Learning to Eliminate Background Effects in Object Recognition

Robin R. Murphy
Department of Mathematical and Computer Sciences
Colorado School of Mines
Golden, CO 80401-1887
rmurphy@mines.colorado.edu

Introduction

Model-based vision relies on the interpretation of observations of component features in order to identify an object. Computer vision systems observe features, which provide evidence for the existence of the whole object. However, events external to the vision sensor can produce misleading evidence which in turn can lead to false recognition. For example, in military target acquisition, the combination of the weather and the terrain frequently creates a reflection of the sun off a lake resulting in an observation with properties similar to a real target and leading to a false recognition. These "background effects" can be eliminated by the addition of contextual knowledge to enable the interpretation process to identify and eliminate false observations. Unfortunately, recognition systems which exploit contextual knowledge typically have two limitations. First, the manpower and expertise involved in anticipating and encoding the possible types of confounding effects is a major obstacle in building the requisite knowledge base. Second, the knowledge is usually embedded in the object model, resulting in an object and environment specific system that cannot be readily extended or transferred to new domains.

One approach to overcoming these limitations is to develop a generic model-based recognition system which is capable of *learning* a significant portion of the contextual knowledge needed to eliminate background effects. This paper describes work in progress at the Colorado School of Mines in conjunction with IBM Federal Systems Corporation on such an object recognition program which identifies deficiencies in its own knowledge, notifies the operator who will construct an explanation for the occurrence of the background effect, and use explanation-based learning to generalize the result for other situations and/or objects. The intended application is target acquisition, however the results of the project are expected to be extensible to other domains.

This work is being supported by a grant from the Colorado Advanced Software Institute.

Our Approach

A twofold approach is taken to learning contextual knowledge for discerning background effects. First, a distinction is made between the role of contextual knowledge in *interpreting* evidence and the *generation* of evidence from observations of features in the object model. This distinction is formally expressed by context-free perceptual and observational grammars. The grammars are expected to simplify what has to be learned, aid in abduction, and to permit the object model database to be extended. The second aspect of our approach is to utilize the principles of explanation-based learning (EBL), while taking advantage of the availability of an operator to aid in constructing the explanation of why certain observations are misleading. Operator assisted EBL is a natural fit for this project because the learning system is expected to both update the knowledge base with the cause of a specific background effect from the explanation given by the operator, as well as to generalize it to other situations. The grammars and how they provide a structure for learning contextual knowledge in perception is the focus of this paper.

Perceptual Grammar

The perceptual and observational grammars developed in [10] formalize the composition of the percept and how to interpret evidence, thereby supplying a structure suitable for learning contextual knowledge. They are extensions of traditional model-based representations used in computer vision. Traditional techniques typically attempt to model a percept in terms of component features. The composition of the percept is expressed with some variant of a directed, acyclic graph (DAG), most notably Bayesian belief nets [2], influence diagrams [7], or relational graphs [4, 5]. The desired percept is the root of a graph with component features as the children. DAG models exploit the advantages associated with graphs: compact storage, ease of traversal, and the ability to append information (such as evidential contribution) to vertices. DAGs also have advantages specifically for modeling: they represent percept compositions, support object instances as a "copy"

$p = s_p$

p ::= percept_connective d d^+

d ::= description_connective ⟨component⟩⟨component⟩$^+$

⟨component⟩ ::= ⟨feature⟩|d|p

⟨feature⟩ ::=

primitive|⟨extended⟩|⟨relational⟩|⟨compound⟩|⟨functional⟩

⟨extended⟩ ::= extending_connective ⟨component⟩

⟨relational⟩ ::= relational_connective ⟨component⟩⟨component⟩$^+$

⟨compound⟩ ::= compound_connective ⟨component⟩

⟨functional⟩ ::= functional_connective ⟨component⟩⟨component⟩$^+$

Figure 1: Production rules for G_p.

of a DAG for a class of objects, and allow links to viewpoints.

