Action, Representation, and Purpose: Re-evaluating the Foundations of Computational Vision

Michael J. Black (Chair)
Department of Computer Science
University of Toronto
Toronto, Ontario M5S 1A4

John (Yiannis) Aloimonos
University of Maryland at College Park
Center for Automation Research
College Park, MD 20742–3411

Christopher M. Brown
Department of Computer Science
University of Rochester
Rochester, New York 14627–0226

Ian Horswill
MIT AI Lab
545 Technology Square
Cambridge, MA 02139

Jitendra Malik
University of California, Berkeley
Computer Science Division
Berkeley, CA 94720

Giulio Sandini
DIST University of Genova
via Opera Pia 11 A
16145 Genova – Italy

Michael J. Tarr
Department of Psychology
Yale University
New Haven, CT 06520–7447

Abstract

The traditional goal of computer vision, to reconstruct, or recover properties of, the scene has recently been challenged by advocates of a new purposive *approach in which the vision problem is defined in terms of the goals of an active agent. In the starkest light the debate can be characterized as one about the role of explicit representations. The extreme traditionalists strive for a detailed representation of the 3D world while the other extreme adopts a strict behaviorist stance which eschews representations in favor of "direct sensing." This panel will explore the roles of action, representation, and purpose in computer vision and, in doing so, will hopefully discover areas of agreement.*

1 Panel Summary

What should be the goal of computer vision? The traditional view (as exemplified by the work of Marr [1982]) poses the problem as:

> The description of the three dimensional world in terms of the surfaces and objects present and their physical properties and spatial relationships.

The alliance between vision and artificial intelligence is founded on this view of vision as a black box providing symbolic scene descriptions. This traditional *recovery paradigm* has recently come under attack for failing to produce accurate and robust descriptions of the world. These failures have prompted critics to propose a new behaviorist paradigm for vision research which emphasizes task-driven perception. From this perspective, the goal of vision becomes:

> The development of fast visual abilities which are tied to specific behaviors and which access the scene directly without intervening representations.

Within this *purposive paradigm*, general scene reconstruction is abandoned in favor of task specific modules that are fast, simple, robust, and often qualitative.

Partisan Purposivism

Advocates of the purposive paradigm see vision in the context of the tasks which an organism must accomplish. In restricted domains, the visual information that supports a particular behavior is identified. Robust procedures are then devised to extract the necessary information directly from the scene and implement the behavior. More complex behaviors are seen as emerging from collections of simpler behaviors.

To support their view, the "purposivists" have leveled the following criticisms at the recovery paradigm:

- current vision systems lack both robustness and accuracy,
- by ignoring the purpose of vision, current systems have failed to generate interesting robotic applications,
- simple purposive systems have been built which exhibit complex behavior that has eluded robots equipped with traditional vision systems.

Central to this approach are the following ideas:

- vision must be considered in the context of the tasks that an organism must perform,
- vision is not isolated but, rather, functions as part of a complex system,
- by having fast and simple processes, the world can act as its own representation, obviating the no need for intermediate, or shared, representations,
- complex behavior emerges from layers of simpler behaviors,
- the purposive view is consistent with the evolution of biological organisms.

Religious Reconstructionism

On the other side of the issue are those who argue that is too soon to abandon the goals of the recovery paradigm and that, in fact, the recovery approach, with its emphasis on representations, provides the best hope for modeling and understanding "general purpose vision" in humans and machines [Tarr and Black, 1991]. While recognizing that the purposive paradigm may be appropriate for describing low-level, or reflexive, behaviors they level the following criticisms at the approach:

- the purposive solutions will not "scale up" to more sophisticated problems,
- the purposive approach cannot account for the complex visual abilities of humans.

Proponents of the recovery paradigm argue that vision requires general flexible representations that can support a variety of tasks. This leads to the following tenets of the recovery paradigm:

- intermediate representations are necessary to reduce the computational cost of vision,
- complex behavior in unknown environments is made possible by powerful representations,
- the notions of action, attention, and purpose are entirely consistent with the recovery paradigm and do not entail a purposive approach,
- the recovery paradigm provides the best hope for understanding human perception,
- the recovery paradigm is consistent with the evolution of biological organisms.

