Hybrid Neuro-Bayesian Spatial Contextual Reasoning for Scene Content Understanding

Denis Garagic*, Majid Zandipour*, Frank Stolle*, Matthew Antone†, and Bradley J. Rhodes*

Abstract – Geospatial scene content understanding facilitates a large number of increasingly important applications. These range from tools to help intelligence analysts perform rapid, high-precision identification of urban scene content to other civilian and military security applications such as geospatial queries, functional object level change detection, and mission planning. In this paper, we present initial research results from a multi-faceted approach for determining and understanding scene content. Our approach performs context-dependent probabilistic reasoning on a set of object hypotheses obtained from a suite of individual object detection algorithms. This neurally-inspired reasoning approach improves the quality of object detections within a given scene and enhances scene content understanding by fusing low-level features, identified objects, high-level context, and spatial constraints to more accurately determine the nature of specific scene level targets. We present results from application of our hybrid spatial contextual reasoning approach to a set of objects automatically obtained from an urban scene by a suite of state-of-the art detection algorithms. We demonstrate that reasoning on the individual detector outputs produces improved precision-recall performance over using the detector outputs alone.

Keywords: Scene understanding, complex object recognition, associative learning, Bayesian networks, probabilistic reasoning.

1 Introduction

Military operations in urban areas require detailed understanding about the location and classification of common objects and spatial features. Learning spatial context for scene content understanding has been an enduring goal of computer vision researchers since the late 1960’s [7], and at present there is no single approach that can achieve this objective. In recent years, the role of contextual influences in object recognition has become an important topic, due both to the psychological basis of context in the human visual system [6] and to the object recognition algorithmic improvements that contextual influences have provided [12]. One of the most straightforward forms of representing the context of an object is in terms of its co-occurrence with respect to other objects. The work in [10] and [9] demonstrates the use of this context, where the presence of a certain object class in an image probabilistically influences the presence of a second class. While these methods achieve good results when many different object classes are labeled per image, they are unable to leverage unsupervised data for contextual object recognition. In addition to co-occurrence context, many approaches take into account the spatial relationships between objects. At the descriptor level, Wolf et al. [13] detect objects using a descriptor with a large capture range, allowing the detection of the object to be influenced by surrounding image features.

We have previously reported a successful approach to learning contextual influences on object recognition and scene understanding [14], [10]. In [14] we proposed a biologically-inspired algorithm for automatic scene understanding and complex object recognition which does not require any handcrafted a priori knowledge. It incrementally learns, with or without a priori knowledge, the associations and interdependencies between compound objects and their primitive components. In addition, the spatial relationships among the simple constituents and their probabilities of pair-wise occurrences are learned incrementally. Based on some of the conclusions drawn from our prior results, we have pursued several enhancements, including the complementary use of associative learning and Bayesian network paradigms to encode the preferences for certain spatial and co-occurrence relationships as well as providing greater capacity for scene interpretation and detail examination. This hybrid contextual reasoning approach improves the quality of object detections within a given scene and enhances scene content understanding, while offering all standard benefits of a graphical model formulation (e.g., well-known learning and inference techniques) [1], [3]. In this paper we present a hybrid spatial contextual reasoning approach, and demonstrate its performance on a set of individual object detections obtained from an urban scene by a suite of
state-of-the-art object detectors. We also report improvements in precision-recall obtained when our approach reasons over the raw detector outputs, as compared with the performance of the individual detectors alone.

2 Methodology

In this section, we present experiments and initial results from a multi-faceted approach for determining and understanding scene content. Our reasoning solution operates on results from state-of-the-art scene segmentation and object recognition algorithms, combining their detections through complementary use of associative learning and Bayesian networks. This neurally-inspired hybrid learning technique fuses low-level features, identified objects, high-level context, and spatial constraints to more accurately determine the presence and identity of specific scene level objects.

The first component of the proposed contextual reasoning algorithm, Object Probabilistic Associative Learning (OPAL) [14], is based on Probabilistic Neural Associative Incremental Learning (pNAIL) [10] and automatically discovers the conditional probabilities and hierarchical structure of elements comprising a scene. OPAL’s output is a belief map of possible scene types based on the set of simple objects present in a given scene. The second component of our hybrid contextual reasoning algorithm, a Bayesian network, uses this belief map to initialize learning of spatial relationships between constituent objects over the space of possible relationship networks using standard structure learning algorithms [3]. Since a Bayesian network constitutes a complete probabilistic model of the variables in a domain, the network contains all information needed to answer any probabilistic query about these variables. One of the benefits of this top-down reasoning strategy, which utilizes scene and spatial context information, is to provide a regularization mechanism to reduce ambiguity and false positives (misclassifications) that arise when classifying objects or scenes individually. The system iterates as information from below and above is propagated through the representation hierarchy and constraint networks as illustrated in Figure 1. Each pass of the algorithm updates the hypotheses that best explain the complete scene given the available information; thus, queries at any point during algorithm iteration will retrieve the best available information at that time. Details of the learning and inference modes of our hybrid spatial contextual reasoning approach are presented in sections 2.1 and 2.2.

