Context-aware Classification for Incremental Scene Interpretation

Arne Kreutzmann
Cognitive Systems Laboratory
Department of Informatics
Vogt-Koelln-Str. 30
Germany, 22527 Hamburg
Kreutzmann@informatik.uni-hamburg.de

Kasim Terzić
Cognitive Systems Laboratory
Department of Informatics
Vogt-Koelln-Str. 30
Germany, 22527 Hamburg
Terzic@informatik.uni-hamburg.de

Bernd Neumann
Cognitive Systems Laboratory
Department of Informatics
Vogt-Koelln-Str. 30
Germany, 22527 Hamburg
Neumann@informatik.uni-hamburg.de

ABSTRACT

Appearance-based classification is a difficult task in many domains due to ambiguous evidence. Knowledge about the relationships between objects in the scene can help resolve this problem. In this paper, we present a new probabilistic classification framework based on the cooperation of decision trees and Bayesian Compositional Hierarchies, and show that introducing contextual knowledge in the form of dynamic priors significantly improves classification performance in the façade domain.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Algorithms, Theory

Keywords

context-driven event interpretation, guided vision based on high-level reasoning

1. INTRODUCTION

Rapid, comprehensive and accurate recognition and understanding of complex visual scenes, while seemingly effortless for humans [4], remains difficult for computers. One of the reasons appears to be that object recognition in Computer Vision is predominantly conceived and performed for objects in isolation, neglecting contextual information and reasoning mechanisms. But the appearance of isolated objects is often too noisy and ambiguous to permit reliable classification. Scene interpretation methods may help here, because they aim at inferring assertions about a scene from many objects interpreted together. Consider the example of an image of a person in front of an upper-floor window (see Figure 1). From this evidence a human would expect the presence of a balcony and could predict the position of a railing. Rather than pure deductions, scene interpretation consists of educated guesses, or, as Max Clowes (1971) has put it, scene interpretation is "controlled hallucination". On the other hand, if it is known that there is no balcony, then we can reason that the person must be a detection error. This illustrates a possible interaction of high-level context and low-level image analysis results which is the topic of this contribution.

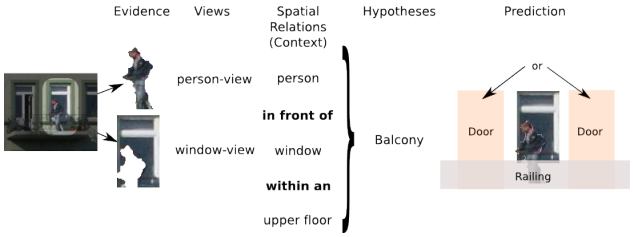

Figure 1: A person and a window in an upper floor create a strong context for a balcony, which in turn allows the prediction of a possible railing as well as a door.

In this paper, we will present a method for improving object classification using the evolving context of stepwise scene interpretation within the SCENIC framework [6]. Here, scene interpretation is formulated as a stepwise search for a configuration satisfying the evidence and a scene model. Hence we see the context-aware classification also as a stepwise process: The context which we will exploit during the classification process is created step by step from the evolving scene interpretation based on the objects classified so far. For example, by classifying evidence as a `Door` and assigning it a role in the scene model as part of a specific `Balcony`, a context is established for yet unclassified evidence. A conceptual representation of contextual relations is provided by aggregates which are the building blocks of a compositional hierarchy in the high-level knowledge base. The notion of aggregate is used in the sense that it describes an attributed object with has-part relationships to other objects. A `Balcony`, for example, may have attributes position and size and be composed of a `Window`, a `Door` and a `Railing` which are

UCVP '09, November 5, 2009 Boston, MA

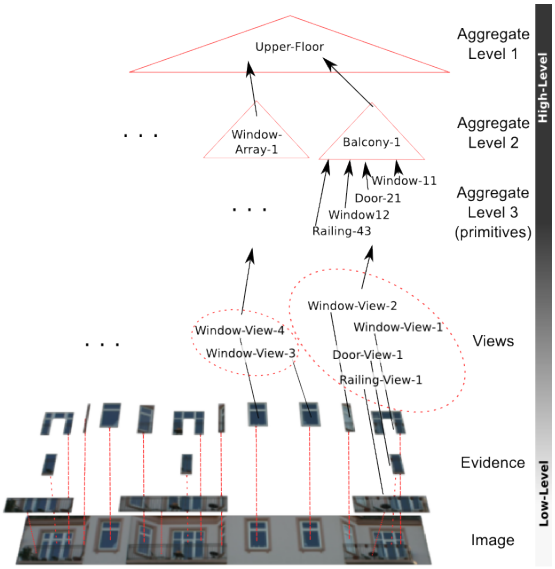

Figure 2: Here the layers/levels are shown. The high-level processes perform scene interpretation and exploit context. Low-level processes detect evidence and create views from this evidence with the contextual help of the high-level.

aggregates themselves. Aggregates without parts will also be called primitives.

