Inferring Context through Scene Understanding

Michael Walton and **Doug Lange, Ph.D.**
Space and Naval Warfare Systems Center Pacific

Song-Chun Zhu, Ph.D.
University of California Los Angeles

Abstract

In this work we propose the application of scene understanding methods using probabilistic graphs to the problem of learning and representing computational context. Many machine learning methods treat their input space as a 'bag of features'; that is, inference is accomplished by taking some global aggregation which may be engineered or learned. A limitation of this approach is that many models lack an explicit representation of semantic and physical relationships such as the spatial, temporal, semantic and fluent context of objects and events in a scene. In this work, we relate the problem of capturing computational context to the scene parsing literature. We then propose a method of representing and learning contextual relationships from data using stochastic grammars implemented as And-Or Graphs with applications to the Naval Tactical C2 Domain.

Introduction

Command and Control (C2) systems must effectively represent information to a mission commander, leading to rapid and correct decisions. Key goals of such systems include providing an interface to force {projection, readiness, employment}, intelligence, and situational awareness. Often, the approach to these interfaces is to present information in a manner that is natural for humans to interpret, that is, visual displays of information (Bemis, Leeds, and Winer 1988) (John et al. 2004) (Smallman et al. 2001). Such displays apply across a variety of domains, from mission planning to tactical command and control, because not only are they intuitive for a human to understand but also because such displays easily represent contextual relationships among objects (Smallman et al. 2001). The challenge moving forward is the increasing complexity of the battle space increases the difficulty of creating effective windows to the required information for a commander (John et al. 2004). Because the information density is increasing, important information and relationships hide from human view and integrating information for planning becomes difficult. The question is, can we augment or replace human capability with computer systems that can discover the relevant, highly context-sensitive relationships and information that a commander requires?

Scene understanding is a sub-field of computer vision which is concerned with the extraction of entities, actions and events from imagery and video. In many ways, scene understanding may be seen as a holistic or gestalt approach to computer vision which incorporates multiple vision tasks into a common pipeline. Traditional computer vision challenges include low-level feature extraction [1], image segmentation [2] and object classification [3]. Additionally, scene understanding is concerned with the semantic relations between visual objects. Each of these objectives shares direct analogs to the Naval tactical domain and may be extended far beyond the domain of images and video. We posit the idea that many of challenges faced in scene understanding are shared with the problem of efficient contextual inference and representation in C2. Therefore, we propose that adaptations of the same algorithms and data structures which have been proven effective for scene understanding applied to imagery may be similarly applied to representing Naval tactical context.

Methods

Scene Parsing

A particularly well-studied approach to the problem of natural scene understanding is scene parsing (Yao et al. 2010). This area of research strives to encode the semantic information of an image as a parse tree, or more generally a parse graph. Associated algorithms for manipulating and traversing these structures enable sophisticated capabilities such as causal inference (Fire and Zhu 2015) (Pei, Jia, and Zhu 2011), natural language text generation (NTG) and answering queries about the content of an image or video

[1] The representation of raw, high dimensional input images in a feature space via either learned or engineered features. Common engineered methods include SIFT descriptors, various color and gradient histograms. Feature learning has been broadly explored in diverse research domains and a fair discussion is beyond the scope of this proposal

[2] For example, partitioning the set of pixels which comprise an image of a bicyclist into two unlabeled groups

[3] In which a segmented object in a scene is mapped to some high-level semantic concept. For example consider the labeling of pixels which compose a bicycle and cyclist with distinct classifications

(Tu et al. 2013). Working from the analogy to natural language processing initially drawn in (Yao et al. 2010), image parsing computes a parse graph which represents the most likely interpretation of an image. Extending this analogy into the naval tactical domain, tactical scene parsing computes the most likely interpretation of the available Intelligence, Surveillance, Reconnaissance (ISR) information. Notionally, a tactical parse tree is a structured decomposition of the totality of a ship or battlegroup's ISR datasources [4] such that all input feeds are explained. A "tactical parse graph" is subsequently augmented with lateral edges allowed at all levels of the hierarchy which specify spatial and functional relations between nodes.

