

Probabilistic Scene Models for Image Interpretation

Alexander Bauer¹,

¹ Fraunhofer Institute for Optronics, System Technologies
and Image Exploitation IOSB, Fraunhoferstr. 1, 76131 Karlsruhe, Germany
alexander.bauer@iosb.fraunhofer.de

Abstract. Image interpretation describes the process of deriving a semantic scene description from an image, based on object observations and extensive prior knowledge about possible scene descriptions and their structure. In this paper, a method for modeling this prior knowledge using probabilistic scene models is presented. In conjunction with Bayesian Inference, the model enables an image interpretation system to classify the scene, to infer possibly undetected objects as well as to classify single objects taking into account the full context of the scene.

Keywords: Image Interpretation, Image Understanding, High-level vision, Generative Models, Bayesian Inference, Relaxation Labeling, Importance Sampling

1 Introduction

Many applications of computer vision aim at the automated interpretation of images as a basis for decision making and planning in order to perform a specific task. Image interpretation summarizes the process of creating a semantic scene description of a real-world scene from single or multiple images. A scene represents a spatio-temporal section of the real-world in terms of its physical objects, their properties and relations. The corresponding semantic scene description contains all task-relevant objects, properties and relations described at a task-relevant abstraction level.

In many applications of image interpretation, it is not sufficient to detect and classify objects based on its appearance alone (e. g. the existence of a building in the scene). Rather, higher-level semantic descriptions (e. g. the function of the building being a workshop) have to be inferred based on the spatial configuration of multiple objects and prior knowledge about possible scenes. Prior knowledge can also be useful to improve results of purely appearance-based object recognition methods by ruling out unlikely detections and focus the attention on likely occurring, but undetected objects.

The image interpretation problem has drawn scientific interest since the 80s in the fields of artificial intelligence and computer vision. The main challenges have been identified early [1], yet their solution has not been ultimately determined:

- **Knowledge representation** – How to model prior knowledge about possible scene descriptions?

Bauer, A., "Probabilistic Scene Models for Image Interpretation, In: Communications in Computer and Information Science, vol 81, Springer, (2010). DOI: 10.1007/978-3-642-14058-7_58

The original publication is available at
<http://www.springerlink.com/content/gn92324j2146g406/>

- **Hypotheses Matching** – How to match possible scene descriptions against incomplete and erroneous detections of objects in the image?
- **Inference** – How to derive inferences from prior knowledge in order to improve and complete the scene description?

Probabilistic models, becoming more and more popular in computer vision, provide an intuitive way to model uncertainty, and the Bayes' Theorem provides a consistent framework for inference based on incomplete evidence. These properties of probabilistic methods have motivated the development of probabilistic scene models for image interpretation, meant to model prior knowledge about possible scenes using probability theory. The presented model design is targeted to the assisted interpretation of infrastructure facilities from aerial imagery [2], but it potentially generalizes to other image interpretation applications as well. This paper describes the scene model structure and how the main challenges of knowledge representation, hypotheses matching and inference can be tackled in a probabilistic framework for image interpretation. For better understanding, the contribution is illustrated on the application for the interpretation of airfield scenes.

2 Related Work

Early approaches to image interpretation aiming for the description of complex scenes were inspired by the advances in artificial intelligence in the 80s in the field of rule-based inference [1],[4],[5],[6]

From the 90s until today, probabilistic approaches and Bayesian inference have drawn attention from cognitive psychology as well as from the computer vision community. Since then, several probabilistic approaches for high level image interpretation have been proposed, of which only a few can be mentioned here. Rimey and Brown developed a system to control selective attention using Bayesian Networks and Decision Theory [7]. Lueders used Bayesian Network Fragments to model taxonomy and partonomy relations between scene objects to compute the most probable scene interpretation based on perceptive information [8]. A stochastic graph grammar in conjunction with a Markov Random Field has been used by Lin et al. to recognize objects which are composed of several parts with varying alignment and occurrence [9]. In [10] it was also applied to aerial imagery.