In Figure 2, the different processing layers used in this paper are shown and ordered from low-level processes — which extract evidence from an image and associate it with possible object views — to high-level processes — which assign these views to objects, integrate these objects into aggregates using role assignment, and create predictions for other objects or aggregates. The concept of views allows for distinguishing between the expected image analysis result for an object present in the scene (a view) and the possibly corrupted evidence which has been actually generated. Since the experiments described in Section 4 are based on annotated images and to simplify the presentation, we will omit the difference between views and evidence throughout this paper. In the following, we will concentrate on the interaction between the high level, which creates the context and the low level, which provides the visual clues.

In Section 2, we will discuss related and previous work and highlight some shortcomings which we will try to address. Our novel approach is presented in Section 3 and the main parts are explained in Sections 3.1 and 3.2. This approach is then tested in Section 4 with images of floors of a building as examples, and a conclusion is drawn in Section 5.

2. RELATED WORK

Scene interpretation as understood in this paper is based on the conceptual framework presented by the work of Neumann [19, 13, 6, 7], but we focus on a new approach which extends this framework with a high-level probabilistic model.

Early work on integrating scene interpretation with a probabilistic model was done by Rimey [17]. In his work, he used Bayesian Networks (BNs), first introduced by Pearl [14], to model the parts of an aggregate *as caused* by the aggregate and hence independent from each other given knowledge about the parent aggregate. Koller and Pfeffer [8] extend BNs in an object-oriented manner for the representation of structured objects. Gyftodimos and Flach [5] introduce hierarchical BNs to allow multiple levels of granularity. While these contributions improve the expressive power of BNs, they do not specifically support compositional hierarchies of aggregates as required for context modeling in scene interpretation based on on ontologies understandable by humans. In this paper, Bayesian Compositional Hierarchies (BCHs) will be introduced for this purpose in Section 3.2.1.

Context through visual clues has been in the focus of recent work [15, 22]. Nearby visual clues are used to reduce the number of false positives. However other high-level knowledge can not be integrated directly, and visual context will usually be too weak to provide conclusions like the one shown in Figure 1.

Li et al. [10] and Murphy et al. [11] presented a way to incorporate a global context: If the type of picture is known (e.g. the type of sport shown in the picture), then the priors for the classes can be changed correspondingly to increase the chance of correct classification. However, details of the contextual relationships between the objects within a scene are not exploited.

3. CONTEXT-AWARE CLASSIFICATION

Evidential information can often be decomposed into a context-independent and a context-dependent part. In Computer Vision the former will generally be the local visual appearance and the latter the influence of its surrounding area or the attributes which depend on the relationships to other objects in the scene. The underlying assumption is that the local appearance is independent of the context of the object. The context on the other hand influences the expectation for the existence of an object at a certain position but does not influence its appearance. Henceforth every evidence

$$E = \begin{pmatrix} E_\mathrm{p} \\ E_\mathrm{a} \end{pmatrix}$$

will be partitioned into E_a, which denotes the local appearance features (in our experiment these are: aspect ratio, rectangularity, compactness and size), and E_p, the position of the evidence as the contextual part of the evidence. Due to this assumption, the following formula holds:

$$P(C, E) = P(E_\mathrm{a}|C) \cdot P(C, E_\mathrm{p}) \tag{1}$$

where C denotes the class. In order to classify evidence E given other already classified evidences $e_1, \ldots, e_n$ and their respective classes $c_1, \ldots, c_n$, the following formula needs to be maximized:

$$P(C|E, e_1, \ldots, e_n, c_1, \ldots, c_n) = \frac{\overbrace{P(E_\mathrm{a}|C)}^{\text{local features}} \overbrace{P(C|E_\mathrm{p}, e_1, \ldots, e_n, c_1, \ldots, c_n)}^{\text{context}}}{\underbrace{P(E_\mathrm{a})}_{\text{normalize}}}. \tag{2}$$

We will use decision trees (DT) for the *local feature* part, as described in Section 3.1, and Bayesian Compositional Hi-

erarchies (BCHs) for the *contextual* part, as described in Section 3.2. Evidence is classified in a stepwise process as described in Algorithm 1.

