Context-based Probabilistic Scene Interpretation

Bernd Neumann, Kasim Terzic

Department of Informatics, University of Hamburg
{neumann, terzic}@informatik.uni-hamburg.de

Abstract. In high-level scene interpretation, it is useful to exploit the evolving probabilistic context for stepwise interpretation decisions. We present a new approach based on a general probabilistic framework and beam search for exploring alternative interpretations. As probabilistic scene models, we propose Bayesian Compositional Hierarchies (BCHs) which provide object-centered representations of compositional hierarchies and efficient evidence-based updates. It is shown that a BCH can be used to represent the evolving context during stepwise scene interpretation and can be combined with low-level image analysis to provide dynamic priors for object classification, improving classification and interpretation. Experimental results are presented illustrating the feasibility of the approach for the interpretation of facade images.

Keywords: Scene interpretation, probabilistic scene models, context-based interpretation

1 Introduction

Scene interpretation has the goal to understand visual scenes beyond single-object recognition. The basic architecture of a knowledge-based scene interpretation system is shown in Fig. 1.
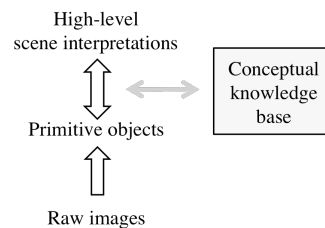

Fig. 1: Basic architecture for scene interpretation

Image analysis procedures first determine a low-level description of the image (or image sequence) in terms of patches or segments which at best correspond to meaningful primitive objects. The task of high-level scene interpretation is to determine meaningful object configurations, situations and activities based on the primitive objects provided by low-level analysis, and on models in a conceptual knowledge base. For many application domains, *compositional hierarchies* provide a

natural structure for the conceptual knowledge base. A compositional hierarchy consists of *aggregates* which describe a scene in terms constituents at a lower hierarchical level, meeting certain conditions. Typical structures of a compositional hierarchy are shown in Fig. 2 for the facade domain. Although the facade domain is used for examples in this contribution, the approach is generic in nature. Note that the conceptual knowledge base distinguishes between models of physical objects (e.g. *B-Door*) and models of object appearances in images (e.g. *B-Door-View*).
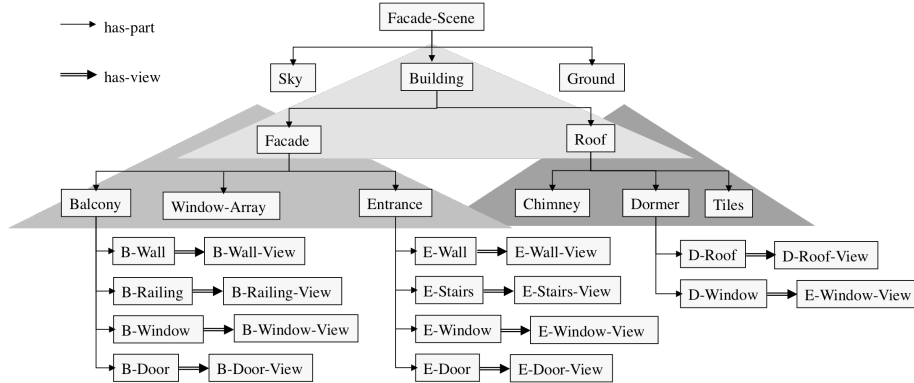

Fig. 2: Compositional hierarchy for facade scene (not all components shown). Triangles mark aggregate structures.

In this contribution, we present a new approach to uncertainty management for stepwise high-level scene interpretation. As evidence is incorporated step by step, alternative interpretation possibilities arise which can be ranked effectively by the probabilistic high-level context based on preceding interpretation steps.

The paper is structured as follows. In the next section, we discuss related work. In Section 3, we present the conceptual probabilistic framework developed for high-level scene interpretation. This framework has been implemented in an operational interpretation system called SCENIC. System details and experimental results are described in Section 4. We conclude with a summary of the results.

