EVIDENTIAL COMBINATION OF ONTOLOGICAL AND STATISTICAL INFORMATION FOR ACTIVE SCENE CLASSIFICATION

Thomas Reineking, Niclas Schult
Cognitive Neuroinformatics, University of Bremen, Enrique-Schmidt-Straße 5, 28359 Bremen, Germany

Joana Hois
Research Center on Spatial Cognition SFB/TR 8, University of Bremen
Enrique-Schmidt-Straße 5, 28359 Bremen, Germany

Keywords: Ontology, Statistics, Dempster-Shafer theory, Scene classification, Information gain, Knowledge representation, Domain analysis and modeling.

Abstract: We introduce an information-driven scene classification system that combines different types of knowledge derived from a domain ontology and a statistical model in order to analyze scenes based on recognized objects. The domain ontology structures and formalizes which kind of scene classes exist and which object classes occur in them. Based on this structure, an empirical analysis of annotations from the LabelMe image database results in a statistical domain description. Both forms of knowledge are utilized for determining which object class detector to apply to the current scene according to the principle of maximum information gain. All evidence is combined in a belief-based framework that explicitly takes into account the uncertainty inherent to the statistical model and the object detection process as well as the ignorance associated with the coarse granularity of ontological constraints. Finally, we present preliminary classification performance results for scenes from the LabelMe database.

1 INTRODUCTION

One of the reasons why humans are so successful at solving complex problems is that they are capable of utilizing different forms of knowledge. Here, we focus on two distinct aspects of an agent's knowledge representations in particular. On the one hand, formal domain ontologies provide background knowledge about the world by expressing necessary and general logical relations between entities. These relations reflect the definitions or inherent determinations of the entities involved, or they reflect general rules or laws effective in the domain under consideration. Degrees of belief, on the other hand, account for the fact that perception of the world is intrinsically uncertain. They are usually obtained from empirical data and thus belong to the realm of statistics. These two forms of knowledge are complementary in nature since ontological models generally abstract from the uncertainty associated with perception while statistical models do not explicitly represent the semantics of variables in addition to being restricted to encoding low-order relations due to the resulting combinatorial complexity. We argue that both forms should be used when reasoning about problems that involve uncertainty and that can not be accurately modeled by statistics alone.

A good example of a difficult reasoning problem where both ontologies and statistics can offer valuable information is that of vision-based scene classification. Distinguishing between semantic classes of scenes is essential for an autonomous agent in order to interact with its environment in a meaningful way. While coarse perceptual classes (e.g., coast, forest) can be recognized based on low-level features (Oliva and Torralba, 2001), classifying scenes which differ mainly with respect to activities an agent could perform in them (e.g., mall–shopping, kitchen–cooking) requires an object-centric analysis. The problem therefore consists of detecting objects and combining this information with knowledge about the relations of object and scene classes. These relations can either be of statistical nature (e.g., 70% of all street scenes contain cars) or they can be categorical (e.g., all bedrooms contain a bed). The former can be naturally represented by object-scene co-occurrence probabil-

Reineking T., Schult N. and Hois J. (2009).
EVIDENTIAL COMBINATION OF ONTOLOGICAL AND STATISTICAL INFORMATION FOR ACTIVE SCENE CLASSIFICATION.
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 72-79
DOI: 10.5220/0002304300720079
Copyright © SciTePress

ities whereas the latter can be expressed by domain ontologies.

The problem of assigning a semantic class label to a scene on the basis of sensory information has been discussed in several works recently. In (Schill et al., 2009), a domain ontology is used for inferring room concepts from sensorimotor features associated with objects. Objects are analyzed via saccadic eye movements which are generated by an information maximization process. In contrast to our approach, the system does not distinguish between categorical and statistical knowledge, but rather uses expert knowledge for expressing degrees of belief about relations of room concepts and objects. In (Martínez Mozos et al., 2007), semantic scene classification is used for enriching the spatial representation of a mobile robot. The focus of their paper is to apply boosting to the classification of laser range scans as well as Haar-like features extracted from camera images. However, the system exclusively uses statistics for modeling the relation of features and scene classes without utilizing explicit knowledge about objects.

