Information Processing in the Primate Visual System: An Integrated Systems Perspective

DAVID C. VAN ESSEN, CHARLES H. ANDERSON, DANIEL J. FELLEMAN

The primate visual system contains dozens of distinct areas in the cerebral cortex and several major subcortical structures. These subdivisions are extensively interconnected in a distributed hierarchical network that contains several intertwined processing streams. A number of strategies are used for efficient information processing within this hierarchy. These include linear and nonlinear filtering, passage through information bottlenecks, and coordinated use of multiple types of information. In addition, dynamic regulation of information flow within and between visual areas may provide the computational flexibility needed for the visual system to perform a broad spectrum of tasks accurately and at high resolution.

V ISUAL IMAGES PROVIDE AN IMMENSELY RICH SOURCE OF information about the external world. We use this information so effortlessly and efficiently that it is easy to underestimate the computational complexity of ordinary visual routines. For instance, driving a car on a busy street requires coordination of a number of ongoing visual tasks: reading traffic signs, recognizing familiar landmarks, localizing and tracking vehicles, and being alert to pedestrians and other sources of sudden danger. The results of these ongoing analyses must be appropriately routed to brain regions involved in perception, motor control, and planning. Our ability to perform such tasks accurately, rapidly, and reliably in the natural environment requires an extremely sophisticated and well-engineered visual system.

The challenge of understanding vision has prompted widespread interest in interdisciplinary approaches that attack the problem from complementary viewpoints. In this article we discuss a combined neurobiological and systems engineering approach to studying the primate visual system. The neurobiological approach provides a detailed anatomical and physiological description of the visual system and suggests a number of key principles, including modular design, hierarchical organization, and the presence of distinct but intertwined processing streams. The systems engineering perspective provides a framework for analyzing and interpreting these and other aspects of visual system organization. It emphasizes the need for computationally sound models that are grounded in basic principles of signal processing and respect both the power and the limitations of the underlying neural circuitry.

D. C. Van Essen is a professor in the Biology Division, California Institute of Technology, Pasadena, CA 91125. C. H. Anderson is a senior member of the technical staff, Jet Propulsion Laboratory, Pasadena, CA 91109. D. J. Felleman is an assistant professor in the Department of Neurobiology and Anatomy, University of Texas Medical School, Houston, TX 77030.

Anatomical Overview

The anatomy of the primate visual system has been intensively studied in the macaque monkey, whose visual system is similar in many ways to that of humans. Figure 1 provides an overview of the major components of the macaque visual system, as seen in the intact right hemisphere (upper left) and in unfolded two-dimensional maps of the cerebral cortex (center) and of subcortical visual structures (lower left). All structures are drawn to scale, so their sizes reflect the amount of neural machinery available for processing in different centers.

Areas. The major retinal output goes via the lateral geniculate nucleus (LGN) to the striate cortex (area V1) and from there to a mosaic of extrastriate cortical visual areas shown in various colors in Fig. 1. Thirty-two distinct cortical areas associated with visual processing have been described on the basis of anatomical, physiological, and behavioral information (1, 2). Twenty-five areas are primarily visual in function; the remaining seven are also implicated in other functions such as polysensory integration or visually guided motor control. The identification of some areas is questionable or controversial, though, and alternative partitioning schemes have been suggested, especially for the temporal lobe (3). Altogether, visual cortex occupies about half of the 100 cm^2 extent of each hemisphere. V1 and V2 each exceed 10 cm^2, but most areas are less than one-tenth this size.

Hierarchies. To date, 305 pathways interconnecting the 32 cortical visual areas have been identified with modern pathway-tracing techniques (1). This constitutes nearly one-third of the number there would be if the network were fully interconnected. Hierarchical relations between areas have been assessed by the use of information about the cortical layers in which pathways originate and terminate (1, 4). For some pathways the laminar pattern suggests ascending (forward) information flow from a lower to a higher area. These are generally paired with reciprocal pathways that have patterns suggesting feedback from a higher to a lower area. Other pathways have patterns suggesting lateral connections between areas at the same level. Systematic application of these criteria leads to a hierarchy containing ten levels of cortical visual processing plus several additional stages of subcortical processing (Fig. 2). The visual hierarchy is extensively linked to centers associated with motor control, other sensory modalities, and cognitive processing (1, 5), only a few of which are shown here. Different pathways vary greatly in strength, and there are alternative schemes for cortical connectivity that emphasize mainly the robust connections (6, 7).