AOG Knowledge Representation In statistical natural language processing (NLP), a stochastic grammar is a linguistic framework with a probabilistic notion of grammatical wellformedness and validity (Manning and Schtze 1999). In general, a grammar is a collection of structural rules which may generate a very large set of possible configurations from a relatively small vocabulary. AOG is a compact yet expressive datastructure for implementing stochastic grammars. It is observed in (Zhu and Mumford 2006) that a stochastic grammar in the form of an AOG is particularly well suited to scene parsing tasks. An AOG is constructed by nodes where each Or node has child nodes corresponding to alternative sub-configurations and the children of an And node correspond to a decomposition into constituent components. Intuitively, this recursive definition allows one to merge AOGs representing a multitude of entities and objects into larger and increasingly complex scene graphs. Theoretically, all possible scene configurations could be represented by composition of all observed parse graphs in a dataset. Therefore, the AOG is a compact formalization of the set of all valid parse graph configurations that may be produced by the corresponding grammar.

The lateral edges in an AOG correspond to relations which allow the graph to encode contextual information between entities at all levels of the hierarchy and-or tree subgraph. These edges form subject predicate object (SPO) triples that are suitable for extraction from, or decomposition to RDF triplestores. The relations may be distance based, geometric or semantic. Distance based relations may be of the form A near B; similarly geometric relations may span large distances but encode complex relationships regarding the arrangement of entities in a scene (for instance C collinear with D, E concentric to F). Semantic or functional relations encode abstract information about entities in the scene. Examples from the imagery domain include "boat carrying covered cargo" or "person holding firearms"; in the tactical domain these could represent essential contextual details that could be easily lost in the face of overwhelming information density such as "ship traffics drugs" or "agent is hostile". Development of efficient means of context-sensitive

[4]For now, the reader may conceive of these data as the extracted entities and relations that may be rendered in some form of C2 display which directly interfaces with an operator or analyst. Further discussion of the proposed input sources can be found in Section 1.2.1

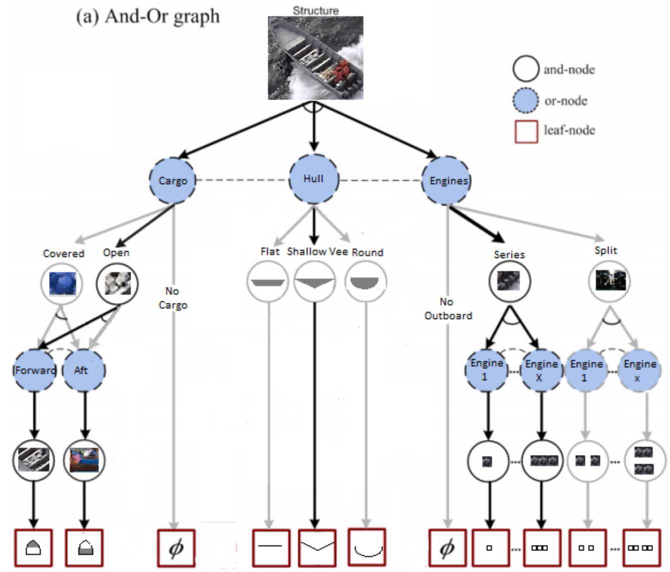

Figure 1: (a.) Example of an And-Or Graph expressing a stochastic grammar for boats (Yao et al. 2010). The And nodes in the graph specify the feature dependencies to satisfy a particular sub-grammar; for example a two handed clock must have cargo in the Forward and aft sections of the boat. The Or nodes indicate alternate sub-configurations for a particular attribute, for example a boat may have a round hull or a shallow vee hull.

inference, anomaly detection and operational summarization on these data for C2 is an open applied research question that would benefit greatly from the application of tactical scene parsing to Navy datasources.

Parse Graphs In the image parsing framework proposed by Zhu (Zhu and Mumford 2006), AOG stochastic grammars may be learned from data represented as parse graphs. Zhu and colleagues assign energy terms to each node of the AOG which may be drawn from expert knowledge or learned from data.[5] Intuitively, the energy of each Or node is inversely proportional to the likelihood of each alternative configuration; similarly, the energy associated with compositional And nodes with respect to their children captures the uncertainty associated with these relations. A parse graph is a labeled directed graph where each node corresponds to an entity with some semantic attribute indicating its type. In addition to discussing AOGs as a compact representation of the space of valid configurations, the graphical model specifies a prior probability distribution over parse graphs. Inferred parse graphs may then be derived by maximizing a posterior which is proportional the prior defined by the AOG and the uncertainty of generated candidate scene parses (Tu et al. 2005).