2   Related Work

Compositional hierarchies have been employed for high-level scene interpretation by many researchers [1, 2, 3, 4, 5] with basically non-probabilistic (crisp) frame-based representations, as commonly used in AI. Rimey [6] was the first to model compositional hierarchies with tree-shaped Bayesian Networks (BNs), requiring parts of an aggregate to be conditionally independent. Koller and coworkers [7, 8] extended BNs in an object-oriented manner for the representation of structured objects. Their Probabilistic Relational Models allow augmenting a crisp relational structure with an arbitrary probabilistic dependency structure. Gyftodimos and Flach [9] 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 modelling in scene interpretation. For this purpose, Bayesian Compositional Hierarchies (BCHs) were developed [10]. This structure mirrors the formal logic-based ontologies of [5] but allows for a probabilistic ranking of alternative interpretations. An interesting alternative approach is provided by Mumford and Zhu [11] where a grammatical formalism takes the place of hierarchical knowledge representation and parsing algorithms are applied for scene interpretation, leading to efficient processing but complicating the integration with large-scale knowledge representation.

Context through visual clues has been in the focus of recent work [12, 13]. 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. Li et al. [14] presented a way to incorporate a global context: If the type of picture is known (e.g. the type of sport shown), then the priors for the classes can be changed correspondingly to increase the chance of correct classification. However, details of the contextual relationships between scene objects are not exploited. This is possible in our approach where dynamic context exploitation is integrated into a general probabilistic framework for scene interpretation.

3   Probabilistic Models for High-level Interpretation

In this section, we will describe the probabilistic framework developed for scene interpretation. We will first give a formal specification in a general form (Section 3.1) and suggest beam search as a suitable interpretation algorithm. We then show how probabilistic scene interpretation can be realized efficiently with a probabilistic model organized as a Bayesian Compositional Hierarchy (BCH) (Section 3.2). We finally present a further specialization where probabilistic information is separated into a structural and an appearance part. This way, object classification can be based on an appearance-based low-level component and a structure-based high-level component which provides dynamic expectations based on the evolving scene context.

3.1   General Probabilistic Framework

In a general form, probabilistic scene interpretation can be modelled as evidence-based reasoning with large joint probability distributions (JPDs). Let $\underline{R}_{all} = \underline{R}_1 .. \underline{R}_N$ be vector-valued random variables describing *scene objects* of a conceptual knowledge base (properties of primitive objects and aggregates of a compositional hierarchy), and let $\underline{V}_{all} = \underline{V}_1 .. \underline{V}_K$ be random variables describing *view objects* associated with the scene objects (see Fig. 2), i.e. entities which can be observed in an image, such as groups of SIFT features or other low-level descriptors. For the purposes of this contribution, we will assume that each primitive scene object is related to exactly one view object which describes its possible appearances in an image. Different objects may look alike, especially under noisy conditions, hence observed evidence may not be unambiguously assigned to a view object, causing alternative interpretation candidates.

A probabilistic model for a scenario with a particular set of objects is given by a JPD $P(\underline{R}_{all}\ \underline{V}_{all})$ where $\underline{R}_{all}$ are hidden variables and $\underline{V}_{all}$ are observables. This JPD can be used to express all relevant probabilistic dependencies:

– how the properties of observations are related to properties of scene objects,
– how the properties of a specific scene object are related to each other,
– how properties of one scene object are related to properties of other scene objects.

To represent scenarios which may occur with different object configurations (sets of objects), we combine the corresponding JPDs as alternative models using a random variable S with domain 1..M and distribution $P_S(1)=p_1$ .. $P_S(M)=p_M$ to describe the selection of an alternative model:

$$P_{scenario} = P_S(S)\ P^{(S)}(\underline{R}_{all}^{(S)}\underline{V}_{all}^{(S)}) \tag{1}$$

Typically, the objects of a scenario will only be a subset of all objects present in a realistic scene, hence we have to account for "noise objects". To simplify the following presentation, we will only consider scenes where the observed evidence exactly matches one or more of the alternative models in $P_{scenario}$. A treatment of noise objects and missing evidence can be added without affecting the basic procedure.