Intelligence and Representation

The purposive movement in computer vision parallels similar movements in the rest of AI; for example, Rodney Brooks has been a strong proponent of a behavior-based view of AI consistent with the purposive approach. The popularity of these approaches has been increasing as evidenced by recent IJCAI awards. At the last IJCAI in Australia, Brooks was presented with the "Computers and Thought Award" for his paper "Intelligence Without Reason" [Brooks, 1991], while at the 1989 IJCAI, Dana Ballard, a proponent of "Animate Vision", received the award for best paper (see [Ballard, 1991]).

Within vision, the increased emphasis on behavior can also be seen in work on visual attention and active vision.[1] The popularity of these behavior-based approaches has grown with their success, but it is time to take a critical look at the goals of the purposive paradigm, decide wherein its contribution lies, and re-evaluate the goals of the field.

The Panel Discussion

This panel provides a critical examination of the significant questions surrounding the purposive paradigm:

- Does the "purposive" view represent a new paradigm for computer vision?
- Is vision ever non-purposive?
- Is purposive vision simply good engineering?
- Can complex visual behavior arise from a collection of task-specific modules or are there powerful underlying representations supporting a variety of flexible behaviors?
- What is the biological and evolutionary evidence supporting the various paradigms?
- Are the purposive and recovery paradigms mutually exclusive or is there some common ground where they can coexist?

The panel discussion is intended to search for the common ground that exists between these approaches, make clear the relationships, and point to future research directions. To achieve this synthesis, the panel consists of researchers spanning the representation–action spectrum.

2 Commentaries

John (Yiannis) Aloimonos

[1] There are numerous competing and similar phrases in use, including: active vision, active sensing, animate vision, purposive vision, selective perception, and visual attention. Due to a lack of precise definitions, these terms are often confused or used interchangeably.

Purposive and Qualitative Active Vision

The purposive paradigm is a fundamentally new way of examining problems of visual perception. Up to now, vision was regarded as a recovery problem, i.e., as the problem of reconstructing an accurate representation of the 3-D scene and its properties from image cues such as shading, contours, motion, stereo, color, etc. This approach has contributed many theoretical results and has led to new mathematical techniques, e.g. related to regularization and discontinuities. But what is vision for? Why do animals have it and why do we want to understand it? The answer is, of course, that we need vision in order to accomplish visual tasks. In the biological world, organisms need vision in order to recognize their friends and enemies, avoid danger, find food and in general survive. In the world of robots, vision is needed to make them capable of performing various tasks while interacting with their environment. However, recovering the scene and its attributes is not a necessary condition for the accomplishment of visual tasks. Many such tasks can be achieved visually without reconstruction but through the recognition of patterns, objects or situations.

"What to recognize" is concerned with the questions we pose. The purposive paradigm calls for formulating questions that are directly related to visual tasks, i.e. that have a purpose. Knowledge of 3-D motion is much more than we need to answer the purposive question: Is this moving object coming closer to the observer? Purposive thinking leads us to pose questions whose answers will only help to solve the particular task at hand, and will not be of general use. This level of the paradigm is parallel to Marr's computational theory and makes sure (or rather, tries to insure) that the resulting algorithms will be of minimal complexity.

"How to recognize" (patterns, objects or situations) is related to the algorithmic level of Marr's paradigm. Qualitative vision calls for the development of algorithms that are simple, robust and based on qualitative techniques, such as comparisons of quantities or discrete classifications. Qualitative vision, which in the past has been wrongly called inexact, makes sense here because it is coupled with purposive vision, which formulates questions for which qualitative solutions are possible.

To demonstrate the usefulness of the approach we consider visual motion (or navigation) problems, and assume that the observer is active. We describe the preliminary design of Medusa, a purposive and qualitative visual motion machine that can robustly solve many navigational problems without reconstructing the scene (for details see [Aloimonos, 1990]).

Christopher M. Brown

General Vision:

General vision is defined as that form of vision that reliably, quickly, and compellingly links us to an outside world that is structured but whose behavior is not predictable in detail. When our eyes are open, we (and other animals) can see what is out there, including unexpected things, well enough and reliably enough to survive.

