Sensible Scenes: Visual Understanding of Complex Structures through Causal Analysis*

Matthew Brand, Lawrence Birnbaum, and Paul Cooper
Northwestern University
The Institute for the Learning Sciences
1890 Maple Avenue, Evanston IL 60201
brand@ils.nwu.edu

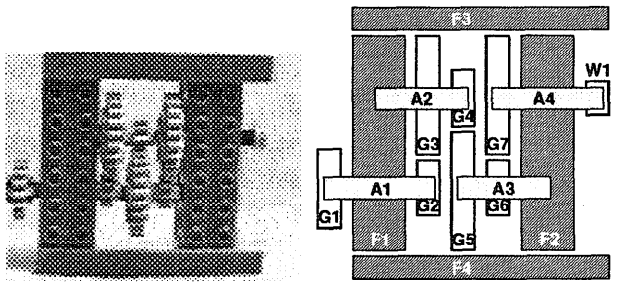
Abstract

An important result of visual understanding is an explanation of a scene's causal structure: How action—usually motion—is originated, constrained, and prevented, and how this determines what will happen in the immediate future. To be useful for a purposeful agent, these explanations must also capture the scene in terms of the *functional* properties of its objects—their purposes, uses, and affordances for manipulation. Design knowledge describes how the world is organized to suit these functions, and causal knowledge describes how these arrangements work. We have been exploring the hypothesis that vision is an explanatory process in which causal and functional reasoning plays an intimate role in mediating the activity of low-level visual processes. In particular, we have explored two of the consequences of this view for the construction of purposeful vision systems: Causal and design knowledge can be used to 1) drive focus of attention, and 2) choose between ambiguous image interpretations. Both principles are at work in SPROCKET, a system which visually explores simple machines, integrating diverse visual clues into an explanation of a machine's design and function.

Visual understanding

A fundamental purpose of vision is to relate a scene to the viewer's beliefs about how the world ought to be—to "make sense" of the scene. Understanding is the preparation we make for acting, hence our beliefs are fundamentally causal in nature; they describe the world's capacity for action and change. "Making sense" of a scene means assessing its potential for action, whether instigated by the agent, or set in motion by forces already present in the world.

*This work was supported in part by the National Science Foundation, under grant number IRI9110482. The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting, part of The Arthur Andersen Worldwide Organization. The Institute receives additional support from Ameritech, North West Water Plc, Institute Partners, and from IBM.



a b

Figure 1: A scene explored by SPROCKET, and a schematic representation of its explanation.

In the physical world of everyday experience, action is usually motion. One way scenes make sense is that they organize potential motions in a way that addresses some function. At the very least, scene elements must preserve their integrity: Plants, mountains, buildings, and furniture are all structured to address the forces of gravity. Often, scenes are structured for motion: Mechanical devices are structured to contain, direct, and transform motion. Ritual spaces, control panels, and computer interfaces are structured to organize users' gestures into meaningful messages. In every case, the way we visually make sense of these scenes is by explaining their configuration relative to a theory of how the world works. Because we are purposeful, we organize that causality into theories of how scenes are *designed*—how structures are organized to address functions.

We are investigating the role that knowledge of causality and design play in the perception of scenes. In particular, we are building systems for explanation-mediated vision—vision systems whose output is not feature lists, shape descriptions, or model classifications, but high-level explanations of why the scene makes sense: How it addresses its function, how it might causally unfold in the future, how it could be manipulated, or how it might have come to be.

In previous papers, we have shown that the explanation of static scenes in terms of their stability and structural integrity is an important aspect of

understanding. In particular, causal analysis of this sort can be used to address such issues as image segmentation [Cooper *et al.* 1993] and focus of visual attention [Birnbaum *et al.* 1993]. Previous systems have worked in the domains of children's blocks and tinkertoys. In this paper, we extend the scope of the approach to encompass more complex mechanical structures. This paper describes a system currently under development—SPROCKET—which explores and explains the design of simple machines. Our previous systems—BUSTER, which explains the stability of stacked blocks towers, BANDU, a variant which plays a rudimentary game of blocks, and FIDO, which resolves occlusions in tinkertoy structures by causal analysis—are discussed briefly at the end of the paper.

