Maintaining Visual Models of a Scene Using Change Primitives

Andrew Fano
Paul Cooper

The Institute for the Learning Sciences
Northwestern University
Evanston, Illinois 60201
fano@ils.nwu.edu
cooper@ils.nwu.edu

Abstract

The traditional goal of vision - scene model construction from an image - is inappropriate for vision systems that interact continuously with a dynamic world. In this paper, we present our initial explorations of an alternative approach which holds that a primary purpose of vision is to maintain an extant model. We argue that a model may be maintained by incorporating changes in the scene that can be characterized at a high level of abstraction yet manifest themselves at relatively low levels of analysis. We develop classes of qualitative change primitives that are easily detectable in the image. Existing task-relevant models and the associated domain knowledge are used to disambiguate the interpretation of these changes, allowing them to modify the existing model.

1. Introduction

The basic goals of computer vision are often assumed to be well-defined. For example, Charniak and McDermott claim that "unlike many problems in AI, the vision problem may be stated with reasonable precision: given a two-dimensional image, infer the objects that produced it, including their shapes, positions, colors, and sizes." [Charniak & McDermott, 1985]. A similar view of the aims of vision calls for techniques to be developed which can recover descriptions of physical properties of visible surfaces such as their distance and the presence of edges from inherently ambiguous and noisy primary image data [e.g. Poggio, et. al., 1987].

It has become clear, however, that vision is not a photo analysis task. In practice, vision systems interact continuously with the world. This insight has become popular as the paradigm of active vision [Bajscy 1988, Ballard 1989], which views visual perception as a process dynamically interacting with its environment, integrated with action, and responsive to cognitive goals of the agent. To date, active vision research has addressed primarily issues of what new constraints are offered by a moving camera [e.g. Alloimonos et. al. 1987, Propokopwicz & Cooper 1992], and how real-time computations of simple goal-directed properties can be achieved [Swain 1990, Coombs & Brown 1992]. But the assumptions about the higher level cognitive goals of vision must be re-examined as well.

The ability to generate a representation of a scene depicted in an image is, of course, of great value and ultimately necessary. We believe, however, that often a more modest aim will prove useful and feasible in many situations. That is, we suggest that the central high-level task for an active vision system is *maintaining* a model of a scene over time (and multiple visual samplings of the scene). The assumption of a stable scene justifies the further assumption that the relevance of a model of a scene will persist over time. If we must go through the difficult task of generating a model it seems sensible to try to use the fruits of the labor to their fullest potential before reengaging the process. As we sit in our living room, a room with which we are presumably intimately familiar, we do not continuously regenerate new models of the room. Reusing existing models of a scene would therefore appear to be an approach worth considering.

Scenes, of course, do change. This suggests that a central task for vision should be identifying and incorporating changes in the scenes to the existing model. This paper reflects our initial explorations of this idea. Our emphasis is the extraction of *qualitative* scene features at a high level of abstraction, that can be used to update a task-relevant model. There is currently great interest in developing qualitative visual properties that can be extracted from images [Kahn 1993, Mundy & Zisserman 1992]. We suggest that the identification of qualitative change primitives for model maintenance may be a useful approach.

Following a discussion of the potential advantages and disadvantages of the basic proposal, we discuss a task in detail, along with some task-relevant qualitative descriptors that can be simply derived from visual data. We also outline the beginnings of a set of domain independent qualitative visual change primitives, to illustrate the form that such an approach might take. Finally, we discuss commonalities between model maintenance approaches in tasks as diverse as language understanding and social simulation.

2. Feasibility

The initial issue to be addressed is whether the task of model maintenance is feasible at all, and if so, to identify potential areas of leverage and difficulty.

2.1. Ambiguity

Our approach can address the basic visual problem of resolving interpretive ambiguity. Regions of an image may be consistent with a variety of interpretations. However, in general, few of these interpretations are likely to be consistent with an existing model. Disambiguation can therefore proceed by requiring consistency with a complete, albeit possibly partially obsolete model, instead of one being built from scratch that is likely to be incomplete and contain other ambiguous regions. Moreover, since entirely new models of the scene are not being generated, the ambiguities which arise are largely restricted to the interpretation of changes.

2.2. Complexity

The most serious potential problem with this approach we can see is if noticing and incorporating changes to an existing model turns out to be less efficient than generating one from scratch. This seems unlikely. Animals, for example, are very efficient at noticing visual change. The architecture of animal vision systems, which provides the well-known property of motion sensitivity but not acuity in the periphery, also supports the argument. Moreover, the detection of changes in a scene even with a moving camera has been demonstrated to be tractable [Prokopowicz 1994].

