$

where the A_j^i stand for classes of objects belonging to $V_N \cup V_T$. Let us define the precise meaning of these production rules.

Let R be a region of an image and Y an evidential variable describing the class contained inside this region. R is supposed to be segmented in k regions R_1, R_2, $\dots$, R_k and the class contained in each region R_i is defined by the evidential variable X_i (see Fig 2). The use of evidential variables makes it possible to deal with the uncertainty about the class of the regions. However, even though one might have uncertainty on the value of the class inside these regions, the (strong) assumption is made that every region contains one and only one instance of a class. For example, if it is certain that $Y \in \{car, pedestrian\}$ (that is to say, $m_Y(\{car, pedestrian\}) = 1$), this means that the region contains either one (full) car or one (full) pedestrian, but this does not mean the region can contain a pedestrian **and** a car.

The spatial relationship between the regions R_1, R_2, $\dots R_k$ is described by a deterministic variable Ξ (since this spatial layout in an image between a set of regions is always observable). Ξ can describe a global layout between the regions (the regions are "aligned" or "radial") as well as a set of pairwise relationships.

The production process defined by the production rules $\Gamma_1, \dots, \Gamma_{n_{k,\xi}(A)}$ is associated to a conditional mass function $m_{X_1,\dots,X_k|Y=A,\Xi=\xi}$. To be consistent with the production rules, the only possible focal sets of the latter mass function are the $\Omega_i = \{A_1^i, A_2^i, \dots, A_k^i\}$ and the sets defined as unions of Ω_i. Let us emphasize two particular cases:

1. The case when the focal elements are the Ω_i is equivalent to the production process of a stochastic grammar.
2. The case when $m_{X_1,\dots,X_k|Y=A,\Xi=\xi}(\cup_{i=1}^{n_{k,\xi}(A)} \Omega_i) = 1$ is an assumption of complete ignorance about the relative frequencies of output of the production process.

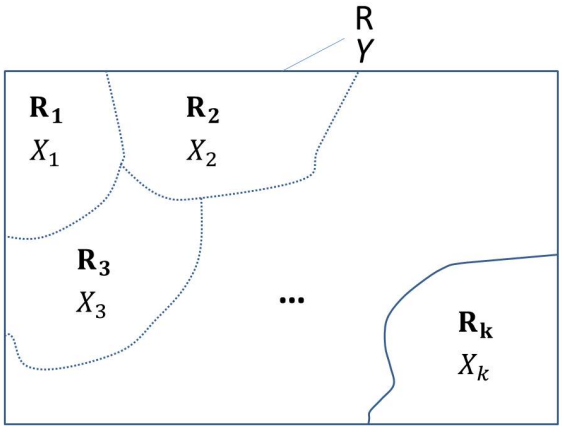


Fig. 2. Illustration of the derivation process: the class of the region R (the whole rectangular area) is described by an evidential variable Y. The region R is then segmented in k regions R_1, R_2, ... R_k the classes of which are also decribed by evidential variables.

The conditional mass $m_{X_1,...,X_k|Y=A,\Xi=\xi}$ describes the derivation process of A into k components with spatial relation ξ. Consequently, conditional mass has to be defined for every possible number of components and every spatial layout in which the class of object A can be derived.

3 Model of an image interpretation

Parse hypergraphs will be introduced in this section as the structure for image interpretation. Given an evidential grammar $\{V_N, V_T, S, \Gamma, \mu\}$, we define $\Omega = V_N \cup V_T$ as the set of labels which will be used to annotate the image.

3.1 Parse hypergraphs

Let us consider a test image I which is supposed to be parsed hierarchically in l layers ($l \geq 2$) of a parse hypergraph. A given layer $i \in \{1, \ldots, l\}$ is composed of a set of n_i regions which are denoted $\{R_{i1}, \ldots, R_{in_i}\}$. The layer l, at the top of the hierarchy, contains a single region which is the image itself I, we can thus write $R_{l1} = I$. If the layer is not the bottom one (ie, $i > 1$) a region R_{ij} can be partitioned in a set of regions of the lower layer (see Fig 3).

Moreover, an evidential variable denoted X_{ij} taking its value in Ω is introduced to describe the content of every region R_{ij} and is associated to the belief function $m_{X_{ij}}$.