Let $\underline{e}_{all}$ be all evidence about a scene, then scene interpretation amounts to determine the most probable instantiation of one of the alternative models based on this evidence. This can be achieved by two maximization steps.

1. Determine the model m with JPD $P^{(m)}$ and the view objects $\underline{V}_{all}^{(m)}$ which are most probable given the observed evidence.

$$\max_{m,\underline{V}_{all}^{(m)}} P_{scenario}(m\underline{V}_{all}^{(m)}\mid \underline{e}_{all}) = \max_{m,\underline{V}_{all}^{(m)}} p_m P^{(m)}(\underline{e}_{all}) \tag{2}$$

2. For the winning model m and view objects $\underline{V}_{all}^{(m)}$, determine the instantiations of hidden variables $\underline{R}_{all}$ which maximize

$$\arg\max_{\underline{r}_{all}} P^{(m)}(\underline{R}_{all}^{(m)} = \underline{r}_{all}\mid \underline{V}_{all}^{(m)} = \underline{e}_{all}) \tag{3}$$
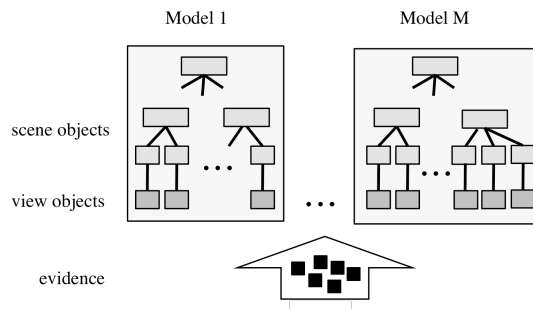

Fig. 3: Evidence assignment to view objects of alternative models

The underlying assignment problem is formidable as illustrated in Fig. 3. In a stepwise process, noisy and ambiguous evidence must be assigned to one of several models in one of a large number of permutations, therefore a global maximization as

defined by Eqs. 2 and 3 is rarely feasible for real scene interpretation tasks. Even in restricted domains such as the facade domain, the probabilistic model $P_{scenario}$ may consist of hundreds of alternative models, each comprised of hundreds of random variables representing scene object and evidence properties with potentially large value ranges. Hence the search space may be very large and approximate methods are required.

As a general strategy for approximate probabilistic scene interpretation we propose beam search [15], where only promising alternative partial interpretations are kept (in the "beam"), while improbable ones are discarded. There are three reasons for this choice. First, beam search is based on incremental evidence assignment which allows to generate dynamic expectations, indispensable for the analysis of time-varying scenes and real-time recognition tasks. Second, beam search permits scalable parallelism by selecting an appropriate beam width. Third, beam search corresponds well to our formulation of probabilistic scene models in terms of alternatives.

For scene interpretation with beam search, the following steps have to be executed:

A   Initialise the beam with all alternative models given by Eq. 1.
B   Select unassigned evidence $\underline{e}$.
C   Determine possible assignments of $\underline{e}$ for each model in the beam, clone models in case of multiple assignments, discard models where no assignment is possible.
D   Rank models using Eq. 2 (where $\underline{e}_{all}$ now stands for all evidence observed so far), discard unlikely models from beam.
E   Repeat B to E until all evidence is assigned.
F   Select highest-ranking model, determine final interpretation using Eq. 3.

Note that evidence selection in Step B can be crucial in the initial phase where non-discriminating evidence could give rise to a possibly large number of alternative models. A selection which maximizes information gain is advisable, as in decision tree methodology [16].