Algorithm 1 The iteration loop for classification and context exploration for a set of evidences $\mathfrak{E}$.

```
while 𝔈 ≠ ∅ do
  for all E ∈ 𝔈 do
    C*_E ← arg max_C P(C|E) {using equation (2)}
  end for
  C*, E* ← arg max_E Confidence(E, C*_E)
  𝔈 ← 𝔈 \ {E*}
  e_{n+1} ← (E*, C*)
  update context {including role assignment}
end while
```

3.1 The Appearance Model via Decision Trees

The probabilistic object appearance model $P(E_a|C)$ in Equation 2 can be obtained from a decision tree. Decision trees are commonly used for all kinds of classification and PDF estimation tasks, as explained in [16, 21] and used recently in [9]. Decision trees generally approximate $P(C|L)$, where L are the leaves representing regions of the feature space. Since $P(C|L)$ is an approximation of $P(C|E_a)$, one can obtain $P(E_a|C) = P(L|C)$ using the Bayes rule and the $P(L)$ and $P(C)$ determined from the training dataset. The underlying assumption is that the appearance doesn't change based on the context.

We learn the decision tree based on an annotated training set which relates each sample, described by an n-dimensional feature vector f, to a class C from our ontology. Starting with a root node containing all samples, the feature space is iteratively subdivided by searching through all possible splits and choosing the one that minimizes the *Gini coefficients* of the new sub nodes:

$$G(t) = \sum_{i \neq j} P(c_i|t)P(c_j|t)$$

where $P(c_i|t)$ is the probability that the node t represents an instance of class c_i. This is repeated until all leaves are pure (each leaf contains samples of only one class).

Since decision trees learned in this way are known to overfit, CART pruning [2] is performed. For each node in the tree, the strength of the link of the node to its leaves is given by

$$g(t) = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}$$

where $R(t)$ is the misclassification rate at node t, $R(T_t)$ is the estimated misclassification rate of a sub-tree T with node t as the root node, and $|\tilde{T}_t|$ is the number of leaves in T_t. The node with the lowest $g(t)$ is made into a leaf, and pruning performed again on the new tree. The result is a succession of trees, starting with the original tree and ending with a tree consisting only of the root node. Each of these trees is evaluated as a classifier on an unseen validation set and the best classifier is picked as the final tree.

The resulting tree has impure leaves, with samples from several classes. The probability $P(C|L)$ of a leaf l can be estimated for each class c as $P(c|l) = \frac{N_c(l)}{N(l)}$, where $N_c(l)$ is the number of samples in l belonging to class c, and $N(l)$ is the number of all samples in l. Bottom-up classification without context thus amounts to choosing the class C for a feature vector f belonging to leaf L, such that $P(C|L)$ is maximum. The context can be integrated by determining $P(L|C)$ and using it in Equation 2.

3.2 Context via Bayesian Compositional Hierarchies

In this section we first describe the high-level probabilistic structure which is used to generate the context-dependent class estimate $P(C|E, e_1, \ldots, e_n, c_1, \ldots, c_n)$ in Equation (2). We then explain the assignment of evidence to a specific random variable of the high-level model if more than one qualify. Finally, we present the learning procedure by which the high-level model is obtained from annotated training data.

3.2.1 Bayesian Compositional Hierarchies

As indicated above, probabilistic context information is provided by the compositional structure of scene models represented in the high-level knowledge base. In this section we shortly describe Bayesian Compositional Hierarchies (BCHs) which have been developed for this purpose [12].

A BCH is essentially a tree-shaped Bayesian Network isomorphic to the compositional hierarchy of a domain, with aggregates as nodes. An aggregate description consists of a vector of random variables comprising all context-relevant properties of its parts (the internal description of the aggregate) and of the aggregate as a whole (its external description). Unlike [17], the joint probability distribution (JPD) of all random variables of an aggregate is unrestricted, so that probabilistic dependencies can be modeled freely, providing a true probabilistic context.

Since aggregate parts may be aggregates by themselves, an aggregate is not only represented as a part of the internal description of the higher-level aggregate but also as the external description of an aggregate one level down the compositional hierarchy. This preserves object-oriented descriptions of aggregates at the expense of overlapping descriptions between parent and child.

The tree structure of the BCH allows probability updates based on message passing as presented in [14]. In addition, a BCH is based on the assumption that an aggregate conditioned on its external description in the parent node is independent of all other random variables in the parent node. This simplifies the propagation mechanism as only changes of subsections of an aggregate description need to be passed on through the tree.

