The Timing of Information Transfer in the Visual System

LIONEL G. NOWAK and JEAN BULLIER

1. Introduction

The mammalian visual cortex is composed of a constellation of cortical areas that are interconnected by a dense network of corticocortical connections. Among those connections, it is usual to distinguish between feedforward and feedback connections. Feedforward connections carry information away from area V1 toward the parietal and the temporal lobes, whereas feedback connections carry impulses in the reverse direction (Salin and Bullier, 1995). It is becoming increasingly apparent that, despite its complexity, the visual system processes information very rapidly. The delay imposed by neuronal processing in the correction of visually guided movements is of the order of 100 msec (Rossetti, 1997). Recent results also suggest that visual recognition of complex scenes is possible within 100–200 msec (Thorpe et al., 1996).

The question we address is how such a complicated network of interconnected cortical areas can process information so rapidly. We first review the evidence concerning two major constraints for processing speed by neural net-
works: the neuronal integration time and the axonal delays. We then present evidence for the rapid responses of neurons to visual stimulation at different stages of the visual system. We conclude by discussing how different models of the visual system are compatible with such a rapid processing time.

2. Structural Constraints: Axonal Conduction and Integration Times

2.1. Axonal Conduction Time

2.1.1. Minimizing the Sizes of Cortical Axons

In large-brained animals, white matter takes up a large proportion of the brain volume. For example, in humans this proportion is estimated at 40–60% (Hofman, 1985; Murre and Sturdy, 1995). In order to maintain brain volume within reasonable bounds, it is therefore important to minimize the space taken by axons. This is particularly relevant for corticocortical axons since they make up 90% of the white matter underlying the cortex (Abeles, 1991).

Minimizing axonal volume is probably the major constraint that pushes highly interconnected neurons to group together in the same cortical areas and modules, since these are highly efficient strategies to conserve axonal volume (M. E. Nelson and Bower, 1990; Mitchison, 1992; Murre and Sturdy, 1995). Another way to reduce the volume of white matter is to reduce axon diameter. This probably explains why, with a few exceptions noted below, corticocortical axons are generally small compared to axons of the peripheral nervous system. Within the gray matter of the visual cortex, most axons are smaller than 1 μm in diameter; for example, in cat and monkey area 17, the mode of the distribution of axon diameter is 0.6–0.7 μm (Haug, 1968; Peters and Sethares, 1996). The small size of cortical axons is also maintained in the white matter. This can be observed in the corpus callosum, which provides a good sample of corticocortical axons. In the macaque monkey, the mean diameter of callosal axons is less than 1 μm, and in the rostral part of the corpus callosum, all axons have diameters less than 1 μm (LaMantia and Rakic, 1990). This contrasts with the monkey optic nerve, in which most axons are larger than 1 μm (Reese and Ho, 1988).

A small number of giant axons (larger than 2.5 μm) is observed in the part of the corpus callosum containing axons between the extrastriate areas and parietal cortex. In contrast, the region of the corpus callosum containing axons linking together both temporal cortices does not appear to contain giant axons (LaMantia and Rakic, 1990). This suggests that giant axons are limited to specific networks for which high conduction velocity is necessary. We will see below that indeed processing speed is higher in dorsal extrastriate areas and parietal cortex than in temporal cortex.

Since there is a linear relationship between conduction velocity and fiber diameter (see Section 2.1.2), minimizing axon size for conserving space has important consequences on delays due to conduction along axons.
2.1.2. Methods for Estimating Conduction Delays

There are several ways of estimating delays due to conduction of the action potential along the axon: calculating it from the distribution of axon sizes, measuring the position of the peak with respect to the origin of the time difference axis in cross-correlation histograms (CCH), measuring the latencies of excitatory postsynaptic potentials (EPSPs) recorded with spike-triggered averaging, or measuring latencies of spikes evoked by antidromic activation using electrical stimulation. In addition, in some cases, it is possible to use latencies of spikes or EPSPs evoked by orthodromic electrical stimulation.

The morphological method consists in estimating the distribution of fiber diameters and deducing that of axonal conduction velocities. The diameter of the axon with its myelin sheath is called the fiber diameter. Axon and fiber diameters are linearly related with a coefficient of 0.77 (Waxman and Bennett, 1972) (0.64–0.87 for rabbit callosal axons; Waxman and Swadlow, 1976). Using a coefficient of 5.5 between conduction velocity and fiber diameter (Hurst, 1939; Waxman and Bennett, 1972), one gets 7.14 as the coefficient between conduction velocity and axon diameter. Note, however, that this linear relationship applies only for myelinated fibers. For nonmyelinated axons, conduction velocity varies as the square root of the axon diameter (Rushton, 1951).

Deducing conduction velocity from axon diameter has the advantage of giving an estimate of the conduction delays of all the axons of a given pathway. However, to get a proper estimate of axon diameters, it is necessary to measure them on electron micrograph (EM) images. The EM technique is best used on well-identified tracts such as the interhemispheric commissures, which can be cut in a plane orthogonal to the direction of the axons. In the case of axons between two cortical areas, for example, it is necessary to stain fibers with axonal tracers such as biocytin or Phaseolus vulgaris leucoagglutinin. These techniques usually produce a small number of well-stained axons and it is difficult to know to what extent they do not favor the largest axons. Also, estimation for axon diameter from EM images is more difficult to use in that case, because of the uncertainty concerning the angle between the section plane and the axis of the axon. Finally, it is known that large differences in diameter can exist between the main axon trunk and its cortical ramifications (Honzel et al., 1994), which are often nonmyelinated. Measuring only the axon trunk diameter may therefore lead to a serious underestimate of the conduction time unless the entire axon is reconstructed.

Cross-correlation is a technique that gives the temporal relationship between the firing times of two neurons (A and B) recorded extracellularly. The action potentials of neuron A are used as time markers to trigger the averaging of the activity of the other neuron (B). The cross-correlation histogram is a measure of the probability of observing a spike in neuron B as a function of the time before or after the occurrence of an action potential in neuron A (Perkel et al., 1967). It can also be seen as an estimate of the firing rate of neuron B as a function of the time before or after a spike in neuron A (Abeles, 1982). When the firing times of the two neurons are independent, the cross-correlation histogram is flat. If the neuron A activates neuron B with a delay \( t \), the cross-correlation histogram shows a peak with a delay \( t \), indicating that the firing probability...
of the second neuron increased after neuron A fired an action potential. When the two neurons are simultaneously activated by a third group of neurons, the cross-correlation histogram presents a peak centered at the origin of time (due to the common input, the discharge probability of neuron B increases at the same time as that of neuron A).