The idea of utilizing co-occurrence statistics for modeling context has been proposed in (Kollar and Roy, 2009). The authors obtained their model via dense sampling over annotations in the Flickr image database. Using this information, a robot is then able to predict the locations of objects based on previously observed objects. (Maillot and Thonnat, 2008) introduce an object recognition system for classifying pollen grains. The classification is based on a training set of sample images supported by a domain ontology. Based on this training set, classification-relevant dependencies between pollen grain classes and their features are extracted, and this information is added to the ontological information. Although their ontology provides a vocabulary for the domain, it does not itself, i.e., without the training set, indicate such classification-relevant dependencies.

In this paper, we propose an information-driven architecture that combines a domain ontology with a statistical model which we apply to the problem of scene classification based on observed objects. An overview of the system architecture is given in the next section. The knowledge representation consisting of the domain ontology and the statistical model is described in section 3. Section 4 explains the belief update and the system's information gain maximization strategy. Object detection is discussed in section 5 along with preliminary results for classifying scenes from the LabelMe database. The paper concludes with a discussion of the proposed architecture and future work is pointed out.

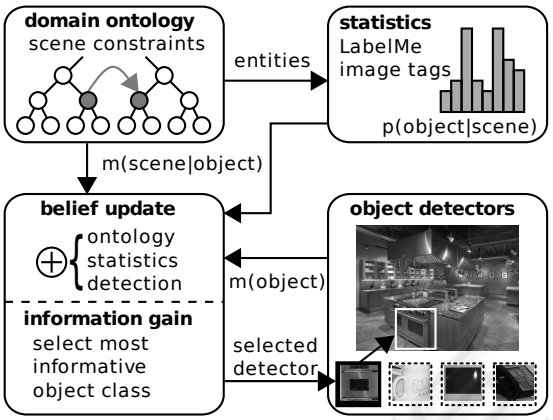

Figure 1: Sketch of the system architecture showing the four main components of the scene classification system.

2 SYSTEM ARCHITECTURE

The proposed scene classification system is composed of four main components: a domain ontology for scenes, a statistical model, a reasoning module for updating the scene belief via information gain maximization, and an image processing module consisting of class-specific object detectors (see figure 1). The domain ontology defines the vocabulary of object and scene classes as well as occurrence constraints between the two (e.g., *kitchens contain cooking facilities*). The statistical model, on the other hand, provides co-occurrence probabilities of object and scene classes which are estimated from the large annotated image database LabelMe (Russell et al., 2008).

Based on the current scene belief, the system computes the most disambiguating object class according to the principle of maximum information gain. The gain is defined as the expected decrease in uncertainty between the current and the predicted scene belief. It depends on the discriminatory power of the object class (e.g., 'stove' when having to distinguish between kitchens from offices) as well as the system's confidence in being able to recognize objects of that class (if stoves are hard to recognize, their discriminatory power is of little use). Once the most informative object class (i.e., the one minimizing the expected resulting uncertainty) is determined, the system invokes a vision-based search for this class in order to support or reject the current scene hypotheses.

The search is modeled by an object detection mechanism which determines whether an instance of this object class is present in the image corresponding to the current scene. Since the main focus of our work is the interplay between ontology-based and statistic-based knowledge representations, the system's vision

part is currently modeled as a simplified object detection mechanism consisting of a set of class-specific binary Support Vector Machines (SVMs). Each SVM is evaluated on a separate data set in order to obtain a probabilistic model, which reflects the average error rate during classification. This model is later used to quantify the level of uncertainty associated with the detection of a specific object class, which allows the system to judge the reliability of detection results.

In order to combine the uncertainty resulting from the object detection and from the statistical model with the set-based propositions obtained from the ontology, we use Dempster-Shafer theory. Representing the scene belief within this framework allows the assignment of belief masses to sets of propositions. The framework is thus suited for expressing the non-specificity associated with knowledge obtained from the ontology.

Overall, the architecture analyzes a scene in a cycle of bottom-up object detection followed by a belief update based on statistical and logical inferences, and top-down information gain maximization for selecting the next object class for detection. In order to classify a scene, the system first computes the expected scene information gain associated with the action of looking for a specific object class. After selecting the most informative object class, the vision module invokes the corresponding object detector and updates the current scene belief depending on the detection result, the constraints defined by the ontology for this object class and the co-occurrence probabilities of the statistical model.