Processing streams. Two major processing streams originate within the retina. About 80% of retinal ganglion cells are parvocellular (P) cells projecting to the P layers of the LGN, whereas 10% are magnocellular (M) ganglion cells projecting to the M layers of the LGN (8). In V1 and V2, these are reorganized into a tripartite arrangement (7), the so-called P-B, P-I, and M streams (6). In V1

the compartments are patchy and specific to particular cortical layers, whereas in V2 they form stripes parallel to the surface and extending through all layers (7, 9). Additional areas associated with the M stream include V3, MT, MST, and some posterior parietal areas. The P-B and P-I streams include separate subregions of V4, and this segregation may persist into inferotemporal (IT) areas (5, 10). Despite the specificity of connections within each stream, substantial cross talk occurs at many levels, as demonstrated anatomically (1, 6) and physiologically (11).

Information Processing Strategies

Within this anatomical framework, it is important to understand how visual information is represented by individual neurons, how it is transformed and discarded at successive levels, how it is distributed among different processing streams, and how it contributes to the performance of specific visual tasks. We will discuss recent progress on five topics related to these issues.

Information bottlenecks and scale invariance. Physical images on the retina contain far more information than can be efficiently handled by the brain. An initial stage of data reduction occurs in the retina, where the nonuniform distribution of $\sim 10^6$ ganglion cells leads to a variable resolution representation that is approximately scale-invariant (12, 13). To illustrate the significance of scale invariance, imagine looking at a colleague's face across a table and fixating on the tip of her nose. Resolution is very high at the fixation point ($\sim 10^4$ pixels per degree squared) but declines sharply at progressively more peripheral locations. If the friend moves closer (while

one still fixates the nose), the image as a whole becomes larger, and components of the image that are off-center (say, the ears) shift peripherally on the retina as they enlarge. Outside the central 2°, the loss in spatial resolution resulting from the peripheral shift almost exactly counteracts the increase in image size. The net result is that the amount of information transmitted about the colleague's face is nearly independent of viewing distance, except in the central 2°, where there is a gain or loss of information as the face approaches or recedes.

Filters, not feature detectors. Neurons at each stage of processing are best described as filters that are selective along multiple stimulus dimensions. Retinal ganglion cells and lateral geniculate cells have concentric, antagonistically organized receptive fields and broad spatial frequency tuning, allowing them to carry some low spatial frequency information while nonetheless emphasizing higher frequencies (14). Transmitting the difference between a central value and the local mean luminance is an effective data reduction step; for natural images it increases the statistical independence of activity in neighboring cells (that is, decreases their correlation).

The primary role of P cells is to encode information about luminance contrast. They respond to low and moderate temporal frequencies (mainly 1 to 20 Hz) over the full range of spatial frequencies visible at a given eccentricity (14, 15). M cells are optimized for higher temporal frequencies (mainly 5 to 40 Hz), but they provide only about one-third the spatial resolution because of their lower sampling density. Thus, the P and M systems cover partially overlapping regions of the spatiotemporal information space illustrated at the bottom of Fig. 3 (12, 14–16).

A secondary function of P cells, the transmission of spectral information, is achieved by having photoreceptors with different spectral sensitivities subserve the center and surround mechanisms. However, despite the vividness of color perception, color makes only a minor contribution to the information content of natural images. Visual acuity is many times worse for patterns differing only in spectral composition than for patterns differing in luminance (17). Like the P system, the M system also conveys a second type of information: many M cells have nonlinearities that encode the presence of fine-grained details, without representing precise spatial relations (that is, phase information) (14, 15, 18).

Filtering in both spatial and temporal domains continues in area V1. A unifying descriptive framework (on the basis of cat and monkey studies) is that cortical receptive fields are oriented in both spatial (x-y) and space-time (x-t) coordinates. Spatial orientation gives rise to conventional orientation selectivity for elongated stimuli and to sharpened spatial frequency tuning, whereas spatiotemporal orientation gives rise to velocity tuning (19).