Following this current, the presented approach contributes to the efficient application of probability theory for the interpretation of images depicting complex scenes, such as they appear in remote sensing and aerial reconnaissance. In contrast to previous approaches, it is focused on the classification of objects at the functional level (see Section 1), rather than on detection and classification on the appearance level. It is able to improve and interpret results acquired from low-level methods such as automated object recognition algorithms and can be used to control their execution.

3 Bayesian Inference Applied to Image Interpretation

The Bayes' theorem provides a sound formalism to infer the distribution of a random variable given evidence in terms of uncertain observations and measurements. Everything that is required for Bayesian inference is to model the prior distribution of the random variable and to define a conditional probability of the observations given each realization of the variable. Applied to the image interpretation problem, the unknown random variable S represents the correct semantic description of the scene. Evidence collected from the image as a set of object observations is described by the random variable O . According to the Bayes' theorem, the updated posterior probability distribution can be calculated using

$$P(S = s_i | O = o_k) \propto P(O = o_k | S = s_i) \cdot P(S = s_i). \quad (1)$$

For brevity, the term for the normalization of the distribution to 1 is omitted in (1). The prior distribution $P(S = s_i)$ is defined by a probabilistic scene model, modeling possible scene realizations and their typical object configurations in terms of their probability of occurrence. The conditional probability $P(O = o_k | S = s_i)$ models the uncertainty in the recognition of objects. Using the posterior distribution, several useful inferences can be calculated to direct the iterative development of the scene description, which will be explained in Section 6.

4 Representing Possible Scene Descriptions

A scene description describes a real-world scene in terms of its task-relevant physical objects. In real-world scenes, functionally related objects are often arranged in spatial relation to each other to form an object composition which is described as a new object.

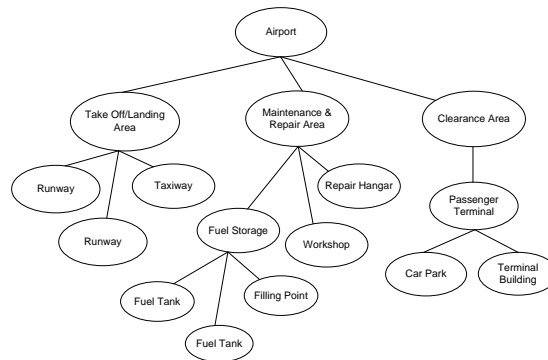


Fig. 1. Example of an interpretation tree for an airfield scene

For example on an airfield, buildings which are dedicated to the maintenance and repair of aircraft are composed to a “Maintenance & Repair Area”. This leads to a natural description of a scene as a tree of objects, in which the edges define functional composition, as illustrated using an example in Fig. 1.

The resulting *interpretation tree* $s=(\Omega,F)$ describes all objects of a possible scene description in terms of *object nodes* $\omega_i \in \Omega$. The set of edges F represents their functional composition. Object nodes are associated with a particular *object class* Φ , written as $\omega_i \sim \Phi$. Object classes are concepts such as “Airport”, “Runway”, etc.

The number of interpretation trees possible to occur can be very large, due to the high variability of real-world scenes. A lot of variations results from different numbers of occurrence of objects of a single object class. Variations in the structure of real-world scenes are also likely, even if the objects in the scene serve a similar function. For example in the case of airfields, regional variations and the advances in design of airfields over decades resulted in very different object configurations. However, object class occurrence and scene structure are important characteristics of a scene and therefore have to be taken account for in the model. To tackle the complexity problem, an approximation method is proposed for inference in Section 6.

5 Modeling Prior Knowledge in a Probabilistic Scene Model

As mentioned in Section 3, the probabilistic scene model must provide a prior probability for each possible interpretation tree, i.e. a distribution of the random variable S , which represents the correct scene description of the currently investigated image. A second requirement on the scene model is that the acquisition of the model parameters must be tractable and comprehensible. As sufficient training data is hardly available for a complex domain, in most cases it will be necessary to consult a human expert to establish a comprehensive and useful scene model. Therefore a modeling syntax is chosen, which is inspired by the verbal description of object classes by a human expert. Nevertheless, learning can be implemented by estimating the conditional probabilities of the model from training data.