3 KNOWLEDGE REPRESENTATION

The underlying knowledge representation of the system comprises (i) statistical and (ii) ontological information on the domain. Statistical information results from empirical data, in our case from the LabelMe database. It is obtained by averaging over occurrences of objects (scene entities) in certain scenes. Statistics are generally subject to noise and they depend on the availability of sufficient sample data. Furthermore, they are restricted to representing low-order relations since the complexity of representation and data acquisition increases exponentially with the number of variables.

While statistical information reflects the probability of relations between objects and scenes, ontological information reflects logically strict constraints and relations between them. A domain ontology for scenes primarily has to formalize the kind of scenes and objects that exist as well as their relationships. In contrast to statistics, it does not rely on a sample set of data, but on expert knowledge and general common-sense knowledge of the domain. It may especially be formalized on the basis of foundational developments in ontological engineering, as outlined below. In essence, (i) statistics contribute a probabilistic correspondence between objects and scenes obtained from a (finite) data set, while (ii) the domain ontology contributes a formalization of entities and relations that exist in the domain, which also provides the vocabulary for the statistics.

We introduce the domain ontology for scene recognition in the next section, and the statistical analysis of the domain subsequently.

3.1 Ontology

The domain ontology for visual scene recognition provides the system with information on scenes, objects in scenes, and their relations. Although ontologies can be defined in any logic, we focus here on ontologies as theories formulated in description logic (DL) (Baader et al., 2003). DL is supported by the web ontology language OWL DL 2[1]. Even though ontologies may be formulated in more or less expressive logics, DL ontologies have the following benefits: they are widely used and a common standard for ontology specifications, they provide constructions that are general enough for specifying complex ontologies, and they provide a balance between expressive power and computational complexity in terms of reasoning practicability (Horrocks et al., 2006). Our scene recognition system uses a domain ontology for specific scenes (SceneOntology[2]) , which is formulated in OWL DL 2. Furthermore, the system uses Pellet (Sirin et al., 2007) for ontological reasoning.

In our scenario, the ontology provides background knowledge about the domain to support scene classification. Its structure adopts methods from formal ontology developments. In particular, it is a logical refinement of the foundational ontology DOLCE (Masolo et al., 2003). For practical reasons, we use the OWL version DOLCE-Lite. The domain ontology for scenes conservatively extends DOLCE-Lite, i.e., all assertions made in the language of DOLCE-Lite that follow from the scene ontology already follow from DOLCE-Lite. Essentially, this means that the scene ontology completely and independently specifies its vocabulary, i.e., it can be seen as an ontological module (Konev et al., 2009).

[1] http://www.w3.org/TR/owl2-semantics/

[2] http://www.ontospace.uni-bremen.de/ontology/domain/SceneOntology.owl

Reusing DOLCE ensures that the domain ontology is based on a well-developed foundational ontology. Certain types of classes and relations can be reused and their axiomatizations can be inherited. Particular ontological classes specified in the scene ontology that are involved in the scene recognition process are SceneClass, SceneEntity, and Scene. Their refinements of DOLCE-Lite are defined as follows:

$$\text{Scene} \sqcap \text{SceneEntity} \sqsubseteq \text{dolce:physical-object}$$

$$\text{SceneClass} \sqsubseteq \text{dolce:non-physical-object}$$

The class dolce:physical-object is a subcategory of physical-endurant in DOLCE, which represents those entities that have a physical extent and which are wholly present in time (Masolo et al., 2003). Scene and SceneEntity are subclasses of this dolce:physical-object. The class SceneEntity represents physical entities that occur in spatial scenes and that correspond to segmented objects in scene images. These entities are determined by their intrinsic (inherent) properties. Examples are Furniture, Refrigerator, Chair, Appliance, Tree, and Plant, namely entities that are contained in indoor and outdoor scenes. These scenes are represented by the class Scene. The relation contain specifies precisely the relation that certain instances of SceneEntity are contained in a certain Scene. The class Scene can be informally described as a collection of contained SceneEntities. In practice, it is related to a specific view on the environment that is perceived by the visual system.