A further reason to be optimistic about the feasibility of noticing and incorporating changes is that in many domains the types of visual changes which occur are often indicative of a small set of well defined changes in a description of the scene. Crudely, scenes are usually stable and when they're not, they change in predictable ways. Noticing and incorporating changes, therefore, may be facilitated by using knowledge of the types of changes that occur in the scenes of certain domains. Often changes which may be characterized at relatively low levels of description correspond with little ambiguity to changes in the model at a relatively high level. Furthermore, knowledge of the particular domain in which the scene occurs can additionally constrain the interpretation of a change. So for example, if we are looking at a still lake, or a pool, and there is a sudden change in texture of the water, it can only mean that an object has moved in the water.

Changes in a scene, of course, can only be due to the movement of the agent or other objects in the scene. However, our approach differs from traditional motion research by focusing on the incorporation of qualitative changes to a persisting model as opposed to the detection and parameterization of motion.

It is important to note that not all of the changes which are used to elaborate a model need be of a visual nature. So, for example, if we see a speeding car honking its horn, both auditory and visual information can be used to specify and maintain a single model of the car. According to this model, what it means to understand an input is largely independent of the modality of the input. It doesn't matter, for example, whether we see a red traffic light, or are told of its presence. Understanding a traffic light means creating an explanation of its relationship to our plans and goals. Traditional vision systems whose intended output is a model of a scene seem ill-suited for continuous use in this way since the models they produce would need to be constantly re-elaborated with other forms of information.

In short, the most important thing worth computing from moment to moment are the *changes* to the existing model as opposed to a restatement of the entire model.

3. Qualitative Changes to Task-relevant Models

If a central purpose for vision is change detection for model maintenance, the nature of important visual primitives will be drastically different from the traditional candidates. Under this view, the question to be answered by a given input is not "what is this?" If we assume an extant model, we largely know the answer. Instead the question to be answered is "what does this change mean?" Possible answers include identifying opportunities, threats, failures, constraints on alternatives, progress toward a goal, and so on. In each case, the meaning of the input will vary somewhat with the currently active plans and goals. Since the model is already known, features can be extracted that are directly relevant to modifying it in some way. Instead of extracting primitives such as edges or parameterized motions, we can find low-level visual evidence for the specific kinds of change that are

relevant at a high-level of abstraction. Clearly, the exact form that the model takes will play a large role in this framework.

We begin by assuming that models are task relevant. That is, our initial development assumes the seeing agent is executing a task, and that there exists a model of the surroundings sufficient to support the execution of this task. We are interested solely in incorporating into the model task-relevant changes evident in the scene. (Of course, the assumption that only one task is being pursued in unrealistic).

3.1. Qualitative Change Primitives for Driving

Consider the task of driving. Currently, there is much interest in automating the visual control of driving [e.g. Pomerleau 1991, Dickmanns 1992]. Most of this research is dedicated to developing real-time control responses for the simpler situations. But what makes the driving task relatively complicated (and in turn makes these approaches inadequate) is the types of changing situations that can arise.

Our model begins instead with a high-level specification of the overall task. In driving, the task, briefly, is to make progress toward a desired destination while avoiding collisions with obstacles. A useful model of the situation should contain representations corresponding to the location and nature of nearby obstacles, their velocities, and, if they are agents (e.g. cars and pedestrians) their intentions (e.g. are they aware of me? are they stopping? turning? cutting me off?). A model of the road will typically record boundaries such as the curbs, road surface (e.g. gravel vs. asphalt, potholes) and impending landmarks that require a response (e.g.. intersections and road signs). Finally, of course, a model requires a representation of one's self, including the direction, velocity, desired direction, etc.

Given this relatively complicated set of phenomena, what should vision do? Clearly, one task is to support the moment-to-moment control of steering, the focus of Pomerleau's work, for example. Our central point is that *given the circumstances, some visual inputs can be immediately understood as relevant changes and incorporated into the model.* We provide two examples.

First, consider the visual control of highway driving. If the model we start with is one in which the observer is driving on a highway at 60 miles per hour facing forward, and a line is introduced into the field of vision from the side, then there are very few things to which this change can plausibly correspond. With almost no exceptions it will be a car, truck, or motorcycle passing. What else could it be?