The scene model is defined as the set of interrelated *object class models* $M(\Phi)$, from which all possible interpretation trees can be generated. As an illustrative example, Fig. 2 shows some of the object class models and their relations necessary to model possible interpretation trees of an airfield. Three types of object class models are defined:

- *Composition models* (C-Models $M_C(\Phi)$) describe an object class in terms of a composition of other objects (e. g. the ‘Airport’ model in Fig. 2). Such object classes occur at the upper levels of the interpretation tree (such as “Runway Area”, see Fig. 1). To represent the probability of all possible compositions, the distributions of the number of occurrences of each subordinate object model is defined in the compositional model. Assuming independence on the occurrence of different object models, it is sufficient to define the distribution for each single object class and to establish the joint probability distribution by multiplication. In the example shown in Fig. 2 the distributions are chosen to be uniform inside a reasonable

interval, which simplifies the acquisition process in cooperation with an expert by using statements such as: “Airports have at least one runway, up to a maximum of 5 runways”. However, more informative distributions can be used to incorporate more detailed prior knowledge, also taking into account dependencies between the occurrence probabilities of different object classes.

- *Taxonomy models* (T-Models $M_T(\Phi)$) define abstract object class models, which summarize different realizations of an abstract object class. For example the object class “Airfield” is further specified by the discrimination of disjunctive subtypes of that object class, as depicted in Fig. 2. For each subtype, a probability is defined which represents the conditional probability $P(\omega_i \sim \Phi_j | \omega_i \sim \Phi_T)$ of an object node ω_i to be associated with the subtype discriminations Φ_j , given it is associated with the abstract object class Φ_T .
- *Atomic models* (A-Models $M_A(\Phi)$) define object classes which can be neither further discriminated by more specific object classes nor divided into sub-parts. In Fig. 2, all objects which are not described by a box are represented by an A-model.

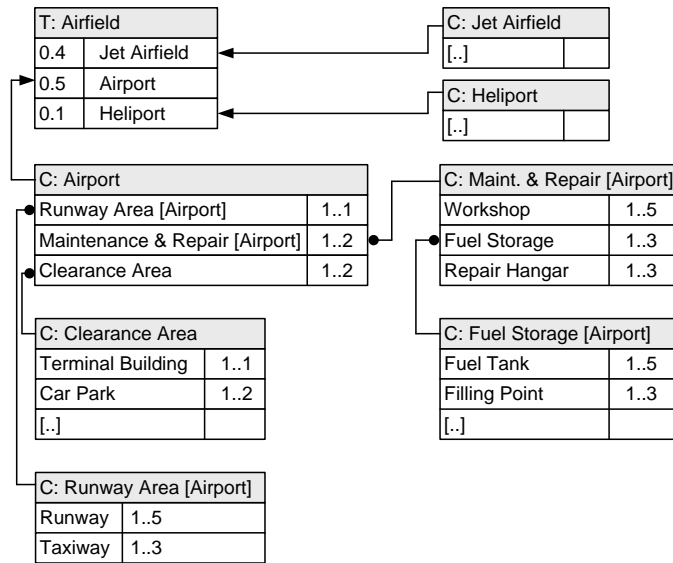


Fig. 2. Section of a scene model for airfield scenes. Abbreviations: T: Taxonomy Model, C: Composition Model. Figures in T-Models represent probabilities of different realizations, intervals in C-Models stand for uniform distributions of the number of occurrences for each subordinate object model.

