PERSPECTIVES: COGNITION

The Manifold Ways of Perception

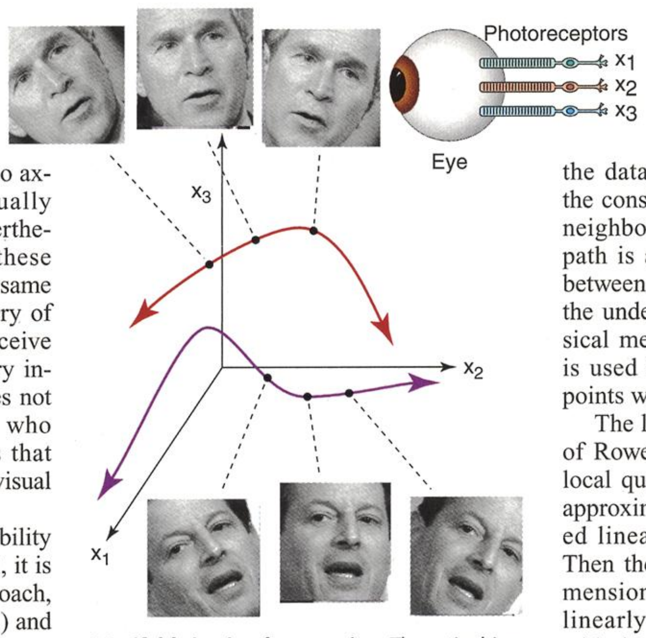
H. Sebastian Seung and Daniel D. Lee

Two-and-a-half millennia ago, the Greek philosopher Heraclitus, observing that the world is in eternal flux, wrote that you can never step in the same river twice. If he were alive today and working as a psychologist, he might say that you can never see the same face twice. Indeed, faces can grow hair, acquire wrinkles, or be surgically enhanced. But facial images also vary from moment to moment, as you can demonstrate at home while watching television. Make a small aperture in a piece of paper, and place it over a face on the screen. The light coming through the aperture will vary with time, mostly as a result of changes in the location and orientation of the face.

The aperture might show a tooth at one instant, and a nostril at the next, crudely simulating the fluctuations in light incident on a single retinal photoreceptor cell. This illustrates that the signals carried from the eye to the brain by the million or so axons in the optic nerve are perpetually changing as we look at a face. Nevertheless, we are able to perceive that these changing signals are produced by the same object. This is the fundamental mystery of perception: How does the brain perceive constancy even though its raw sensory inputs are in flux? The mystery intrigues not only scientists but also engineers, who yearn to construct vision machines that equal the performance of humans at visual object recognition.

To precisely characterize the variability of images and other perceptual stimuli, it is essential to take a mathematical approach, which is just what Tenenbaum *et al.* (*1*) and Roweis and Saul (*2*) have done on pages 2319 and 2323 of this issue, respectively. An image can be regarded as a collection of numbers, each specifying light intensity at an image pixel. But a collection of numbers also specifies the Cartesian coordinates of a point with respect to a set of axes. Therefore, any image can be identified with a point in an abstract *image space*.

Now consider a simple example of image variability, the set M of all facial images generated by varying the orientation of a face (see the figure). This set is a continuous curve in the image space. It is continuous because the image varies smoothly as the face is rotated. It is a curve because it is generated by varying a single degree of freedom, the angle of rotation. In other words, M is intrinsically one-dimensional, although it is embedded in image space, which has a high dimensionality equal to the number of image pixels. If we were to allow other types of image transformations, such as scaling and translation, then the dimensionality of M would increase, but would still remain far less than that of the image space. In this generalized case, M is said to be a *manifold* embedded in the image space. A curve is an example of a one-dimensional manifold, whereas a sphere is an example of a two-dimensional manifold (*3*).



Manifolds in visual perception. The retinal image is a collection of signals from photoreceptor cells. If these numbers are taken to be coordinates in an abstract image space, then an image is represented by a point. Only three dimensions of the image space are depicted, but actually the dimensionality is equal to the number of photoreceptor cells. As the faces are rotated, they trace out nonlinear curves embedded in image space. If changes in scale, illumination, and other sources of continuous variability are also included, then the images would lie on low-dimensional manifolds, rather than the simple one-dimensional curves shown. To recognize faces, the brain must equate all images from the same manifold, but distinguish between images from different manifolds. How the brain represents image manifolds is as yet unknown. According to one hypothesis, they are stored in the brain as manifolds of stable neural-activity patterns.