The cross-correlation method has the major advantage that it gives the actual average delay between the production of one spike in a neuron and a resultant spike in the target neuron, including axon conduction time and synaptic delay. Furthermore, measures derived from cross-correlation histograms are made from the activity of neurons activated by relatively “natural” stimuli, in contrast to electrical stimulation that artificially synchronizes the spikes of a large number of neurons. Calculating cross-correlation histograms is therefore the method of choice for measuring delays between two neurons. Its use is limited, however, by the difficulty of isolating more than a small number of pairs of interconnected neurons in two different structures.

Spike-triggered averaging is related to the calculation of cross-correlation histograms. Two neurons are recorded, one using intracellular and the other extracellular recording. The membrane potential of the intracellular neuron is averaged using the spikes of the other neuron as triggering events. In the averaged intracellular records, one can sometimes identify small synaptic potentials that are interpreted as resulting from the monosynaptic connection between the two neurons. Measuring the EPSP latency provides a good estimate of conduction time in the axon and synaptic delay between the two neurons. Another method with an even lower yield consists in recording intracellularly two neurons at the same time and recording the latency of the EPSP evoked in one neuron by stimulation of the other neuron. The advantages (reliability of measures and lack of artificial synchronization) and the disadvantage (low encounter rate) of these techniques are similar to those of cross-correlation.

Electrical stimulation can be used to stimulate axons in the orthodromic or antidromic direction. When orthodromic activation is used, latencies can be measured between the stimulus and the resulting spike in the postsynaptic neuron in case of extracellular recording, or the EPSP if the recording is intracellular. This technique should be used with caution with neural populations which are interconnected by reciprocal connections. This is due to the fact that electrical stimulation activates axons and not cell bodies (Nowak and Bullier, 1997a,b) and that, because of the presence of recurrent collaterals, it is difficult to know whether one neuron is activated by true orthodromic activation or whether electrical stimulation elicited a spike which traveled in the antidromic direction and provided a synaptic input through recurrent collaterals of neighboring neurons.

These difficulties disappear when antidromic activation is used. Measuring antidromic latencies with electrical stimulation is relatively easy and gives conduction times between structures that can be identified with electrophysiological recordings. Besides the possibility that electrical stimulation may be biased toward large neurons, which are easier to record from, and toward large axons, which are more easily activated by electrical stimuli (Ranck, 1975), one of the difficulties related to this method is that of knowing the actual site of stimulation. Unless small stimulating currents are used, it is difficult to avoid the possibility of activating axons in the white matter below the stimulated area. One way
to reduce this risk is by verifying that the laminar distribution of activated neurons conforms to that predicted by the anatomy (Swadlow and Weyand, 1981).

Assessing the validity of the electrical stimulation method can be done by comparing the distribution of conduction velocities calculated from antidromic latencies and fiber length and that calculated from the axon sizes. When this is done on callosal connections of the macaque monkey using a coefficient of 7.14 between conduction velocity and axon size (Swadlow et al., 1978; LaMantia and Rakic, 1990), the match is actually quite good, suggesting that antidromic latencies to electrical stimulation give a proper estimate of the conduction times of all the axons between the two structures.

2.1.3. Specificity of Conduction Times in Different Connections and Species

Each set of connections possesses a characteristic pattern of conduction delays. This is well illustrated in Fig. 1, which presents the distributions of antidromic latencies of neurons in cortical area V1 of the awake rabbit following electrical stimulation in several afferent structures (Swadlow and Weyand, 1981). The figure shows that there are important differences between the conduction times in corticogeniculate, corticocollicular, callosal, and association corticocortical axons.

It is remarkable that the ordering of conduction times between different targets of cortical neurons (corticotectal short, association intermediate, and corticotalamic long) is conserved in different species such as rabbit, monkey, and cat (see below). It is also observed for different sensory modalities; one can find a similar ordering of latencies for the efferent connections of area V1, area S1 and the motor cortex of the rabbit (Swadlow, 1992, 1994). This suggests that speed of information transfer in different corticofugal pathways is subjected to the same constraints across different sensory modalities and different species.

![Diagram](image)

**Figure 1.** Distributions of antidromic latencies in rabbit area V1 after electrical stimulation in the LGNd, area V2, visual cortex of the other hemisphere (CC), and superior colliculus (SC). The small vertical lines represent the 10%, 50% (median), and 90% centiles. Redrawn from Swadlow and Weyand (1981).
Contrary to what one would expect, conduction time along a given fiber tract is not governed mainly by fiber length. Comparison of antidromic latencies of callosal axons in different species (Fig. 2) shows that conduction delays are much shorter in the cat than in rabbit, despite the fact that they cover a longer distance. Conduction times are similar in rabbit and monkey despite the much longer paths of callosal axons in monkey. In the corticocortical pathway, a similar interspecies order can be observed, with axons in the cat displaying shorter conduction times than in the rabbit or the monkey (which do not differ significantly; see references in Section 2.1.4f). Higher conduction speed is also found in the retina-geniculo-striate pathway of the cat compared to that of the monkey (Stone, 1983). This suggests that brains of different species function at different speeds: cat brains are faster than monkey and rabbit brains, at least for the visual system.

2.1.4. Axon Conduction Delays in Different Sets of Connections

2.1.4a. Three Classes of Axons in the Retinogeniculate Pathways. The relatively large sizes of axons in the optic nerve and the optic tract in cats and monkeys (Hughes and Wässle, 1976; Reese and Guillery, 1987; Reese and Ho, 1988) suggest that transfer time between the retina and the cortex must be rapid. Approximately half of the conduction delay is due to the slow unmyelinated intraretinal part of the axon (Stanford, 1987). From the optic disk to the dorsal lateral geniculate nucleus (LGNd), axons are myelinated and conduct much faster. There appears to be a compensation of the intraretinal delay by the conduction velocity of the myelinated part of the axon so that the conduction delay between the retinal ganglion cell and the LGN does not depend on the position of the ganglion cell in the retina (Stanford, 1987).