The complete *scene model* $M(S)$ is defined by the tuple $M(S) = \langle M_C, M_T, M_A, \Phi_0 \rangle$ in terms of the sets of the three kinds of object class models and the root object class

model Φ_0 . From the scene model, all possible interpretation trees $s \in S$ and their corresponding prior probability $P(S = s)$ can be generated using the following algorithm:

1. Create object node ω_0 as the root node of the interpretation tree s and associate it with object class Φ_0 . Initialize $P(S = s) := 1$
2. For every object node ω_i associated with a T-Model, choose a subtype object class Φ_j and update $P(S = s)$ with $P(S = s) \cdot P(\omega_i \sim \Phi_j | \omega_i \sim \Phi_T)$ as defined in the T-model.
3. For every object node ω_i newly associated with a C-Model, choose a composition according to the C-Model description and create the object nodes of the composition. Update the scene prior probability $P(S = s)$ by multiplying with the composition probability according to the C-Model.
4. Repeat step 2 and 3 until no T-Model remains and all C-models have been treated in step 3.

Using the algorithm, for any given interpretation tree s , the corresponding prior probability $P(S = s)$ can be determined by choosing the respective subtype classes in step 2 and the respective compositions in step 3.

If the decisions in step 2 and 3 on the T-model subtype or the C-model composition are chosen randomly according to the corresponding discrete probability distribution defined in the models, the scene model draws samples from the prior probability $P(S = s)$. This fact is exploited for Monte-Carlo approximation in Section 6.

6 Matching Object Observations to Interpretation Trees

Object observations in the image can be either made by a human interpreter or by a computer vision system in a bottom-up process. In order to apply prior knowledge defined in the scene model of Section 4, it is necessary to determine a likelihood probability $P(O / S)$ for a set of observations given a candidate interpretation tree. To express the probability of mismatch in terms of the number of unmatched object observations n and the number of unmatched object nodes p , a heuristic likelihood function is chosen:

$$P(O = o_k | S = s_i) \propto \exp(-[\lambda \cdot n(o_k, s_i) + q(o_k, s_i)]) . \quad (2)$$

The parameter λ controls the balance of influence of both counts on the inference result. To determine the counts however, a matching between object observations and the object nodes of the interpretation tree has to be established. This has to be done based on the features of the observed objects and their spatial relations.

To approach this problem, the object nodes of the interpretation tree are rearranged as nodes of a graph with the connecting arcs representing their expected spatial relations. Expected features are represented as node attributes. Accordingly, object observations, their observed features and spatial relations are represented as a graph as well, based on a probabilistic object-oriented representation as described in [11]. This formulation relates the matching problem to general graph matching problems,

which have been extensively studied in literature [12]. In many cases, good and efficient approximations to the NP-complete matching problem have been achieved using relaxation labeling [13]. It is an iterative graph matching method using heuristic compatibility coefficients. One of the most appealing reformulations of relaxation labeling has been presented by Christmas, Kittler and Petrou [14] by deriving compatibility coefficients and update function in accordance with probability theory. However, their formulation is only suitable for continuous spatial relations such as distance, but not on discrete locative expressions such as nearness and adjacency. If a formally derived probabilistic relaxation scheme can be found, it might be possible to derive a formal definition of the likelihood probability, for example based on graph-edit distance [15].

In the context of this paper, relaxation labeling using heuristic compatibility coefficients and the likelihood function (2) has been used.

7 Inference

If the posterior distribution $P(S = s_i | O = o_k)$ is established, manifold inferences can be calculated. A specific class of inferences can be expressed as the expectation of an indicator function of the scene description variable S

$$E_{S|o_k} \{I_{\Psi}(S)\}. \quad (3)$$

The indicator function takes the value 1 if a specific condition on the interpretation tree, the object observations or the corresponding matching holds true; otherwise it is defined to be zero.

Selecting the next object class to search for in the image is a task, which can be supported by defining the indicator function $I_{\Phi}(S)$ to represent the condition that an unmatched object instance of object class Φ exists in the interpretation tree S . The expectation of that indicator function is equal to the probability of occurrence of the object class. The occurrence probabilities can be used to guide the image interpretation process for efficient establishment of a complete scene description.

The same indicator function is useful to determine the distribution of the root node of the interpretation tree, representing the overall classification of the scene, for example in the case of airfields, if it is a military airfield or a civil airport.

