Seeing is Believing: Why Vision Needs Semantics

Matthew Brand, Lawrence Birnbaum and Paul Cooper
Northwestern University
The Institute for the Learning Sciences
Evanston, Illinois
brand@ils.nwu.edu

Abstract

Knowledge about the functional properties of the world constrains and informs perception. For example, looking at a table, chair, a building or a sculpture, we are able to resolve occluded attachments because we know that in order to stand, an object's center of gravity must lie within its footprint. When when we see a floating wheel in the interior of a vehicle, we know that it is probably the means by which the driver communicates steering information to the chassis. Movable handles imply input to machines; fixed handles imply an upside and a downside to any object they grace.

We are constructing a machine-understanding machine with which to explore the usefulness of semantics in perception. This system will investigate simple mechanical devices such as gear trains, simultaneously building a representation of the structures and functions of parts, and using that representation to guide and disambiguate perception. In this paper we discuss how this work has led to an understanding of perception in which a semantics of structure and function play a central role in guiding even the lowest level perceptual actions.

Vision is Cognition

We distinguish *visual understanding* from *visual recognition* by the central questions that drive the two activities. For recognition, the question is "What is out there?" For understanding, the question is "What is happening/can happen in this scene?" or more specifically, "How can I interact fruitfully with the scene?"

Humans see and understand the world in terms of its affordances [Gibson 66], which signal the potential for function and for interaction. To see and act purposefully, robots must likewise be designed with a capacity for the visual understanding of the affordances of their worlds [Brand & Birnbaum 92].

Visual understanding is, firstly, explaining the scene with regard to the goals and causal knowledge of the viewer, and secondly, explaining the image with regard to the scene[1]—what is known as image understanding[Birnbaum et al. 92]. An explanation should capture the why and how of a scene: the causal relations between objects, the sources of motion and stability, and the potential uses of these causal properties for the viewer. For example, in the reduction engine pictured in figure 1, two gears are causally related to each other in that they will transmit (and reverse) rotational motion. The handle is causally related to the viewer in that it affords the viewer an opportunity to inject motion into the situation. A visual understanding of figure 1 will include the assessment, "This is a device which, when powered from one handle, causes the opposite wheel to rotate at a much higher torque and slower speed."

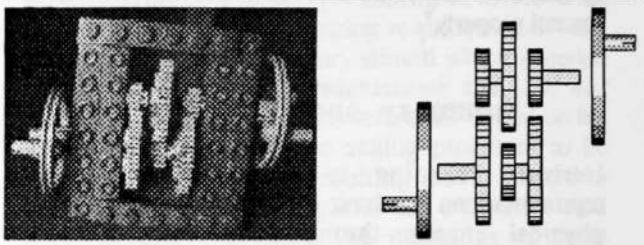

Figure 1: A reduction engine as it appears to the camera, and an exploded schematic of its drivetrain. To understand it, we piece together a coherent explanation of the what, why, and how of this drivetrain, using clues from the image, from knowledge of function and structure, and from new views procured by perceptual acts.

Our prescription for visual understanding dispenses with the conventional notion of an order of processing (e.g. [Marr 82, Barrow & Tanenbaum 78][2]). Instead of a visual front end which outputs image segmentation descriptions for a back end to use, understanding is a matter of negotiation between the constraints and hypotheses of a generative semantics of function and the activity of low-level visual routines. The semantics guide the activity of the visual hardware through queries: in the course of building an explanatory model

[1] I.e. what is traditionally called computer vision: grouping high-contrast gradients into edges, finding flow boundaries, matching to models, etc. [DARPA 92]

[2] This filter-then-analyze approach has has previously been called into question by, among, others, Tanenbaum himself; see [Witkin & Tanenbaum 83]

from visual reports of clues to function, gaps and inconsistencies in the model to are used to general queries for the visual routines, which cue various perceptual acts that result in new reports.[3] The visual processes answer queries by testing for features and tracking invariants in the scene that have functional significance, such as tracking parallel lines (generated by the edges of a rod) to find the end of the rod. We are in the process of constructing such a system and have analyzed several image sequences in the manner we suggest. This papers describes aspects of a number of these analyses.

Why and how are functional properties detectable in an image? Scenes are structured, and the causally 'loaded' regions of a scene tend to be where parts interface: where they are joined or where there are contrasts in motion. This means that the parts of the scene where change is most likely and most significant—the parts of interest to a robot—often have characteristic manifestations at predictable locations in the image. For example, meshed gears produce adjacent regions of optical flow with opposite curl. At the junction itself, the flow will converge, then diverge. Statically, a gear meshing introduces characteristic textures into the image because of the toothing, and this manifests itself as a local peak of a high-frequency component in the image. This is what robots would look for if they were made to fix car transmissions.