Reconstructionist Vision:

Much work and thought has gone into extracting physical property images supporting quantitative punctate visual judgements (depth, reflectance, motion, slant, tilt), and as stepping stones to qualitative judgements and actions (segmentation, recognition, manipulation). Punctate reconstruction ("Physics-based vision") has been for some time the dominant paradigm in computer vision. Reconstructionism does not rule out active techniques, and the work is growing ever more sophisticated (e.g. [Aloimonos and Schulman, 1989]).

Reconstructionism usually comes with a bottom-up, data-driven approach to general vision. This is good: high-level knowledge (or wishful thinking, or probability) has nothing to do with low-level vision. Methodologically, however, reconstructionism can lead to unjustified assumptions of linearly interacting, modular subsystems and thus to potentially irrelevant research on such decoupled phenomena, outside the context of a working system.

Neurophysiological and Psychophysical data is complex and open to many conflicting interpretations.

Purposive Vision:

One current shortcoming of purposive, animate vision is that its scientific and technical claims are not agreed on. Purposivists may claim that general vision will emerge from an organization (a hierarchy, an interacting set) of special-case vision solutions. How? How many purposive abilities are necessary? Are there primitive purposes, and if so what are they?

Whatever it is, purposive vision often comes with a top-down, goal-oriented, task-driven approach to general vision. This is good in that it encourages integration of visual systems. It makes some design suggestions about integrated vision architectures, such as hierarchies of purposive behaviors. However, the purposive approach can discourage thinking about general vision. How do we see unexpected or unwelcome or improbable things? How does what we see affect our goals (as opposed to vice versa)? Postulating a "general, reliable, visual awareness" purpose begs the question.

Purposive vision places emphasis on dynamic, multi-resolution vision algorithms, and interaction with the world. This is good.

Computer Vision:

Both sides cite complexity arguments in their favor. However, until more details of a theory of general vision appear, it is difficult to gauge the potential contributions and costs of task-independent and task-dependent visual mechanisms.

At this point, both reconstructive and purposive theories largely ignore learning. A computer is certainly a relative *tabula rasa*, and it seems practically advisable to provide machines with a way of evolving, developing, or learning what they need to know rather than to try to wire or program it in ourselves.

Still unaddressed are the problems of dynamically prioritizing among possible tasks, and of picking an appropriate level of visual awareness (e.g. segmentation, or "what is the relevant 'object'?"). General vision does not mean doing all possible visual processing; that is technically infeasible. Thus the control of selective perception becomes a central issue.

I shall present recent results (algorithms, data) in computational models of intelligent control of perception in static and dynamic scenes (videotape). It is likely that I will have some psychophysical demonstrations (if not "experiments") to share as well, created using our eye-tracker with subjects performing search tasks.

Ian Horswill

Representational Agnosticism

Suppose we need to build a doorknob detector. One way of doing this is to build a model of all visible surfaces and their markings, then look at the model to decide if there are any doorknobs in it. Whether our system computes a model or not, the model is not its goal–doorknob detection is. If finding doorknobs directly in the image is more efficient than building the model, then we may not want the model.

One objection to this reasoning is that it assumes that we only use vision for one thing when, in fact, we use it for many things, perhaps doorknob detection at one moment and finding prey the next. There seem to be an unbounded number of visual tasks and we don't want to build a different box to compute each one from the grey levels.

So suppose now that we want the system to answer a number of questions: "is s/he looking at me?", "am I in a corridor?", "which way is the restroom?", *etc.* We can build a model of the surfaces and their markings, but then rather than having a system which answers only one of our questions, we have a system which doesn't answer any. We *still* need task-specific processing to answer each question, even with the model. We'd like to think that an image is like a picture and a model like the object itself, but the model is really more like a hologram–it's just another picture which needs to be interpreted, only it's in 3D.