The evidential variable X_l is constrained to a single value which is the "Scene" label. "Scene" corresponds to the S symbol of any formal grammar, which is the starting symbol from which all the sentence is derived up to the leaves. The higher-level labels are expected to correspond to more complex objects, like "car" or "pedestrian", than the lower-level labels, which are expected to correspond to parts of objects like "wheel", or "pedestrian head". However, the scale issue implies that it is possible for any object to be located in a small region of an image, making it impossible to parse into its components. Consequently,it must be possible for any symbol (except S) to terminate

at any level of the labels hierarchy. Thus, the evidential grammar $\{V_N, V_T, S, \Gamma, \mu\}$ is defined with $V_N = V_T$.

A *Node* N_{ij} is defined as the pair (R_{ij}, X_{ij}), that is to say a region and the description of its content. The nodes will be used to represent the oriented hypergraph structure of the image interpretation with two kinds of edges being considered:

1. An edge "is part-of/is composed-of": $\{N_1, \ldots, N_p\}$ are part-of N, or reciprocally N is composed-of $\{N_1, \ldots, N_p\}$ if the regions covered by $\{N_1, \ldots, N_p\}$ define a partition of the region covered by N.
2. An edge describing the spatial relationship between a set of nodes at the same level.

The first type of edge defines a parse tree since it represents the decomposition of the scene into objects and parts of objects. By adding the second type of edge between nodes of the same layer, the parse tree is augmented to a parse hypergraph. Indeed, to describe complex relations between a set of nodes (radial rays of a bicycle), edges linking several nodes should be considered. The set of nodes and edges defining an image interpretation is called a parse hypergraph of the image I and is denoted $P_h(I)$

3.2 Evidential network corresponding to an image interpretation

Let $U = \{X_{11}, X_{12}, ..., X_l\}$ be the set of all the variables used for the image interpretation and p the number of edges in a given parse hypergraph $P_h(I)$. A set of masses $\mathbb{M} = \{m_1, \ldots, m_p\}$ is introduced where every mass m_i is defined on the discernment frame $U_i \subset U$ where U_i is a set of variables whose corresponding nodes consist in one node which is linked to several nodes with the "part-of" link (cf Fig 3). Let $O = \{U_1, \ldots, U_p\}$ be the set of all the discernment frames. The 3-tuple $\mathbb{E} = \{U, O, \mathbb{M}\}$ defines an evidential network and matches the structure of the image interpretation $P_h(I)$.

As for the Bayesian network where the graphical model corresponds to a set of conditional independences between random variables, the evidential network facilitates the inference of the unknown belief functions since the masses included in $\mathbb{M}$ provide an expression of the joint mass of the subsets of variables U_i independently of the rest of the evidential variables of U. The global joint mass function can be computed as:

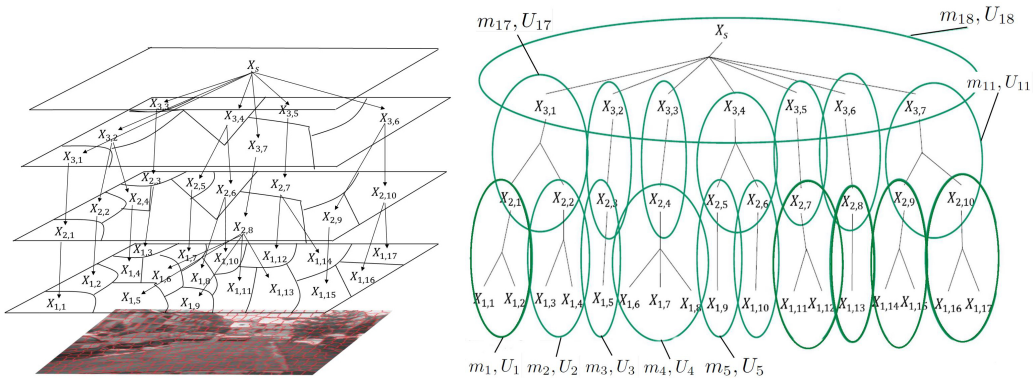
$$m_U = \bigoplus_{i=1}^{p} m_{U_i \uparrow U},$$

where $m_{U_i \uparrow U}$ stands for the vacuous extension in the product space of U.