Understanding just a small part of a picture—even a single component of a structure—immediately yields a rich set of expectations about neighboring part boundaries, structural concomitants, typical axes of motion, and so forth, and these in turn have characteristic manifestations in the image. This is because most things in our visual experience have the *quality of design*: their construction reflects a host of functional constraints. Even the simplest functional constraint—resistance to the pull of gravity—profoundly influences design and appearance, and generates for us many expectations that guide visual cognition. This is equally true within and outside the realm of man-made objects: The world is pervaded by function.

In this paper we present the beginnings of a generative functional semantics for vision, with enough detail to account for example scenes ranging in complexity from sticks and strings to common machines.

The Importance of Being Connected

It is generally understood that the causal properties of the scene are usually mediated by physical connections between the parts it contains. Understanding an object or scene requires visually tracing through the causally most "loaded" connections between subparts.[4] Toward this end, we have been developing a catalog of connection types, in which each connection is indexed along with a description of function, typical structural correlates, and characteristic visual manifestations. We now have a rich catalog of mechanical connections ranging from E-clasp fasteners to gear meshings to hub-axle interfaces—nearly 40 connection types at time of writing. The descriptions of function and structural correlates provide great leverage in visual search, generating hypotheses about neighboring parts, as-yet-unperceived assemblies, and the relative locations of parts.

Knowledge about connections provides a reasoner with a special and highly useful set of expectations about the world. In order to use this knowledge, we also need to have good theories of how and why parts are put together, and of the capabilities of our vision system to extract useful features and invariants from the image. This requires a large rule base which expresses the principles of rational design, and which describes—in terms of the visual routines—the perceptible artifacts of design. Design semantics tell us a good deal about what kinds of image processing we need. This is true of both abstract and specific constraints. At the abstract level, for instance, we have a constraint such as the following:

> A drivetrain assembly has function if it transduces, regulates or switches motion. In visual terms, this function is manifuest in the following rule: *A patch of the scene is explained if it connects to two patches of differing motion (transduction); if it connects to just one patch of motion but appears to have significant mass (regulation); or if over time, its position relative to connecting patches changes so that their optical flow is no longer related (connection and disconnection)*

Similarly, at the specific level, we have a rule such as the following:

> Most axis- and rail-mounted machine parts have some symmetry with respect to their axes of motion so to reduce vibration (and simplify manufacture). This includes gears, carriages, and pistons. In visual terms, this functional constraint implies the following rule: *For most moving parts, there is a way to orient the camera so that motion of the part causes a minimal change in its visual profile.*

We have developed a set of such rules sufficient to produce explanations for the objects pictured in the paper. This knowledge combines with an explicit (if somewhat simplified) theory of the image-processing and camera-orienting subsystems to make predictions about which visual routines (e.g. [Ullman 84]) to engage and what misclassifications they can make about features in the image. To describe the vision subsystem, we identify the assumptions and strategies built into its camera-orienting and feature-extracting processes, and then produce characterizations of when and how various low-level routines will produce spurious reports:

- A change of perspective usually suffices to distinguish adjacency from occlusion. *A report of non-*

[3] This is similar in spirit to work in text-understanding by [Ram 89]

[4] Indeed, when we ask colleagues to look at the objects and pictures in this paper, we see them visually trace out the "functional drivetrain" of an object.

adjacency from the visual system is reliable; reports of adjacency can be mistaken if the parts are close.

- A gradual dip and then recovery in the frequency of the strongest signal component taken along a line through the image implies a periodic texture mapped onto a curved object, such as a gear. [Bajcsy & Lieberman 76] *If the camera is not oriented in the plane of a gear before using the visual toothed-wheel detector, there may be spurious negative reports.*

In sum, not only do we need knowledge of the causal structure of the world, but we need knowledge of how that causal structure is revealed (and sometimes mistaken) by perceptual actions. One kind of knowledge tells us what is missing or wrong in our explanation of a scene; the other kind tells us how to find missing information in the image, or where to find mistaken interpretations in the explanation.

