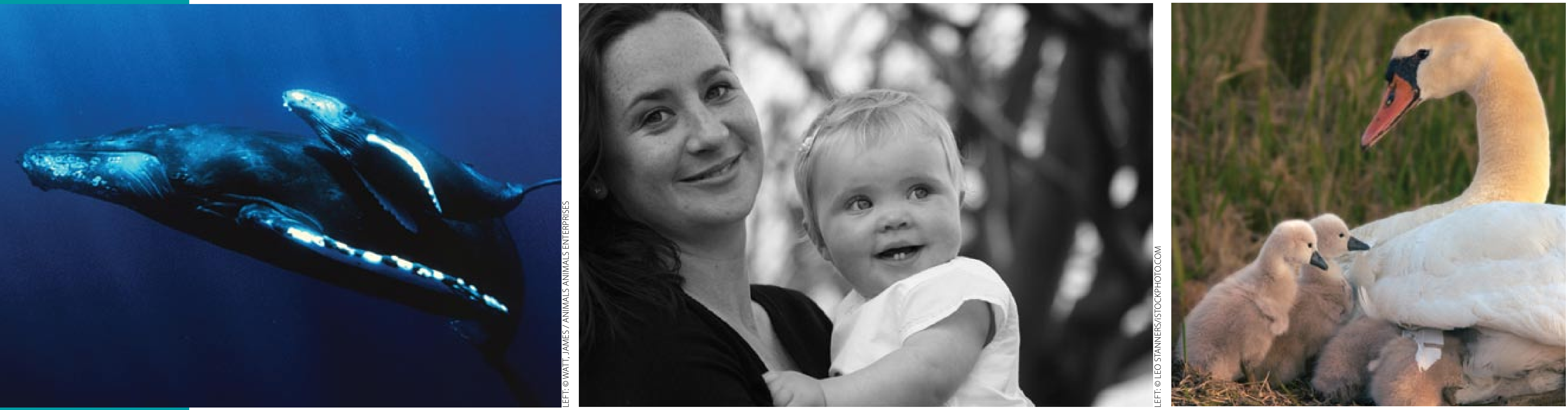
VISUAL UNDERSTANDING

By Melanie Mitchell



LEFT: © WATT, JAMES / ANIMALS ANIMALS ENTERPRISES
LEFT: © LEO STANNERS/iSTOCKPHOTO.COM
LEFT: © BIGSTOCKPHOTO.COM

The mind quickly translates an image into abstract meaning: These pictures of a humpback whale with her calf, a human mother and child, and a swan and her cygnets easily conjure the concept of "mother and child, or children."

Consider the pictures above. What concept do all three images represent?

Of course there are any number of different concepts these images represent, such as "animals," "mammals," "things with limbs," and "multiple objects," but most people would very quickly answer "mothers and babies," or something similar, perceiving that more abstract concept to be the intended meaning of this juxtaposition.

How does the mind so quickly translate an array of pixels of different light intensities and colors into an abstract *meaning?* Visual understanding of this kind is a great mystery for neuroscientists and psychologists studying how the brain accomplishes this translation, as well as for computer scientists who are trying to build programs that automatically determine semantics from pictures. Although science has uncovered many of the brain and perceptual mechanisms underlying low-level vision, very little is known about how the brain accomplishes higher-level, more abstract perception. Similarly, while modern computer vision systems have impressive performance in some specific domains, there are no systems able to recognize instances of visual categories or understand the contents of visual scenes with anywhere near the generality and robustness of human perception. Our current lack of understanding of how low-level visual inputs can be translated into high-level conceptual descriptions has been termed the "semantic gap."

Garrett Kenyon, a theoretical neuroscientist at Los Alamos National Laboratory (LANL), and I, a computer and cognitive scientist, recently organized a workshop on this topic at the Santa Fe Institute. The workshop, entitled "High-Level Perception and Low-Level Vision: Bridging the Semantic Gap," brought together a small group of prominent neuroscientists, psychologists, computer scientists, and theoretical biologists to discuss the semantic gap from an interdisciplinary perspective and to see if a set of common principles of visual understanding could be discovered from the collective knowledge of the participants.

Visual perception can be roughly categorized into different levels of information processing. In *low-level vision*, primitive visual features such as color, contrast, texture, edges, and contours are extracted from the collection of photons that impinge upon the retina. Subsets of the image are identified as separate "segments" and relationships among these segments (e.g., spatial adjacency or color similarity) are encoded.