In more recent literature, Zhu and colleagues have begun to pioneer extensions to this framework to incorporate

[5]The generalization of these methods to Naval data is an essential research question

multiple datasources as well as forming augmentations to their models for representing temporal relations as well. For instance, the Temporal And-Or Graph (T-AOG) represents a stochastic event grammar in which the terminal nodes are atomic actions and the vocabulary consists of a set of grounded spatial relations such as positions of agents and interactions with the environment. The associated algorithms have been used to successfully detect actions and fluent [6] transitions (Pei, Jia, and Zhu 2011) as well as infer causal relationships as suggested in (Fire and Zhu 2015).

Causal Inference & Intent Prediction Understanding perceptual causality is a crucial capability of humans which enables us to infer hidden information from sensory inputs as well as make predictions of future events in the environment. This is further compounded in the tactical domain in which a commander must make correct, efficient C2 decisions from the estimated tactical scene. As a motivating example, consider a case where we can directly observe some container but not its contents. Although the content of the container may not be directly observable, we can infer its contents through observing interactions with the object: perhaps in a video, a person recovers a metallic cylinder from inside; in the context of a pleasure craft, this may drive conclusions that the cylinder is a soda can, the container is a cooler, which likely contains other food and beverage items etc. Independently, a spatial parse of this scene may draw conclusions regarding 'person' 'box' 'drink'; similarly, a temporal parse may detect the action 'opening box'. However from the causal joint of these perspectives, richer associations may be made (eg. 'person has drink'). Further, this Pei et. al (Pei, Jia, and Zhu 2011) demonstrate that parses of this form also enable the inference of the entirely unobservable *intent* governing an agents' actions (such as 'person is thirsty').

Spatial AOGs (S-AOG) (Zhao and chun Zhu 2011) and temporal AOGs (T-AOG) model the spatial decomposition of scenes and the temporal decomposition of events into atomic actions respectively. Similarly, a causal AOG (Fire and Zhu 2015) (C-AOG) models the causal decomposition of events and fluent changes. Correspondingly, a STC-AOG jointly models all three perspectives in an interconnected graph. In much of Zhu and collaborators work on STC-AOGs a taxonomy is formed with a universal node type as the root and all considered objects, events and fluents defined by their respective ontologies as subtrees. This shared structure is crucial for computing semantic similarity between concepts in the case of joint inference.

Joint Parse Graphs & Inference In joint parsing tasks across parse graphs generated from multiple data sources, three types of challenges arise and criteria must be derived for resolving: coreference, deduction and rejection (Tu et al. 2013). Coreferrence refers to the procedure by which multiple references to a singular entity are associated across multiple parse graphs. In the case of an identified `smuggler_ship` and a separate identification of `recreational_vessel` , both of these should be associated with references to the higher-level classification `ships`. Related semantics for singular entities are detected and resolved by their ontological similarity. For example, the `Ship` type is a parent of `Smuggler_ship` and `recreational_vessel` and entities of these types should possess strong semantic similarity; weighted similarity measures may be defined for each edge in the ontology. For real world scenarios, in which multiple text, video, and picture inputs must be incorporated, the treatement of correferrance is done post parsing of the inputs and involves finding nearest common parent nodes.

Scene parsing methods make the open world assumption (Russell and Norvig) and are not constrained to inference based solely on the current state of the environment, rather these models enable complex deduction by incorporating a probabilistic notion of actions and outcomes. Concretely, deduction is accomplished by inserting candidate subgraphs into the joint parse graph created by the STC-AOG. We only consider inserting subgraphs that increase the likelihood of the joint parse graphs' occurrence. Inserting subgraphs also increases the energy of the joint parse graph, by applying an energy threshold that can be added from a given deduction constrains the amount of deduction that can be performed. At times, several possible deduced parse graphs will fall within our energy threshold, in this case we need to limit the number of deductions performed by limiting the total entropy change of the parse graph deduced given an initial parse graph energy. Equation 1. forms the basis for constraining the iterative process of deduction.

$$H(pg_{de}|pg_{jnt}) = -\sum_{i=1}^{N} P(pg^{i}_{de}|pg_{jnt}) logP(pg^{i}_{de}|pg_{jnt}) > \frac{logN}{c} \quad (1)$$

Here, the entropy of a particular deduction pg_{de} given a joint parse graph pg_{jnt} represents the parse graph before the insertion of the deduced subgraph is bounded by the number of candidate subgraphs N and a hyperparameter c. Scene parsing algorithms will continue inserting low energy deductions until this threshold is satisfied.