To classify an observed object based on its features and taking into account the occurrence of other object observations and their spatial relations, the indicator function can be designed to resemble the condition that the object of interest has been matched to an interpretation node associated to a specific object class model Φ . This way, the probability for the object to be interpreted as being of object class Φ is determined.

If the number of possible interpretation trees is large, such as in a comprehensive model of airfield scenes, calculation of expectations becomes intractable. However, using Importance Sampling, a Monte-Carlo estimation method, approximations can be calculated [16]. As the generation algorithm described in Section 2 is able to generate samples from S according to the prior distribution, the prior distribution is

used as proposal distribution for Importance Sampling. Respectively, the estimator for the expectation of a function $g(S)$ given the posterior distribution is defined as

$$\tilde{E}_{S|O}\{g(S)\} = \frac{\sum_{i=1}^n \omega(s_i) g(s_i)}{\sum_{i=1}^n \omega(s_i)} \quad (4)$$

using the weights

$$\omega(s_i) = \frac{P(S = s_i | O = o_k)}{P(S = s_i)} = P(O = o_k | S = s_i) . \quad (5)$$

The estimator (4) is independent of the normalization of the weights, so the definition of the observation likelihood probability (2) does not need to be normalized.

By previous experiments, it was found that a sample size of 10.000 is sufficient to estimate the expectations at a reasonable accuracy [3]. The Java™ implementation of the estimator is able to generate and process 10.000 samples per second on an Intel™ 2.1 GHz Core 2 CPU. As the sampling distribution is independent on the observations, samples can be reused for different sets of observations and do not have to be redrawn for each recursion step. Therefore, after the initial generation of samples, the calculation time is well below one second, which is acceptable for the application in decision-support systems.

8 Experiments

To study the feasibility of the presented method, scene models for the interpretation of airfield and harbor image have been developed in cooperation with image interpretation experts, each involving about 50 different objects classes. As ground-truth for the evaluation, airfields and harbors scenes labeled from aerial images. As a first step, to compare the benefit of different modeling aspects (unary features, global scene context, local object context and spatial relations), the respective classification accuracy has been determined for objects in 10 different airfield scenes. The unary feature in this experiment was the appearance-level class (building, paved area or antenna). In order to incorporate spatial relations, the relaxation labeling of Rosenfeld et al. [13] was used and the compatibility coefficients were chosen to be 1 in the case that the object's distance was below a fixed threshold, zero in all other cases. Figure 3 displays the results, which show that the classification accuracy is significantly improved when taking into account prior knowledge about different scene realizations and using additional binary features such as spatial nearness for the association of objects nodes in the model and objects observations in the image.

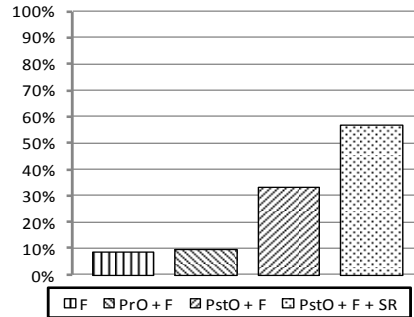


Fig. 3. Experimental result for the classification accuracy of single objects in airfield scenes using different levels of prior knowledge modeling. F: unary features and uniform prior occurrence probability, PrO+F: prior occurrence probability from scene model and unary features, PstO+F: posterior occurrence probability taking into account other objects in the scene and unary features, PstO+F+SR: Like PstO+F but using spatial relations (nearness) to objects in the local neighborhood in the relaxation scheme.

9 Conclusion and Outlook

For automated image interpretation, the problems of knowledge representation, hypotheses matching and inference have to be addressed, especially if higher-level semantic descriptions have to be extracted from the image. In this paper, probabilistic scene models are suggested to model prior knowledge about possible scene descriptions and their application in image interpretation using Bayesian inference is explained. Probabilistic scene models are defined in a human understandable way, allowing a human expert to determine the required parameters using compositional and taxonomic models even in the absence of training data. To match possible scene descriptions against object observations, a heuristic likelihood function is proposed and the use of relaxation labeling is suggested to establish a correspondence between object observations and a candidate scene description. Three exemplary inferences, which can be derived from the posterior distribution of scene descriptions given incomplete object observations, are proposed and their approximate calculation in the context of high-dimensional scene models using Importance Sampling is suggested. The evaluation of the object classification accuracy shows that using the proposed method, object classification significantly benefits from the consideration of prior knowledge about possible scene realizations and spatial relations between objects.