These vision systems have several interesting properties. First, they are vertically integrated, i.e. they encompass both low-level and high-level visual processing. Second, the output of these systems is in the form of meaningful explanations about the internal dynamics of the scene. These explanations could potentially be used to make predictions, to plan to enter and manipulate the scene, or to reason about how the scene came to be. Third, these systems employ causal theories which, although relatively small, are sufficiently powerful to generate explanations of highly complex structures. Fourth, they share an explicit model of what is interesting in a scene: They use functional and causal anomalies in ongoing explanations to generate visual queries, thus providing control of focus of attention in low-level visual processing. Fifth, causal and functional explanation forms a framework for the integration of evidence from disparate low-level visual processes. We suggest that these properties are natural consequences of building perception systems around explanation tasks.

Explanation-mediated vision

The goal of explanation-mediated vision is to assign each part in the scene a function, such that all spatial relations between parts in the scene are consistent with all causal relations between functions in the explanation [Brand *et al.* 1992; Birnbaum *et al.* 1993]. In previous papers, we have shown how this constraint can guide the visual perception and interpretation of complex static structures—structures where the possible motions of each part must be arrested. The same basic insight applies to structures in which possible motions are merely constrained. Thus a logical extension for explanation-mediated vision is the visual perception and interpretation of complex dynamical structures.

Visual cognition of machines

The visual understanding of simple machines is an approachable task because (1) causal relations are mediated by proximal spatial relations, and (2) the domain supports a very rich causal/functional semantics. This extends from simple principles of structural integrity to sophisticated knowledge about the design and layout of drivetrains. Design principles for kinematic machines are abundant generators of scene expectations (see [Brand 1992; Brand & Birnbaum 1992]). For example, a function of a gear is to transmit and alter rotational motion. To find a gear in the image is to find evidence of a design strategy which requires other torque-mediating parts in certain adjacencies and orientations, specifically an axle and one or more toothed objects (e.g., gears or racks).

Understanding machines is a matter of apprehending their design and function. Designs are strategies for decomposing functions into arrangements of parts; thus the overall strategy of SPROCKET is to reconstruct the designer's intent by applying the principles and practicum of machine design to hypotheses about the configuration of parts. As they are discovered, each part is assigned a function, such that (1) causal relationships are consistent with spatial relationships, and (2) the functions of individual parts combine to give the machine a sensible overall purpose.

Gearbox design

To bridge the gap between structure and function, we have developed a surprisingly small qualitative design theory for gearboxes (18 rules, not including spatial reasoning and stability rules). It begins with axioms describing the overall functional mission of a machine, and progresses through design rules to describe the way in which gears, axles, rods, hubs, and frames may be assembled to make workable subassemblies. The rules ground out in the predicates of rigid-body physics and spatial relations: adjacent parts restrict each other's translational motion; varieties of containment limit rotational degrees of freedom; etc. The design theory, adequate for most conventional spur-gear gearboxes, incorporates the following knowledge:

1. Knowledge about explanations:
   (a) An explanation describes a structure of gears, rods, frame blocks, and a ground plane.
   (b) A structure is explained if all parts are shown to be constrained from falling out and are functionally labelled.
   (c) A scene is considered explained if all structures in it are explained and all generated regions of interest have been explored.
2. Knowledge about function:
   (a) A moving part must transduce motion between two or more other parts.
   (b) A singly-connected moving part may serve as input or output.
   (c) A machine has one input and one output.
   (d) A fixed part serves to prevent other parts from disengaging (either by force of gravity or by action of the machine).
3. Knowledge about gears:

(a) In order to mesh, two gears must be coplanar and touching at their circumferences.
(b) Meshed gears are connected, and restrict their rotation to opposite directions and speeds inversely proportioned to their radii.
(c) A gear may transduce motion from a fixed axle (rod) to a meshing gear.
(d) A gear may transduce motion between two meshing gears.

