# Seeing and Understanding: Representing the Visual World

Y. ALOIMONOS, C. FERMÜLLER, AND A. ROSENFELD

*Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, MD 20742-3275 ⟨{yiannis,ar}@cfar.umd.edu⟩*

Most research in computational vision has concentrated on efforts to recover three-dimensional (3D) metric structure from images. This may be because research communities are influenced by fashion. Our views about the architecture of an intelligent system possessing perception, that is, its components and their relationships, are influenced by the current dominant views regarding intelligence, the mind, and brain. One view about the brain, which has been and still is widely held by neurologists, assumes that there is a separation between the "two causally linked faculties of seeing and understanding, the former a passive and the latter an active process" [Zeki 1993]. As a consequence of this view, computational vision is commonly treated as a discipline whose goal is the recovery of metric descriptions of the scene that can be utilized for reasoning.

After considerable effort and many theoretical results about scene recovery, computational vision is now moving into a relatively mature stage. It is now well understood that when we make assumptions about the scene, that is, impose specific models on the visible world, recovery of the model is usually possible provided the assumptions actually hold. For simple models of the geometry and the physical properties of the scene, we already have working systems that are finding their way into industry. But developing vision systems for particular environments, although useful, does not illuminate basic questions regarding the process of vision.

Perhaps the most important lesson learned from all the research on computational vision has been the realization that recovery of a complete metric description of extrapersonal space from images is very difficult. This, along with recent results in neurobiology and technological advances that allow us to "look at" brains "seeing" and "thinking" [Posner and Raichle 1994] is slowly changing our view about the architecture of vision. The view of neurologists is that "it is no longer possible to separate the process of seeing from that of understanding" [Zeki 1993]. An emerging theme is that an intelligent system with vision may create several representations of its spatiotemporal environment that it can use to accomplish particular tasks.

Clearly, these representations can be less intricate than metric descriptions; they may even be "qualitative" rather than quantitative. The understanding of these representations, their robust extraction in real time, their placement into memory, and the creation of an indexing mechanism based on them are fundamental research questions. Further thought about the nature of these representations leads to the conclusion that the tasks in which the system will use them are a major factor in their definition. The purpose of a visual computation determines to a large extent the nature

of the representations involved in it [Aloimonis 1993]. It would, however, be of little use to develop a different representation for every single task. Luckily, many visual tasks, involving navigation or manipulation, share various environmental invariances.

If $R(x, y)$ is the range to every point in the scene measured from the eye and indexed in retinal coordinates (a metric description), we may call $f(R(x, y))$ a qualitative description if $f$ is some function that is not known exactly. Clearly, we do not have much freedom in choosing the function $f$; it will be determined by the geometry and physics of the particular problem we are addressing. Action, the ultimate goal of the computation, can also provide constraints on $f$. Knowledge of general facts about and properties of $f$ might be enough to allow the map $f(R(x, y))$ to be used to accomplish certain tasks. We might know the general behavior of $f$ as regards singularities, monotonicity, and so on. For example, if $f$ is a monotonic function of $R$, then its values can be used to order the range values and create an ordinal depth map. For a stereo system, without complete knowledge of its extrinsic parameters, we can develop a function of range of the form $f(R) = a/R + b$, around the fixation point, which allows the creation of an ordinal depth map [Fermüller and Aloimonos 1995a].

The models of space that we use in our heads are probably too complex for us to determine. We can, however, come close to them by developing a large set of qualitative descriptions. Thus an interesting and feasible research program would be to redo all the computer vision we have learned—this time using more sophisticated models of the image-formation process and no specific assumptions about the scene in view—and search for useful functions $f$ of the distance from the eye to the world.

These ideas fit well with provisional models of vision and space such as Feldman's [1985] "four frames" model. In the retinotopic frame, information is extracted about the gradients of the spatio-temporal image in many directions. Then, through a process of matching patterns of local measurements as in Fermüller and Aloimonos [1995b], the system could develop a number of representations of space-time associated with each view. These representations could be used in the stable feature frame to stabilize the image and build a stable frame through a process like mosaicing, but with the matching taking place not at the retinal level but at the levels of the representations. Also, the representations associated with a view could be used to index into memory. Space could then in principle be represented as a collection of views [Rosenfeld 1987; Poggio and Edelman 1990; Ullman and Basri 1991] along with their associated representations. These might range from pure 2D representations (images) to pure 3D representations. Mumford's 2.1D sketch [1990] is an example of such a representation. Other examples include hazard maps, ordinal maps, and maps containing affine or projective structure.

In spite of all the work in this field, "we really have no clear idea how we see anything. This fact is usually concealed from students.... Surely after all that careful work and after all those elaborate arguments it would be bad form to suggest that we still lack any clear scientific understanding of the process of vision. And yet, by the standards of the exact sciences (such as physics, chemistry, and molecular biology), we do not yet know, even in outline, how our brains produce the vivid visual awareness that we take so much for granted." [Crick 1994]. We expect that the study of qualitative representations of visual space will contribute significantly to our understanding of perception.

## REFERENCES

ALOIMONOS, Y 1993. Active vision revisited. In *Active Perception*, Y. Aloimonos, Ed, Lawrence Erlbaum Associates, Hillsdale, NJ, 1–18.

CRICK, F 1994. *The Astonishing Hypothesis* Macmillan, New York.

FELDMAN, J. 1985. Four frames suffice: A provisional model of vision and space. *Behavioral Brain Sci. 8*, 265–289.

FERMÜLLER, C. AND ALOIMONOS, Y. 1995a. Representations for active vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

FERMÜLLER, C. AND ALOIMONOS, Y. 1995b. Qualitative egomotion. *Int. J. Comput. Vision 15*, 7–29.

MUMFORD, D. 1990. The 2.1D Sketch. In *Proceedings of the 2nd International Conference on Computer Vision*.

POGGIO, T. AND EDELMAN, S. 1990. A network that learns to recognize three-dimensional objects. *Nature 343*, 263–266.

POSNER, M. AND RAICHLE, M. 1994. *Images of Mind*. Scientific American Library.

ROSENFELD, A. 1987. Recognizing unexpected objects: A proposed approach. *Int. J. Pattern Recogn. Artif. Intell. 1*, 71–84.

ULLMAN, S. AND BASRI, R. 1991. Recognition by linear combinations of models. *IEEE Trans. PAMI 13*, 992–1005.

ZEKI, S. 1993. *A Vision of the Brain*. Blackwell Scientific, Cambridge, MA.