In the case of purely multivariate normal distributions (as used in the experiments described in Section 4), change propagation upwards or downwards in a BCH can be computed efficiently as follows. Let G be a multivariate normal distribution

$$G \sim N(\mu_G, \Sigma_G) \text{ with } \mu_G = \begin{pmatrix} \mu_C \\ \mu_D \end{pmatrix}, \Sigma_G = \begin{pmatrix} \Sigma_C & \Sigma_{CD} \\ \Sigma_{CD}^T & \Sigma_D \end{pmatrix}$$

where D is the subsection of variables whose distribution must be updated. If $(\mu_D, \Sigma_D) \mapsto (\mu'_D, \Sigma'_D)$, then the resulting distribution $N(\mu'_G, \Sigma'_G)$ is calculated as following:

$$\Sigma'_C = \Sigma_C - \Sigma_{CD}\Sigma_D^{-1}\Sigma_{CD}^T + \Sigma_{CD}\Sigma_D^{-1}\Sigma'_D\Sigma_D^{-1}\Sigma_{CD}^T \quad (3a)$$

$$\Sigma'_{CD} = \Sigma_{CD}\Sigma_D^{-1}\Sigma'_D \quad (3b)$$

$$\mu'_C = \mu_C + \Sigma_{CD}\Sigma_D^{-1}\left(\mu'_D - \mu_D\right). \quad (3c)$$

For a more detailed description see [12].

Often aggregates may have a variable number of parts, e.g. a `Window-Array` may contain 2 to 10 `Windows`. Since the JPDs for aggregates with distinct part cardinalities may be quite different, each cardinality receives its own aggregate model. This leads to a large number of alternative high-level models, each initially weighted by its likelihood. As new evidence is incorporated, these weights are adjusted according to the probabilities of alternative evidence assignments.

3.2.2 Classification and Role Assignment

We now show how the most probable class C for evidence E is determined if there are several alternative models and a single model may contain several variables of a given class. Let X_c be the set of random variables for class c. To calculate the probability for a class at a given position, one has to check each variable associated with that class and calculate the probability of the given position. In every model M the sum for every class is calculated and then weighted by the likelihood w_M of the model when adding up the sums. This is then normalized and yields the desired result.

$$P(C|E_{\mathrm{p}},\ldots) = \alpha \cdot \sum_{M} w_M \sum_{X \in X_C} P_M(X = E_{\mathrm{p}}|\ldots) \qquad (4)$$

After evidence has been classified, it needs to be integrated into the alternative models. This integration is similar to classification, but instead of summing the probabilities in each model, the maximum is selected. The variable corresponding to the maximum is selected as the role for the evidence and the evidence is assigned to that variable. Every evidence is assigned to every model if possible and the models to which the evidence could not be assigned are discarded. The weights for the models are adjusted according to the probabilities of the role assignment.

3.2.3 Learning Compositional Hierarchies

It is difficult to estimate probabilistic relations between objects in complex scenes, and handcrafting a BCH may also bias an evaluation. We have obtained the BCH for our experiments automatically from annotated images. The annotations are transformed into graphs representing the compositional hierarchies, the structure of and the parts of each aggregate are sorted (e.g. from left to right). Each (sorted) graph structure results in one alternative model, and when two or more annotations have the same graph structure, they are regarded as drawn from the same distribution. For each aggregate, a single multivariate normal distribution is learned and then the BCH is constructed from the learned distribution and the given part-of relationships. The weights for each model are created by counting the number of samples and normalized.

4. EXPERIMENT AND RESULTS

The system was evaluated on a database of floors from the façade domain, which is available as an outcome of the eTRIMS project (see [1] for more information). 393 floors were extracted from the eTRIMS annotated image database. Each object in the image is annotated with a bounding polygon and a class label. Figure 3 shows several examples from our database. Due to annotation errors and in order to generalize, each sample was drawn several times, and each time uniform noise was applied. This also addresses the problem that several models have only few examples, and produce singular covariance matrices.

Table 1: A representative distribution of the classes within our data of the floors within an façade.

class	relative frequency
`Balcony`	8.038 %
`Canopy`	0.629 %
`Door`	9.697 %
`Entrance`	1.287 %
`Gate`	0.114 %
`Ground-Floor`	1.831 %
`Person`	0.029 %
`Railing`	9.153 %
`Sign`	1.802 %
`Stairs`	0.486 %
`Upper-Floor`	9.411 %
`Vegetation`	0.057 %
`Wall`	0.057 %
`Window`	52.489 %
`Window-Array`	4.920 %

Description of Data.