High-level vision is the process of translating these lower-level perceptions into the recognition of objects and their conceptual relationships, yielding a coherent description of the image as a whole. This information processing is not wholly one way: it is generally believed that the different levels of processing continually communicate with one another in both a feed-forward and feedback manner. Not only does information from lower level processing influence higher levels, but higher-level processing, involving stored knowledge, can guide lower-level perception in tasks such as segmentation. The ability to fluidly integrate the different levels is a major source of visual understanding in humans and other animals.

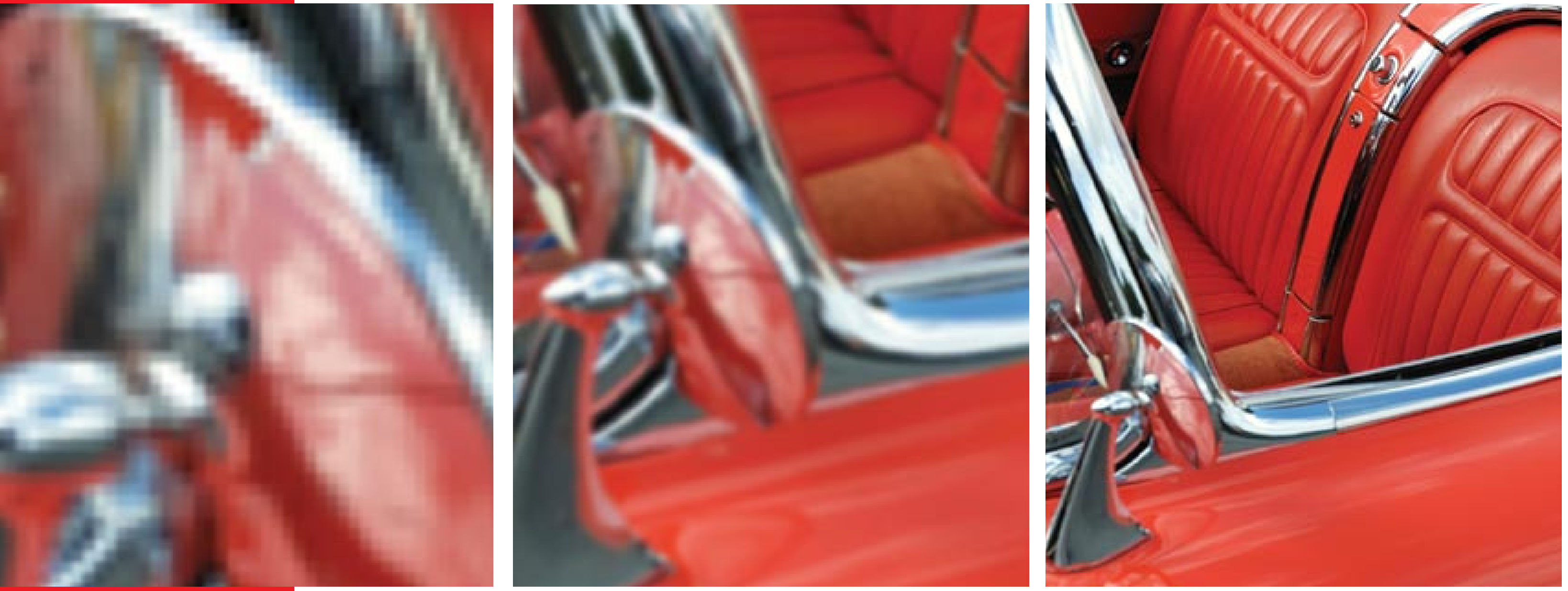
The talks and discussions at the workshop concerned scientific and technological enigmas at different levels of visual processing. The first set of speakers—mainly computational neuroscientists—described recent research on how the brain performs low-level vision. Bartlett Mel of the University of Southern California and Ilya Nemenman from LANL each spoke about the lowest level of the visual system: individual neurons. Their common message was that neurons are surprisingly complex, both with respect to the information encoding and processing they (and their subunits, such as dendrites) can perform and with respect to how information is communicated between individual neurons. In light of data recently obtained from detailed probes into individual cells, novel brain imaging techniques, and realistic computer models, classical theories in neuroscience are being radically redrawn. Neuroscientists are now rethinking paradigms that were once generally accepted; results from recent experimental studies are undermining notions such as the static "receptive fields" of neurons, the transmission of information among neurons via the simple counting of spikes, and the role of dendrites as "passive" conductors of electrical signals. As Tom Stoppard wrote in his play *Arcadia*, "It's the best possible time to be alive, when almost everything you thought you knew is wrong."