The belief function of the leave nodes of the parse tree are assumed to be known as an input data and we wish to marginalize U to compute the belief functions on all the other variables. At this point, two questions clearly stand out:

1. Given the hierarchy of regions, how to propagate the belief function from the leaves to the upper nodes?
2. What could be a criterion which could lead us the most realistic interpretation of an image?

These questions will be addressed in the two next sections.



(a) (b)

Fig. 3. (a) Hierarchy of regions for the interpretation of a test image (b) Evidential network corresponding to this hierarchy. A joint mass is defined for all the sets of nodes linked by a "part-of" edge (the sets are surrounded by green circles.)

4 Belief propagation

4.1 Bottom-up inference

Given an interpretation $P_h(I)$ of a test image I and the corresponding evidential network $\mathbb{E} = \{U, O, \mathbb{M}\}$, every mass m_i is defined on a set of variables U_i, the corresponding nodes of which are composed k nodes of a layer $j \in \{1, \ldots, l-1\}$ linked together by a spatial relationship ξ, and linked to one node in the layer $j+1$ by a "part-of" link. The belief functions of the leaves are assumed to be known as the input of the present algorithm. The scheme which is proposed here consists in computing the belief functions iteratively from layer 2 to layer l: the nodes of layer 1 are used to compute the belief functions of the nodes of layer 2, and so on up to the root.

In a first step, the vacuous extension is applied to the functions m_{X_1}, m_{X_2}, $\ldots$, m_{X_p}. The resulting functions are denoted $m_{X_1 \uparrow X_1, X_2, \ldots, X_p, Y}$, $m_{X_2 \uparrow X_1, X_2, \ldots, X_p, Y}$, $\ldots$, $m_{X_p \uparrow X_1, X_2, \ldots, X_p, Y}$. 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:

$$m^1_{X_1, X_2, \ldots, X_p, Y} = \bigoplus_{i=1}^{p} m_{X_i \uparrow X_1, X_2, \ldots, X_p, Y}.$$

In a second step, all the N conditional belief functions corresponding to grammar rules involving the rewriting of a symbol into p symbols under relation ξ are deconditioned and a set of N functions defined on the product space $\{X_1, \ldots, X_p, Y\}$. 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:

$$m^2_{X_1, X_2, \ldots, X_p, Y} = \bigoplus_{k=1}^{N} m^k_{X_1, X_2, \ldots, X_p, Y | \Xi = \xi},$$

where Ξ is the observable variable defining the spatial relation between the regions. This function is then combined with the previous one, to get a belief function taking into account all the available information:

$$m_{X_1,X_2,\ldots,X_p,Y} = m^1_{X_1,X_2,\ldots,X_p,Y} \oplus m^2_{X_1,X_2,\ldots,X_p,Y}.$$

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

4.2 Top-down propagation

Once all the mass functions have been computed through the previous inference mechanism, a top-down scheme can be performed to disambiguate the mass functions at the lower layers of the hierarchy. For every subset of variables $U_i = \{X_1, X_2, \ldots, X_p, Y\}$ and after the joint mass has been computed, k marginalizations are applied to compute $m_{X_1}, m_{X_2}, \ldots, m_{X_k}$. This process is performed from the root down to the leaves for every subset of variables. This scheme results in a disambiguation of the variables by re-injecting the information gathered through the combinations performed during the inference down to the variables corresponding to the segment level of the image.

5 Search for the optimal interpretation

The groupings of nodes with the "part-of" links define a hierarchy of regions and this hierarchy define an evidential network. By using the scheme detailed in the previous section, the belief functions are computed from the lower levels 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. Using traditional stochastic visual grammars, we build the parse tree of a test image with maximal posterior probability. For evidential visual grammars, we propose to use the minimal conflict accumulated at the root node as the optimization criterion to select the parse hypergraph.

5.1 Optimization criterion

The key value that will be used to measure how relevant it is to group a set of nodes is the conflict located in the newly constructed node. By conflict, we mean here the value of the mass function of this node on the empty set.