Conduction times between the retina and the LGN in the cat have been measured by cross-correlation studies and shown to be around 4 msec on average, with a range of 1.8–7 msec for X and Y cells (Cleland et al., 1971; Mastronarde, 1987). Delays between the LGN and the cortex appear shorter, on the order of 1–6 msec as found in cross-correlation studies (Tsumoto et al., 1978; Tanaka, 1983). A heroic cross-correlation experiment between retina and cortex performed by Lee et al. (1977) showed that the delay between a spike in a retinal ganglion cell and its cortical target is of the order of 4.5–7.5 msec for X or Y
cells. Such a short conduction time is remarkable given the long distance between the eye and the cortex of area 17. It is interesting to note that it takes a similar amount of time to get from the eye to area 17 than between two adjacent cortical areas (see Section 2.1.4c).

Since the early recordings of George Bishop in the 1930s, it has been known that the transfer of information in optic nerves of mammalian species is made along three groups of axons with different conduction velocities (Stone, 1983). These have been called X, Y, and W in the cat (Stone, 1983) and M, P, and K in the monkey (Casagrande, 1994; Hendry and Yoshioka, 1994). Differences between these groups in axon conduction times from retina to cortex can be sizable because there is a positive correlation between the conduction speeds of axons of geniculate cells and those of their retinal afferents (Cleland et al., 1976; Marracca, 1976; Wilson et al., 1976; Schiller and Malpeli, 1978). In the cat, for example, if we add the retinogeniculate conduction times (Kirk et al., 1975) to the geniculocortical delays (Wilson et al., 1976), we find that, on the average, Y axons transfer information to cortex 3.5 msec faster than X axons, and X axons are faster by 6 msec than W axons. In the monkey, a similar calculation can be made for M and P pathways by adding conduction times from the retina to the optic chiasm (Gouras, 1969; Schiller and Malpeli, 1977b), from the chiasm to the LGN (Dreher et al., 1976; Schiller and Malpeli, 1978), and from the LGN to cortex (Schiller and Malpeli, 1978; Bullier and Henry, 1980). This shows that, on average, M axons take 5.5 msec less time to transfer impulses to cortex than P axons. In keeping with that conclusion, the study of Müldorf and Singer (1979) shows that, after electrical stimulation of the optic nerve or optic chiasm, the onset latency of synaptic potentials in area V1 is shorter by 6 msec in layer 4Ca, which receive M inputs, than in layer 4Cβ, which is activated by P afferents.

2.1.4b. Intra-Area Axonal Conduction Times. Once afferent messages have reached the cortex of area 17, they are processed by interneuronal communication within the area and between different cortical areas. Little is known concerning the axon conduction delays in these two groups of pathways. For intra-area conduction, most data come from intracellular recordings in cortical slices of rat visual or somatosensory cortex, using spike-triggered averaging of membrane potential or double intracellular recordings. The values are remarkably long (several milliseconds), given that the recorded neurons are distant by only a few hundred micrometers (Thomson et al., 1988; Mason et al., 1991; Nicoll and Blakemore, 1993). These correspond to very slow conduction velocities (of the order of a few tenths of m/sec). With antidromic activation techniques, the axon conduction velocity for axons of intrinsic connections ranged between 0.26 and 0.47 m/sec within rat area 17 (Nowak et al., 1997) and averaged 0.28 m/sec in rat area 18a (Lohmann and Röhrig, 1994).

Conduction delays were also measured in cat visual cortex with spike-triggered averaging (Komatsu et al., 1988). These are also long (e.g., 2.8 msec for 720 μm distance), corresponding to conduction velocities of the order of 0.3 m/sec, after subtraction of the synaptic delay. Interestingly, the conduction velocity for the intrinsic collaterals of an axon may be very different from that of the main trunk projecting elsewhere. Kang et al. (1988) used spike-triggered averaging to measure the latencies of intracortical EPSPs evoked by pyramidal
tract cells in cat motor cortex. When these latencies are compared with the antidromic latencies evoked by stimulation of the pyramidal tract, it appears that conduction delays within cortex are of the same order of magnitude (a few milliseconds) as transfer time in the pyramidal tract axon, despite the enormous difference in conduction distances (a few hundred micrometers in intracortical connections versus several centimeters for pyramidal tract axon). This confirms the slow nature of conduction in intracortical collaterals of cortical axons.

These very slow conduction velocities for intra-area communication in rat and cat cortex fall below the range usually associated with myelinated axons (down to 1 or 2 m/sec; Rushton, 1951; Waxman and Bennett, 1972). This suggests that intrinsic collaterals of pyramidal cell axons that make up most local connections are usually not myelinated along their whole length. It is also in keeping with the observation that many cortical axons are not myelinated in the cortical gray matter (Peters and Sethares, 1996).

The low speed of information transfer within a given cortical area raises interesting questions concerning the temporal aspects of intra-area information processing. Local ramifications of cortical collaterals can reach several millimeters (Gilbert and Wiesel, 1981; Martin and Whitteridge, 1984). With conduction velocities of the order of 0.3 m/sec, this means that spikes will reach the target neurons located 3 mm away 10 msec later. This may explain the observation in in vivo intracellular recordings of cat cortical neurons of long delays (10–30 msec) of synaptic activity triggered by visual stimulation in peripheral parts of the receptive field (Fregnac and Bringuier, 1996). Slow conduction velocities of intrinsic connections may also be the reason for the slow (0.1–0.2 m/sec) spread of synaptic activity across monkey visual cortex revealed by optical imaging (Grinvald et al., 1994).

2.1.4c. Association Corticocortical Axons. Corticocortical connections between areas of the same hemisphere are usually reciprocal and it is possible to identify feedforward and feedback connections (Felleman and Van Essen, 1991; Salin and Bullier, 1995). Besides anatomical and functional differences (Salin and Bullier, 1995), it is possible that these two types of connections show differences in conduction delays. Such differences have been found in the thalamocortical connections between LGN and cortex. For example, in the cat, most LGN axons conduct impulses in less than 4 msec in the feedforward direction (Wilson et al., 1976), whereas numerous corticogeniculate axons take more than 4 msec to transfer information in the feedback direction (Harvey, 1980; Grieve and Sillito, 1995).