Efficient storage of $P_{scenario}$ and computation of the marginalisations in Eq. 2 may easily become a bottleneck for realistic tasks. Therefore, Bayesian Network technology is required and the dependency structure of object properties plays an important part. In the following section, we will present Bayesian Compositional Hierarchies [10] which allow arbitrary dependencies within an aggregate but provide a tree-shaped dependency structure between aggregates and thus efficient computational procedures in tune with compositional hierarchies.

3.2 Bayesian Compositional Hierarchies

Probabilistic dependencies in compositional hierarchies for high-level scene interpretation have characteristics which can be exploited for efficient storage and computation. For one, strong mutual dependencies are often confined to the parts of an aggregate and less stringent between parts of different aggregates. Second, higher-level representations often abstract from lower-level details. We therefore propose a probabilistic model, called Bayesian Compositional Hierarchy (BCH), where aggregates (i) are modelled by arbitrary JPDs, (ii) possess an external description

abstracting from details about their parts, and (iii) depend on other aggregates only via the part-of relations of the compositional hierarchy.
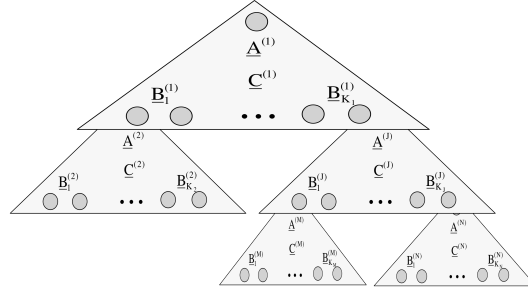

Fig. 4: Structure of a BCH. Triangles represent aggregates, circles represent parts.

Fig. 4 shows the schematic structure of a BCH. Each aggregate is described individually by a JPD $P(\underline{A}\ \underline{B}_1..\underline{B}_K\ \underline{C})$ where $\underline{A}$ is the external description, $\underline{B}_1..\underline{B}_K$ are descriptions of the parts, and $\underline{C}$ expresses conditions on the parts. For example, a *Balcony* can be described by bounding boxes $\underline{B}_1$, $\underline{B}_2$, $\underline{B}_3$ for *B-Door*, *B-Window* and *B-Railing*, distances $\underline{C}$ between these parts, and a bounding box $\underline{A}$ for the *Balcony* as a whole. The hierarchy is constructed by taking the external aggregate descriptions at a lower level as part descriptions at the next higher level, hence $\underline{B}_i^{(1)} = \underline{A}^{(2)}$ etc.

The JPD of the complete hierarchy is given by

$$P(\underline{A}^{(1)}..\ \underline{A}^{(N)}) = P(\underline{A}^{(1)}) \prod_{i=1..N-1} P(\underline{B}_1^{(i)}..\ \underline{B}_{K_i}^{(i)}\ \underline{C}^{(i)} \mid \underline{A}^{(i)}) \qquad (4)$$

This remarkable formula shows that the JPD of a BCH can be easily constructed from individual aggregate representations, and belief updates can be performed by propagation along the tree structure. Let $P'(\underline{B}_i)$ be an update of $P(\underline{B}_i)$, by evidence or propagation from below. Then the updated aggregate JPD is

$$P'(\underline{A}\underline{B}_1..\ \underline{B}_K\underline{C}) = P(\underline{A}\underline{B}_1..\ \underline{B}_K\underline{C})P'(\underline{B}_i)/P(\underline{B}_i) \qquad (5)$$

A similar equation holds when $P(\underline{A}^{(m)})$ is updated by propagation from above.

If the aggregate JPDs are Multivariate Gaussians, propagation can be performed by closed-form solutions [10]. Multivariate Gaussians are also very convenient for implementing the interpretation by beam search described in Section 3.1. The marginalisations required for ranking alternative interpretations according to Eq. 2 are directly available from the aggregate covariances, and the final maximising interpretation according to Eq. 3 can be given in terms of the mean values.