One might reply that interpreting the hologram is simpler–that basic preprocessing has been done which is needed for any visual task. Certainly a system which can perform many visual tasks will need to reuse modules for many different tasks and share preprocessing steps between large numbers of tasks, but it's not clear that all visual tasks require reconstruction as a preprocessing step; surface models are useful for many tasks, but are they so useful for so many tasks that we should *define* vision to be surface reconstruction? Can we really expect to build good surface models of trees, ponds, clouds, waterfalls, or mountains? Even if we could, how would the extra dimension help us distinguish between a tree and a cloud? One might say that trees are green and bushy on top and brown and relatively thin on the bottom, but I doubt that it's easier to measure bushiness in 3D than in 2D.

I feel the recovery paradigm as outlined in the introduction is misguided; no single representation is the goal of all of vision. However, the introduction's purposive paradigm errs the other way. I agree that vision needs to provide fast visual abilities, but that says nothing of whether to use intermediate representations. (I have serious doubts whether there is an objective definition of representation, intervening or otherwise, and have watched intelligent people have heated arguments over whether something was a representation.) I think it's best to remain representational agnostics and to decide what computations are best for a given problem, with or without mediating representations.

Let's worry about what to see before deciding how to see it.

Jitendra Malik

Let us start with a general and broad enough view of vision that we can all agree on–vision is about starting from a stereo pair of spatiotemporal image sequences $I_l(x, y, t)$ and $I_r(x, y, t)$ gathered by a (possibly moving) observer and the extraction of information adequate to subserve the needs of manipulation, navigation and recognition. Clearly there was an evolutionary pressure for this in biological vision and we can argue from an engineering point of view that most of our vision applications fall into one or more of these categories.

There is nothing particularly original or novel about this view, contrary to the claims of the active or purposive vision partisans. For a fairly eloquent statement, I recommend Gibson's books (eg. [Gibson, 1979]) – his notion of affordances is very much the kind of thing that is being implemented in active vision groups around the world.

The key question of scientific interest is how do we go from the image level to a representation of the information useful for the task/s. Gibson had the peculiar notion of 'direct pickup' which it is fair to say is discredited by what we know from neurobiology. There are clearly intervening stages–in primate vision there are a number of different areas in the visual cortex each with its own complete retinotopic map and with a hierarchy of information flow with a notion of 'earlier' and 'later' stages. It is quite natural to regard these as intermediate representations. While the number of stages is greater in higher animals, there is a primitive notion of hierarchy even in the simplest of vision systems, such as the housefly much studied by the late Reichardt and his collaborators in Tubingen. There one can think of the different layers of a multilayer neural network as constituting the intermediate representations.

From a computational point of view, arguments for the need for intermediate representations can be made on the basis of computational complexity (see e.g. Tsotsos [1992]). If you consider very simple tasks, such as demonstrations of tracking white balls on dark backgrounds, the intermediate representations are trivially simple but they nevertheless exist. If one wants to deal with the rich complexity of natural scenes, all evidence suggests one would have to have ever more sophisticated intermediate representations, or more of them, or more stages of them.

The useful discussion is therefore on the nature of these intermediate representations. Marr had a specific proposal for what these intermediate stages might be–the two key ones were (1) the primal sketch, based on detecting, localizing and grouping edge tokens, and (2) the 2.5 D sketch based on a reconstruction of depth and/or surface orientation from various shape-from-X modules. Most of the arguments coming from the active vision/qualitative vision/task-based vision camp are criticisms of the 2.5 D sketch representation. I happen to agree with many of them, but to me these arguments merely suggest that alternative representations should be sought. It is also reasonable to be agnostic about whether these intermediate representations are common for various tasks, or specialized.