Examples

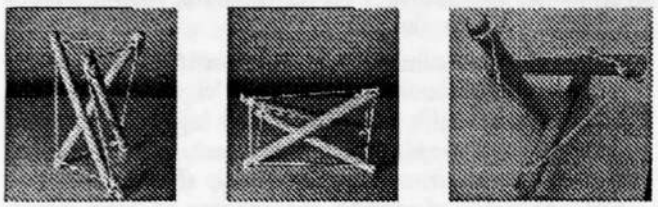

Figure 2: Views of a tensegrity object standing up, on its side, and from above. Different views lead to different explanations.

Tensegrity

A tensegrity [Fuller 75] is a rigid structure of rods and cables. The simplest possible construction, consisting of three rods and nine cables is pictured in figure 2. None of the rigid elements touch each other, yet the whole structure stands. People find tensegrity constructions fascinating because a very basic assumption of visual explanations fails to apply, namely that a rigid object is decomposable into substructures that support each other [Birnbaum et al. 92]. The only means of connection in the tensegrity is tethering; there is no support and only the illusion of suspension. In fact, gravity plays no role in its stability.

It is, however, the illusion of suspension that allows the tensegrity to be explained. A first view of the structure will reveal a large part (a rod) which looks as if it should be falling. To explain its stability, one scans up the rod, looking for an attachment which prevents it from falling in the direction that it leans. Near the top, a cable from another rod prevents this collapse. However, this does not explain why the rod doesn't pitch in a direction perpendicular to the cable, and a further scan reveals a nearly perpendicular cable which partially fulfills this function: it keeps the rod from falling "outwards." To explain why it does not fall inward, we look for a third fixating element, and find a third cable attached to the endpoint that has a small vector component contrary to inward motion. The rod endpoint is now considered stable, as all motions are restricted (some, apparently, by gravity). The rod as a whole, however, may not be. Thus, to explain why the rod doesn't slide out from underneath itself, a similar set of scans discovers three more tethers. Now, the rod is provisionally considered stable. Yet all the cables need to be explained, and this leads to similar explanations of the other two rods. At the end, every part has been assigned a function, and every force appears to have been countered.

However, this results in a circular explanation, where each part is, ultimately, held up by itself. In order to "ground" this explanation, we must invoke the principle of symmetry. Symmetry is a design stratagem for canceling out all forces. It is necessary to know about symmetry, and how to look for some kinds of symmetry in an image, if one is to explain why static objects stand up. Symmetry is the most common form of balance, which is often the ultimate explanation of stability.

Symmetry is also a way of resolving explanatory loops. For the tensegrity object we propose a threefold rotational symmetry around a vertical axis, and orient the camera above the structure. The endpoints are used to estimate where the symmetrical axis is, and once the camera is collinear with this axis, a visual routine processes the image to find evidence of rotational symmetry. Finding symmetry completes the explanation.

Reduction Engine

A reduction engine works on the principle that a small gear connected to a large gear will reduce speed and increase torque. To explain such a machine, the input and output must be found, the drivetrain must be traced, and the parts that serve to frame and stabilize the object must be identified. The order of discovery of all these assemblies is not important—finding any one or part of any one produces many functional clues about where and how to look for other parts.

For example, finding a protuberance from the face of a wheel (an ellipse in the image) is a good indicator of an axis or handle. An ellipsis-finding Hough transform will tell us where to expect the axis. If the protruberance is off-center, then it is a handle, which indicates that the part is an input or output to the machine.

In the reduction engine, a wall lies directly behind the wheel, so the axis is invisible. However, it is reasonable to expect that the axis is fixed in place by the frame, so the wall is hypothesized to be part of the frame, and the axis is hypothesized to pass through it. Scanning along the line of the hypothesized axis brings a toothed-texture into center view, which can be verified as a gear with the appropriate curvature for the axis. The axis is now provisionally explained.

To explain the gear, it must mesh with at least one other gear (or a chain). [Brand & Birnbaum 92]

describes a system for scanning a camera across a train of meshed gears, reporting when a bounding wall has been hit or no more gears have been found. When this finds a meshing gear, the first gear is explained. Explaining this second gear requires looking for an axis to carry along the motion to another part, since no other connecting gears can be found. The axis is almost entirely hidden, so the same strategy that verified the first axis is used. At this point, the operations just described repeat to explain the remainder of the mechanism.

Other Examples

Even without the ability to move the camera to scan for new information, functional expectations will resolve ambiguities in the interpretation of a scene. [Halabe 92] has implemented a program with a modest semantics of attachment and stability that will "reattach" legs of Tinkertoy constructions that have been "severed" by occlusion.