Information received by an operator may also be conflicting, for instance when initially classified smugglers boats adjust course and speed toward the strike group. Revision is performed to resolve conflicts in the STC-AOG. In this example, multiple reports may place the same object as neutral in some parse graphs and foe in others. This violation of their tactical AOG renders this an impossible occurrence necessitating revision of one or more sub parse graphs. Changes to each parse graph will increase its associated energy, therefore scene parsing methods typically enforce minimal revisions by setting a threshold.

Probabilistic Modeling The prior probability of a parse graph is inversely proportional to the energy present in that parse graph. In the following discussion, we will freely change between probability and energy to simplify the mathematical expressions, but equation 2 shows us that the prob-

[6] A logic and artificial intelligence term indicating some condition or attribute of agents or environment which can change over time

ability of a parse graph and its energy can be easily interchanged using the relation:

$$P\left(pg\right) = \frac{1}{Z}e^{-E_{STC}\left(pg\right)} \tag{2}$$

Where Z is the normalization factor, and $E_{STC}\left(pg\right)$ is the energy of that parse graph in the STC-AOG. To calculate the energy of a parse graph for an STC-AOG we sum up the energy of each individual graph and the energy incurred by joint interactions.

$$E_{STC}\left(pg\right) = E_S\left(pg\right) + E_T\left(pg\right) + E_C\left(pg\right) + \sum_{r \in R^*(pg)} E_R\left(r\right) \tag{3}$$

Here the terms $E_S\left(pg\right), E_T\left(pg\right)\ and\, E_C\left(pg\right)$ are the energy terms defined by the spatial, temporal and causal AOGs respectively. In Zhu's notation $R*$ is the set of relations across the spatial, temporal and causal domains. Each AOG has a parse graph energy defined by the sum of the energy associated with the configuration selected at the or node $E_{or}\left(v\right)$ and the energy associated with a relation between and nodes $E_R\left(r\right)$.

$$E\left(pg\right) = \sum_{v \in V^{or}(pg)} E_{or}\left(v\right) + \sum_{r \in R^*(pg)} E_R\left(r\right) \tag{4}$$

From this general definition of total energy for a parse, we may specialize these models by defining the energy of their constituent nodes uniquely for the in the spatial (Zhu and Mumford 2006)(Tu et al. 2005), temporal (Pei, Jia, and Zhu 2011), and causal domains (Fire and Zhu 2013).

Answering Questions A joint parse graph is a structured, semantic representation of the objects, events and fluents present in a data set. These joint parse graphs can be used in semantic queries in order to answer natural language questions about the state of the tactical scene. These questions may vary in complexity and include dependencies on scene parsing algorithms' unique capabilities for entity resolution, joint inference of partially observable information, deduction and prediction. The joint parsing methods developed in (Pei, Jia, and Zhu 2011) are capable of generating multiple parses of a single scene. The multiple parse graphs correspond to different interpretations of the same scene; therefore to answer a question accurately, multiple interpretations may be combined to determine the probability of a particular interpretation $P(a)$ by summing the posterior probability $P(pg)$ of each parse graph where pg implies interpretation a; here this is denoted with an indicator function $\mathbb{1}(pg \models a)$

$$P(a) = \sum_{pg} P(pg)\mathbb{1}(pg \models a) \tag{5}$$

In (Pei, Jia, and Zhu 2011) Pei et al. propose and demonstrate a user-facing query engine for interacting with and extracting critical information from parse graphs. A natural language query is composed by the user and entered into a web-application GUI front-end. Text input is parsed into SPARQL queries which are compared against RDF decompositions of joint parse graphs. Responses to the queries are then presented to the user in the GUI with the associated marginal probability, interpretable as a confidence, of the response.

Quantitative metrics such as ROC (Receiver Operating Characteristic), and associated precision and recall measures are suitable and commonly used in measuring performance in prediction or query answering problems. Precision is proportional to the number of relevant objects and relations indicated by the scene parse with respect to ground truth. This quantity degrades when superfluous information is included in the graph. Recall is proportional to the number of objects and relations present in both the ground truth and the scene parse. Questions may be of the form of 'Who What When Where or Why' and answers may be boolean valued or return the empty set or elements of the STC-AOG ontology dependent on the query.