In contrast to Scene and SceneEntity, SceneClass formalizes the type of the scene, i.e., it indicates the category of collections (Scene) of entities (SceneEntity). Examples are Kitchen, Office, ParkingLot, and MountainScenery. SceneClass is a subclass of dolce:non-physical-object, which is an endurant that has no mass. It constantly depends on a physical endurant, in this case, it depends on the collection of entities that are physically located at a certain Scene or that are 'commonly' perceivable in this scene. In the scene ontology, SceneClass therefore defines the DOLCE-relation generically-dependent-on to one Scene, which is defined by a conjunction of disjunctions of restrictions on those SceneEntity that may occur in the scene. Specific subclasses of SceneClass and SceneEntity are taken from the database of LabelMe.

Hence, for a specific SceneClass s_i, there is a number of subclasses t_k of SceneEntity that necessarily have to occur at the Scene r_j of the SceneClass s_i. Specific subclasses of SceneEntity are taken into account by the following conjunction, with K_{s_i} indicating the index set of SceneEntity t_k, which constrain s_i as defined by (2), and N indicating the total amount of subclasses of SceneEntity:

$$\xi_{s_i} = \bigwedge_{k \in K_{s_i}} t_k \quad \text{with } K_{s_i} \subseteq \{1,\ldots,N\}. \tag{1}$$

Each SceneEntity t_k is taken from constraints of the SceneClass s_i as follows:

$$\textit{generically-dependent-on}(s_i, (\textit{contain}(r_j, t_k))). \tag{2}$$

The distinction being drawn between SceneEntity and SceneClass is based on an agent-centered perspective on the domain of possible scenes from the LabelMe database. While instances of SceneEntity (e.g., chair, refrigerator, or sink) are on the same level of granularity, instances of SceneClass (e.g., kitchen, street corner, or warehouse) are on a broader level of granularity and they particularly depend on a collection of the former. The levels of granularity depend on the agent, i.e., in our case the vision system, that perceives its environment, i.e., a specific scene. The ontological representation of entities differing in granularity aspects is grounded in this agent-based (embodied) vision, as outlined, for instance, in (Vernon, 2008). Note, however, that although an 'open world' assumption underlies the ontological representation, the ontology takes into account precisely the objects that are classifiable by the system. Currently, the scene ontology distinguishes between 7 different scene classes and 24 scene objects.

Constraints of specific SceneClasses, such as Kitchen, are given by the scene ontology on the basis of SceneEntities. A sample of such constraints is illustrated in the following example (formulated in Manchester Syntax (Horridge and Patel-Schneider, 2008)):

Class: Kitchen
SubClassOf: SceneClass,
generically-dependent-on only (Scene and (contain some Oven)),
generically-dependent-on only (Scene and (contain some Sink)), ...

Queries over such constraints using the reasoner Pellet support the scene recognition process by providing general background knowledge about the domain. Given a request for a specific scene class, the reasoner returns the constraints given by ξ_{s_i} as formulated in (1).

3.2 Statistics

The statistical model represents the relation of a scene class s_i and an object class t_k by their co-occurrence probability $p(t_k|s_i)$. These conditional probabilities are estimated by computing relative scene-object tag

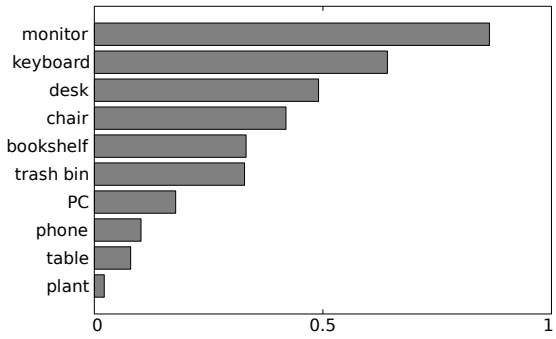

Figure 2: Conditional probabilities of object occurrences for a given scene. In this example probabilities of object classes in scene class Office are shown.

frequencies from the LabelMe database. We restrict our model to these second-order relations since higher-order models exhibit combinatorial complexity, even though this implies ignoring possible statistical dependencies between object classes. After excluding scenes not containing any known object classes, 9601 scenes along with 28701 known object classes remain for the statistical analysis. An example of the co-occurrence distribution for the scene class Office is shown in figure 2.