For corticocortical connections, this has been directly addressed in the rat between areas 17 and 18a using antidromic activation of single neurons. The results led to the conclusion that conduction times are similar in feedforward and feedback connections (Nowak et al., 1997). Antidromic latencies of feedforward axons ranged between 3.7 and 7.9 msec and those of feedback between 4.1 and 8.9 msec. Differences in mean values of latencies (3.7 msec for feedforward and 6 msec for feedback) and conduction velocities (0.41 and 0.38 m/sec) were not statistically significant. In the monkey visual cortex, comparison of the diameter of axons in feedforward and feedback connections between monkey areas
V1 and V2 (Rockland and Virga, 1989, 1990) also leads to the conclusion that, in contrast to thalamocortical connections, feedforward and feedback corticocortical connections involve similar conduction times.

Whether measured with antidromic activation or by the onset latency of evoked potentials (Nowak et al., 1997), it appears that conduction times in axons between areas 17 and 18a in the rat are long (several msec), given the short distance between these areas. The onset latency of synaptic responses obtained in rat extrastriate cortex after electrical stimulation in area 17 ranged between 3 and 10 msec, as measured by voltage-sensitive dyes in vivo (Orbach and Van Essen, 1993).

Long conduction times were also observed in the rabbit visual cortex. According to Swadlow and Weyand (1981), the majority of antidromic latencies of area 17 neurons are between 2 and 12 msec (median at 6–7 msec) after electrical stimulation in area 18. The corresponding conduction velocities range between 0.23 and 5.74 m/sec (median at 0.64 m/sec), values that are only marginally larger than in the rat.

Very little is known concerning the conduction times in connections between cortical areas in the cat and monkey. Recording of antidromically activated neurons in cat area 17 after stimulation in areas 18 and 19 shows that some feedforward corticocortical axons conduct rapidly (less than 2 msec delay; Toyama et al., 1974; Bullier et al., 1988), while others are slow (2–12 msec delays; Bullier et al., 1988). Feedback connections from these areas onto V1 tend to involve small axons (below 1 μm; Henry et al., 1991) and are therefore likely to conduct impulses slowly. Orthodromic and antidromic latencies of area 17 neurons activated by electrical stimulation in PMLS tend to be shorter (mostly below 3 msec; G. H. Henry, personal communication) than those from areas 18 or 19. This is an interesting observation given that PMLS is often considered to be the homologue of area MT in the macaque (Shipp and Grant, 1991; Payne 1995) and that connections between areas 17 and MT in the macaque involve very rapidly conducting axons (see below).

In the monkey, by subtracting the latencies of the current sinks evoked in areas V1 and V2 by electrical stimulation of the optic radiation (Mitzdorf and Singer, 1979), it is possible to deduce that it takes several milliseconds to transfer information between these two areas. The short distance between areas V1 and V2 (a few millimeters) and the fact that axon diameters are of the order of 1 μm (Rockland and Virga 1990) suggest that axon conduction times are of the order of 1 msec. Therefore, most of the delay between areas V1 and V2 probably corresponds to synaptic delays and integration times.

In contrast to the connections between areas V1 and V2, at least some of the neurons projecting from V1 to MT are characterized by the presence of giant axons, 3 μm or larger (Rockland, 1995). This predicts that very short latencies should be measured between areas V1 and MT. Indeed, despite the large separation (greater than 1 cm) between corresponding regions in areas V1 and MT, measurements of antidromic latencies of area V1 neurons after electrical stimulation in area MT gave very low values, between 1.1 and 1.7 msec (Movshon and Newsome, 1996). Beyond MT, there are several areas of the parietal cortex, such as MST and VIP, that show dense myelination in the gray matter. If we assume
that this is a characteristic feature of terminals of rapidly conducting axons by analogy with area MT, we can conclude that the pathway to parietal cortex through areas MST and VIP must involve very short conduction delays. This is in keeping with the very short latencies to visual stimulation that have been reported in this region (see Section 3).

There are no direct measurements of conduction delays in other corticocortical connections in the monkey. Measurement of axon size in the connection between V2 and V4 (Rockland, 1992) and between TEO or V4 to V2 and V1 (Rockland et al., 1994) gave values between 0.5 and 1.5 \( \mu \text{m} \). Because of the deep sulci, the distance between interconnected regions in V2 and V4 must be of the order of 2 cm. This suggests that it takes several milliseconds to conduct impulses from V2 to V4.

More information is needed concerning the conduction delays between cortical areas in the monkey in order to understand the large range of latencies to visual stimulation (see Section 3). From what we know, it appears that most interarea axonal transfers take several milliseconds, with the exception of the connection between V1 and MT, which appears remarkably fast.

2.1.4d. Callosal Axons. Conduction times in the callosal pathway of rabbits has been extensively studied by Swadlow and his collaborators using electrical stimulation. The shortest antidromic latency was 2.4 msec, but most latencies were long, with a maximum value of 39.8 msec (Swadlow, 1974). Swadlow and Weyand (1981) obtained similar results, with conduction velocities ranging between 0.5 and 8.25 m/sec (median 2.8 m/sec). The large range of conduction velocities and the very low values for some axons correspond to the morphological characteristics of the callosal axons in the rabbit. In the posterior part of the splenium, 45\% of the fibers are not myelinated, with diameters between 0.08 and 0.6 \( \mu \text{m} \); the myelinated axons range between 0.3 and 1.85 \( \mu \text{m} \) in diameter (Waxman and Swadlow, 1976). A few very large fibers (2–3 \( \mu \text{m} \) diameter) were also observed.

Conduction times in cat callosal axons have been measured by several groups (Tovar et al., 1974; Harvey, 1980; Innocenti, 1980; McCourt et al., 1990). All these studies returned short conduction times, most of them being between 1 and 5 msec, corresponding to relatively rapid conduction velocities (9–17 m/sec; mean 13 m/sec). This is clearly much faster than in the rabbit (Fig. 2). The difference can be explained by the high proportion of large-diameter myelinated fibers in the cat corpus callosum (Naito et al., 1971).

Fewer studies have been made of interhemispheric conduction delays in the macaque monkey. As mentioned above, Swadlow and his colleagues (Swadlow et al., 1978) concluded that callosal axons of neurons in the prelunate gyrus of the monkey are relatively fast conducting (median 7.4 m/sec), leading to conduction delays between areas V4 of both hemispheres of 2.8–22.5 msec (median 7 msec). To estimate delays in other regions of the corpus callosum, one can use the comprehensive morphological study of LaMantia and Rakic (1990). Most axons must conduct impulses more slowly than 7 m/sec and very few giant axons conduct faster than 20 m/sec. Assuming an average axon length of 50 mm, this correspond to interhemispheric delays longer than 7 msec for most axons and
below 2.5 msec for the fastest axons. In the rostral part of the corpus callosum, it is likely that the numerous nonmyelinated axons (LaMantia and Rakic, 1990) conduct more slowly than 1 m/sec, which would give interhemispheric delays larger than 50 msec.