3.3 Separating Appearance and Structural Context

In the probabilistic framework presented so far, evidence assignment is governed by a distribution which captures appearance properties of objects as well as the influence of the structural context, e.g. position and orientation of other objects. Considering the rich methodology for appearance-based object classification and the

need to modularize, it is useful to separate the two knowledge sources. In this section, we present an approach where a high-level probabilistic model is used to represent the structural context providing structure-based dynamic priors for an appearance-based probabilistic classifier.

We assume that each evidence $\underline{e}$ can be decomposed into two parts $\underline{e} = [\underline{e}_a \; \underline{e}_s]^T$ where $\underline{e}_a$ describes appearance properties (e.g. size, shape, colour, texture) and $\underline{e}_s$ describes structurally relevant properties (e.g. position, distance to reference objects). Furthermore, we assume that both components are statistically independent, given knowledge about which view object (and hence which scene object) has caused this evidence. Hence

$$P^{(m)}(\underline{e}_a \underline{e}_s \mid \underline{V}_i^{(m)}) = P^{(m)}(\underline{e}_a \mid \underline{V}_i^{(m)}) \; P^{(m)}(\underline{e}_s \mid \underline{V}_i^{(m)}) \tag{6}$$

In beam search, let $\underline{e}$ be the evidence to be assigned to some view object $\underline{V}_i^{(m)}$ and let $\underline{e}_{past}$ be past evidence already assigned to view objects $\underline{V}_{past}^{(m)}$. Then to perform the ranking in Step D, we have to compute

$$P(m\underline{V}_i^{(m)} \mid \underline{e}_a \underline{e}_s \underline{e}_{past} \underline{V}_{past}^{(m)}) = \frac{P(\underline{e}_a \mid m\underline{V}_i \underline{e}_{past} \underline{V}_{past}^{(m)}) P(m\underline{V}_i^{(m)} \underline{e}_s \mid \underline{e}_{past}^{(m)} \underline{V}_{past}^{(m)})}{P(\underline{e}_a \underline{e}_s \mid \underline{e}_{past} \underline{V}_{past}^{(m)})} \tag{7}$$

$$= \frac{P^{(m)}(\underline{e}_a \mid \underline{V}_i^{(m)}) p_m P^{(m)}(\underline{V}_i^{(m)} = \underline{e}_a \mid \underline{e}_s \underline{e}_{past} \underline{V}_{past}^{(m)})}{\sum_{m, \underline{V}_i^{(m)}} P^{(m)}(\underline{e}_a \mid \underline{V}_i^{(m)}) p_m P^{(m)}(\underline{V}_i^{(m)} = \underline{e}_a \mid \underline{e}_s \underline{e}_{past} \underline{V}_{past}^{(m)})} \tag{8}$$

The enumerator of Eq. 8 specifies the probability that a specific view object (and hence a specific scene object) in model m has the appearance $\underline{e}_a$. This is a likelihood function which can be provided by low-level image analysis, for example, based on a probabilistic decision tree. The second term of the enumerator specifies the probability of observing $\underline{e}_a$ given the structural information $\underline{e}_s$ and past evidence assignments. This is information which can be provided by high-level interpretation, for example, in terms of a BCH. It provides a dynamic prior for the appearance-based likelihood function of the first term. The denominator, finally, is a normalizing factor.

In the following section, we will report about scene interpretation experiments where Eq. 8 is used to rank evidence assignment to primitives of a structural facade model represented by a BCH. It is shown that this improves classification and interpretation performance as compared to single-object classification based only on appearances.

4 Evaluation of Context Influence

In this section, we present two experiments illuminating the probabilistic interaction of low-level and high-level interpretation processes. In the first experiment, described in more detail in [17], the effect of probabilistic context on the classification rate was tested in the domain of building floors. In the second

experiment, probabilistic models of entire facades were used to interpret and combine conflicting and error-prone segmentations from several low-level algorithms.