4 REASONING

In order to compute the belief about the current scene class, knowledge about scene-object relations from the statistical model and the domain ontology are combined with the object detection beliefs from the vision module. While the statistical model and object detection can be accurately described by Bayesian probabilities, the constraints defined by the ontology result in propositions about sets of scene classes without any form of belief measure assigned to the single elements within these sets. We therefore use Dempster-Shafer theory (Shafer, 1976) throughout the architecture since it generalizes the Bayesian notion of belief to set-based propositions, thus making ignorance explicit and avoiding unjustified equiprobability assumptions. In particular, we use a variant of Dempster-Shafer theory known as the transferable belief model (Smets and Kennes, 1994), which is based on an open world assumption accounting for the fact that not all scenes can be mapped to the modeled classes.

4.1 Belief Update

Let Θ be a finite and exhaustive set of mutually exclusive hypotheses. The belief induced by a piece of evidence can be expressed as an (unnormalized) mass function $m : 2^{\Theta} \rightarrow [0,1]$ that assigns belief values to arbitrary subsets $A \subseteq \Theta$ (including $\emptyset$ (Smets, 1992)) such that $\sum_A m(A) = 1$. Combining two pieces of evidence which induce a belief m_1 and m_2 respectively is done by applying the conjunctive rule of combination denoted by $\textcircled{\cap}$:

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B)\, m_2(C). \qquad (3)$$

If there is a prior belief m and the truth of a hypothesis A can be established with certainty, m can be conditioned on A yielding the conditional belief $m(\cdot \,|\, A)$.

Let $\widehat{T}$ be the set of indices k corresponding to all object classes t_k for which a visual detection was performed up to this point. The scene belief $m_{\widehat{T}}$ is then computed by conjunctively combining the detection beliefs $m(t_k)$ of each object class t_k with the conditional scene class belief derived from the statistical model and the domain ontology, which can be written as (Dubois and Prade, 1986):

$$m_{\widehat{T}} = \bigcirc\!\!\!\!\!\!\cap_{k \in \widehat{T}} \sum_{t_k} m(\cdot \,|\, \{t_k\})\, m(t_k). \qquad (4)$$

This is the basic update equation of the system. It is important to note that it can be computed recursively, i.e., the scene belief is updated based on the prior belief and the information from a new detection.

The change of the scene belief depends on the belief about the object class' presence $m(t_k)$ (see section 5) as well as on the conditional model $m(S \,|\, \{t_k\})$ that assigns mass values to sets of scene classes S given the (non-)presence of object class t_k. As mentioned above, the latter belief reflects two sources of knowledge and can therefore be expressed as a conjunctive combination of a statistical part m^{sta} and an ontological part m^{ont}:

$$m(\cdot \,|\, \{t_k\}) = m^{sta}(\cdot \,|\, \{t_k\}) \textcircled{\cap} m^{ont}(\cdot \,|\, \{t_k\}). \qquad (5)$$

The statistical model m^{sta} is obtained by applying Bayes' rule (without normalization) to the conditional probability $p(t_k|s_i)$, which is generated from training data (see section 3.2):

$$m^{sta}(s_i \,|\, \{t_k\}) = p(t_k|s_i)\, p(s_i), \qquad (6)$$

$$m^{sta}(s_i \,|\, \{\neg t_k\}) = (1 - p(t_k|s_i))\, p(s_i). \qquad (7)$$

The ontological model m^{ont} does not yield any information if t_k is true because the presence of an object in a scene is never impossible. In this case, mass 1 is assigned to the whole hypothesis space Θ_S, expressing a state of total ignorance. However, if a scene class s_i requires an object class t_k according to the domain ontology, the non-presence of t_k implies the rejection of s_i. Therefore, $\neg t_k$ rules out all those scene

classes s_i for which the index set K_{s_i} belonging to the conjunctive constraint ξ_{s_i} defined by (1) contains the object class index k:

$$m^{ont}(\Theta_S \,|\, \{t_k\}) = 1, \tag{8}$$

$$m^{ont}(\Theta_S \setminus \{s_i | k \in K_{s_i}\} \,|\, \{\neg t_k\}) = 1. \tag{9}$$