2.1.4e. Slow Corticogeniculate Axons. As illustrated in Fig. 1, corticothalamic axons are characterized by long conduction delays due to very slow conduction velocities (mean 0.64 m/sec in the rabbit). Studies in the cat have also identified a population of slow corticogeniculate axons (Harvey, 1980; Grieve and Sillito, 1993). On average, it takes several milliseconds to transfer impulses from LGN to cortex and 10 msec to return to the LGN. It is therefore likely that the initial 20 msec of responses of rapidly activated neurons in the LGN (Y cells) is not affected by cortical feedback. Such slow conduction of corticothalamic axons suggests that either this pathway is used to modify the long-term membrane properties of relay neurons (McCormick and Van Krosigk, 1992) or that it acts preferentially on neurons that respond late to visual stimulation, such as lagged cells (Mastroberdey, 1987; Saul and Humphrey, 1990).

2.1.4f. A Fast Cortical Pathway to the Superior Colliculus. Latencies measured in the cortex following electrical stimulation of the tectum are short (Fig. 1), due to the high conduction velocities of corticofugal axons (average 8.4 m/sec in the rabbit). Similar rapid transfer times in the corticofugal pathway have been demonstrated in the cat (Palmer and Rosenquist, 1974; Toyama et al., 1974; Harvey, 1980; Thalluri and Henry, 1989; Casanova, 1993) and in the monkey (Finlay et al., 1976).

The superior colliculus is innervated by W and Y types of retinal axons in the cat (Hoffmann, 1973) and the monkey (Schiller and Malpeli, 1977b). It also receives in addition a cortical input from neurons located in layer V of various strata cortical areas (Gilbert and Kelly, 1976; Fries, 1984). This indirect input appears to be under the control of the M system in the monkey (Schiller and Malpeli, 1978) and the Y system in the cat (Hoffmann, 1973). It is interesting to note that the fastest set of output axons from visual cortex is the corticofugal pathway (Fig. 1). There seems therefore to be a fast corticofugal circuit that is driven by fast retinal axons and provides rapid activation to the superior colliculus. Such an organization gives support to the idea that fast conducting axons are associated with specific roles requiring rapid reactions, such as activating eye movements.

2.2. Neuronal Integration Time

The axonal conduction delay is not the sole parameter that determines the time it takes to transfer information from one structure to another. Once the spikes have reached the axon terminals, one must consider the length of time it takes for a depolarizing event evoked by the incoming spike to bring the membrane potential of the target cell to spike threshold. This amount of time, which
we will name the “integration time,” can be relatively long and can vary depending on the state (e.g., the membrane potential) of the neuron.

2.2.1. Chronaxie Measurements

In order to characterize quantitatively the integration time in neurons, the simplest experiment situation consists in injecting depolarizing current pulses into intracellularly recorded neurons. It is then possible to determine when the threshold for action potential is reached as a function of the current intensity. Such studies have been performed in cat motor cortex neurons in vivo and rat visual cortex in vitro. The results showed what can be expected, namely that the threshold is reached faster for higher current intensities. A good linear fit of the relationship can be obtained by plotting the inverse of the firing time as a function of current intensity, such that \( I = a/t + b \), with \( I \) the current intensity and \( t \) the time of spike occurrence. This relationship was established by Weiss (1901) for the extracellular stimulation of frog and toad motor nerves.

The relationship \( I = a/t + b \) can be written as \( I = I_0 G/t + I_0 \) (e.g., Ranck, 1975). The constant terms of the equation can be determined from the linear regression between \( 1/t \) and \( I \). The parameter \( I_0 \) correspond to the rheobase, which is the current for which a spike would be obtained at a time equal to infinity (when \( 1/t \to 0 \)). The constant \( G \) is the well-known chronaxie. It corresponds to the latency of the spike that is obtained with a current intensity set to twice the rheobase value.

The average values of chronaxies obtained in cat motor cortex neurons are 9.8 msec for fast pyramidal tract neurons and 27.1 msec for slow pyramidal tract neurons (Koike et al., 1970). In rat visual cortex, the average value obtained for neurons of the supragranular layers is 15 msec (Nowak and Bullier, 197a). In other words, the delay between the onset of a depolarization and the spike that results from this depolarization varies from 1 to several tens of msec when the current intensity is twice the threshold value. This is much larger than the chronaxies of axons, which are usually measured in tenths of a millisecond (Ranck, 1975; Nowak and Bullier, 197a).

The chronaxies of neurons are likely to be influenced by their membrane properties. This includes the resting membrane potential, the passive time constant, and different conductances that are activated in the subthreshold range of membrane potential. Such mechanisms include the potassium current \( I_K \), which acts as a brake and delays the time at which the threshold for the action potential is reached (Segal and Barker, 1984; Eder et al., 1994), or the nonactivating sodium current \( I_{NaP} \), which induces an anomalous rectification in cortical neurons which increases the apparent time constant (Stafstrom et al., 1985; Sutor and Hablitz, 1989). Thus, the chronaxie, and therefore the integration time, is likely to be influenced by passive and active membrane properties, to an extent that remains to be established. In addition, it is possible that both network activity and neuromodulatory substances acting on these properties affect the integration time of cortical neurons. Integration time is therefore a parameter that is determined by multiple factors.
2.2.2. Integration Time at Resting Membrane Potential

It is difficult to extrapolate the results of chronaxic measurements to the integration time of cortical neurons when they are activated by natural stimuli. In these conditions, single EPSPs are too small to depolarize neurons from rest to spike threshold. Neuronal depolarization requires the temporal summation of synaptic potentials. The number and the degree of synchronization of these synaptic potentials determine the speed of the temporal summation. Altogether, the shape of the underlying synaptic current differs from the square pulse used for chronaxic measurements. Nonetheless, chronaxic measurements tell us that integration times in neurons must be of the order of several to several tens of milliseconds.

The integration time that can be measured for natural stimulations corresponds to the difference between the latency of a spike and the onset latency of the triggering compound synaptic potential. Despite its fundamental importance in understanding the temporal aspect of cortical processing, integration time has been measured only in a small number of studies, because such measurements require the use of the difficult technique of intracellular recordings in vivo.