Collective information processing by groups of neurons was a key topic for other participating brain scientists, including Garrett Kenyon, Pam Reinagel (University of California, San Diego), John George (LANL), Fritz Sommer (Redwood Center for Theoretical Neuroscience), and David Field (Cornell). The question of how such neuronal groups represent information is currently being hotly debated in neuroscience. Particularly significant controversies that were discussed concern (1) the existence and role of correlated spiking in neural populations; (2) the nature and advantages of distributed and sparse representations, in which sensory information is encoded with only a small number of active neurons at any given point in time; (3) methods for measuring statistical and information-theoretic properties of natural visual inputs and inferring their implications for how the brain efficiently encodes visual information; and (4) the role of feedback connections from higher brain levels.

FOR THE THEORISTS IN THE GROUP, the experimentalists' talks drove home the difficulty of obtaining and interpreting data about the brain, which is perhaps the most complex of all natural systems. The theorists, on the other hand, underscored the increasing importance of computer modeling in neuroscience.

With respect to understanding sensory information processing or any other brain process, computational neuroscientist Tony Bell (Redwood Center for Theoretical Neuroscience) asked, "What level of description should we be focusing on?" He challenged the group with his assertion that "the fact that information flows all the way up and down the reductionist hierarchy…has significant implications for the way neuroscience (and presumably other areas of biological information processing) should be done." The group discussed some examples of information flow to and from levels as low as genetic regulation, and several in the group strongly disagreed that it is important to include such levels when studying information processing in the brain.

BELOW: © ISTOCKPHOTO.COM



Visual perception can be categorized into different levels of information processing: Low-level vision takes in primitive visual features such as color and texture, while high-level vision translates these perceptions into recognition of objects and their conceptual relationships, yielding a coherent description of the whole.