There is a basic distinction in V1 between simple and complex cells. Simple cells act as quasilinear filters (20) and are analogous to the filters in wavelet representations, whose sensitivity profiles are restricted in both space and spatial frequency. Wavelet representations have become popular in image processing because they provide a statistically independent and complete visual repre-
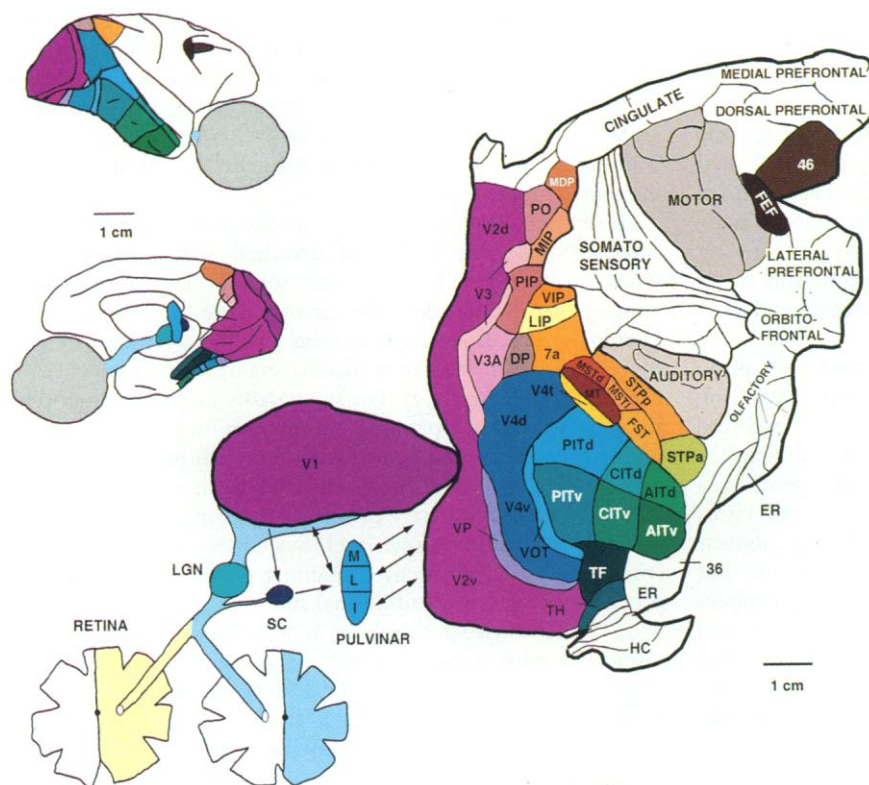
Fig. 1. An overview of the macaque visual system, as seen from lateral and medial views of the right hemisphere and from unfolded representations of the entire cerebral cortex and major subcortical visual structures. The cortical map contains several artificial discontinuities (for example, between V1 and V2). Minor retinal outputs (~10% of ganglion cells) go to the superior colliculus (SC), which projects to the pulvinar complex, a cluster of nuclei having reciprocal connections with many cortical visual areas (37). All structures (except the much thinner retina) are ~1 to 3 mm thick. [Modified, with permission, from (1), with subcortical structures based on (12) and (38).]

sentation (21). Complex cells have pronounced spatial nonlinearities and provide a measure of image power within a restricted range of orientations, spatial and temporal frequencies, and space (22). Like the subpopulation of nonlinear M cells in the retina, complex cells convey information about the statistics of images (for example, texture and "motion energy"), while discarding information about spatial details. This may help, for example, in distinguishing the roughness of a rocky terrain without the need for pebble-by-pebble scrutiny; such strategies have important applications in image processing (23). This perspective differs from the notion promoted by "feature detector" advocates (24) that early cortical processing is mainly devoted to the explicit detection of features such as edges or zero-crossings of intensity derivatives. These are essentially logical, all-or-none decisions that are better postponed to later stages of analysis (25).

Information divergence and convergence. Besides the aforementioned spatial and temporal characteristics, many neurons in V1 and in extrastriate areas are selective along other dimensions, including wavelength and binocular disparity. Figure 3 addresses how these diverse types of selectivity, discernible among cortical neurons (middle row), are related to the P and M streams (bottom row) and

to different visual tasks (top row). Two basic tasks of vision are object recognition (what it is) and the determination of spatial relationships among objects in a dynamic three-dimensional world (where it is, and where it is going). Corresponding to this functional distinction is an anatomical dichotomy: object recognition is selectively impaired by lesions of IT areas, whereas judgments of spatial relations are impaired by lesions of area 7a and other posterior parietal areas (26).