Greutzfeldt and Ito (1968) used small visual stimuli flashed inside the receptive field of neurons intracellularly recorded in cat area 17. They found that the onset latency of the EPSPs was between 30 and 40 msec after stimulus onset. The spike latency was between 35 and 100 msec. In other words, the delay between the onset of the synaptic potentials and the action potential it elicited (the integration time) was between 5 and 60 msec. Integration time was also reported recently by Volgushev et al. (1995) in cat area 17 using whole-cell patch-clamp recording. The stimuli were bars flashed in the receptive fields of the neurons. The integration time was 5–12 msec for optimally oriented stimuli (mean 7.6 msec) and 6–15 msec (mean 10.6 msec) for orientation close to the optimal (nonoptimal orientation, but sufficiently close to it) to elicit action potentials.

These two studies show that the integration time is much longer than what would be predicted from the results of orthodromic electrical activation of the retinotectal pathway. Several studies showed that there is only a fraction of a millisecond between the arrival of afferent spikes in geniculate axons and the triggering of a spike in first-order neurons (Singer et al., 1975; Bullier and Henry, 1979). These results indicate that cortical neurons can react with extremely fast integration times in the artificial situation of electrical stimulation of their afferents, when resulting EPSPs are presumably very numerous and synchronized by the electrical pulse. The relationship between current intensity and chronaxie (see above) suggests that such fast integration times require very high synaptic currents that are probably outside the physiological range.

However, as mentioned above, in response to visual stimuli, neurons reach their firing threshold quite slowly. This slowness suggests that either a small number of functional inputs contribute to the beginning of the response, and/or there is a rather coarse synchronization of the afferent activity. This would be the consequence of the small number of individual synapses established between a thalamocortical axon and a given cortical neuron (Freund et al., 1985, 1989;
Peters et al., 1994), combined with the latency scatter reported for thalamic relay cells in response to visual stimulation (see below).

Matsumura (1979) recorded intracellularly from the motor cortex of awake behaving monkeys. He noticed that the visual cue used to trigger the behavioural response of the monkey elicited an EPSP with a latency between 80 and 180 msec. This EPSP preceded the movement by 70 to 180 msec. More interestingly, the delay between the onset of this EPSP and the first spike was long; 39 ± 29 msec (mean ± S.D.) in pyramidal tract neurons, 77 ± 63 in nonpyramidal tract neurons. This again points to the fact that, in normal conditions, action potentials of cortical neurons may appear only very late after the first EPSPs. In that case, however, it is possible that the visually elicited synaptic response required the addition of synaptic potentials provided by other sources of afferents to lead to spike threshold.

2.2.3. How EPSPs Are Transformed in Spikes when Neurons Are Close to the Firing Threshold

The preceding section concerned the time it takes for the membrane potential of a neuron to increase from rest to firing threshold. This part of the discussion deals with the way and the time it takes for an EPSP to be transformed into a spike when the membrane potential is already close to firing threshold. For this, we need a tool that gives the delay between the onset of an EPSP and the onset of the resultant spike. One way is to relate the shape of the cross-correlation histogram (CCH) peak to that of the underlying EPSP.

In the case of a monosynaptic connection, the shape of a cross-correlation histogram peak indicates at what time and with what dispersion the spike was triggered from the underlying EPSP. In a modelization study, Knox (1974) established that, theoretically, the shape of a CCH peak must reflect that of the rising phase of the EPSP. This is easy to understand; the rising part of the EPSP is that part of the EPSP that will cross the firing threshold. Several studies have tested the prediction of Knox's modelization in the spinal cord (Kirkwood and Sears, 1978; Fetz and Gustafsson, 1983; Gustafsson and McCrea, 1983; Cope et al., 1987). One general outcome from these studies was that the shape of a CCH peak corresponds to the temporal derivative of the underlying EPSP when the EPSP is of large amplitude. In other words, the time window during which an EPSP is effective to trigger an action potential (which corresponds to the integration time when the neuron is sufficiently close to firing threshold) corresponds to the duration of the rising part of the EPSP. However, for EPSPs of small amplitude, and for EPSP in a noisy neuron, the relationship between CCH shape and EPSP derivative no longer holds. In these conditions, the time window during which the EPSP would be effective in triggering an action potential can be as long as the whole EPSP duration.

In neocortex, the relationship between the shape of an EPSP and the timing of the action potential it triggers may be more complex: due to the presence of an active regenerative conductance ($I_{Na}$) that is activated below spike threshold (Stafstrom et al., 1985), the time lag between the onset of an EPSP and the resulting spike can be long, up to several tens of milliseconds (Reyes and Fetz, 1993a,b). This "delayed crossing" effect is less prominent for larger EPSPs.
Among in vivo studies of neocortical neurons, none has directly established the relationship between EPSP shape and spike output timing. However, comparing the rise time of EPSPs reported from in vitro studies and the width of CCH peaks obtained for monosynaptic connections reported from in vivo cross-correlation study would allow one to determine if, indeed, the integration time is bounded by the duration of the EPSP rise time. The rise time of monosynaptic EPSPs has been determined in several studies relying on spike-triggered averaging or on dual intracellular recordings in brain slices. The reported rise times range from less than 1 msec to a few, at most 5 msec (Komatsu et al., 1988; Thomson et al., 1988, 1993; Mason et al., 1991; Nicoll and Blakemore, 1993; Thomson and West, 1993). The width of cross-correlation histogram peaks corresponding to monosynaptic connections has been reported in some studies for neurons recorded within the same cortical area. Egermont (1992) reported peak widths less than 5 msec. Toyama et al. (1981) found peak durations of 5–7 msec, with some departure in a small number of cases (up to 15 msec). Michalski et al. (1983) found peak widths between 0.5 and 6 msec (2.22 msec on average). In the study of Gochin et al. (1991) the majority of monosynaptic peaks have a width close to 1 msec.

Therefore, there seems to be a good match between CCH peak width and EPSP rise time, although some CCH peaks are wider than expected. In any case, the CCH peak width is less than that of the EPSP width measured at half-height (an underestimate of the real EPSP duration), which is between 4 and 20 msec (Komatsu et al., 1988; Thomson et al., 1988; Mason et al., 1991; Nicoll and Blakemore, 1993).