CREDITS: (BUSH) HARRY CABLUCK/AP PHOTO; (GORE) HILLERY SMITH GARRISON/AP PHOTO

Although the preceding discussion is biased toward vision, manifolds are also relevant to other types of perception. Furthermore, scientists in many fields face the problem of simplifying high-dimensional data by finding low-dimensional structure in it. Therefore, the manifold learning algorithms described by Tenenbaum *et al.* (*1*) and Roweis and Saul (*2*) are of potentially broad interest. The goal of the algorithms is to map a given set of high-dimensional data points into a surrogate low-dimensional space. Both start with a preprocessing step that decides for each data point which of the other data points should be considered its neighbors. Then both compute measures of the local geometry of the manifold, after which the original data points are no longer needed.

In the Isomap algorithm of Tenenbaum *et al.*, the local quantities computed are the distances between neighboring data points. For each pair of nonneighboring data points, Isomap finds the shortest path through the data set connecting them, subject to the constraint that the path must hop from neighbor to neighbor. The length of this path is an approximation to the distance between its end points, as measured within the underlying manifold. Finally, the classical method of multidimensional scaling is used to find a set of low-dimensional points with similar pairwise distances.

The locally linear embedding algorithm of Roweis and Saul computes a different local quantity, the coefficients of the best approximation to a data point by a weighted linear combination of its neighbors. Then the algorithm finds a set of low-dimensional points, each of which can be linearly approximated by its neighbors with the same coefficients that were determined from the high-dimensional data points. Both algorithms yield impressive results on some benchmark artificial data sets, as well as on "real world" data sets. Importantly, they succeed in learning nonlinear manifolds, in contrast to algorithms such as principal component analysis, which can only learn linear manifolds.

Because manifolds are fundamental to perception, the brain must have some way of representing them. Clues to the nature of this representation may come from studies of how information is encoded in large populations of neurons. Population activity is typically described by a collection of neural firing rates, and so can be represented by a point in an abstract space with dimensionality equal to the number

H. S. Seung is at the Howard Hughes Medical Institute and Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. D. D. Lee is at Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, USA.

of neurons. Neurophysiologists have often found that the firing rate of each neuron in a population can be written as a smooth function of a small number of variables, such as the angular position of the eye (*4*) or direction of the head (*5*). This implies that the population activity is constrained to lie on a low-dimensional manifold.

What is the connection between such neural manifolds and the image manifolds we have just discussed? According to a well-known idea, memories are stored in brain dynamics as stable states, or dynamical attractors (*6*). Because the possible images of an object lie on a manifold, it has been hypothesized that a visual memory is stored as a manifold of stable states, or a continuous attractor (*7*). Recent studies of neural manifolds suggest that continuous attractors actually do exist in the brain (*8*, *9*). Whether they are the basis of visual and other types of perception remains to be resolved. If the answer is affirmative, then manifolds will prove to be crucial for understanding how perception arises from the dynamics of neural networks in the brain.

References

1. J. Tenenbaum, V. de Silva, J. C. Langford, *Science* **290**, 2319 (2000).
2. S. Roweis, L. Saul, *Science* **290**, 2323 (2000).
3. K. Devlin. *Mathematics: The Science of Patterns* (Scientific American Library, New York, 1997).
4. J. L. McFarland, A. F. Fuchs, *J. Neurophysiol.* **68**, 319 (1992).
5. J. S. Taube, *Prog. Neurobiol.* **55**, 225 (1998).
6. J. J. Hopfield, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554 (1982).
7. H. S. Seung, *Adv. Neural Info. Proc. Syst.* **10**, 654, (1998).
8. H. S. Seung, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13339, (1996).
9. K. Zhang, *J. Neurosci.* **16**, 2112 (1996).