4. Knowledge about axles and hubs:

(a) If a rod intersects an object but does not pass through, then the object (implicitly) contains a hub, which penetrates it. The rod penetrates the hub, restricting the rotational axis of the object. The object and the rod restrict each other's axial translation.
(b) If an object penetrates another object, it eliminates freedom of non-axial rotation, and freedom of translation perpendicular to its axis.
(c) If a rod goes through an object, then it passes through an implicit hub.
(d) If a non-round rod penetrates a non-round hub and the inscribed circle of the hub isn't larger than the circumscribed circle of the rod, then the hub and the object share the same axis and rate of rotation.
(e) If a hub penetrates a rod and either is round or the circumscribed circle exceeds the inscribed circle, then the rod restricts the hub to its principal axis of rotation.

5. Knowledge about frames and stability:

(a) Frames are stable by virtue of attachment to the table or to other frame pieces.
(b) Objects are stable if all of their motions downward are arrested:

Perception and pathologies

Beyond the design theory, much of the domain knowledge necessary to properly interpret mechanical devices resides in the practicum of the mechanical design: common part configurations, design pathologies, and knowledge of typical shapes and textures. This knowledge forms the basis for strategies for scene inspection, hypothesis formation, anomaly detection, and hypothesis revision.

Design pathologies are particularly important in our approach, since evidence gathering and hypothesis revision are driven by anomalies in the explanation [Schank 1986; Ram 1989; Kass 1989]. Anomalies are manifest as gaps or inconsistencies in the explanation. In machine explanations they take the form of inexplicable gears, assemblies that appear to have no function, or assemblies that appear to defeat their own function. These anomalies reflect underlying design pathologies in the system's current model of the machine, if not in the machine itself. We are currently developing a catalogue of gearbox design pathologies. Each pathology is indexed to a set of repair hypotheses. A repair hypothesis describes previous assumptions that may be suspect, and proposes scene inspections that may obtain conclusive evidence. The example below illustrates at length a repair hypothesis which compensates for a known weakness of one of our visual specialists: Gear inspections will sometimes construe two meshed gears as a single large gear. This generally leads to design pathologies which make the perceived structure dysfunctional or physically impossible.

Visual specialists

Figure 2a shows the output of a visual specialist built to look for groupings of short parallel contrast lines. This "tooth specialist" is used to look for mechanically textured surfaces, which are usually gears. The specialist uses a simple syntactic measure for gear candidates: groups of four or more short parallel contrast edges in the image. Like most specialists, it is meant to be deployed in limited regions of interest in the image to answer spatially specific queries such as, "Is there a gear oriented vertically and centered on this axle?" These queries may include initial guesses as to position, orientation, and scale hints. If a specialist succeeds, it will return more precise information on the location of a part, along with a confidence measure. Other specialists scan regions of interest for shaped areas of consistent color, for example rectangles, parallelograms, and ellipses (see [Brand 1993] for details).

Example

For the purposes of this exposition, the "tooth specialist" has been applied to the entire image (figure 2a). The specialist correctly identified 4 gears, made two spurious reports (that there is a horizontal gear and that there is a small vertical gear in the lower right corner), fused the two middle gears into one large one, and made a partially correct report of small gear in the lower right. To simplify this example, we will ignore reports to the right of the middle and concentrate on the diagnosis and correction of the fused gears.