Unfortunately, current instantiations of DAGs are ad hoc and have other disadvantages. The DAG models are rigid; they are either developed for a single perceptual task ignoring the need to model different uses (for example, a model sufficient for recognizing a coffee cup may not be sufficient for grasping the cup) or monolithic ignoring the role of selective perception. The models tend to be committed to a particular set of primitives (such as generalized cylinders). Many modeling techniques favor strictly geometric features and restrict relationships to IS-A and PART-OF hierarchies, excluding other types of features (e.g., qualitative, functional).

The graph representation was extended and encapsulated into a context-free perceptual grammar in [10], leading to a DAG capable of supporting a broader feature palette, selective perception, and sensor fusion. The underlying concept of the perceptual grammar, G_p, is that a percept can be expressed in terms of *descriptions*. A description is a collection of features which are necessary and sufficient for the accomplishment of a particular sensing objective; therefore each description is selectively perceptive. A feature is not limited to a particular type; the grammar recognizes primitives, such as edges and regions; relational features, such as *bigger, adjacency, distance, compactness, above,* and *beside* [1]; compound features, including grouping by perceptual organization techniques [8]; and functional features, for example "sitability" [12]. The family of descriptions of an object for all sensing objectives is called the *percept model.* More than one description can be used to represent the percept for a particular sensing objective.

The grammar is given by $G_p = \langle N_p, \mathsf{T}_p, P_p, s_p \rangle$ where: N_p, is the set of nonterminal elements or features; T_p, the set of terminal elements, are the connectives (or sensing algorithms), and the basic feature, primitive; P_p, are the production rules; and s_p, the start element, $s_p \in N_p$, is the percept: p. The production rules, P_p, are given in extended Backus-Naur Form using prefix notation in Figure 1.

Observational Grammar

G_p must be supplemented because it suffers from the same difficulties as traditional DAG based methods in representing the transformation of evidence for features in a model to evidence for a percept. DAG-based techniques such as [2, 7] attempt to embed the evidential contribution of a feature to the overall belief in a percept directly at each feature node in the DAG. This is challenging because the evidential contribution of a feature may depend on the *context.* The description of the sensing objective affects the belief in the percept because some models are "better" than others. A description of a MY COFFEE CUP as a set of 3D surfaces may be sufficient for recognition but a description using 2D features may be more distinctive. The choice of sensors and algorithms will also influence the belief in the percept. For example, the Sobel and Marr-Hildreth operators are both edge detectors but give different results when applied to the same image. The surrounding environment also plays an important role in determining the evidential contribution of a feature. Consider recognizing MY COFFEE CUP, which is the only cup in the office that particular shade of blue with a cracked handle. There may be many coffee cups the same shape and size in my office. Since it known to be the only cup that color, blue contributes more evidence that it is MY COFFEE CUP than other features in the description such as size, shape, and cracked handle. On the other hand if MY COFFEE CUP is at home where there are other cups the same color blue, the evidential contribution of the color diminishes while the contribution of observing the cracked handle increases.

Traditional DAGs attempt to accommodate the impact of the perceptual context on evidence by adding new nodes and links to the model to reflect the probabilistic conditioning of the evidence on the evidence for the current context. Unfortunately this has many undesirable side-effects. More nodes clutter the model and add computational complexity to the evaluating the graph. Also, it is impractical to determine the conditioning effects in advance and generate the requisite probability functions. More relevant for this paper is that attempting to insert context knowledge at the feature level complicates learning because it merges the contributions of observation and context for recognition, obscuring the source of any errors in perception.

The shortcomings in the perceptual grammar were addressed in [10] which developed an *observational grammar* isomorphic to the perceptual grammar but capable of evidential management and propagation. The observational grammar produces a DAG whose structure corresponds to the composition of the percept given by the G_p by limiting the introduction of contextual influences.