Consistency measure of an interpretation The dependencies between the variables induced by the grouping of nodes with the "part-of" link imply two remarks:

1. If several nodes N_1, N_2, $\ldots N_p$ are linked to an upper node M by a "part-of" edge in the parse tree, and if their mass functions don't match any derivation rule, the upper node M will contain a conflict equals to 1. More generally, a node containing a high conflict indicates high inconsistency in the subtree.
2. When several nodes N_1, N_2, $\ldots N_p$ are linked to an upper node M by a "part-of" edge in the parse tree, the combination formula implies that the conflict of M is bigger than the maximum conflict of N_1, N_2,$\ldots N_p$. Thus, any conflict appearing during the bottom-up belief propagation process in the evidential network will appear on the root node.

These two remarks drive us to consider the conflict in the node N_{l1} as the main optimization criterion, since as the root node of the evidential network, it aggregates all the conflict contained in the evidential network and it gives a measure of the quality of the hierarchy. The lower is $m_{X_{l1}}(\emptyset)$ (or the higher is $m_{X_{l1}}(S)$ since $m_{X_{l1}}(S) = 1 - m_{X_{l1}}(\emptyset)$), the more consistent is the hierarchy.

Remarks on the number of optimal interpretations The principle of consistency maximization does not ensure that only one optimal interpretation exists. In some cases, many alternative interpretations of an image can be found which have no conflict at the root node. Actually, two factors impact the number of interpretations that can be expected:

1. The quantity of information carried by the belief functions of the input nodes.
2. The number of production rules in the model.

The less informative the input belief functions are and the more numerous are the production rules, the more possible combinations there are to form different consistent interpretations. The production rules being a part of the model, they have to fit correctly the database: too restrictive rules will lead to no available interpretation since any interpretation will have a high level of conflict. On the contrary, defining too many configurations or allowing too many possible decompositions of the objects will lead to an output containing many unsatisfactory interpretations.

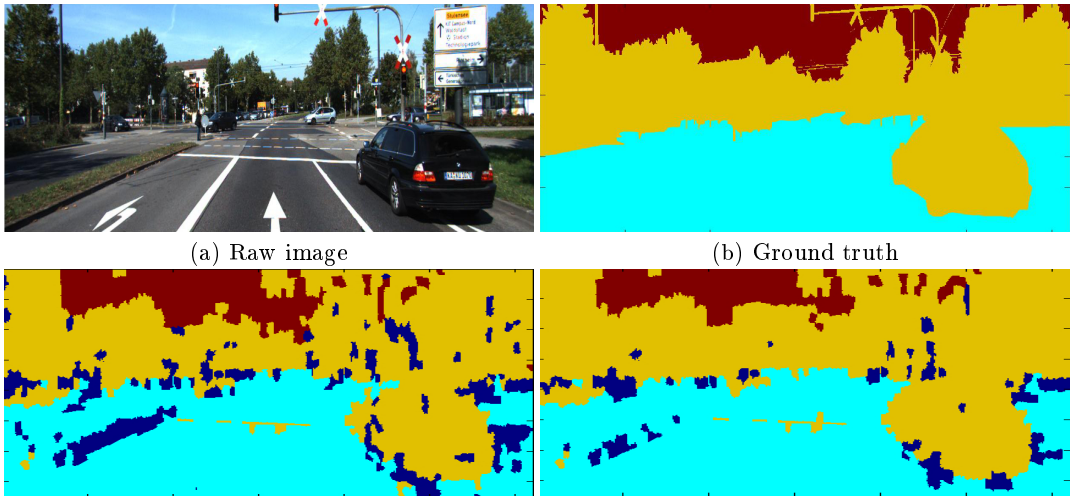
5.2 Optimization algorithm

The number of possible parse hypergraphs is very large and exploing the whole space is untractable. A greedy algorithm is thus introduced here to obtain a relevant interpretation of a test image 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 by the "part of" relationship 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.

6 Experiments

The KITTI dataset [6] was used to validate our approach. The stereo color camera system and Velodyne LIDAR were used as sensors. Using the algorithms presented in [11], the images of the



(a) Raw image (b) Ground truth

(c) Before processing (d) After processing

Fig. 4. Three classes $\Omega = \{Ground, Sky, Obstacle\}$ classification result. The deep blue color indicates that the image segment has not been classified by the system.

dataset were first oversegmented and a combination of modules was used and merged locally at the segment level to provide a belief function for every segment of the image.