There are 15 object classes in the ontology:

- the primitives: `Door`, `Person`, `Sign`, `Wall`, `Window`, `Gate`, `Railing`, `Canopy`, `Stairs`, `Vegetation`
- functional entities consisting of several parts: `Entrance`, `Window-Array`, `Balcony`, `Ground-Floor`,`Upper-Floor`

Table 1 shows the relative frequencies of the classes in the extracted floors. The annotation includes a rough estimate of the image scale, which makes it possible to estimate the size of the objects. This might seem unreasonable since such information is not usually present, but recent work shows that scale can often be estimated from a single image [18].

Benchmark Definition.

The input is a set of polygons, based on manual annotations, which need to be classified by a combination of local (visual) and global (contextual) features. The task is to classify each polygon by assigning it a label from our ontology.

Using the polygons from the annotations as evidence eliminates errors due to *false positives* and allows clear statements about the helpfulness of context.

Used Features.

The local features E_a used in the experiments were area, aspect ratio, compactness and rectangularity. They were chosen to be very simple here for two reasons. First of all, they are fast to compute and, as shown in [20, 3], the façade domain is a difficult classification problem even with much more complex features. Secondly, using weak features helps to illustrate the strength of probabilistic modeling and the usefulness of context. The global features E_p are the horizontal and vertical position.

Figure 3: Examples of the variety of floors with the domain. Because the floor where automatically extracted using the annotations also floor that are only partially visible, such as the floor with only one window visible are in the training and testing data.

Table 2: Classification rate determined by 10-fold cross-validation

fold	BCH	DT	BCH & DT
0	0,65	0,70	0,75
1	0,66	0,65	0,78
2	0,62	0,64	0,77
3	0,69	0,72	0,78
4	0,68	0,68	0,75
5	0,62	0,65	0,76
6	0,67	0,71	0,83
7	0,63	0,7	0,83
8	0,66	0,68	0,76
9	0,63	0,71	0,79
Average	0,652	0,685	**0,780**
Std. Dev.	0,025	0,029	**0,029**

Results.

The performance of the combined system is compared to the performance of the Bayesian Compositional Hierarchies (BCH) alone and the decision tree (DT) alone to evaluate the improvement obtained by exploiting context. The evaluation is based on ten-fold cross-validation with the results shown in Table 2. The average improvement of the classification rate observed over all testing data is 13%, from 0.685 of the decision tree (local visual appearance only) to 0.780 with the help of the contextual model. This shows that scene context can significantly improve classification.

One observed problem was that incorrect classifications introduced early on, result in incorrectly established context, which often leads to further incorrect classifications. Although Algorithm 1 introduced in Section 3 tries to minimize this effect by starting with the most reliable evidence, incorrect initial classifications still occur, and the classification performance on such images is very poor. These few images drag the classification rate down.

An excerpt from the confusion matrix over all of the ten folds is shown in Table 3. It can be seen that the ground floor is often confused with the upper floor. Since the difference between the two is almost purely contextual (as opposed to visual), it might be better to delay the classification until other objects provide a strong enough context.

5. CONCLUSIONS

In this paper, we have presented a probabilistic framework which combines bottom-up appearance-based classification with *dynamic* priors determined from the evolving global scene context. This not only helps to reduce complexity, but is also suitable for temporal scenes, where not all evidence is simultaneously available.

The experiments performed on a real-world dataset have shown that classification results can be significantly improved by using our contextual high-level model. An advantage of the presented approach is that it does not depend on parameters, and both the decision trees and the BCH are automatically learned from the training data. Another important aspect is that non-visual structural information can be used to determine the context, e.g. part-of relations, as long as such information is provided in the learning data.

While context helps in general, we also noticed that incorrect context can make things worse, as incorrectly classified evidence sometimes worsens the classification of succeeding evidence. This only occurs with a fraction of the images.

In order to reduce the effects of early misclassifications, we would like to explore the use of parallel interpretations. This can be done within the current framework by cloning the appropriate models, but will result in increased computation time.

Another aspect worth investigating is the detection and learning of context-dependent and -independent features. A floor might be best classified after a number of other pieces have been found as it depends more on the parts it contains than on its visual appearance.

The next step towards full scene interpretation is the use of *real* image analysis procedures instead of the annotations. At this time, we will need to cope with false positives and negatives which do not play a part in the current experiments. We have also started evaluating the performance compared to competing methods (such as Markov Random Fields and other BN formalisms).