Let me conclude with the working hypothesis on the nature of the intermediate representations that we have adopted in our research in computational vision at Berkeley. I believe that the use of a first stage of edge detection as the precursor to all the early vision modules such as stereo, texture and motion analysis is a mistake. This might have been a reasonable thing to do in indoor scenes of smooth, non-textured polyhedral objects (Roberts's original domain), but for outdoor scenes, the fraction of detected brightness edges that correspond to geometrically significant object boundaries is small. I believe that instead one should use the result of convolving the image with a bank of first or second Gaussian derivative filters at different orientations and scales [Perona, 1991]. This particular representation was suggested by biological evidence but can be prefectly well motivated by computational criteria. We have shown how this representation can be used as a precursor to stereopsis, texture and motion processing. The result of this processing is a set of retinotopic maps of disparity, optical flow etc. along with boundaries defined by one or more of the attributes of color, brightness, texture, optical flow and disparity. 2D maps of optical flow, disparity etc. carry in them information that could directly be used for guiding tasks. Or one could imagine constructing additional intermediate representations. At this stage we are pursuing both lines of research. There are advantages for both. Specializing for a task enables one to incorporate additional constraints e.g. in driving a car on a road we need to be concerned with only 2 degrees of freedom instead of 6 for the general structure-from-motion problem. Generic processing steps may be useful from the point of view of being good for multiple purposes and thus avoiding a combinatorial explosion.

It seems it would be more fruitful if we concentrated our attention on what intermediate representations are best, developing algorithms for computing them, showing that they were robust, and demonstrating that these representations were useful for particular tasks.

Giulio Sandini

The advancement of science is characterized by debates, and intellectual battles between researchers seeing "things" from different perspectives and having different beliefs. Fortunately computer vision is not different. On one side, in fact, it demonstrates the complexity of the problems that the community is trying to understand (and to some extent to solve), and on the other, these shifts in interest demonstrate that we are learning from past errors and successes: we know better what we can achieve and what is unfeasible. For this reason I believe that the on-going debate between "reconstructionists" and "purposivists" should be considered not as a war of religion, but as a natural consequence of the lively research activity in the field, and, to a limited extent, of the wide research efforts devoted to "reconstruction" in the last few years.

One of the key issues in this debate is that of *complexity* and the related issue of "how to build a complex system". The first observation is that "complexity" does not mean "generality": one could build a complex system without aiming at general purposiveness. On the contrary I believe that such a generality is not achievable and not even reached by the human visual system. This is the major objection I have with respect to the arguments put forward in the introduction of the reconstructionist approach, namely to think that it is possible to aim at a *general purpose* system. Even the human visual system is not general purpose but it is required to operate in a very specific environment and to perform a limited number of tasks: it is not of much use underwater, it only detects a limited number of wavelengths, it does not allow to measure volumes in metric terms, it does not give an absolute measure of color, it cannot perceive a flying arrow, it can't even distinguish between a "real" scene and mirrored one.

Conversely the major criticism raised with respect to the purposivism is the fact that the approach is too fragmented and that the number of functionally independent, task specific modules, may be too large to represent efficiently. Currently there is no existence proof against this criticism and, as such, it has to be considered carefully: does the purposive approach scale up? The reason why we are confident that this can be achieved is that behaviors can be composed (much like reconstruction approaches) and therefore complex behaviors are not completely "functionally independent" on the contrary they could be *dynamically compiled* according to purpose. Following Dana Ballard's ideas of "RISC models of visual behaviors", the debate between "reconstructionists" and "purposivists" is similar, as far as the complexity issue is concerned, to the on-going confrontation between RISC and CISC computers: so-called general-purpose computers can, in fact, be efficiently realized using RISC CPUs. The ultimate demonstration shall be the building of complex machines and, to this aim, behaviors can be used to identify not only what to compute but also how to use it. This may, in fact, lead to behaviors sharing the same computational processes (instructions?) even if they may be acting entirely independently. The current state of the art, in this respect, is to study simple purposive behaviors to find practical solutions and to help identify processing commonalities.

In order to better explain this concept a few examples derived from the recovery of the optical flow field or measures derived from it will be presented in some task-driven (purposive) visually guided behaviors.

Michael J. Tarr

Wither Representations?

There has been growing interest in the idea that intelligence can be modeled, implemented, and understood without representation. In particular, researchers in computer vision and robotics have vocally supported the notion of "purposivism" – that is the development of AI systems through task-specific encapsulated modules, each offering only local functionality in a restricted domain. The stated assumption is that so-called "higher-level" behaviors will emerge from the combinatorial properties of many task-specific modules, e.g., without the need for representational structures more complex than the local input and output (generally successful task completion). In computer vision this view has been contrasted with the more traditional "reconstructive" approach that takes as its goal the derivation of functional descriptions of the visible world that are useful for general vision. However, because of supposed inadequacies, failures, and design incompatibilities, supporters of the purposive approach argue that reconstruction will neither lead to successful machine vision systems, nor to a better understanding of biological vision. Consequently they suggest that current approaches to AI are flawed and should be discarded in favor of their new, more practical alternative.