For accurate and robust performance, each of these tasks needs access to many types of information. For example, identification of a particular object (say, a sphere or a cylinder) can be based on binocular disparity cues (when viewing a random-dot stereogram), on velocity cues (in a structure-from-motion task), or on perspective cues that use orientation and spatial frequency information. However, ordinary visual tasks typically involve the coordinated use of several types of information. In this respect, it is reasonable that each processing stream is associated with two or more types of selectivity, rather than just one type (Fig. 3, middle row) (6, 7). Each stream has a distinctive constellation of selectivities, and at higher levels each stream contains cells that also have more complex receptive field properties, suggestive of intermediate stages of cortical pro-

cessing (2, 27). The specific characteristics of the M stream make it well suited for various tasks requiring motion information, but motion is not the only type of information it represents. The P-B stream is suited for tasks needing spectral information and the P-I stream for tasks needing detailed spatial information, but other types of information are multiplexed into these streams as well.

To take advantage of the full spatiotemporal frequency spectrum transmitted by the optic nerve, higher levels of the visual system need access to information from both the P and the M systems during the performance of many tasks. The anatomical cross talk described above provides a neural substrate for this convergence, and results from selective lesion studies indeed indicate that either the P or the M system alone can mediate a wide range of tasks (for example, depth perception, speed, and direction discrimination) as long as stimuli are presented in the appropriate spatiotemporal range (16, 28).

Dynamic aspects of information flow. The recognition of complex patterns is too computationally demanding to be carried out at maximal resolution across the entire visual field. However, the problem is manageable because pattern recognition occurs mainly within a restricted window of attention that can be rapidly shifted in position and changed in spatial scale. Several lines of psychophysical evidence suggest that the window of attention has an effective spatial resolution about 20 to 30 sampling elements across, independent of shifts in position and scale (29). Because this is only ~1% of the capacity of the optic nerve, it represents a major additional bottleneck in information flow.

Several models for the window of attention have been proposed, which differ in the strategies used to regulate the information
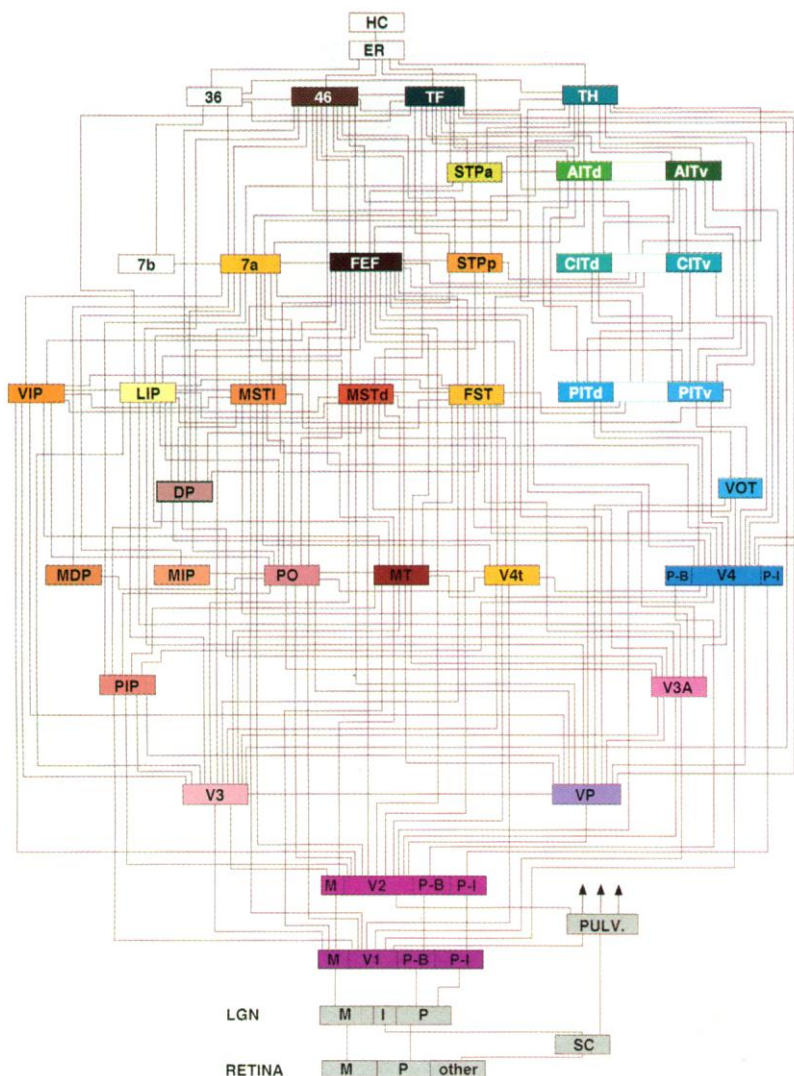