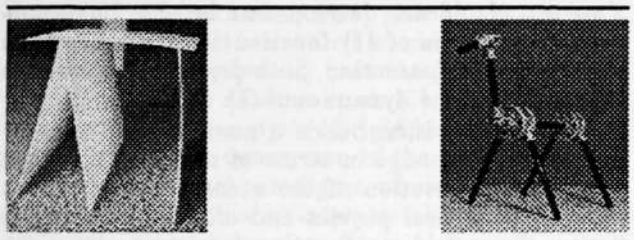

Figure 3: A house of cards and a toy horse. Stability constraints make it possible to reason about occlusions such as the horse's hind leg and the obscured card.

Similar analyses have been done with various houses of cards (e.g. figure 3), in which connections mediate support or friction, but there is no attachment.

A Functional Analysis of Vision

We believe that the use of functional semantics in the design of vision systems and visual primitives applies to the whole range of systems that process images and/or perform visual reasoning. Whether designing a hard-wired animat, or compiling the knowledge to be used in a mechanical reasoning system, a functional analysis will outline the kinds of features that are needed for choosing actions [Brand 91], the kinds of image-predicates that are necessary to support those features, and the kinds of ambiguities that the system will face.

Semantic constraints are pervasive in the world, thus we need a functional semantics for every kind of scene. We outline below the main functional relationships inherent in different kinds of scenes, to show the basic building blocks of visual semantics for different tasks. We identify three general types of scenes, briefly sketch the fundamental questions that drive explanation in each, outline the basic principles used in these explanations, and describe how they may be detected in an image.

Static Objects

- Will it fall apart? (How is the motion of all parts constrained?) The means of static constraint are blockage (support and containment are special cases), attachment, and tethering (again, suspension is a special case). Blockage is detectable though adjacency in the image, especially in the horizontal plane, where one part is supporting another. Attachment can be inferred from off-horizontal adjacency, partial containment, and characteristic attachment artifacts such as screw and rivet heads. Tethering can be inferred from adjacency to the end of a long thin object such as a cable.
- Will it fall on me? (How is the center of gravity placed within the footprint? Or how is the object affixed to something heavier?) Typical means of standing include spread feet (or narrow tops), counter-balancing, and anchoring. Spread feet are often visible as protuberance on the ground plane diverging from the object-image center. Narrowing can be calculated, for example, as a gross geometric predicate, or by looking for finer detail higher up in the image (e.g. a greater proportion of high-frequency components). Counter-balancing and symmetry are profoundly difficult to find in an image; we are compiling a host of methods, including looking for anomalously thick or long projections to diagnose counter-balancing. Anchoring often requires projections into the ground plane, often accompanied by bumps in the plane (e.g. tree roots).
- What can it hold up? (What devices of support, attachment, etc., does it have that are not used in its own skeletal integrity?) This is often a matter of identifying objects which afford support or attachment but do not participate in the explanation of the object's stability. Unused high horizontal surfaces (tabletops), hook shapes or vertical points (coatstands), and regions of concavity (bowls) are good indicators of overall function.

Objects with internal motion (Machines)

- How is motion constrained and channelled[5]? In machines, the means of constraining motion always leave a dimension or two of freedom. This is principally achieved by partial containment (eyes, hubs, sockets, etc.) in the man-made world, and by flexion in the natural world. This is a difficult problem for us, since most of a containment device is obscured from view. At present, we plan to simply infer containment devices from the limited motions of parts. There is some potential in developing a library of visual signatures for containment devices, much as the screw-head is a signature for a largely invisible part.

[5]This is very similar to the question asked of static objects. In fact, we had analyzed several machines before realizing that static objects are a special case, in which *all* motion is restricted.

- Why are all these parts moving? (How is motion communicated? What kinds of connections are there?) The principal means for communication of motion are attachment and friction. Communication produces characteristic patterns of flow in adjacent regions. Optical flow algorithms may only suffice to reveal regions of varying motion, requiring other visual processes to close in on and resolve details of how motion is communicated.
- What kind of motion does this produce? Classification of motion into rotation, translation, lifting, swinging, hammering, etc., provides a useful index to function, and often suggests a likely mechanism. For example, repeated translational motion along a line almost always requires an associated rotational motion.
- How do I connect with it? (What is the interface to the rest of the world?) This is similar to the use question asked of static objects. There is a fairly limited range of control devices which specifically interface to the human hand, and which have characteristic shapes: handles, buttons, dials, and steering wheels, for example. These will have to be resolved by local searches in the image for characteristic shapes.