In response to these arguments, several observations may be made. In terms of possible goals for AI, it is whether the objective is to develop systems on a par with human performance, to develop systems that mimic less complex organisms, or to develop systems that accomplish a given task regardless of the relationship to biological intelligence. While the latter two objectives may play an important role in AI research, they do not satisfy the sometimes unstated assumption that a basic goal of AI is to accomplish the first objective – the development of systems that can perform perceptual, cognitive, and motor tasks comparable to humans. If this aim is to be met, there must exist reasons to believe that the purposive approach offers an adequate research paradigm. On the contrary, an analysis along psychological, evolutionary, and computational dimensions suggests that the purposivism *alone* is insufficient to understand and emulate complex behaviors, and, in particular, many of the behaviors associated with human intelligence.

One reason for this skepticism is doubt as to whether behaviors more complex than those specific to each module will emerge from their assembly. For it is not simply the gathering of large numbers of domain-specific mechanisms that produces "intelligence," but rather the ability of such modules to share their output in a manner that facilitates inferences about the structure of the world. Thus, while modules may be task-specific in their implementation, they are general in their ability to contribute to more complex representations. However, if this interaction is removed, it is unclear as to whether flexible processing mechanisms will emerge or even be attainable – in particular, because a purposive system is faced with the burden of implementing modules for *every* distinct behavior, often with a significant duplication of effort.

Another reason to question whether the purposive approach offers a panacea is that purposive systems are not really devoid of representations, but rather that the representations are not explicit. Indeed, many purposive systems implement constraints and algorithms quite similar to those used in traditional approaches to AI. Moreover, it is often the case that such systems, explicitly representational or otherwise, have been successful to the extent that they include suitable representations. Thus, because representations often provide a much clearer understanding of the problems at hand, there are few reasons to abandon this kind of approach.

In summary, these and other pieces of evidence lead to the conclusion that the representational approach offers a viable framework for understanding both human and machine intelligence, and, moreover that there already good examples where the approach has already been successful. However, it is not that AI should pursue traditional representational approaches at the exclusion of purposivism, but rather that a blending of the task-specific purposive paradigm with traditional representational approaches will prove more fertile than either approach alone.

3 Conclusion

In these commentaries we are beginning to see a softening of the extreme positions of the "Partisan Purposivists" and "Religious Reconstructionists". The community is realizing that the extreme behaviorist and reconstructionist views are both untenable, and moreover, that there is much to be gained from bringing the camps together in what might more appropriately be described as *Pragmatic* Purposivism or *Revisionist* Reconstructionism.

References

[Aloimonos and Schulman, 1989] J. Aloimonos and D. Schulman. *Integration of Visual Modules: An Extension of the Marr Paradigm*. Academic Press, Boston, 1989.

[Aloimonos, 1990] J. Aloimonos. Purposive and qualitative active vision. In *Proc. Int. Conf. on Pattern Recognition*, volume 1, pages 346–360, Atlantic City, NJ, June 1990.

[Ballard, 1991] D. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[Brooks, 1991] R. A. Brooks. Intelligence without reason. *IJCAI*, vol. 1, pp. 569–595, Aug. 1991.

[Gibson, 1979] J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA, 1979.

[Marr, 1982] D. Marr. *Vision*. W. H. Freeman and Company, New York, NY, 1982.

[Perona, 1991] P. Perona. Deformable kernels for early vision. *CVPR-91*, pp. 222–227, Maui, June 1991.

[Tarr and Black, 1991] M. J. Tarr and M. J. Black. A computational and evolutionary perspective on the role of representation in computer vision. Technical Report YALEU/DCS/RR-899, Yale University, October 1991.

[Tsotsos, 1992] J. Tsotsos. On the relative complexity of active vs. passive visual search. *IJCV*, 7(2):127–141, Jan. '92.