Generating Text Summaries In (Yao et al. 2010), (Tu et al. 2005) scene parsing methods are used to summarize and annotate natural images using natural language text generation. In this scenario, a parse graph is mapped to natural language sentences that maximally explain the input. This differs from question answering scenario in that the models' outputs are non-interactive and express the full content of the parse graph. Such human readable summaries may be appropriate for rendering directly in C2 displays, or for reporting and automated brief generation. Evaluation of these text summaries may proceed similarly to the metrics proposed for evaluating query responses. Tu et al. (Tu et al. 2005) collected a set of human annotations and used the most frequently occurring relations as ground truth. Similarly, we may generate experimental ground truth parses of tactical inputs where suitable. This methodology will enable the possibility for incorporating human expert generated context annotations as ground truth.

Tactical Scene Classification It is readily apparent the a traversal of an AOG from the leaves to the root correlate with higher levels of semantic abstraction. An open question for more fundamental investigation is how higher-level information may be incorporated as indications and warnings into C2 interfaces. For example, it may be possible to deduce agent intent, threat assessment or other entirely unobservable information from joint parse graphs. Deductive reasoning capabilities enabled by scene parsing plays a key role in these use cases, for instance:

- Parses of text reports may indicate the exchange of funds for weapons and explosives with an individual
- This individual may be an owner of, or in some way associated with a small boat located nearby
- Various electronic intelligence may indicate the same vessel is inbound
- Imagery may indicate this vessel has large containers onboard.

In isolation, none of this information is particularly useful. However, jointly parsing across these otherwise disjoint analytic pipelines could produce a very clear indication which could be displayed directly to the operator without explicit queries. By selecting and incorporating the contextual information surrounding a particular ship, one may infer its intent and deduce likely future behaviors. Furthermore, using the scene parsing methods we propose, all logical premises leading to a conclusion would, by design and technical necessity, be associated with a conditional likelihood.

Conclusion

Methods of inferring the context of events and objects from sensor measurements are critical in the Naval tactical command and control domain. In the field today, much of this inference is left to human operators which is largely dependent on their training, experience and intuition. In future systems, we hope to provide a means of assisting and automating this process to inform operators' assessment of increasingly complex and rapidly changing battlespace. Scene parsing using And-Or graphs provides a promising and efficient means of representing, querying and summarizing complex contextual relationships from ISR inputs.

References

Bemis, S. V.; Leeds, J. L.; and Winer, E. A. 1988. Operator performance as a function of type of display: Conventional versus perspective. *Hum. Factors* 30(2):163–169.

Fire, A. S., and Zhu, S.-C. 2013. Using causal induction in humans to learn and infer causality from video. In *COGSCI*.

Fire, A., and Zhu, S.-C. 2015. Learning perceptual causality from video. *TIST* 7(2):1–22.

John, M.; Manes, D. I.; Smallman, H. S.; Feher, B. A.; and Morrison, J. G. 2004. Heuristic automation for decluttering tactical displays. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 48(3):416–420.

Manning, C. D., and Schtze, H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Pei, M.; Jia, Y.; and Zhu, S.-C. 2011. Parsing video events with goal inference and intent prediction. In *2011 International Conference on Computer Vision*. Institute of Electrical & Electronics Engineers (IEEE).

Russell, S., and Norvig, P. Artificial intelligence: a modern approach (2nd edition).

Smallman, H.; John, M. S.; Oonk, H.; and Cowen, M. 2001. Information availability in 2d and 3d displays. *IEEE Comput. Grap. Appl.* 21(4):51–57.

Tu, Z.; Chen, X.; Yuille, A. L.; and Zhu, S.-C. 2005. Image parsing: Unifying segmentation, detection, and recognition. *Int J Comput Vision* 63(2):113–140.

Tu, K.; Meng, M.; Lee, M. W.; Choe, T. E.; and Zhu, S. C. 2013. Joint video and text parsing for understanding events and answering queries. *CoRR* abs/1308.6628.

Yao, B. Z.; Yang, X.; Lin, L.; Lee, M. W.; and Zhu, S.-C. 2010. I2t: Image parsing to text description. *Proceedings of the IEEE* 98(8):1485–1508.

Zhao, Y., and chun Zhu, S. 2011. Image parsing with stochastic scene grammar. In Shawe-Taylor, J.; Zemel, R. S.; Bartlett, P. L.; Pereira, F.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc. 73–81.

Zhu, S.-C., and Mumford, D. 2006. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision* 2(4):259–362.