As skilled agents in the driving domain, drivers *expect* to see other vehicles attempting to pass. It seems quite plausible that these expectations will be partially expressed in terms of the change primitives most likely to uniquely identify them. Therefore, when this change is encountered, and this expectation is satisfied, the model is updated by introducing a new vehicle. Repercussions from the incorporation of this change include determining if the vehicle poses a threat (e.g. are they too close to me? could they stop in time?), ascertaining the degree to which one's options have been changed (e.g. can I still change lanes?) and plans have been facilitated or impaired (e.g. can I still get over to the exit? can I still see the car I'm following?).

In addition to updating the model and performing the inferences demanded by the task, new expectations can be specified, again, partly in terms of change primitives reflecting new possibilities in the situation. The new vehicle might slow down, for example. This change will be manifest, among other ways, by a sudden brightening at the center of the rear surface (the brake light activating). Sudden increases in light intensities can occur for many reasons, but, at this location in this situation, interpretation of the change can again be heavily constrained.

To reiterate the crucial point, a central difficulty for vision is bridging levels of abstraction from image input to high-level representations. But viewing vision as model maintenance via qualitative change primitives provides a simple and direct mechanism for bridging this representational gap. While we have, as yet, no implementation to support our claims here, it is obvious enough that the early vision portion of the system will be relatively trivial. It is equally clear, on the other hand, that substantial research in the representation of tasks must take place before a fully convincing demonstration can be built.

Consider a second specific example from the task of driving: Perhaps the most important obstacle to avoid when driving are pedestrians. One of the most closely monitored aspects of a driving scene are the curb boundaries. The curb boundaries demarcate the region containing the bulk of the information to which drivers must attend. Consequently, occlusions of the curb line often correspond to changes with important repercussions for the task of driving. New occlusions of the boundary line, evident from the appearance of a new angle with the curb line, will often signify the introduction of a new object onto the road, such as a pedestrian or car. Details of the occlusion such as the angle of intersection can help provide information such as the direction the object is heading. Of course many curb occlusions will be due to other cars between the driver and the curb. Since these cars are presumably part of the existing model, accounting for occlusions in

the vicinity of known objects should be relatively feasible. The monitoring of changes to the curb line would help explain why we are likely to immediately notice a pedestrian stepping onto the road, but not take note of the same pedestrian stepping onto the sidewalk from a store (intersecting a similar line formed by the sidewalk and buildings) moments earlier. As with cars passing, the appearance of pedestrians on the road is a predictable event in the course of driving. The expectation of this event can be expressed in terms of qualitative change primitives enabling the interpretation of these changes to be greatly constrained.

3.2. Domain Independent Change Primitives

Clearly, change primitives are not defined for a single domain. Instead, a set of generic change primitives must be developed that is useful in multiple domains. Domain-specific expectations may then be expressed in terms of these change primitives, whose satisfaction imply specific modifications to a task model.

In this section, we outline our initial development of change primitives based on two widely useful aspects of the spatial and visual world: *geometry* and *causality*. Our primitives are not meant to be exhaustive, or even completely correct. Instead, they are meant to suggest classes of such primitives that could be developed.

3.2.1. Geometric Change Primitives

As is widely understood, motion and change in the world cause changes in the projected geometries in an image. We outline some simple qualitative changes of this kind here, in particular involving *image angles* formed between linear scene features in projection. Extracting our change primitives does *not* require determining veridical scene angles. Instead, qualitative changes are observed in *low-level image data*, and combined with expectations deriving from the existing model to effect qualitative changes in the model *at high levels of abstraction*.

Movement of observer or object:

If an angle on a horizontal plane, straddling your line of vision becomes more acute (see figure 1)

- You are moving toward the angle (or the angle is moving toward you...) or:
- You are moving up

To demonstrate the point that early visual processing is not difficult, particularly since recovering scene characteristics is not necessary, we have processed the two images in Figure 1 to recover the actual angles. Robust determination of the fact that some qualitative change has occurred (e.g. "increase in obtuseness of an angle") is relatively straightforward.