A major problem in this domain is the similar appearance of objects which play different roles in structural models and hence must be distinguished, e.g. balcony doors vs. entrance doors. This makes structural context a very important factor for disambiguation and complete understanding of a scene, while motivating appearance models to be oriented at coarser groupings than the roles in structural models. For our experiments, we therefore developed an appearance model in terms of a probability distribution for classes C such as doors and windows instead of balcony doors or balcony windows. The probabilities $P(\underline{C}|\underline{E}_a)$ were estimated by decision trees learnt from a training set of annotated facade images, as described in [18]. Using Bayes' rule with priors determined from the dataset, $P(\underline{E}_a|C)$ can be obtained as required for Eq. 8. A BCH modelling facade floors was learnt from the same training examples.

In the first experiment, the system was evaluated on a database of 393 floors from the facade domain, which were extracted from the eTRIMS [19] annotated image database. In this database, each object in the image is annotated with a bounding polygon and a class label. Aggregates above the floor level, such as *Facade* and *Building* were not used for this experiment. Fig. 5 shows several examples from the set of floors used.

Fig. 5: Examples of the variety of floors within the domain

The appearance features extracted from segmentations were area, aspect ratio, rectangularity and compactness. The context was determined based on the position of the object in the image and its probabilistic relation to the positions of previously integrated objects. The task was to classify the annotated polygons.

The performance was evaluated using ten-fold cross-validation. The results are shown in Fig. 6 left. An overall 13% increase in the classification rate shows that scene context can significantly improve classification. The beam search algorithm was used with beam width 1, i.e. only the best alternative was kept at each step. This led to some compounded errors when wrong context was established early, following a wrong initial classification. We expect that the results will be further improved with a beam keeping several alternatives.

In the second experiment, the system was evaluated as a source of context for segmenting a complete facade image. In this instance, the BCH contained six different probabilistic models of facade structures. The goal was to recognise the structure of a facade image and to use the evolving context to guide the segmentation

process. Three separate segmentation algorithms were used for this experiment (i) a trainable window detector based on Adaboost [20], (ii) an interest region detector [21], and (iii) a detector of stable regions in scale space [22]. Each one of them was run on 190 images, and the results compared to the hand annotations. Each detected polygon was labelled with the correct class from the corresponding annotation in the case of sufficient overlap (50% in this case), or as a "False Positive" if there was no correspondence. Then a 10-dimensional feature vector consisting of the four shape descriptors from the previous experiment and the means and standard deviations of the three colour channels (RGB) was computed for each detected region. Then the class likelihood distribution was modelled by a decision tree for each detector separately, including the likelihood that a false positive produced the detection.

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	0.70	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**

Fig. 6 left (Exp. 1)**:** Classification rate obtained by 10-fold cross-validation. "BCH" shows the probabilistic context based on the position alone, "DT" the class probability based on appearance features estimated by the decision tree, the last column shows the combined result.

Fig. 6 middle and right (Exp. 2)**:** Facade image (middle) and result of context-based segmentation (right). The BCH used six structural models, out of which the most fitting one was automatically selected and used to interpret the polygons provided by the low level.

As in the first experiment, the region with the highest classification confidence was classified and the context propagated, changing the class posteriors of the remaining regions. Although the combination of the three detectors found many false positives (over 80%) in addition to correct detections, the strong probabilistic context managed to enable a complete segmentation of the image, as shown in Fig. 6 right. Most regions were discarded as false positive detections, and the remaining detections interpreted according to the most fitting of the six structural models.

5. Conclusion

A new general framework for stepwise probabilistic scene interpretation has been presented, where evidence classification is supported by dynamic expectations generated from the evolving high-level context. Experiments with learnt models of facade images demonstrate significant improvements over context-free classification and show that weak low-level results can be compensated by high-level context.