Figure 2b illustrates the state of the explanation after a first-pass application of the design rules to the candidate gears found by the "tooth specialist." This preliminary model assumes that:

1. An axle `A1` has been surmised to support and connect gear `G1` and gear `G2` from frame piece `F1`.
2. Gear `G1` must be fixed to axle `A1` and apparently lacks a meshing gear.
3. Gear `G2` meshes with gear `G3` and must be fixed to axle `A1` (otherwise it rotates freely on `A1` and a gear must be found below it).
4. An axle `A2` has been surmised to support gear `G3` from frame piece `F1`.
5. An axle `A*` has been surmised to support the large gear `G*`. The axle must either connect to frame piece

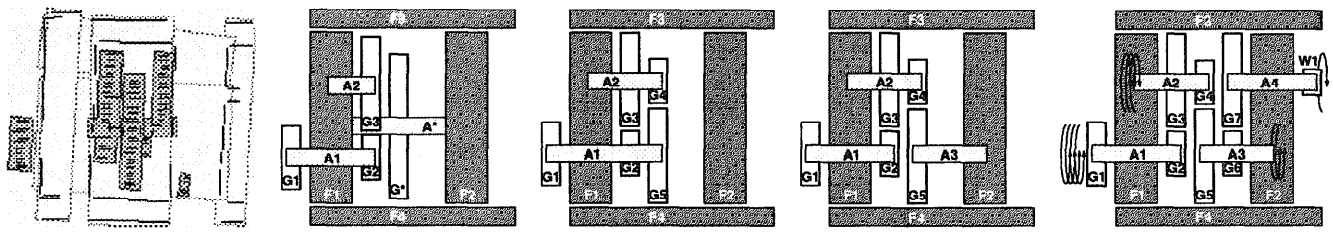


Figure 2: Three successively more sensible interpretations of the gear specialist's reports.

F1, in which case it runs in front of or behind gears G1 and G2, or to frame piece F2, or to both.

6. Gear G* must be meshed with a two gears above and below, or one gear above or below plus a fixed axle A* which must then connect to some other part of the drivetrain, or gear G* may also be for interface.
7. Frame pieces F1, F2, F3, and F4 are all attached.

These hypotheses are established by explaining how each part satisfies as mechanical function, using the above-described rules about function and configuration. The most productive constraint is that every moving part must have a source and a sink for its motion. Gears, for example, must transduce rotational motion, either between two meshing gears, between one meshing gear and a fixed axle, or between the outside world and a meshing gear or fixed axle (as an interface).[1] Given this constraint, two anomalies stand out immediately: gears G1 and G* are not adequately connected to be useful.

Visual inspection determines that G1 is indeed unmeshed, and spatial reasoning determines that it lies outside of the frame. As a matter of practicum, this suffices to support the hypothesis that G1 is indeed for interfacing, and the anomaly is resolved.

To resolve the anomaly in hypothesis number 6, the system must find at least one meshing gear above or below G*. However, there is no room between G* and frame piece F4 for such a gear, and there is no room for an axle to support a gear above G*. Thus G* is reclassified as an inexplicable gear anomaly.[2]

Since we know that fusion is one of the typical errors made by the gear-finding specialist, we have written a repair method for this anomaly which attempts to split fused gears. The method looks for nearby axles to support the gears resulting from the split, and if present, uses these axles as guides for the division. In this case, G* is split into gear G4, which is put on axle A2, and gear G5, which is put on axle A1. Conjectured axle A* is discarded at the same time. This state of affairs is illustrated in figure 2c. The new gears mesh with each other, and in order to transmit motion, both are assumed to have fixed axles. This is necessary because there is still no room to place additional meshing gears. It also makes all four gears G2-5 appear functionally viable; each gear will transmit motion to and from another part.

Unfortunately, there is an anomaly in this configuration: gears G2-5 are now in a locked cycle, and will not turn. This is detected when propagation of constraints reveals that each gear is required to spin at two different speeds simultaneously. The three elements of the new hypothesis—that 1) G* is split into G4 and G5, 2) G4 is fixed on axle A2, 3) G5 is fixed on axle A1—must be re-examined. Retracting (1) returns to the original anomaly, so this option is deferred. Retracting (2) deprives G4 of its axle and leaves gear G2 dangling without a sink for its motion. Retracting (3) merely leaves G5 without an axle. This is the cheapest alternative, since, as a matter of practicum, it requires only a single alteration and axle-additions are generally low-overhead alterations. The required axle (A3) can be surmised coming from frame piece F2. This leaves gears G2-4 properly explained, unlocks the drivetrain, and eliminates all design pathologies except for a yet undiscovered sink for G5, which will lead to the discovery of the rest of the mechanism. The repair produces a model illustrated in figure 2d.