Fig. 2. A hierarchy of visual areas in the macaque, based on laminar patterns of anatomical connections. About 90% of the known pathways are consistent with this hierarchical scheme; the exceptions may reflect either inaccuracies in the reported anatomical data or genuine deviations from a rigid hierarchical scheme. [Modified, with permission, from (1), with subcortical connections based on (37)]
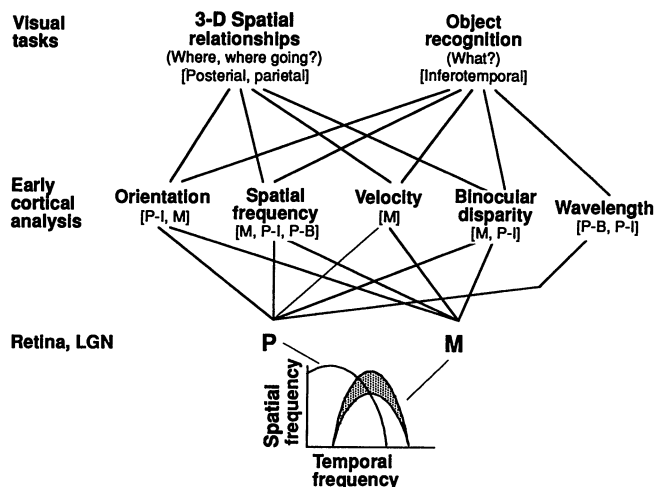
Fig. 3. Convergence and divergence in visual processing. Arrows represent major lines of information flow from subcortical P and M streams (bottom) to the selectivities represented among neurons at early stages of cortical analysis (middle) and from there to two general tasks of vision (top level). The hatched portion of the M cell curve represents their nonlinear component of processing. The processing streams associated with each property in the middle row are assigned on the basis of a high incidence of selectivity recorded physiologically (6, 7).

reaching pattern recognition centers (12, 30). The model we prefer (12, 31) is based on selective gating of neural inputs that is coordinated across a series of processing stages, including areas V1, V2, and V4, as well as IT. This model accounts for the preservation of information about detailed spatial relations within the window of attention, which we consider crucial to any comprehensive model of attention. It makes specific predictions about dynamic effects of attention on receptive field properties, which are supported by physiological recordings from V4 and IT of alert monkeys (32). The control mechanism for initiating and directing attentional shifts may involve the pulvinar as well as the posterior parietal cortex (33), but its neurobiological implementation remains poorly understood.

Modularity and computational flexibility. Complex problems are often best solved by breaking them into smaller components. In considering how this principle might be reflected in the design of the brain, numerous engineering issues and trade-offs arise, just as in computer design (34). (i) Creating separate modules for different subtasks allows neural architecture to be optimized for particular types of computation. It also allows different types of information to be represented explicitly, in ways that can simplify later stages of analysis. However, it is important to avoid overly specialized modules that lack flexibility, are rarely used, or are inordinately complex to construct. (ii) Replicating the same module many times over, as exemplified by the internal structure of V1 (35), allows stereotyped computations to be carried out massively in parallel. However, high-level tasks such as pattern recognition, which are computationally expensive and require large amounts of stored data, should not be duplicated unnecessarily. (iii) In principle, having both types of modularity offers great computational flexibility. However, to capitalize, it is critical to maintain coordinated and efficient routing of information between modules.

In a highly modular system, many important subtasks may involve functions (for example, controlling information flow) that are peculiar to the specific architecture of the system. These may appear obscure or arcane when analyzed only in terms of the inputs and outputs of the whole system. This suggests a need for caution in interpreting cortical function as a one-to-one mapping between individual visual areas or streams and the perceptions of color, form, or motion (7, 36). Instead, we emphasize a task-oriented perspec-

tive, keyed on understanding how overall tasks are broken into subtasks and on elucidating the computational strategies needed for their implementation (6).