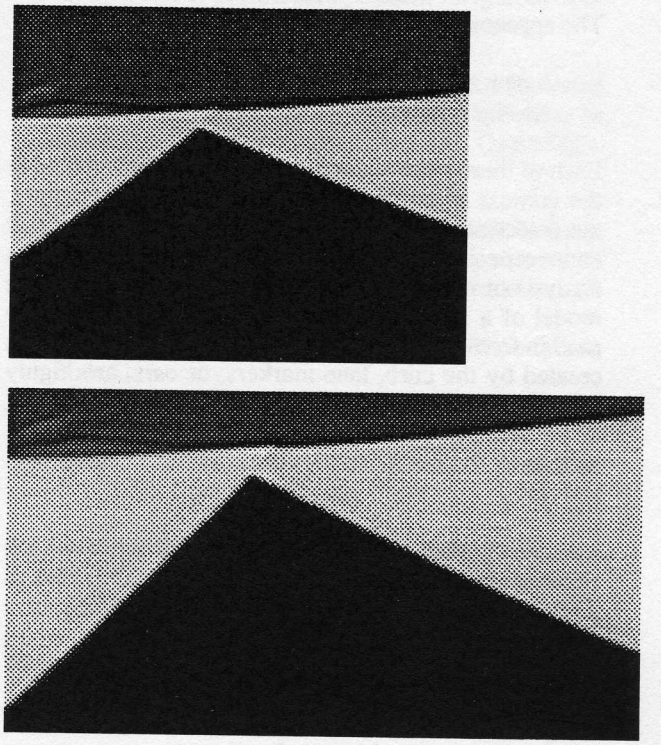

Figure 1.

Top: Far View
Bottom: Near View

As the observer approaches the sheets of paper, the angle they form becomes measurably more acute, as is evident in the near image.

If an angle on a horizontal plane, NOT straddling your line of vision becomes more acute

- You are moving away from the angle
- You are moving down

If the angle is on a vertical plane flip the above rules by 90 degrees.

Changes in angles formed by the edge of an object and a background line:

If the angle moves left and the vertex point on the edge is moving and the object edge is above the background edge:

- The object is moving down

If the angle moves left and the vertex point on the object edge is stable:

- The object is moving left

The appearance of a new angle can mean:

- A new occlusion
- An object is being rotated

Each of these rules becomes useful when instantiated in the context of an existing task model. For example, the detection of many of the most crucial changes that can occur in a driving scene may be aided by an analysis of the changes of angles involved. A driver's model of a scene is dominated by the activity on the road in front of him. In particular, changes to the lines created by the curb, lane markers, or cars, are highly significant. Focusing one's attention to changes that occur to these lines allow one to maintain the relevant aspects of the model and ignore irrelevant background activities.

The detection of some of these change primitives, such as a new angle, in the region of the image identified as a car, for example, would indicate the opening of a door or the turning of the car.

3.2.2. Causal Change Primitives

Looking beyond image geometry, it is productive to consider the scene in terms of the physical causality of the objects in it [Brand, Birnbaum, & Cooper 1993]; that is, why objects are interacting or moving. The initiation of motion, for example, is clearly important -- - a force is being exerted. Given an existing model, just noticing motion initiation should provide important evidence about how to update the model. We outline a list of these kinds of causal scene phenomena, and in a few instances, describe how the domain independent interpretation could be disambiguated in a specific domain, in this case (again), the domain of driving.

For example:

Still to Moving: smooth acceleration

- A force has been applied
- A force has been removed
- A machine has been turned on
- Stability is lost
- The viewer is accelerated

Still to Moving: Impulse

- An impact has occurred
- An object has been broken
- You have hit an object

Moving to Still: Smooth Deceleration

- A force has been applied or removed
- A machine has been turned off
- Stability is achieved
- You have accelerated to velocity of objects

As with the geometric changes listed above, the interpretation of these changes can be further disambiguated by knowledge of the task. For driving, the still-to-moving transition is particularly important and simple to interpret. Motion where there was previously none almost certainly suggests a car starting to move. New relative image motion from a moving car suggests a relative acceleration between two cars, etc.

Geometric and causal change primitives are simply two examples of the kinds of changes which can be used to update a model. We have performed similar analyses on changes involving texture, color, brightness, and occlusion. These rules, of course, are far from complete. The point was simply to explore how changes in these attributes might help constrain the maintenance of a model of a scene.

We are aware that many of these features, such as the change of position of an object may be difficult to compute. Traditional model-based vision requires the computation of these attributes in general, with very few applicable constraints. We suspect that the burden of computing these features will be eased by the use of expectations derived from the existing model.