Terrains

- Where are the animate objects? (What's moving and what are our relative positions in the food chain?) This is largely a matter of noticing independent translational motion in the image sequence. Visually, we look for small regions of depth change, as well as texture anomalies.
- Where can I pass or flee? (What part of the terrain is navigable for an agent with legs or wheels like mine?) The most important constraint for land navigation is continuity of ground plane, followed by smoothness. Another important affordance for navigation are things that can be climbed. For this reason, it is useful to look for low-frequency texture on objects that rise out of the ground plane, for example a tree with rough bark. One special case–stairs–adds the constraint that the vertical texture have a single strong frequency component.
- Where can I take shelter or hide? (What part of the terrain has limited accessibility and/or limited visibility?) The key to this function is identifying places in the world where vision itself doesn't work very well. One hides in caves or overhangs, which are bounded regions of relative darkness and low contrast, or one hides in underbrush: areas of omnidirectional high frequency image noise.

Vision Requires Outlook

Vision has long suffered the notion that an artificial visual cortex will be a "front end" for an intelligent system that itself is not necessarily visually sophisticated. A consequence of this view is that much talent and energy has been invested in trying to find an appropriate form of output for vision systems. Once an output representation has been invented, there is the usual struggle of finding a robust algorithm to map images to reasonable (literally) outputs. This has typically resulted in recognition systems, which match the image to a database of models via reverse optics transformations (e.g. [Horn 86]). We have learned from this work that no single algorithm or image transform ever works more than perhaps 80% of the time[6].

Recently, some researchers have given attention to the *use* of visual processing, that is, what happens in the "back end" (e.g. [Ballard 89]). This has led to a reformulation of vision in which processing is specifically aimed at quickly extracting the features that are most decisive for the immediate pursuit of a goal. In the "active vision" paradigm, the back end is reciprocally considerate of the front end, reorienting the camera to procure ever better input for the feature detectors. This is typical in visual navigation systems, which extract surprisingly few topographic features from the image, and then make strikingly good use of them. This is a significant development because it incorporates the notions of (1) functionally derived features and (2) focus of attention, both deployed according to an analysis of the dynamics of the task.

Recognition vision builds a model of the scene by explaining the image in terms of the physics of light and the configuration of the scene. It incorporates analyses of optical physics and of shape, which give it a mathematical, nonfunctional slant. Active vision work tends to be miserly in its representation, but tries to participate directly in the causality of the scene. It incorporates analyses of the task and of visual invariance across motion, which give it a decided functional slant and, significantly, a good measure of robustness[7]

What is missing in vision, though hinted at by active vision, is a functional analysis of the *world*—of what is being looked at. The purpose of vision is not to describe the image in terms of segmentation candidates, but to explain the scene in terms of what we believe about the world. The primary visual belief that humans enjoy is the dictum that "form follows function." The world that we see is one of design, everywhere imbued with function, and interesting mainly because we have to interact with it.

The questions we ask of our eyes are: "Will it fall on me?" "Will it support my weight?" "Where can I pass?" "What does it do?" These functional questions lead straight to structural questions: "Does the center of gravity lie outside the footprint?" "What are the load-bearing lines?" "Where is the ground plane navigable by foot?" "How does its motion relate to a human activity?" The structural questions in turn lead to questions posed of the world (of the image or of an image stream): "Where above the ground plane is the visual centroid?" "How thick is the train of connect-

[6] Minsky, personal communication

[7] Active vision aims to reduce uncertainty through tracking; thus the importance of invariance across motion.

ed substructures that rises from the ground plane to carry my weight?" "Where is the illumination gradient smooth or striped (steps)?" "Where is a handle-shaped object and the drivetrain that it moves?"

One might object to our emphasis on questions such as, "Why is this part here?" and, "How do these things relate?" when humans seem able to answer, "What is out there?" so effortlessly. Humans have prodigious visual memories, and equally uncanny powers of recognition. However, it is not recognition we are trying to explain; it is the original cognition. Given the amount of work this takes, it is not surprising that we are equipped with a caching mechanism which uses the memory of the first cognition to speed perception of the same object later on.