The computations carried out within any given cortical area are anatomically constrained by its inputs and its intrinsic synaptic circuitry. There may be a fundamental distinction between ascending pathways in the hierarchy, whose circuitry dictates the basic classes of analysis carried out within an area, and the modulatory influences exerted by feedback pathways from higher centers and perhaps by the pulvinar. These modulatory influences may allow for several types of computational flexibility: adjusting the exact nature of the transformation carried out within an area; regulating the inputs on which the computation is made (input gating); or switching the targets to which the results are transmitted (output gating). We suspect that all three of these dynamic control processes are important throughout the cerebral cortex. Collectively, they may allow the brain to reorganize its computational structure adaptively, on a rapid (~100 ms) time scale, for optimal utilization of the incoming data and of the available neural resources. Just as the brain controls other bodily functions, it may exert explicit control over its own computations. Translating this general hypothesis into specific, neurobiologically plausible models and into critical experimental tests will be a challenge for the future.

REFERENCES AND NOTES

1. D. J. Felleman and D. C. Van Essen, *Cereb. Cortex* **1**, 1 (1991).
2. D. C. Van Essen, in *Cerebral Cortex*, A. Peters and E. G. Jones, Eds. (Plenum, New York, 1985), vol. 3, pp. 259–329; J. H. R. Maunsell and W. T. Newsome, *Annu. Rev. Neurosci.* **10**, 363 (1987); R. Desimone and L. B. Ungerleider, *Handb. Neuropsychol.* **2**, 267 (1989).
3. B. Seltzer and D. N. Pandya, *Brain Res.* **149**, 1 (1978); D. Boussaud, C. F. Ungerleider, R. Desimone, *J. Comp. Neurol.* **296**, 462 (1991).
4. K. S. Rockland and D. N. Pandya, *Brain Res.* **179**, 3 (1979); J. H. R. Maunsell and D. C. Van Essen, *J. Neurosci.* **3**, 2563 (1983).
5. D. C. Van Essen, D. F. Felleman, E. A. DeYoe, J. F. Olavarria, J. J. Knierim, *Cold Spring Harbor Symp. Quant. Biol.* **55**, 679 (1990).
6. E. A. DeYoe and D. C. Van Essen, *Trends Neurosci.* **11**, 219 (1988).
7. M. Livingstone and D. Hubel, *Science* **240**, 740 (1988); S. M. Zeki and S. Shipp, *Nature* **335**, 311 (1988).
8. V. H. Perry, R. Oehler, A. Cowey, *Neuroscience* **12**, 1101 (1984).
9. R. B. Tootell, M. S. Silverman, E. Switkes, R. L. DeValois, G. H. Jacobs, *Science* **220**, 737 (1983).
10. D. J. Felleman and E. McClendon, *Soc. Neurosci. Abstr.* **17**, 1282 (1991); E. A. DeYoe and L. C. Sisola, *ibid.*, p. 1282.
11. J. G. Malpeli, P. H. Schiller, C. L. Colby, *J. Neurophysiol.* **46**, 1102 (1981); J. H. R. Maunsell, T. A. Nealey, D. D. DePriest, *J. Neurosci.* **10**, 3323 (1990); V. P. Ferrera, T. A. Nealey, J. H. R. Maunsell, *Invest. Ophthalmol. Visual Sci. (Suppl.)* **32**, 1117 (1991).
12. D. C. Van Essen and C. H. Anderson, in *Introduction to Neural and Electronic Networks*, S. F. Zornetzer, J. L. Davis, C. Lau, Eds. (Academic Press, Orlando, FL, 1990), pp. 43–72.
13. H. Wässle, U. Grünert, J. Röhrenbeck, B. B. Boycott, *Vision Res.* **11**, 1897 (1991).
14. A. M. Derrington and P. Lennie, *J. Physiol.* **357**, 219 (1984).
15. R. T. Marrocco, J. W. McClurkin, R. A. Young, *J. Neurosci.* **2**, 1275 (1982); for a different viewpoint, see: E. Kaplan, B. B. Lee, R. M. Shapley, *Prog. Retinal Res.* **9**, 273 (1990); C. Blakemore and F. Vital-Durand, *J. Physiol. (London)* **380**, 453 (1986).
16. W. H. Merigan, in *From Pigments to Perception*, A. Valberg and B. B. Lee, Eds. (Plenum, New York, 1991), pp. 117–125.
17. C. Noorlander and J. J. Koenderink, *Vision Res.* **23**, 1 (1983); K. T. Mullen, *J. Physiol. (London)* **359**, 381 (1985).
18. L. N. Thibos, D. J. Walsh, F. E. Cheney, *Vision Res.* **27**, 2193 (1987).
19. R. L. DeValois, D. G. Albrecht, L. G. Thorell, *ibid.* **22**, 545 (1982); R. L. DeValois and K. K. DeValois, *Spatial Vision* (Oxford Univ. Press, New York, 1988); J. McLean and L. A. Palmer, *Vision Res.* **29**, 674 (1989).
20. J. P. Jones and L. A. Palmer, *J. Neurophysiol.* **58**, 1233 (1987).
21. S. B. Mallat, *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674 (1989); E. P. Simoncelli and E. H. Adelson, in *Subband Image Coding*, J. W. Woods, Ed. (Kluwer, Norwell, MA, 1991), pp. 143–192; D. J. Field, *J. Opt. Soc. Am.* **4A**, 2379 (1987).
22. R. C. Emerson, J. R. Bergen, E. H. Adelson, *Vision Res.*, in press.
23. C. H. Anderson, P. J. Burt, G. S. van der Wal, *SPIE Proc. Intell. Robots Comput. Vision* **579**, 71 (1985); C. Chubb and G. Sperling, *J. Opt. Soc. Am.* **5A**, 1986 (1988).
24. H. B. Barlow, *Perception* **1**, 371 (1972); D. Marr and E. Hildreth, *Proc. R. Soc. London Ser. B* **207**, 187 (1980).