Three classes of objects were considered for the definition of the grammar: "Ground", "Obstacle" and "Sky". The pairwise links "occlude", "is occluded by", "bordering" and "disjoint" were used to describe the spatial relationships between the regions. The starting node S can produce an arbitrary combination of these objects under a few basic spatial constraints (for example, the sky cannot occlude the other types of objects). At the part of object level, three classes were considered: "Patch of ground", "Patch of obstacle", and "Patch of sky". The pairwise links "neighbouring" and "disjoint" were considered. Each object can be decomposed in an arbitrary combination of the corresponding patches of object under a spatial constraint of neighbourhood. It should be noticed that this model takes only little advantage of the potential of the grammars to decompose complex objects in structured and reusable components. Indeed, no input data for parts-of-objects were available for that purpose. We plan to apply a model at the object part level in future work.

Fig 5 shows the classification results before and after processing of the evidential grammar. Since our system allows us to represent ignorance, it can happen that no decision can be made about some segments, which explains why the recall rate is different from the diagonal of the confusion matrix. A result image is shown in Fig 4. As it can be seen on these results, there is little difference in precision but the evidential grammar provides an important improvement in recall. This comes from the disambiguation provided by the top-down propagation of the belief, which transfers the fusioned information at the region level back to the segment level.

7 Conclusion

We have introduced an original framework for image understanding based on visual grammars and Dempster-Shafer theory. Our method makes it possible to introduce uncertainty in the production

	Ground	Obstacle	Sky	Recall
Ground	96,4	3,6	0	86,1
Obstacle	5,9	94,0	0,1	81,3
Sky	0	33,2	66,8	66,5

(a)

	Ground	Obstacle	Sky	Recall
Ground	97,0	3,0	0	95,7
Obstacle	6,2	93,7	0,1	85,8
Sky	0	33,2	66,8	66,5

(b)

Fig. 5. Confusion matrices, values are given in percentage. (a) Results without processing of evidential grammar (b) Results after processing of evidential grammar

rules of the grammar, which allows us to bypass artificial knowledge when the relative frequencies of the output of some derivation rules cannot be estimated reliably. Moreover, our method can handle input data tainted with uncertainty. We demonstrated the efficiency of our method by post-processing the classification result of an oversegmented image with uncertainty on the class of the segments. We showed that our method provides significant improvement on the recall rate. Future work will take into account information at the part-of-object level in order to fully exploit the strength of the grammars, which lies in their ability to combine visual elements to detect complex structured objects even in highly cluttered environment.

References

1. R. Brehar, C. Fortuna, S. Bota, D. Mladenic, and S. Nedevschi. Spatio-temporal reasoning for traffic scene understanding. In *Proc. of ICCP*, pages 377–384, Pittsburgh, 2011.
2. N. Chomsky. *Syntactic Structures*. Mouton: The Hague, 1957.
3. A. Ess, B. Leibe, K. Schindler, and L. V. Gool. Moving obstacle detection in highly dynamic scenes. In *Proc. of ICRA*, pages 4451–4458, Kobe, 2009.
4. A. Ess, T. Mueller, H. Grabner, and L. J. Van Gool. Segmentation-based urban traffic scene understanding. In *Proc. of BMVC*, pages 84.1–84.11, London, 2009.
5. D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73(1):41–59, June 2007.
6. A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? In *Proc. of CVPR*, pages 3354–3361, Providence, USA, June 2012.
7. B. Julesz. Textons, the elements of texture perception, and their interaction. *Nature*, 290, 1981.
8. M. Marszałek and C. Schmid. Semantic hierarchies for visual object recognition. In *Proc. of CVPR*, pages 1–7, Minneapolis, 2007.
9. G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.
10. C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *Proc. of ECCV (4)*, pages 733–747, Marseille, 2008.
11. 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 MVA*, Kyoto, 2013.
12. B. Yao, X. Yang, and S-C. Zhu. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In Yuille et al., editor, *Energy Maximization Methods in Computer Vision and Pattern Recognition*, pages 169–183. Springer, August 2007.
13. S-C. Zhu and D. Mumford. A stochastic grammar of images. *Found. Trends. Comput. Graph. Vis.*, 2(4):259–362, 2006.