Final analysis

Through this process of exploration, hypothesis, anomaly, evidence procurement, and reformulation, SPROCKET develops a functionally and causally consistent explanation of how the individual elements of the mechanism work together to conduct and modify motion. Our goal is to have SPROCKET analyze the resultant model of the drivetrain to provide a functional assessment of the entire device, e.g.: The machine is for the conversion of high-speed low-torque rotation into high-torque low-speed reversed rotation (figure 2c), or vice-versa.

Precursor systems: Seeing stability

The most ubiquitous aspect of our causal world, at least physically, is the force of gravity. Nearly everything in visual experience is structured to suit the twin constraints of stability and integrity. To understand a static scene such as a bridge or building, one explains

[1] Currently SPROCKET is ignorant of chains, racks, ratchets, and other more sophisticated uses of gears.

[2] Ostensibly, meshing gears could be *behind* G*. This is beyond SPROCKET's abilities, as it is limited to machines where the drivetrain is laid out in a small number of visually accessible planes.

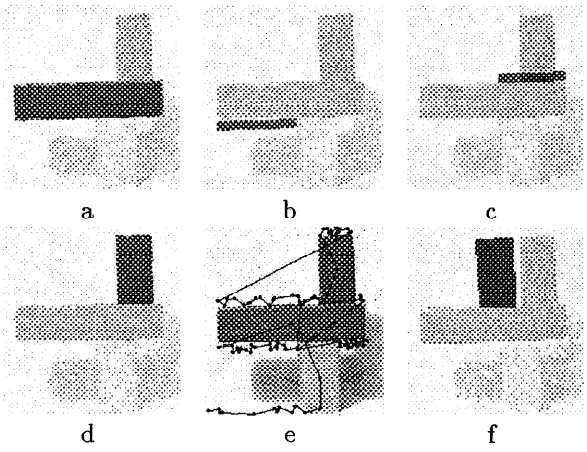


Figure 3: Snapshots of BUSTER's analysis of a three-block cantilever. Regions of interest are highlighted; the rest is faded. See the text for explanation.

how it meets these constraints. We have built a number of vertically integrated end-to-end systems that do this sort of explanation, perusing objects that stand up, asking (and answering) the question, "Why doesn't this fall down?"

Understanding blocks structures

BUSTER [Birnbaum *et al.* 1993] does exactly this sort of visual explanation for structures made out of children's blocks. BUSTER can explain a wide variety of blocks structures, noting the role of each part in the stability of the whole, and identifying functionally significant substructures such as architraves, cantilevers, and balances.

In static stability scenes, the internal function of each part is to arrest the possible motions of its neighbors. BUSTER's treatment of cantilevers provides a simple illustration of this constraint. In figure 3a BUSTER has just discovered the large middle block, and noticed a stability anomaly: It should roll to the left and fall off of its supporting block. To resolve this anomaly, BUSTER hypothesizes an additional support under the left end of the block, but finds nothing in that area (3b). A counterweight is then hypothesized above the block and to the right of the roll point (3c), and a search in that area succeeds, resulting the discovery of a new block (3d). BUSTER thus assesses the structure as a cantilever. Figure 3e shows the attentional trace for the entire scene analysis.