The observational grammar G_o considers context as a mechanism for *interpreting*: given that it is known MY COFFEE CUP is the only blue one in the office and that the region of interest is the office, blue counts more for

⟨percept⟩ = s_o

⟨percept⟩ ::= percept_interpretation⟨description⟩ ⟨description⟩$^+$

⟨description⟩ ::=

description_interpretation⟨evidential_feature⟩⟨evidential_feature⟩$^+$

⟨evidential_feature⟩ ::=

feature_interpretation feature_observed feature_expected

Figure 2: Production rules for G_o.

recognition in that situation. Interpretation of the evidence from the observation of a feature can be viewed as different frames of discernment (i.e., frames of evidential reference). The grammar defines three such frames:

- *Feature: how much belief there is for a feature in the model.* The belief in the expected feature given the physical operating characteristics of the particular sensor and algorithm being used to observed it is computed for each feature according to an feature_interpretation function.
- *Description: how the belief in each feature is transformed into belief for a description of the percept.* The role of each feature in the description may be different from its role in the composition of the percept (e.g., blue may count more). The description_interpretation function factors in the different evidential contributions of each feature.
- *Percept: how the belief in a description translates into belief for the percept and its sensing objective.* Since some descriptions offer more evidence than others, the belief for the description may need to be weighted. This is done through a description_interpretation function.

Essentially the feature_interpretation functions compute how well what is observed matches the expectations set forth by the model, while the description_interpretation and percept_interpretation functions account for the context. It should be noted that the observational grammar is independent of evidential updating technique; the interpretation functions can be implemented with Bayesian updating using causal matrices as per [11] or by enlargement and refinement of frames of discernment with Dempster-Shafer theory as discussed in [9].

The resulting grammar, $G_o = \langle N_o, \mathsf{T}_o, P_o, s_o \rangle$, can be expressed as: N_o, is the set of nonterminal elements, or types of evidence; T_o, is the set of terminal elements, or interpretation functions and the feature_observed and feature_expected; P_o are the production rules; and s_o is the start element, $s_o \in N_o$: *percept* observation. The production rules, P_o, are given in extended Backus-Naur Form using prefix notation in Figure 2.

Advantages for Learning

Using G_p to model a percept has all the advantages previously ascribed to DAG representations in computer vision, plus more. It can support any type of feature through connectives which correspond to sensing algorithms. The grammar is recursive, allowing complex percepts to be constructed from other percepts. Each description is a de facto model of the percept for a sensing objective, therefore it is inherently supports selective perception.

The dichotomy of the representation into composition and interpretation by G_o is expected to be advantageous for learning how to modify the evidential contributions of sensors based on context. Learning background effects should be simplified because only the interpretation functions will be operated on, not the percept models, permitting straightforward credit assignment.

How the grammars and their associated DAGs are actually implemented will also play an important role in automating learning of contextual effects. The implementation is expected to follow [10] where the percept model DAG consists of frames (feature nodes) linked by pointers (edges). The description and percept interpretation functions are a set of rules attached to the percept model itself. An executable sensing plan is built by examining the percept model DAG and then selecting the appropriate description(s) sub-trees to meet the sensing objectives for the projected environmental conditions and sensor availability. The selected description is itself a DAG. The sensor and sensing algorithms are chosen and their sensing procedures and associated feature interpretation functions are placed in slots at the feature nodes in the DAG. The DAG now satisfies G_o. The sensing plan is executed by a depth-first evaluation of the DAG.

The frame implementation for the model and sensing plan makes it easy to examine interpretation functions, change them, and add more. Also, procedures for using active sensing to disambiguate contextual effects can be attached to the frames. Frames are expected to contain slots with the justification for rules which can be examined by the operator and learning system in order to construct explanations. Additional structures will be needed for linking features, sensors, and conditions so that a learning system can generalize the impact of background effects to other percepts and sensors.

EBL

The grammatical representation and supporting data structures described above should enhance learning of contextual knowledge about perception. Operator assisted explanation-based learning was chosen as the learning technique because the data fusion system is expected to both update the knowledge base with the cause of a specific background effect from the explanation given by the operator, as well as to generalize it to other situations.