References

1. Tsotsos, J.K., Mylopoulos, J., Covvey, H.D., Zucker, S.W.: A Framework for Visual Motion Understanding. IEEE PAMI-2, 563-573 (1980)
2. Nagel, H.-H.: From Image Sequences towards Conceptual Descriptions. Image and Vision Computing 6(2), 59-74 (1988)
3. Neumann, B.: Description of Time-Varying Scenes. In: Waltz, D. (ed.), Semantic Structures, Lawrence Erlbaum, 167-206 (1989)
4. Georis, B., Mazière, M., Brémond, F., Thonnat, M.: Evaluation and Knowledge Representation Formalisms to Improve Video Understanding. In: Proc. IEEE International Conf. on Computer Vision Systems ICVS06, IEEE Computer Society, 27 (2006)
5. Neumann, B., Moeller, R.: On Scene Interpretation with Description Logics. In: Cognitive Vision Systems, Springer, LNCS 3948, 247–275 (2006)
6. Rimey, R.D. Control of Selective Perception using Bayes Nets and Decision Theory. TR 468, Univ. of Rochester, Computer Science Department, Rochester, USA 14627 (1993)
7. Koller, D., Pfeffer, A.: Object-oriented Bayesian Networks. In: The Thirteenth Annual Conference on Uncertainty in Artificial Intelligence, 302–313 (1997)
8. Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning (eds.), 129–174. The MIT Press (2007)
9. Gyftodimos, E., Flach, P.A.: Hierarchical Bayesian Networks: A Probabilistic Reasoning Model for Structured Domains. In: de Jong, E., Oates, T. (eds.), Proc. Workshop on Development of Representations, ICML, 23–30 (2002)
10. Neumann, B.: Bayesian Compositional Hierarchies - A Probabilistic Structure for Scene Interpretation. TR FBI-HH-B-282/08, Univ. of Hamburg, Department Informatik (2008)
11. Mumford, D., Zhu, S.-C.: A Stochastic Grammar of Images. Now Publishers (2007)
12. Yu, J., Luo, J.: Leveraging Probabilistic Season and Location Context Models for Scene Understanding. In: The 2008 International Conference on Content-based Image and Video Retrieval (CIVR-08), ACM, 169–178 (2008)
13. Perko, R., Wojek, C., Schiele, B., Leonardis, A.: Integrating Visual Context and Object Detection within a Probabilistic Framework. In: Attention in Cognitive Systems: Int. Workshop on Attention in Cognitive Systems (WAPCV), Springer, 54 (2009)
14. Li, L.-J., Socher, R., L. Fei-Fei: Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework. In: Computer Vision and Pattern Recognition (CVPR) (2009)
15. Lowerre, B.: The Harpy Speech Recognition System. Ph.D. thesis, Carnegie Mellon University (1976)
16. Mitchell, T.M.: Machine Learning. The Mc-Graw-Hill (1997)
17. Kreutzmann, A., Terzić, K., Neumann, B.: Context-aware Classification for Incremental Scene Interpretation. In: Proc. Workshop on Use of Context in Vision Processing (UCVP 2009), Boston (2009)
18. Terzić, K., Neumann, B.: Decision Trees for Probabilistic Top-down and Bottom-up Integration. Technical Report FBI-HH-B-288/09, University of Hamburg (2009)
19. etrims - e-Training for Interpreting Images of Man-made Scenes. http://www.ipb.uni-bonn.de/projects/etrims/
20. Terzić, K., Hotz, L., Šochman, J.: Interpreting Structures in Man-made Scenes; Combining Low-Level and High-Level Structure Sources. In International Conference on Agents and Artificial Intelligence (ICAART 2010), Valencia (2010)
21. Jahangiri, M., Petrou, M.: Fully Bottom-Up Blob Extraction in Building Facades. In: Proc. Pattern Recognition and Image Analysis (PRIA), (2008)
22. Drauschke, M.: An Irregular Pyramid for Multi-Scale Analysis of Objects and Their Parts. In: Workshop on Graph-based Representations (GbR'09), LNCS 5534, 293–303 (2009)