2.1 Object Probabilistic Associative Learning

A fundamental role of this component is to represent objects as hierarchical collections of features, other objects, and scene context information [14]. To embody the diversity of object classes we need to support, for example, objects as specific as a ‘stop sign’ and as general as a ‘processing plant’. In our approach objects are represented as graphs whose nodes are features or other objects, and whose edges represent spatial or other relationships between nodes. Node associations – in the form of likelihoods – are incrementally learned, with or without a priori knowledge for each labeled object, from a small number of truthed data examples. Figure 2 illustrates a set of learned node associations representing the ‘Gas Station’ object as composed of spatial relationships with the ‘Gas Pump’, ‘Roof’, ‘Sign’, and ‘Small Building’ objects in the context of ‘Urban’ terrain. The output of the associative learning process is a belief map for a set of combined objects indicating the likelihood of possible

![Diagram](image_url)
scenes based on the set of simple objects present.

For each observation of a set of simple and scene-level objects, OPAL can make predictions about possible scenes with weights indicating probabilities of simple objects having previously been observed in each scene. However, there may be many non-zero weights emanating from the simple object nodes, resulting in a redundant network for which a closed-form statistical inference solution does not exist. In order to overcome this limitation of our OPAL algorithm, we use a graphical model formulation (e.g., Bayesian networks) to revise a scene’s belief map produced by OPAL. The a posteriori map generated by this network is then used as the final scene belief map.

2.2 Bayesian Networks for Spatial Context Belief Refinement

To learn context-sensitive models of spatial relationships among simple objects observed in a scene, we train a Bayesian network to learn spatial relationships between instances in an object-centric frame of reference. A Bayesian network [8] encodes the joint probability distribution of a set of \( v \) variables (e.g., objects in a scene) \( \{x_1, \ldots, x_n\} \), as a directed acyclic graph and a set of conditional probability tables (CPTs). In this paper we assume all variables are discrete, or have been pre-discretized. Each node corresponds to a variable (i.e., an object), and the CPT associated with it contains the probability of each state of the variable given every possible combination of states of its parents. The set of parents of \( x_i \), denoted \( \pi_i \), is the set of nodes with an arc to \( x_i \) in the graph. The structure of the network encodes the assertion that each node is conditionally independent of its non-descendants given its parents. Thus the probability of an arbitrary event \( X = \{x_1, \ldots, x_n\} \) can be computed as

\[
P(X) = \prod_{i=1}^{n} P(x_i | \pi_i) \prod_{i=1}^{n} P(O_i, f_i) .
\]

Low-level detectors report as detections all locations for which a conditional probability \( P(O_i, f_i) \) (i.e., a conditional probability that \( i \)th object is detected given the feature vector, \( f_i \)) is above a threshold chosen to give a desired trade-off between false positives and missed detections. Because OPAL unrolls into (e.g., initializes the structure and parameters of) a Bayesian network for each scene, we can use standard learning and inference methods. In particular, we learn the parameters of our Bayesian network model using the Expectation-Maximization (EM) [2] algorithm and perform inference using a standard variant of Markov Chain Monte Carlo sampling [1]. Furthermore, we learn the set of active relationships from a large candidate relationship pool using a structure search interleaved with the EM [2]. At test time, our system generates the a posteriori belief map as the final scene belief map.

**Learning Bayesian Networks:** Given a training set

\[
T = \{X_1, \ldots, X_t, \ldots, X_s\} \quad \text{where} \quad X_t = \{x_{1,t}, \ldots, x_{n,t}\},
\]

the goal of learning is to find the Bayesian network that best represents the joint probability distribution \( P(x_{1,t}, \ldots, x_{n,t}) \). One approach is to find the network \( W \) (i.e., the set of scene contextual relationships) that maximizes the likelihood of the data or its logarithm [3]:

\[
LL(W \mid T) = \sum_{t=1}^{s} \sum_{i=1}^{n} \log P_w (x_{i,t} \mid \pi_{i,t}) .
\]

In this paper we assume no known structure of contextual relationships in a scene, and focus on the problem of learning network structure and parameters. To achieve the best combination of accuracy and efficiency, we employ hill-climbing search initialized with a network constructed by OPAL, rather than an empty or random initial network. At each search step, hill-climbing creates all feasible variations of the current network obtained by adding, deleting, or reversing any single arc (i.e., probabilistic dependencies between objects in a scene). The best variation becomes the new current network, and the process repeats until no variation improves the score. We extend the log-likelihood scoring function by adding a complexity penalty. For example, we minimize

\[
MDL(W \mid T) = 0.5m \log n - LL(W \mid T) \quad [3],
\]

where \( m \) is the number of parameters in the network. Our learning process outputs an active set of relationships and the parameters of our model, i.e., learned conditional probabilities that encode the strengths of the probabilistic dependencies between objects.
3 Experimental Results