Related Work

Recent work in the understanding of diagrams indicates that researchers have found it useful to employ a simple semantics in conjunction with a simulated visual search for "regions of interest" in the diagram. [Narayan & Chandrasekaran 91] give an example of a straight flat line that is a shared boundary between two objects, which consequently have the potential to slide against each other. [Forbus et al. 87] provide a model for the qualitative analysis of rigid body interactions, given a qualitative description of the scene. Both are primarily post-visual paradigms, whereas we intend for our semantic analyses to interactively guide and disambiguate visual processes. It is also worth noting that most kinematic analyses of scenes, whether qualitative, diagrammatic, or truly visual, use a semantics of *motion*. In contrast, we are interested firstly in a semantics of *function*; and only secondly in its manifestation as motions, shapes and textures.

The work of the Vision and Modeling Group at the MIT Media Lab is also of note because, in trying to model the objects in the scene in terms of bent and deformed superquadrics [Pentland 90], they are also, in a sense, explaining the scene. This interesting approach differs from ours in that it is functionally neutral; such explanations tell how the scene could be made from simple lumps of clay that are deformed and combined to produce complex shapes. No hypotheses about causal relationships and function are present in these explanations, nor does such knowledge guide explaining at the level of image-processing either. However, their work has interesting possibilities because the models, once constructed, are imbued (via simulation) with a causality which includes rigid and elastic body dynamics, mass, and gravity. This could be used to provide feedback to an image-to-model constructor, by telling it whether or not the model is stable and static, or unbalanced and lacking in structural integrity.

Acknowledgements

Thanks to Ken Forbus, Dan Halabe, Bruce Krulwich, Peter Prokopowicz, and Louise Pryor for many useful discussions. This work was supported in part by the Defense Advanced Research Projects Agency, monitored by the Office of Naval Research under contract N00014-91-J-4092, and 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, an Institute Partner, and from IBM.

References

[Bajcsy & Lieberman 76] R. Bajcsy and L. Lieberman. Texter gradients as a depth cue. In *Computer Graphics and Image Processing* **5**, 1976.

[Ballard 89] Dana H. Ballard. Reference Frames for Animate Vision. *Proceedings of AAAI-89*, 1989.

[Barrow & Tanenbaum 78] Recovering Instricsic Scene Characteristics from Images. *Computer Vision Systems*, A.R. Hanson & E.M. Riseman (eds.), New York: Academic Press. 1978.

[Birnbaum et al. 92] Lawrence Birnbaum, Paul Cooper, and Matthew Brand. *Every Picture Tells a Story.* Forthcoming Technical Report, Northwestern University, The Institute for the Learning Sciences. 1992.

[Brand 91] Matthew Brand. Incorporating Resource Analyses into an Action System. 12^{th} *Proceedings of the Cognitive Science Society*, 1991.

[Brand & Birnbaum 92] Matthew Brand and Lawrence Birnbaum. Perception as a Matter of Design. In *Proceedings the AAAI Spring Symposium on Selective Perception*, 1992.

[DARPA 92] *Proceedings of the 1992 DARPA Image Understanding Workshop.* San Mateo, CA: Morgan Kaufmann Publishers, Inc. 1992.

[Forbus et al. 87] Ken Forbus, Paul Nielsen, & Boi Faltings. Qualitative Kinematics: a framework. *Proceedings of the* 10^{th} *IJCAI*, 1987.

[Fuller 75] R. Buckminster Fuller. *Synergetics.* New York: MacMillan Publishing Co., Inc. 1975.

[Gibson 66] J.J. Gibson. *The Sensense Considered as Perceptual Systems.* Botson: Houghton Mifflin, 1966.

[Halabe 92] Daniel Halabe. *A Leg to Stand On.* Unpublished manuscript, Northwestern University, The Institute for the Learning Sciences, 1992.

[Horn 86] Berthold Klaus Paul Horn. *Robot Vision.* Cambridge, MA: The MIT Press, 1987.

[Marr 82] David Marr. *Vision.* New York: W.H. Freeman and Company, 1982.

[Narayan & Chandrasekaran 91] N. Hari Narayan and B. Chandrasekaran. Reasoning Visually about Spatial Interactions. *Proceedings of the IJCAI*, 1991.

[Pentland 90] Alex Pentland. *THINGWORLD 2.0.* Vision and Modeling Group, The Media Lab, MIT. 1990.

[Ram 89] Ashwin Ram. *Question-driven understanding.* Dissertation, Department of Computer Science, Yale University, 1989.

[Ullman 84] Shimon Ullman. Visual Routines. *Cognition* **18**, 1984.

[Witkin & Tanenbaum 83] Andrew Witkin & Jay Tanenbaum. On the Role of Structure in Vision. In *Human and Machine Vision*, Beck, Hope, & Rosenfeld, (*eds.*), 1983.