25. D. M. MacKay, *Nature* **289**, 117 (1981).
26. L. G. Ungerleider and M. Mishkin, in *Analysis of Visual Behavior*, D. G. Ingle, M. A. Goodale, R. J. Q. Mansfield, Eds. (MIT Press, Cambridge, MA, 1982), pp. 549–586.
27. R. von der Heydt and E. Peterhans, *J. Neurosci.* **9**, 1731 (1989); R. Desimone, T. D. Albright, C. G. Gross, C. Bruce, *ibid.* **4**, 2051 (1984); H. Saito *et al.*, *ibid.* **6**, 145 (1986).
28. P. H. Schiller, N. K. Logothetis, E. R. Charles, *Nature* **343**, 68 (1990); W. H. Merigan, C. E. Byrne, J. H. R. Maunsell, *J. Neurosci.* **11**, 3422 (1991).
29. D. C. Van Essen, B. Olshausen, C. H. Anderson, J. L. Gallant, *SPIE Conf. Visual Information Processing: From Neurons to Chips* **1473**, 17 (1991).
30. C. Koch and S. Ullman, *Hum. Neurobiol.* **4**, 219 (1985); F. Crick, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 4586 (1984).
31. C. H. Anderson and D. C. Van Essen, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6297 (1987); in *Proceedings of the Second International Conference on Visual Search*, D. Brogan, Ed. (Taylor and Francis, London, in press).

32. J. Moran and R. Desimone, *Science* **229**, 782 (1985).
33. R. Desimone, M. Wessinger, L. Thomas, W. Schneider, *Cold Spring Harbor Symp. Quant. Biol.* **55**, 963 (1990); S. E. Petersen, D. L. Robinson, W. Keys, *Neuropsychologia* **25**, 97 (1987).
34. M. E. Nelson and J. M. Bower, *Trends Neurosci.* **13**, 403 (1990).
35. D. H. Hubel and T. N. Wiesel, *Proc. R. Soc. London, Ser. B* **198**, 1 (1977).
36. S. Zeki, *Cold Spring Harbor Symp. Quant. Biol.* **55**, 651 (1991).
37. J. H. Kaas and M. F. Huerta, *Comp. Prim. Biol.* **4**, 327 (1988).
38. T. Kusama and M. Mabuchi, *Stereotoxic Atlas of the Brain of Macaca Fuscata* (Univ. of Tokyo Press, Tokyo, 1970).
39. We thank T. Coogan, J. Gallant, J. Braun, and B. Olshausen for critical comments. The research described in this paper was supported in part by grants from NIH, the Office of Naval Research, and the Sloan Foundation, and was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Climate Forcing by Anthropogenic Aerosols