2.2.4. Conclusion: Coincidence Detection and Temporal Summation

The amount of time required to trigger an action potential in a cortical neuron corresponds to the sum of the axonal delay and the integration time. The integration time itself depends on the state of the neuron. When the membrane potential of a cell is close to firing threshold, the firing of an action potential can occur as soon as the EPSP has appeared and can still be observed a few milliseconds after EPSP onset. In that case, the integration time appears to be very short and the time it takes for information to be transferred from one part of the brain to another would be slightly longer than the axon conduction delay.

If the cell is not already depolarized (i.e., if its membrane potential can be considered at rest or if it has been hyperpolarized by inhibitory inputs), a supplementary delay of at least 5 msec must be added. This situation applies to the onset latency of responses to visual stimulation. In that case, it would be a mistake to equate the physiological transfer time to the axonal delay that can be determined either by antidromic activation or by cross-correlation. Hence, knowing axonal conduction delay and integration time, it is possible to predict the difference in onset latency between two different structures in the brain. In the case of areas V1 and V2 in the monkey, it can be expected that most of the axonal conduction delays would be a few milliseconds. Adding a minimum of 5 msec to this conduction delay, it is possible to predict a latency difference between area V1 and V2 close to 10 msec. A value of 10 msec was indeed reported
in two studies comparing visual latencies in areas V1 and V2 under the same conditions (Raiguel et al., 1989; Nowak et al., 1995). This 10-msec latency difference is much larger than expected on the basis of axonal conduction delays only.

Experimental data show that the integration time when a neuron has to be depolarized from rest to firing threshold is long (5 to several tens of msec). This sluggishness can result from several factors: the small size of the incoming EPSPs, and the possibility that only a fraction of them arrive synchronously at the target cell, because of the dispersion of axonal velocities and visual response latencies in the afferents. The duration of the integration time is close to the membrane time constant. It is also close to the mean interspike interval observed in cortical neurons in response to visual stimulation and, therefore, according to the definition of König et al. (1996), it can be considered that neurons in that regime behave as integrators.

However, when neurons are close to their firing threshold, their integration time appears to be very short, since it can be reduced to a fraction of the rise time (1–5 msec) of an EPSP. Under these conditions, the larger and faster rising EPSPs are more likely to trigger action potentials than small EPSPs with a slow rising phase. Therefore, the neurons would be sensitive to the temporal coincidence of EPSPs within a narrow time window (a few milliseconds) and behave as coincidence detectors under these conditions (following the definition of König et al., 1996).

Whether a neuron functions as integrator or as coincidence detector would be determined by the amount of time it can be considered as close to firing threshold during a physiological stimulation. This has not been quantitatively established. However, the published intracellular recordings show that, in response to visual stimulation, neurons are not simply depolarized in a sustained fashion during the passage of the bar stimulus. Instead, the membrane potential undergoes marked fluctuations due to the mixture of incoming EPSPs and inhibitory postsynaptic potentials (IPSPs) (Douglas and Martin, 1991; Ferster and Jagadeesh, 1992; Pei et al., 1994). Whether these large fluctuations of membrane potential are compatible with a coincidence detector model remains to be established (for recent reviews see Shadlen and Newsome, 1994; Softky, 1995; König et al., 1996).

In addition, it remains to be established to what extent neurons that show synchronized firing also activate their target at the same time; this might be the case when the distance between source and target cells is short, but a certain degree of asynchrony might be introduced in the case of long-range connections like the callosal pathway, given the variability in axonal size and conduction velocity that has been reported in visual cortex (see Section 2.1.4d).

3. Timing of Information Transfer to and between Visual Cortical Areas

With the knowledge of the neuronal integration time and axonal delay, it is possible to interpret the latencies of neurons at different stages of the visual
system to the flashed presentation of a visual stimulus. We will examine successively the latencies of retinal, thalamic, and cortical neurons.

3.1. Latency in the Retina and Lateral Geniculate Nucleus

3.1.1. Latencies in Cat Retina and LGN

Latencies of responses of retinal ganglion cells to visual stimulation provide an estimate of the processing time of the retinal network. In the report by Cleland and Enroth-Cugell (1970), the shortest response latency reported was less than 30 msec. In that of Levick (1973), the visual latency determined for 10 ganglion cells at the highest luminance tested (32 times above threshold) was between 20 and 35 msec. These results were obtained from unidentified cell types.

Boltz et al. (1982) measured the latencies of cat retinal ganglion cells with different types of visual stimulation. They found that the shortest latencies were obtained when the stimulus had the same size as the center region of the receptive field. Such stimuli were also those that yielded the maximal discharge rate. With this type of stimulation, Boltz et al. observed that the latency of X cells was longer than that of Y cells by 10–15 msec. The absolute value of latency depended on the contrast between the stimulus and the background, but the latency difference between X and Y cells remained whatever the contrast. For high-contrast stimulation, the absolute values of latencies were between 31 and 44 msec for brisk sustained (X) cells and between 25 and 38 msec for brisk transient (Y) cells.

Three important conclusions can be drawn from these studies: (1) The visual latency of retinal ganglion cells is not less than 20 msec with the optimal stimulation at the highest contrast. This shows that the intraretinal processing is slow. For comparison, it is roughly twice the time needed for somatosensory information to reach the cortex (Welker et al., 1983). (2) The average latency of X cells is longer than that of Y cells when activated by their respective optimal stimuli. (3) Despite the difference of average latency, there is a scatter of response latency, such that there is an overlap of the latency distribution of X and Y cells.

Given these results, it is not unexpected that many of the studies in which visual response latencies have been measured in the cat LGN showed shorter latencies for Y cells compared to X cells. The latency difference that has been reported in different studies is between 6 and 20 msec (Orban et al., 1985; Sestokas and Lehmkühle, 1986; Saul and Humphrey, 1990; Harteit and Heggenund, 1992; Humphrey and Saul, 1992; Mastronarde, 1992). However, although average response latencies differ for X and Y cells, the distributions of latencies show a large overlap for the two cell types, similar to what has been found for the retina. In other word, the likelihood that a Y input reaches the cortex after an X input is not negligible.

The visual latencies in cat LGN are even more widespread when one considers the categories of lagged cells that have been reported in several studies. The range of latencies for such cells is between 100 and 1000 msec (Saul and Humphrey, 1990). In the report of Humphrey and Saul (1992), the X lagged cell
latency was on average 171 msec. Hartveit and Heggelund (1992) found a mean latency of 101 msec for Y lagged cells and 109 msec for X lagged cells. Such values are extremely long compared to those of cortical neurons (see Section 3.2).