The major issues in creating an effective explanation-based learning system for eliminating background effects are: *How does the system operationalize the oper-*

ator's explanation into an internal representation? The system must be able to transfer the operator's understanding into whatever internal representation is used by the evidential reasoning component. Although a natural language interface is outside of the scope of this project, the choice of interface should not burden the operator with understanding the internal representation. *How does the system insure that a new explanation will not introduce errors in other parts of the knowledge base?* The representation must support discerning the impact of an explanation and ask operator for a further clarification if needed. *How does the operator and system communicate?* Given that the operator should be shielded from having to know the internal workings of the fusion system, how can the operator communicate the explanation? Likewise how does the system express the side effects that will be introduced by an inadequate or ambiguous explanation?

Summary

Learning the role of contextual knowledge in interpreting perceptual information is necessary for eliminating false recognition. Work in progress on a system for learning to eliminate background effects in target acquisition using contextual knowledge is currently concentrating on the appropriate modeling of the percept and related knowledge. The proposed DAG representation is expected to be extensible to other applications of learning in computer vision. Future work will concentrate on building operator assisted explanation-based learning mechanisms to act on these knowledge representations.

The DAG is formally expressed by two context-free grammars. The perceptual grammar G_p is concerned with modeling the percept in terms of abstract descriptions and features. The descriptions and features act as building blocks for sensing plans and are similar to logical sensors [6] and equivalence classes [3]. The observational grammar G_o expresses the evidential contribution of sensor observations of the percept, essentially overlaying the structure given by G_p. It incorporates physical models of imprecision and contextual knowledge as three interpretation functions, rather than as conditional probabilities. Learning will impact the interpretations functions, not the object description and features.

Acknowledgments

This research is being conducted under a CASI Technology Transfer grant in conjunction with IBM Federal Systems Corporation. CASI is sponsored in part by the Colorado Advanced Technology Institute (CATI), an agency of the State of Colorado. CATI promotes advanced technology education and research at universities in Colorado for the purpose of economic development.

References

[1] Ballard, D. H., and Brown, C. M., *Computer Vision*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[2] Binford, T.O., Levitt, T.S., and Mann, W.B., "Bayesian Inference in Model-Based Machine Vision", *Uncertainty in Artificial Intelligence 3*, L.N. Kanal, T.S. Levitt, and J.F. Lemmer ed., Elsevier Science Publishers, 1989, pp. 73-95.

[3] Donald, B., and Jennings, J., "Perceptual, Limits, Perceptual Equivalence Classes, and a Robot's Sensori-Computational Capabilities", *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, Osaka, Japan, 1991.

[4] Faugeras, O.D., and Hebert, M., "The Representation, Recognition, and Locating of 3-D Objects", *International Journal of Robotics Research*, vol. 5, no. 3, Fall 1986, pp. 27-52.

[5] Flynn, J.P. and Jain, A.K., "CAD-Based Computer Vision: From CAD Models to Relational Graphs", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 2, Feb., 1991, pp. 114-132.

[6] Henderson, T. and Shilcrat, E., "Logical Sensor Systems", *Journal of Robotic Systems*, vol. 1, no. 2, 1984, pp. 169-193.

[7] Levitt, T.S., Agosta, J.M., and Binford, T.O., "Model-Based Influence Diagrams for Machine Vision", *Uncertainty in Artificial Intelligence 5*, M. Henrion, R.D. Shacter, L.N. Kanal, and J.F. Lemmer, ed., Elsevier Science Publishers, B.V., 1990, pp. 371-388.

[8] Lowe, D.G., *Perceptual Organization and Visual Recognition*, Kluwer, Boston, MA, 1985.

[9] Murphy, R.R., "An Application of Dempster-Shafer Theory to a Novel Control Scheme for Sensor Fusion", *SPIE Stochastic Methods in Signal Processing, Image Processing, and Computer Vision*, San Diego, CA, July 21-26, 1991.

[10] Murphy, R.R., "An Architecture for Intelligent Robotic Sensor Fusion", PhD thesis, Technical Report #GIT-ICS-92/42, College of Computing, Georgia Institute of Technology, Atlanta, GA, 30332-0280.

[11] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[12] Stark, L. and Bowyer, K., "Achieving Generalized Object Recognition Through Reasoning About Association of Function to Structure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 10., Oct., 1991, pp. 1097-1104.