4.2 Information Gain

Aside from passively updating the scene belief in a bottom-up fashion, the system also utilizes a top-down mechanism for selecting object detectors in order to actively reduce uncertainty. If there is little doubt about a scene's class, then it would be wasteful to apply all possible detectors knowing that it would be unlikely to change the scene belief. For this, we need a measure of uncertainty that is applicable to mass functions. We use the local conflict measure $H(m)$ (Pal et al., 1993) here since it is a measure of total uncertainty that generalizes the concept of information entropy:

$$H(m) = \sum_A m(A) \log_2 \frac{|A|}{m(A)}. \tag{10}$$

Selecting the most informative object class t^* for the subsequent detection is done by maximizing the expected information gain. The gain associated with an object class t_k from the set of previously ignored classes $\widehat{T}^C$ is defined as the difference between the current uncertainty $H(m_{\widehat{T}})$ and the expected uncertainty $E(H(m_{\widehat{T}\cup\{k\}}))$ after applying the detector for t_k:

$$t^* = \max_{k \in \widehat{T}^C} \left[H(m_{\widehat{T}}) - E\left(H(m_{\widehat{T}\cup\{k\}}) \right) \right]. \tag{11}$$

The extent of the decrease according to (4) depends on the a priori presence belief for t_k on the one hand and on the discriminatory power of t_k with respect to the current scene belief $m_{\widehat{T}}$ expressed by the object-scene model $m(\cdot \,|\, \{t_k\})$ on the other hand. Since t_k is not directly observable, the integration for computing the expected value must be performed over the belief space $m(t_k)$:

$$E\left(H(m_{\widehat{T}\cup\{k\}}) \right) = \sum_{m(t_k)} p(m(t_k)) H(m_{\widehat{T}\cup\{k\}}). \tag{12}$$

Most importantly, the belief distribution over t_k is characterized by a single value (t_k is a binary variable) so that the integration can be approximated by summing over a normalized histogram of belief values $m(t_k)$. The histogram provides the probability for the t_k detector returning a belief value in a given interval. In the current implementation, it contains 100 bins and is computed during the classifier performance evaluation described in the following section.

5 SCENE CLASSIFICATION

In this section, we show how the system manages to classify scenes from the LabelMe image database and present preliminary results. Since we are not interested in solving the difficult full object detection problem (Schneiderman and Kanade, 2004) here, we use the object pre-segmentation provided by LabelMe. First, we describe how object detection is performed on these segmented images. Subsequently, we explain the complete scene classification process using an example and present results for a randomly selected set of scenes from the LabelMe database.

5.1 Object Detection

To each object class corresponds one specialized detector that is independently applied to the current scene. Detecting the presence of an object class t_k is realized by checking whether at least one segment contains an object of type t_k. In this way, the object detection problem can be reduced to a set of binary classification problems. Here, we deliberately ignore the fact that using a fixed set of segments contradicts the independent classification of segments (each segment belongs to exactly one class) since using the pre-segmentation is just a means of simplification.

Each segment is scaled to a 128×128 gray-value image before being processed by a class-specific binary SVM which was trained on several thousand positive and negative instances of t_k. Since this classifier can exhibit significant error rates due to the oftentimes suboptimal sample quality, the disjunctive combination of classification results is done in a probabilistic fashion. We define the belief about the presence of object class t_k as the probability of at least one segment l being of type t_k given the classifier response $c_{k;l}$ for each segment:

$$\begin{aligned} m(t_k) &= p(\bigvee_l t_{k;l} | \{c_{k;l}\}) \\ &= 1 - \frac{\prod_l p(c_{k;l} | \neg t_{i;k}) p(\neg t_{k;l})}{\prod_l \sum_{t_{k;l}} p(c_{k;l} | t_{k;l}) p(t_{k;l})}. \end{aligned} \tag{13}$$

The above equation can be derived by applying Bayes' rule along with an independence assumption between segments to the probability of t_k not being present. All involved probabilities are learned from a representative set of sample scenes where $p(c_{k;l} | t_{k;l})$ expresses the classifier's true/false positive/negative rates which are a measure of its reliability.