Future work will address the modeling of spatial relations in more detail for the application in aerial image interpretation of complex scenes such as airfields, harbors and industrial installations. Relaxation labeling methods will be evaluated and the integration of discrete locative expressions in the context of probabilistic relaxation will be investigated, based on a representative set of ground-truth labeled scenes. The benefit of an interactive decision-support system for image interpretation [2] based on the presented probabilistic scene models will then be evaluated on aerial images of airfields and harbors.

References

1. Matsuyama, T. and Hwang, V.: SIGMA: A Knowledge-Based Aerial Image Understanding System. Plenum Press (1990)
2. Bauer, A.: Assisted Interpretation of Infrastructure Facilities from Aerial Imagery. In: Proc. of SPIE, vol. 7481, 748105 (2009)
3. Bauer, A.: Probabilistic Reasoning on Object Occurrence in Complex Scenes. In: Proc. of SPIE, vol. 7477A, 74770A (2009)
4. Russ, T. A., MacGregor, R. M., Salemi, B., Price, K. and Nevatia, R.: VEIL: Combining Semantic Knowledge with Image Understanding. ARPA Image Understanding Workshop (1996)
5. Dillon, C., and Caelli, T.: Learning image annotation: the CITE system. Journal of Computer Vision Research, vol. 1(2), 90-121 (1998)
6. Hanson, A., Marengoni, M., Schultz, H., Stolle, F., Riseman, E. and Jaynes, C.: Ascender II: a framework for reconstruction of scenes from aerial images. Workshop Ascona 2001: Automatic Extraction of Man-Made Objects from Aerial and Space Images (III), 25-34 (2001)
7. Rimey, R. D. and Brown, C. M.: Control of Selective Perception Using Bayes Nets and Decision Theory. International Journal of Computer Vision, vol. 17, 173-109 (1994)
8. Lueders, P.: Scene Interpretation Using Bayesian Network Fragments. Lecture Notes in Economics and Mathematical Systems, vol. 581, 119-130, (2006)
9. Lin, L., Wu, T., Porway, J., Xu, Z.: A Stochastic Graph Grammar for Compositional Object Representation and Recognition. Pattern Recognition, vol. 42(7), 1297-1307 (2009)
10. Porway, J., Wang, K., Yao, B., Zhu, S. C.: A Hierarchical and Contextual Model for Aerial Image Understanding, In: IEEE Conference on Computer Vision and Pattern Recognition CVPR, 1-8 (2008)
11. Bauer, A. Emter, T., Vagts, H., Beyerer, J.: Object-Oriented World Model for Surveillance Applications. In: Future security: 4th Security Research Conference Karlsruhe; September 2009: Congress Center Karlsruhe, Germany. Stuttgart: Fraunhofer IRB Verl., 339-345 (2009)
12. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty Years of Graph Matching in Pattern Recognition. International Journal of Pattern Recognition and Artificial Intelligence, vol. 18(3), 265-298 (2004)
13. Rosenfeld, A., Hummel, R. A., Zucker, S. W.: Scene Labeling by Relaxation Operations. Systems, Man and Cybernetics, vol. 6(6), 420-433 (1976)
14. Christmas, W. J., Kittler, J., Petrou, M.: Structural Matching in Computer Vision using Probabilistic Relaxation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17(8), 749-764 (1995)
15. Myers, R., Wilson, R. C., Hancock, E. R.: Bayesian Graph Edit Distance. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22 (6), 628-635 (2000).
16. Koch, K. R.: Introduction to Bayesian Statistics. Second Edition, Springer (2007)