In this section we report the performance of our hybrid spatial contextual reasoning algorithm on a set of objects automatically detected in an urban scene by a suite of state-of-the-art recognition algorithms. Our system first ingests 3D LIDAR point data and extracts simple features such as color and elevation statistics on a multi-scale 2.5D grid. Layer-based processing on this grid allows characterization and clustering of large and coarsely-defined regions such as vegetation, buildings, pavement, and water, while raw 3D processing partitions the scene into dominant surfaces (e.g., walls, rooftops, and ground) and candidate object segments (e.g., ‘compact objects’ such as vehicles, fire hydrants, and lamp posts). The segments are further analyzed to determine plausible class likelihoods using a set of matching and recognition algorithms specialized according to overall object size and shape. Contextual information guides our overall object recognition strategy, conserving and focusing computational resources by restricting both search regions and plausible category lists. Context feeds down from above (e.g., we search for windows and doors only on walls and search for alleys among buildings) and feeds up from below (for example, the presence of pumps, vehicles, and an awning may indicate a gas station).

The main objective of our current experimental work is to assess the performance of our context-dependent probabilistic reasoning to reduce ambiguity and misclassifications caused by the individual object detection algorithms above. If an object was detected but not assigned a class label, we wish to determine whether contextual scene knowledge and observed evidence of other objects detected nearby can be used to classify this object. Formally, this can be defined as estimating a conditional probability $P(O_i = 1 | f_j)$, where $O_i = 1$ indicates the presence of the $i^{th}$ object and $f_j$ is a set of features extracted from the image. For training and testing we used an urban scene containing several thousand objects that were manually annotated with one of 42 category labels. The outputs of local object detectors, in conjunction with context-related features, served as inputs to the hybrid reasoning component. Here we compare the performance of two methods for the task of reducing the number of false positives: i) using low-level individual object detectors; and ii) using our hybrid contextual scene reasoning.

![Figure 3. Precision-Recall (PR) generated using Hybrid algorithm (dashed line) and Individual Object Detectors (solid line); (Top-left) PR curves for ‘Window’ object class; (Top-right) PR curves for ‘Trees’ object class; (Bottom-left) PR curves for ‘Street Light’ object class; (Bottom-right) POR curves for ‘Sidewalk’ object class.](image-url)
algorithm. As an alternative to ROC curves, we summarize these results using Precision-Recall (PR) curves, which can expose differences between algorithms that are not apparent in ROC space. We plot the PR curves produced by our hybrid algorithm against the curves produced by individual object detectors (see Figures 3–5), and observe that our hybrid contextual scene reasoning algorithm provided an improvement in accuracy for all but the ‘window’ category.

4 Conclusions

An important component of higher level fusion and decision making is knowledge discovery. One form of knowledge is a set of relationships between entities and their likelihoods of co-occurrence. We developed a hybrid scene contextual reasoning approach that enables scene understanding and complex object recognition. The OPAL component of our contextual reasoning approach is a probabilistic associative learning algorithm that automatically discovers the conditional probabilities and hierarchical structure of primitive objects comprising a scene. The Bayesian network component learns spatial relationships between objects in an object-centric reference frame. The discovery of co-occurrence and spatial relationships between objects from the truth data without a priori knowledge is an important characteristic of our approach. This knowledge discovery from data depends on the appearance of scene objects captured in the truth data. Results of our hybrid contextual reasoning approach show that the presence of certain neighboring objects within a learned vicinity of a candidate object in a scene narrows the possible class set and enhances correct predictions. In addition, we demonstrated improvements in object detection accuracy obtained when our approach reasons over the individual detector outputs compared to the accuracy obtained when only outputs from the individual detectors are used.

One immediate future task is to apply this reasoning algorithm to the detection and classification of complex objects with many common features. For instance, a street sign, a traffic sign, and a lamppost all have a thin cylindrical core with structures at the top and can lead low-level detection algorithms to produce ambiguous labels.

Figure 4. Precision-Recall (PR) generated using Hybrid algorithm (dashed line) and Individual Object Detectors (solid line); (Top-left) PR curves for ‘Door’ object class; (Top-right) PR curves for ‘Curb’ object class; (Bottom-left) PR curves for ‘Post’ object class; (Bottom-right) POR curves for ‘Sign’ object class.
Acknowledgements

The material presented in this paper is based upon work supported by the AFOSR under Contract No. FA9550-06-C-018 and by DARPA under Contract No. HM1582-07-C-0016.

References