R. J. CHARLSON, S. E. SCHWARTZ, J. M. HALES, R. D. CESS, J. A. COAKLEY, JR., J. E. HANSEN, D. J. HOFMANN

Although long considered to be of marginal importance to global climate change, tropospheric aerosol contributes substantially to radiative forcing, and anthropogenic sulfate aerosol in particular has imposed a major perturbation to this forcing. Both the direct scattering of short-wavelength solar radiation and the modification of the shortwave reflective properties of clouds by sulfate aerosol particles increase planetary albedo, thereby exerting a cooling influence on the planet. Current climate forcing due to anthropogenic sulfate is estimated to be −1 to −2 watts per square meter, globally averaged. This perturbation is comparable in magnitude to current anthropogenic greenhouse gas forcing but opposite in sign. Thus, the aerosol forcing has likely offset global greenhouse warming to a substantial degree. However, differences in geographical and seasonal distributions of these forcings preclude any simple compensation. Aerosol effects must be taken into account in evaluating anthropogenic influences on past, current, and projected future climate and in formulating policy regarding controls on emission of greenhouse gases and sulfur dioxide. Resolution of such policy issues requires integrated research on the magnitude and geographical distribution of aerosol climate forcing and on the controlling chemical and physical processes.

T HE RESPONSE OF THE EARTH'S CLIMATE TO THE PERTUR-
bation in radiative forcing due to increased concentrations of
infrared-active (greenhouse) gases is the subject of intense
research because of the well-documented increases in concentrations
of these gases over the industrial era and the recognition of the
climatic importance of the radiative forcing associated with these
increases. It is becoming apparent that anthropogenic aerosols exert
a radiative influence on climate that is globally comparable to that of
greenhouse gases but opposite in sign. However, this aerosol
radiative influence has received much less attention than forcing by
anthropogenic greenhouse gases. In view of the magnitude of
aerosol influences on climate, it seems mandatory that these influ-
ences should be included in efforts to obtain accurate estimates of
anthropogenic perturbations to the earth's radiation budget at

R. J. Charlson is with the Department of Atmospheric Sciences and the Institute for Environmental Studies, University of Washington, Seattle, WA 98195. S. E. Schwartz is in the Environmental Chemistry Division, Brookhaven National Laboratory, Upton, NY 11973. J. M. Hales is in the Atmospheric Sciences Department, Pacific Northwest Laboratory, Richland, WA 99352. R. D. Cess is at the Institute for Terrestrial and Planetary Atmospheres, State University of New York, Stony Brook, NY 11794-2300. J. A. Coakley, Jr., is with the Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331-2209. J. E. Hansen is at the Goddard Institute for Space Studies, National Aeronautics and Space Administration, 2880 Broadway, New York, NY 10025. D. J. Hofmann is with the Climate Monitoring and Diagnostics Laboratory, National Oceanic and Atmospheric Administration 325 Broadway, Boulder, CO 80303-3328.

present and over the industrial era. Such estimates are essential for
(i) evaluating climate sensitivity from observed climate change, (ii)
evaluating the performance of climate models, and (iii) reliably
predicting potential future climate changes. In this article we
describe the mechanisms by which anthropogenic aerosols perturb
the global climate, provide estimates of the global-average magni-
tude of the aerosol perturbation in radiative forcing, outline the
information required to describe the spatially nonuniform pertur-
bation in climate models, assess where additional information is
required, and suggest approaches to gaining this information.

Although it has long been recognized that tropospheric aerosols
exert a cooling influence on climate because of their scattering of
shortwave radiation and the resultant increase in planetary albedo
(*1*), this influence has been widely assumed to be fairly uniform
spatially and constant temporally [for example, (*2–4*)], and this
perception has been reflected in most analyses of global climate
change (*5–8*). However, industrial activities, especially emissions of
SO_2, which result in the formation of particulate sulfate (SO_4^{2-})
compounds, contribute substantially to tropospheric aerosol, espe-
cially to submicrometer aerosol, which is effective in the scattering of
shortwave radiation (*9*), and this aerosol is distributed quite non-
uniformly over the earth and has substantially increased in concen-
tration since around 1850 (*10–12*). Thus, there is strong reason to
infer that anthropogenic sulfate aerosol substantially enhances local