Playing with blocks

One of the immediate uses of causal explanations is in reasoning about actions to take within the scene. Depending on the goals of the vision system, an explanation might also answer such questions as, "How did this scene come to be?" or "Is there a safe path to navigate through the scene?" With a child's goals, the robot may also want to know, "Where can I add

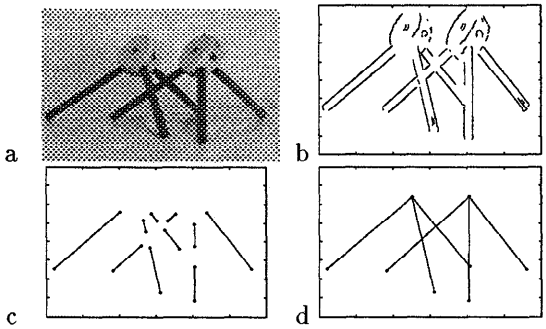


Figure 4: Stages in the explanation of the tinkertoy scene (a). (b) Boundary contours are extracted from stereo pairs. (c) Potential rod segments are extracted from the contours. (d) Rodlets are merged and extended to conform to knowledge about stable structures.

blocks to make the tower more precarious?" " What is the best way to knock it down?" Given that children play blocks partly to learn about structural integrity in the world, these are probably fruitful explanation tasks. BANDU, an augmented version of BUSTER, answers the latter two questions in order to play a rudimentary competitive block-stacking game. The aim is to pile a high and precarious tower, making it difficult or impossible for the opponent to add a block without destabilizing the whole structure. Figure 3f shows an addition that BANDU proposes for the cantilever structure.

Interpreting Tinkertoy assemblies

FIDO [Cooper *et al.* 1993] is a vertically integrated end-to-end vision system that uses knowledge of static stability to segment occluded scenes of three-dimensional link-and-junction objects. Link-and-junction domains are a nice composable abstraction for any complex rigid object. Occlusions tend to be both common and serious in link-and-junction scenes; FIDO uses naive physical knowledge about three-dimensional stability to resolve such problems. Visually, FIDO works with scenes of Tinkertoys assemblies, though its algorithms are generally applicable to a wide variety of shapes.

FIDO's input from early processing is a piecemeal description of the visible scene parts. The footprint of each object is determined to be the convex hull of places where the object touches the ground. If the object's center of mass is in the footprint, it is stable. If a part's object is not stable in this sense, FIDO invokes a series of rules whereby parts try to connect to each other. In this way, invisible connections are hypothesized, parts are extended through unseen regions to touch the ground plane, and entire parts may be completely hallucinated, in order to generate stable subassemblies. FIDO can successfully process relatively complex scenes like that shown in 4a; images b-c show the data this image yielded and the stable structure that was inferred by causal analysis.

Related work

Causal and functional knowledge are enjoying a renaissance in computer vision. Researchers such as [Pentland 1986] and [Terzopoulis & Metaxas 1990] have pointed out that the normal causal processes in the world have consequences for the way objects are shaped. On the functional side, [Stark & Bowyer 1993] have used structural concommitants of function—containment for cups, stability for chairs—to verify category decisions for CAD representations of objects. [Rimey 1992] has developed a system which visually explores table settings, using models of typical arrangements to identify, say, a formal dinner. Although these systems do not have an explicit and generative notion of function—e.g. what cups are for and why they should be cylindrical, or why formal settings have two forks—they do serve as impressive demonstrations of the value of high-level knowledge in visual paradigms.

There is an extensive literature on qualitative causal reasoning about kinematics. [Forbus *et al.* 1987; Faltings 1992] have produced kinematic analyses of ratchets, escapements, and even an entire clock using qualitative reasoning methods. [Joskowicz & Sacks 1991] have developed an algorithm for the kinematic analysis of machines that breaks the devices down into subassemblies, analyzes their configuration spaces, and combines the results into a description of the mechanism's overall kinematic behavior. These approaches feature quantitative shape analysis and rigid-body modelling algorithms that are quite a bit more extensive and more general than we use. Rather than concentrate on the universality of our kinematic models, we have chosen to focus on their compability with perceptual and teleological representations. SPROCKET is limited to machines built of rectangular and cylindrical shapes, with smooth and toothed surfaces, and conjoined by means of attachment, containment, friction, and compression—e.g., Lego machines.