5.2 Results

The test set on which we evaluated the system's performance consisted of 7 scene classes with 12 ran-

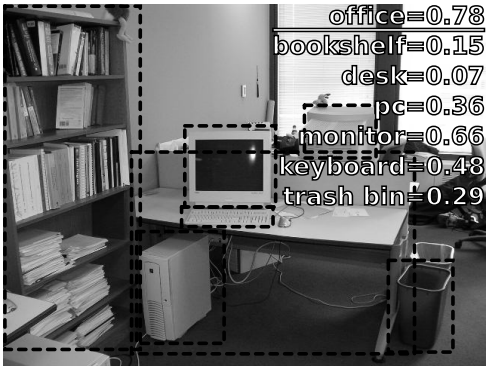

Figure 3: Example scene with marked segments and detection beliefs for present object classes. Even though the detection of single object classes is ambiguous, the system manages to correctly recognize an office scene based on fusing all available information.

domly[3] selected scenes from the LabelMe database each. On average, the classifiers for the 24 object classes exhibited a sensitivity of $p(c_{k;l}|t_{k;l}) = 0.71$ and a specificity of $p(\neg c_{k;l}|\neg t_{k;l}) = 0.79$. A scene counted as correctly classified if the highest belief was assigned to its actual class. Overall, the system correctly classified 61% of all 84 scenes. The recognition rate strongly varied across some scene classes, e.g., offices: 70% and living rooms: 43%. This can be explained by the fact that certain scene classes contained less typical objects as well as by the fact that the classification of certain object classes is more difficult than for others. Considering the simplicity of the detection scheme and the poor quality of many of the displayed objects (caused by occlusion, poor segmentation, low resolution, unusual perspectives, etc.) the overall recognition rate is surprisingly high even though more comprehensive tests are necessary for validation. The information gain strategy for selecting object detectors based on their expected disambiguating influence reduced the average number of detections per scene by 29% compared random selection for reaching the same level of uncertainty.

Figure 3 shows an office scene example that was analyzed by the system. Starting with a vacuous scene belief, the system first computes the expected information gain over all object classes and selects the table detector since tables regularly occur in different scene classes. The table classifier is then applied to all 7 segments, but with no positive responses, the resulting detection probability is only 0.06 (not 0 due to the possibility of false negatives). As a result, the belief for scene classes that often contain tables decreases while the belief for the remaining classes increases. Once office-specific objects are detected (see figure 2), the strong statistical evidence along with the ontological constraints (offices generally contain desks and electronic equipment) cause the office class belief to reach a value of 0.78. At this point, the system has analyzed 13 object classes and the uncertainty level has dropped below the termination threshold of $H(m_{\widehat{T}}) \leq 1.5$.

[3]The only selection criterion was that each selected scene contained at least 3 known object classes.

6 DISCUSSION

In this paper, we showed how scene classification can be performed by combining the complementary information provided by a domain ontology and a statistical model. The consistent propagation of uncertainty from the recognition of local features up to the scene level enables the system to draw inferences even if the input data is very noisy. We chose Dempster-Shafer theory as a framework for representing uncertainty since it can be used for expressing both set-based implications obtained from the domain ontology as well as probabilistic relations. By using unnormalized belief functions, we make an explicit open world assumption which accounts for the fact that not every scene can be accurately mapped to the set of modeled scene classes.

Besides the bottom-up updating of the scene belief, we presented a top-down reasoning strategy for targeted feature selection based on information gain maximization. This selective processing leads to a more efficient analysis of the scene by filtering out uninformative features, which appears to be an important principle of how humans analyze scenes with saccadic eye movements (Henderson and Hollingworth, 1999). Information gain maximization can be interpreted as an attention mechanism, and it reflects findings in neuro-psychology showing that object recognition in humans is not simply a feature-driven process, but rather an interplay of bottom-up processing and top-down reasoning where recognition is influenced by the context (Schill et al., 2001).