4. Model Maintenance and Understanding

This view of vision is consistent with more general approaches to problems in cognition. In natural language processing, for example, a model of the text is not constantly regenerated. Instead it is updated as more text is processed. The existing model is used to help disambiguate references that are otherwise ambiguous in the text currently being processed. DMAP uses a very similar approach to avoid having to build individual representations for each sentence. Instead it directly incorporates information from what it is processing into an evolving model [Riesbeck & Martin, 1985]. Learning too can be viewed this way. We don't learn entire new domains at once. Instead we slowly augment an existing model of a domain with small amounts of new knowledge.

Our work in social simulation uses the basic idea of maintaining a model by incorporating changes in the world to govern the behavior of the simulated agents.

We are extending Hughes' Chimpworld [Hughes 1993], a simulation of chimp politics in which agents maintain models of the aspects of the world that are relevant to their actions. We are developing a general social simulation tool in which agents predict the occurrence of changes in the world. These changes are identified by using knowledge of how they manifest themselves. Presently our agents only operate in a simulated environment but we believe we can begin to extend this work into machine vision by beginning to include information on the visual manifestation of the types of changes which can occur.

5. Conclusion

The work we have presented represents a preliminary exploration of the feasibility of maintaining a model of a scene through time. While we recognize that the ability to generate a model of a scene from an image is of utmost importance, we believe the approach we propose is both feasible and appropriate for systems which are intended to interact continuously with the world. The primary advantage of such a framework is that a persistent visual model is made available for use by an encompassing system which can both be augmented by other forms of information and also used to disambiguate further visual input.

Crucial to the success of such a framework is the identification of changes which imply abstract, qualitative changes to an existing model yet can be detected at low levels of analysis. We have outlined an initial simple set of changes we believe to be promising and shown how they might be specialized to the scenes of particular domains. Equally important to the success of the framework is the further detailed development of task-relevant model representations.

References

Aloimonos, J., Bandapadhay, A., & Weiss, I. (1987). Active Vision. In Proceedings of *First International Conference on Computer Vision.*

Ballard, D.H. (1989). *Animate Vision.* Proceedings, IJCAI-89.

Bajscy, R. *Active Perception.* (1988) Proceedings of the IEEE, 76:996-1005.

Birnbaum, L., Brand, M., & Cooper P. (1993). Looking for trouble: Using Causal semantics to direct focus of attention. In Proceedings of the *Fourth International Conference on Computer Vision ICCV '93*, Berlin, Germany.

Charniak, E., & McDermott, D. (1985). *Introduction to Artificial Intelligence.* Reading, MA: Addison-Wesley. P. 89.

Coombs, D. J. & Brown, C. M. (1992). *Intelligent Gaze Control in Binocular Vision.* Department of Computer Science. University of Rochester.

Dickmanns, E. D. (1992). Expectation-based Dynamic Scene Understanding. In Blake, A. & Yuille, A. (Eds.), *Active Vision* (pp. 303-335), Cambridge, MA: MIT Press.

Feldman, J. A. (1988). *Time, Space and Form in Vision.* Technical Report #244. Department of Computer Science. University of Rochester.

Hughes, L. (1993). Forthcoming Ph.D. Dissertation, Yale University.

Kahn, P. (Ed.) (1993). Proceedings of the *IEEE Workshop on Qualitative Vision*, Los Alamitos, CA: IEEE Computer Society Press.

Mundy, J., & Zisserman, A. (Eds.) (1992). *Geometric invariances in computer vision.* Cambridge, MA: MIT Press.

Poggio, T., Torre, V., & Koch, C. (1987). Computational Vision and Regularization theory. In M. Fischler & O. Firschein (Eds.), *Readings In Computer Vision.* Los Altos, CA: Morgan Kaufman.

Pomerleau, D. A. (1991). Efficient Training of Artificial Neural Networks for Autonomous Navigation. *Neural Networks 3:1*, Terrence Sjnowski (ed.).

Prokopowicz, P. (1994). Forthcoming Ph.D. Dissertation. Northwestern University.

Prokopowicz, P. & Cooper, P. (1993) *The Dynamic Retina: Contrast and Motion Detection for Active Vision.* Forthcoming Technical Report. The Institute for the Learning Sciences. Northwestern University.

Riesbeck, C. & Martin, C. E. (1985). *Direct Memory Access Parsing.* Technical Report #354. Department of Computer Science, Yale University.

Swain, M. J. (1990). *Color Indexing.* Technical Report #360. Department of Computer Science. University of Rochester.