PERSPECTIVES: MICROELECTRONICS

Flip the Chip

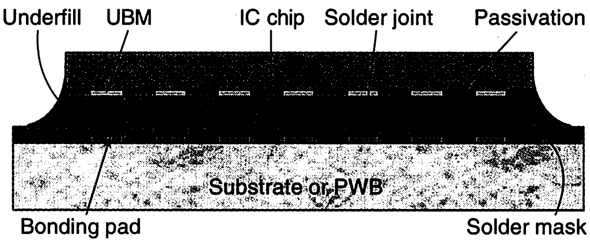
C. P. Wong, Shijian Luo, Zhuqing Zhang

Today's world is filled with fancy electronic products such as laptops, cellular phones, and digital cameras. In all of these systems, electronic packaging plays an important role by supplying power to chips, distributing signals between chips and among devices through interconnects, providing heat dissipation, and protecting components from environmental impact (*1*, *2*). As integrated circuit (IC) fabrication advances rapidly and the market for ever faster, lighter, smaller, yet less expensive electronic products accelerates, electronic packaging faces its own challenges. This is where flip chip packaging comes into play.

Conventional electronic packaging uses wire-bonding technology, in which the active side of the silicon chip faces up and interconnection is created by drawing gold, silver, or copper wires to the substrate. In contrast, in flip chip packaging, the active side of the silicon chip faces down and is directly connected to the substrate or printed wire board (PWB). This technology has many advantages over the conventional approach. It has a much higher input/output (I/O) count because the whole area underneath the chip can be used for interconnection. The shorter signal path between chip and board reduces inductance, thereby increasing signal propagation speed and greatly enhancing electrical performance. Because the chip faces down to the substrate, its backside can be used for heat dissipation. Finally, the entire interconnection on the chip can be made simultaneously in a single step, whereas in wire bonding only one wire is drawn at a time. Flip chip thus offers the possibility of low-cost electronic assembly for modern electronic products.

The authors are at the School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. E-mail: cp.wong@mse.gatech.edu

Since the flip chip was first developed 40 years ago at Bell Labs, many variations of the design have been demonstrated. The most important form of flip chip is the solder bump interconnection or Controlled Collapse Chip Connection (C4) (*3*). In this method, solder bumps deposited on wettable metal terminals on the chip connect with matching bonding pads on the substrate (see the figure). The metal terminal (called under bump metal or UBM) consists of several layers that provide good adhesion to the chip and the solder bump and prevent oxidation of the metal terminal (*4*). Solder bumps can be fabricated through evaporation, electroplating, electroless plating, or screen printing. New technologies include solder jet printing and microball mounting.



Generic configuration of flip chip interconnection with underfill.

After the chip is placed on the substrate, the assembly is subjected to "solder reflow," a heating cycle that melts the solder bump. The surface tension of the molten solder prevents the chip from collapsing onto the substrate. One advantage of this process is that the chip can self-align as a result of the high surface tension of the molten solder. As long as enough solder touches the bonding pad on the substrate, perfect alignment can be achieved.

In the first generation of C4, ceramic substrates with a low coefficient of thermal expansion (CTE) matching that of the silicon were used. However, ceramic substrates are expensive and require high-temperature processing. Furthermore, their high dielectric constant aggravates the signal delay. Organic substrates are favorable because of their low dielectric constant and low cost, but high CTE differences between organic substrates and the silicon chip exert great thermal stress on the solder joints during temperature cycling. The larger the chip, the higher the stress and, hence, the shorter the solder joint fatigue life. This is why organic substrates could not be used in flip chips until underfill was invented in the late 1980s.

Underfill is a liquid encapsulate that is applied between the chip and the substrate. After conversion into a solid material during curing, it exhibits high modulus, high glass transition temperature, low moisture absorption, good adhesion toward chip and substrate, and low CTE matching that of the solder joint. Thermal stresses on the solder joints are redistributed among chip, solders, underfill, and substrate, thereby increasing the solder joint fatigue life 10 to 100 times (*5*).

With increasing application of flip chips on organic substrates, underfill technology becomes the key to achieving highly reliable packaging. However, current underfill materials have several drawbacks that limit their wide application. If an IC is assembled using flip chip without underfill, a failed chip can easily be removed from the substrate by melting the solder joint and replacing the failed chip by a new one. If underfill is applied, this rework process becomes very difficult because most currently used underfill materials are epoxy-based thermosets that go through an irreversible cross-linking (curing). Reworkable underfill materials may solve this problem. An acetal/ketal group has been incorporated into the epoxy resin so that the cured resin can be dissolved in acid and thus reworked (*6*). Epoxy resins containing thermally cleavable linkages