One of the advantages of our system is that it provides an interpretation at each step and can formulate our expectations. We intend to use this in a domain where time is involved and predictions about future events must be made. Such a domain is the focus of an on-going project which involves modeling and recognizing aircraft service operations.

Table 3: A confusion matrix reduced to the most likely classes

	Balcony	Door	Ground-Floor	Railing	Stairs	Upper-Floor	Window	Window-Array
Balcony	151	2	0	1	0	4	83	5
Door	2	139	0	0	0	0	157	1
Ground-Floor	0	0	2	0	0	50	7	0
Railing	3	1	0	181	0	0	94	2
Stairs	0	0	0	4	2	0	10	0
Upper-Floor	0	0	7	0	0	274	12	4
Window	9	33	0	6	0	1	1596	13
Window-Array	1	0	0	4	0	4	57	92

6. ACKNOWLEDGMENTS

The authors Arne Kreutzmann and Kasim Terzić were supported by the EC project FP6-IST-027113 eTRIMS.

7. REFERENCES

[1] etrims - e-training for interpreting images of man-made scenes. `http://www.ipb.uni-bonn.de/projects/etrims/`.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[3] M. Drauschke and W. Förstner. Selecting appropriate features for detecting buildings and building parts. In *21st Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS-08)*, Beijing, China, 2008.

[4] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? *J. Vis.*, 7(1):1–29, 1 2007.

[5] E. Gyftodimos and P. A. Flach. Hierarchical bayesian networks: A probabilistic resoning model for structured domains. In E. de Jong and T. Oates, editors, *Proc. Workshop on Development of Representations*, ICML, pages 23–30, 2002.

[6] L. Hotz and B. Neumann. Scene interpretation as a configuration task. *KI*, 19(3):59–, 2005.

[7] L. Hotz, B. Neumann, and K. Terzić. High-level expectations for low-level image processing. In *The 31st Annual German Conference on Artificial Intelligence (KI-08)*, pages 87–94, Kaiserslautern, September 2008. Springer.

[8] D. Koller and A. Pfeffer. Object-oriented bayesian networks. In *The Thirteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 302–313, August 1997.

[9] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 775–781, Washington, DC, USA, 2005. IEEE Computer Society.

[10] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition (CVPR-09)*, 2009.

[11] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *In NIPS*. MIT Press, 2003.

[12] B. Neumann. Bayesian compositional hierarchies - a probabilistic structure for scene interpretation. Technical Report FBI-HH-B-282/08, Universität Hamburg, Department Informatik, Arbeitsbereich Kognitive Systeme, May 2008.

[13] B. Neumann and R. Möller. On scene interpretation with description logics. *Image and Vision Computing*, 26(1):82–101, 2008.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. The Morgen Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, 1988.

[15] R. Perko, C. Wojek, B. Schiele, and A. Leonardis. Integrating Visual Context and Object Detection within a Probabilistic Framework. In *Attention in Cognitive Systems: International Workshop on Attention in Cognitive Systems (WAPCV-08)*, page 54. Springer London, Limited, May 2009.

[16] D. Poole, A. Mackworth, and R. Goebel. *Computational intelligence: a logical approach*. Oxford University Press, Oxford, UK, 1997.

[17] R. D. Rimey. Control of selective perception using bayes nets and decision theory. Technical Report 468, University of Rochester, Computer Science Department, Rochester, New York 14627, Dec. 1993.

[18] A. Saxena, S. Chung, and A. Ng. Learning depth from single monocular images. *Advances in Neural Information Processing Systems*, 18:1161, 2006.

[19] K. Terzic, L. Hotz, and B. Neumann. Division of Work During Behaviour Recognition-The SCENIC Approach. In *Workshop on Behaviour Modelling and Interpretation, (KI-07)*, Osnabrück, Germany, September 2007.

[20] K. Terzić and B. Neumann. Decision trees for probabilistic top-down and bottom-up integration. Technical Report to be published, Universität Hamburg, Department Informatik, Arbeitsbereich Kognitive Systeme, Jul 2009.

[21] A. R. Webb. *Statistical Pattern Recognition, 2nd Edition*. John Wiley & Sons, October 2002.

[22] J. Yu and J. Luo. Leveraging probabilistic season and location context models for scene understanding. In *The 2008 international conference on Content-based image and video retrieval (CIVR-08)*, pages 169–178, New York, NY, USA, 2008. ACM.