Conclusion

SPROCKET and its sister programs are vertically integrated vision systems that achieve consistent explanations of complex scenes through the application of causal and functional semantics. Using modest generative theories of design and naive physics, these systems purposefully explore scenes of complex structures, gathering evidence to explain the stability, integrity, and functional coherence of what they see. Anomalies in the ongoing explanation drive hypothesis formation, visual exploration, and hypothesis reformulation. Considered as vision systems, they demonstrate the surprising leverage that high-level semantics provide in the control of visual attention, and in the interpretation of noisy and occasionally erroneous information from low-level visual processes. Considered as evidential reasoning systems, they highlight the importance of building content theories that describe not just the possibilities of a domain, but the domain's most likely configurations, the way in which the domain is manifest in perception, and the characteristic errors and confusions of the perceptual system itself.

Acknowledgments

Thanks to Ken Forbus, Dan Halabe, and Pete Prokopowicz for many helpful and insightful comments.

References

[Birnbaum *et al.* 1993] Lawrence Birnbaum, Matthew Brand, & Paul Cooper. Looking for trouble: Using causal semantics to direct focus of attention. To appear in *Proceedings of the Seventh International Conference on Computer Vision*, 1993. Berlin.

[Brand & Birnbaum 1992] Matthew Brand & Lawrence Birnbaum. Perception as a matter of design. In *Working Notes of the AAAI Spring Symposium on Control of Selective Perception*, pages 12–16, 1992.

[Brand *et al.* 1992] Matthew Brand, Lawrence Birnbaum, & Paul Cooper. Seeing is believing: why vision needs semantics. In *Proceedings of the Fourteenth Meeting of the Cognitive Science Society*, pages 720–725, 1992.

[Brand 1992] Matthew Brand. An eye for design: Why, where, & how to look for causal structure in visual scenes. In *Proceedings of the SPIE Workshop on Intelligent Vision*, 1992. Cambridge, MA.

[Brand 1993] Matthew Brand. A short note on region growing by pseudophysical simulation. To appear in *Proceedings of Computer Vision and Pattern Recognition*, 1993. New York.

[Cooper *et al.* 1993] Paul Cooper, Lawrence Birnbaum, & Daniel Halabe. Causal reasoning about scenes with occlusion. 1993. To appear.

[Faltings 1992] Boi Faltings. A symbolic approach to qualitative kinematics. *Artificial Intelligence*, 56(2-3):139–170, 1992.

[Forbus *et al.* 1987] Ken Forbus, Paul Nielsen, & Boi Faltings. Qualitative kinematics: a framework. In *Proceedings of IJCAI-87*, 1987.

[Joskowicz & Sacks 1991] L. Joskowicz & E.P. Sacks. Computational kinematics. *Artificial Intelligence*, 51(1-3):381–416, 1991.

[Kass 1989] Alex Kass. Adaptation-based explanation. In *Proceedings of IJCAI-89*, pages 141–147, 1989.

[Pentland 1986] A.P. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28(3):293–332, 1986.

[Ram 1989] Ashwin Ram. *Question-driven understanding.* PhD thesis, Department of Computer Science, Yale University, 1989.

[Rimey 1992] Ray Rimey. Where to look next using a bayes net: The tea-1 system and future directions. In *Working Notes of the AAAI Spring Symposium on Control of Selective Perception*, 1992. Stanford, CA.

[Schank 1986] Roger Schank. *Explanation Patterns.* L. Erlbaum Associates, NJ, 1986.

[Stark & Bowyer 1993] L. Stark & K. Bowyer. Function-based generic recognition for multiple object categories. To appear in *CVGIP: Image Understanding*, 1993.

[Terzopoulis & Metaxas 1990] Demetri Terzopoulis & Dimitri Metaxas. Dynamic 3d models with local and global deformations: Deformable superquadrics. In *Proceedings of the Fourth International Conference on Computer Vision*, pages 606–615, 1990.