We argue that domain ontologies and statistics can complement each other since ontologies provide a more general description of the world whereas statistics can offer more detailed but noisy information that depends on the availability of suited training data. The LabelMe database is a good example for the problem of partially insufficient data, because for many scene classes it contains only a small number of samples. A more exhaustive domain model in the form of an ontology thus enables the system to draw inferences about classes for which no statistics might be available at all.

A general problem in the context of scene classification is the processing of images that only show parts of a larger scene. Essentially, this means that it is not possible to reason on the basis of an object class' absence. While the explicit representation of uncertainty reduces the severity of the problem in practice, there is always a chance of miss-classifying a scene due to critical objects being out of view. A possible solution to this problem could be to have the system analyze images taken at different view points in the scene.

In the future, we plan to integrate the presented scene classification system into a mobile agent (Zetzsche et al., 2008). Not only does this provide the system with a strong prior due to the agent's past observations, the mobility would also ease the problem of only sensing parts of a scene. In particular, this will require the detection of objects to be performed without any pre-segmentation, which is why we are currently working on providing the system with a more sophisticated vision module. This will also allow us to produce more conclusive experimental results on other data sets. Finally, we think it would be interesting to see whether the generic approach of reasoning based on ontologies and statistics in a belief-based framework could be applied to other domains beyond scene classification.

ACKNOWLEDGEMENTS

This work was supported by DFG, SFB/TR8 Spatial Cognition, projects A5-[ActionSpace] and I1-[OntoSpace].

REFERENCES

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P. (2003). *The Description Logic Handbook*. Cambridge University Press.

Dubois, D. and Prade, H. (1986). On the unicity of Dempster's rule of combination. *International Journal of Intelligent Systems*, 1(2):133–142.

Henderson, J. and Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, 50(1):243–271.

Horridge, M. and Patel-Schneider, P. F. (2008). Manchester OWL syntax for OWL 1.1. OWL: Experiences and Directions (OWLED 08 DC), Gaithersberg, Maryland.

Horrocks, I., Kutz, O., and Sattler, U. (2006). The Even More Irresistible SROIQ. In *Knowledge Representation and Reasoning (KR)*. AAAI Press.

Kollar, T. and Roy, N. (2009). Utilizing object-object and object-scene context when planning to find things. In *International Conference on Robotics and Automation (ICRA)*.

Konev, B., Lutz, C., Walther, D., and Wolter, F. (2009). Formal properties of modularisation. In Stuckenschmidt, H., Parent, C., and Spaccapietra, S., editors, *Modular Ontologies*. Springer.

Maillot, N. E. and Thonnat, M. (2008). Ontology based complex object recognition. *Image and Vision Computing*, 26(1):102–113.

Martínez Mozos, Ó., Triebel, R., Jensfelt, P., Rottmann, A., and Burgard, W. (2007). Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems*, 55(5):391–402.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2003). Ontologies library. WonderWeb Deliverable D18, ISTC-CNR.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Pal, N., Bezdek, J., and Hemasinha, R. (1993). Uncertainty measures for evidential reasoning II: A new measure of total uncertainty. *International Journal of Approximate Reasoning*, 8(1):1–16.

Russell, B., Torralba, A., Murphy, K., and Freeman, W. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1–3):157–173.

Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., and Zetzsche, C. (2001). Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. *Journal of Electronic Imaging*, 10(1):152–160.

Schill, K., Zetzsche, C., and Hois, J. (2009). A belief-based architecture for scene analysis: From sensorimotor features to knowledge and ontology. *Fuzzy Sets and Systems*, 160(10):1507–1516.

Schneiderman, H. and Kanade, T. (2004). Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51–53.

Smets, P. (1992). The nature of the unnormalized beliefs encountered in the transferable belief model. In *Uncertainty in Artificial Intelligence*, pages 292–297.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial intelligence*, 66(2):191–234.

Vernon, D. (2008). Cognitive vision: The case for embodied perception. *Image and Vision Computing*, 26(1):127–140.

Zetzsche, C., Wolter, J., and Schill, K. (2008). Sensorimotor representation and knowledge-based reasoning for spatial exploration and localisation. *Cognitive Processing*, 9:283–297.