RIGHT: © BIGSTOCKPHOTO.COM

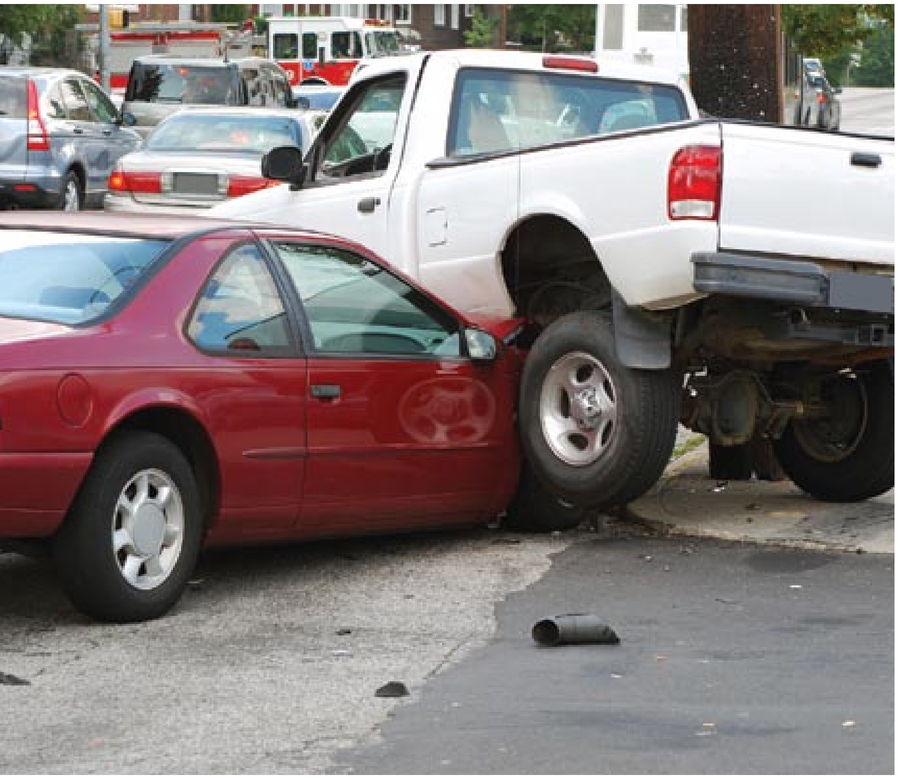
RIGHT: © STEVE GEER/ISTOCKPHOTO.COM

Higher-level visual behavior was a second focus of the workshop. How do humans and other animals use vision in real life, and what can be understood from studying human visual behavior from a psychophysical and psychological perspective? Psychophysicist Simon Thorpe, from the Centre de Recherche Cerveau et Cognition in France, described a surprising set of experimental results showing that humans are able to do some complex visual recognition tasks much faster than had been previously thought. For example, Thorpe's experiments have shown that most people can identify whether or not a picture contains an animal in less than 200 milliseconds. These results call into question the role of feedback connections in visual recognition, since 200 milliseconds is not enough time for signals to propagate to higher brain levels and for those higher levels to send feedback. However, it seems that the brain clearly must use feedback connections for *something* in vision—there are about 10 times as many feedback as feed-forward connections in the visual cortex. Much discussion (some rather heated) concerned the role of feedback information and what kinds of experiments could tease out its function.

Visual *attention* is one area where feedback from higher levels seems essential. Psychologist Todd Horowitz, from Harvard Medical School, addressed the issues of what constitutes visual attention and whether it is needed to bridge the semantic gap. In particular, when viewing a scene, what do we pay attention to and how does that affect our ability to remember the scene later on? Horowitz presented experimental results that indicated that people require focused attention to encode the "gist" of visual scenes in terms of objects present and spatial layout, and in particular, that focusing on the layout is more important than focusing on the objects for remembering a given scene.

Neuroscientist and psychophysicist Peter König (University of Zurich), zeroed in on a particular form of visual attention: the control of eye movements in response to the information content of input stimuli and "top-down" feedback. Computer scientist Dana Ballard (University of Texas at Austin), presented a computer model for "multi-tasking" in visual attention, which his group tested by having it navigate in a virtual reality simulation.

Experiments have shown that people can identify whether or not a picture contains an animal in less than 200 milliseconds.



© BIGSTOCKPHOTO.COM

People require focused attention to encode the "gist" of visual scenes. Interestingly, the scene's layout may be more important to memory than the objects themselves.

Ballard, along with several other participants, argued for the importance of studying vision in the context of the rest of the body. He stressed that the visual system and motor system are tightly interconnected, and constrain one another so as to make the computation of behavior tractable. Psychologist Shimon Edelman, of Cornell University, emphasized that vision scientists, in modeling abilities such as object recognition, should not lose the phenomenological aspect of vision—that is, visual phenomena that do not involve categorization and recognition, including our first-person experience of what it feels like to "see." Psychologist Rob Goldstone, from Indiana University, described a different notion of embodiment: the effects of visual perception on reasoning and abstract problem solving. Goldstone reported on experimental results that indicated that even the most abstract of tasks—mathematical reasoning—is affected by visual input, such as the layout of the problem on the page. His point was that, just as vision scientists need to take the rest of the body into account, psychologists studying abstract reasoning cannot ignore the effects of visual perception that "leak" into such reasoning.

New approaches to classical problems in computer vision and pattern recognition were proposed by computer scientists Lakshman Prasad and James Theiler, both from LANL, and myself. Ideas from "left field" were presented by theoretical immunologist Tom Kepler, of Duke University, who explored possible analogies between the visual system and the immune system, which itself must perform sophisticated, unpredictable, and ongoing tasks of pattern recognition and response. I described how my own work on high-level visual pattern recognition and analogy has been inspired by the immune system and other complex systems with pattern-recognizing abilities.

In the last decade or so, research on both natural and computer vision has become rather narrow and specialized; work on the fundamental problem of how different levels of vision are integrated has been largely neglected. In my view, the main benefit of the workshop was the opportunity for scientists studying widely diverse aspects of vision to communicate to one another the recent major advances and controversies in their areas. On the one hand, learning more about the daunting complexity of the visual system in processing sensory data, juxtaposed with the extraordinary abilities of the mind to glean abstract meaning, only made the gap seem wider. On the other hand, the talks and discussions at the workshop gave all the participants a new appreciation for the scope of the problem and its requirement for interdisciplinary collaboration, a sense of where the most promising directions are, and (at least for me) plans for new collaborations and many fresh ideas to take home and ponder. ◀

Melanie Mitchell is Professor of Computer Science at Portland State University, and External Professor at the Santa Fe Institute.