Sestokas and Lehmkuhle (1986) established how visual latencies in the dorsal lateral geniculate nucleus of the cat are affected by the spatial frequency and contrast of the grating used for stimulation. The onset and peak latencies were found to vary as a function of spatial frequency for both X and Y cells. The shortest onset and peak latencies for X cells were obtained with a spatial frequency of 0.5–0.75 c/deg, and they increased at higher or lower spatial frequencies. For Y cells the shortest onset and peak latencies were obtained at around 0.17 c/deg (the lowest spatial frequency tested) and increased for higher spatial frequencies. The shortest onset latency obtained for X cells with the optimal spatial frequency was 60 msec on average, and about 50 msec for Y cells. Again, with their respective optimal stimuli, the onset latency of X cells was longer than that of Y cells by 10 msec. At low spatial frequency, the onset latency of Y cells was less than that of X cells. At a spatial frequency of about 0.75–1 c/deg, the latencies were similar. At high spatial frequency the onset latency of X cells was shorter than that of Y cells, but the peak latency of Y cells was never longer than that of X cells.

The latency also depends on the contrast of the stimulus. When tested with their optimal spatial frequency, both X and Y cells display their shortest latency at the highest contrast. The relationship between contrast and latency was monotonic. There was no range of contrast for which X-cell latency was shorter than Y latency.

In conclusion, there appears to be both a difference in the average value and an overlap of the latencies of X and Y cells in the cat retina and LGN.

3.1.2. Latencies in Primate Retina and LGN

The visual latencies have not been extensively investigated in the primate retina and LGN. One study performed in the retina (Gouras and Link, 1966) showed that, similar to what has been reported for the cat retina, the intraretinal processing is slow, visual latencies of ganglion cells being rarely less than 20 msec even at the highest contrast. Unfortunately, this study was conducted on a small sample of unidentified cell types, making it impossible to determine differences between different subtypes of primate ganglion cells.

Visual response latencies have been determined in the macaque lateral geniculate nucleus by Marrocco (1976), who found that the median latency of color-opponent LGN cells (presumably parvocellular neurons of type 1 and 2 according to the classification of Wiesel and Hubel, 1966; see also Dreher et al., 1976) is longer than that of non-color-opponent sustained cells (probably type 3 parvocellular cells) by 10 msec and longer by 18 msec than that of transient nonopponent cells (likely to correspond to magnocellular type 3 cells). Notice that this latency difference between visual latencies of magno- and parvocellular neurons is much longer than would be expected on the basis of axon conduction times only (see Section 2.1.4a). This suggests that, similar to what has been reported for the cat, most of the difference in latency originates from different
processing times within the retina. On the other hand, there appears to be an overlap in the distribution of latencies for the different cell types. A similar tendency, concerning both shorter latencies for magnocellular neurons and overlap of latencies for magnocellular and parvocellular neurons, has been reported by Blakemore and Vital-Durand (1986).

Interestingly, Marrocco (1976) observed a good correlation between the latency to visual stimulation and the latency of orthodromic responses after electrical stimulation of the optic chiasm. Neurons with the longest orthodromic latencies are probably those that are innervated by the smallest axons. Since there is a relationship between axon size and soma size (Wässle et al., 1975), retinal ganglion cells that drive these neurons are likely to be small and to have smaller dendritic arbors than other neurons of the same class. Therefore, they may not sample as many inputs as larger neurons, which would lead to slower temporal summation, leading to a longer latency.

The visual latency of neurons in the different subdivisions of the LGN has been established by Irvin et al. (1986) for the galago. The mean latencies they found were 54 msec for neurons of the magnocellular layers, 68 msec for those of the parvocellular layers, 79 msec for those of the koniocellular layers, and 82 msec for the neurons of the interlaminar zone. These last two groups of cells are likely to be the equivalent of the cat W cells.

These studies, therefore, show that a latency difference of 10–20 msec exists between the parvocellular and magnocellular layer neurons. This latency difference is similar to that obtained between X and Y cells in the cat lateral geniculate nucleus and retina. It remains to be established how the latency of primate parvo- and magnocellular neurons is affected by different visual stimuli.
3.1.3. Conclusion

In cat and probably in monkey, there is a difference in the intraretinal processing time, such that the ganglion cells that display the fastest conduction velocity (Y cells or magnocellular neurons) are also those that display the shortest visual latencies for a broad range of visual stimulus parameters.

This latency difference must also appear in the cortex. Indeed, with square stimuli of 1–2 deg width, a latency difference of 20 msec has been observed between layer 4Cα (recipient of magnocellular inputs) and layer 4Cβ neurons (recipient of parvocellular inputs) of macaque monkey area V1 (Nowak et al., 1995). This latency difference is larger than the latency difference that would be expected on the basis of difference in axonal conduction delay only (3–6 msec; see above). To establish the difference of functional delays that can be expected when the different inputs reach the cortex, the intraretinal processing time must be taken into account.

3.2. Visual Latencies in Cat Cortex

Visual response latencies in cat area 17 have been measured in a number of studies. Among the first studies to mention response latencies is that of Creutzfeldt and Ito (1968), who reported visual response latencies for spikes between 35 and 100 msec. These relatively long latencies to visual stimulation can be contrasted with the very short latencies obtained for EPSPs or spikes elicited by electrical stimulation of the subcortical afferent or spike-triggered averaging (see Section 2.1.4a). These long visual latencies have to be related to the long intraretinal processing time, combined with the integration time and the polysynaptic nature of some of the cortical responses.

The observation of long visual latencies in area 17 (compared with other sensory modalities) has been confirmed in a number of studies using extracellular recordings of action potentials. The peak response latencies reported by Ikeda and Wright (1975) averaged 47 msec (range 22–67 msec) for their transient cells and 57 msec (range 36–83 msec) for their sustained cells. Interestingly, Ikeda and Wright reported that simple and complex cells could not be distinguished on the basis of their visual latency (mean latencies 50.4 and 51.3 msec, respectively), another argument to be added to the numerous other arguments showing that, in cat primary visual cortex, simple and complex cells do not represent successive steps in a hierarchy of information processing, but most likely function as two parallel systems. Eschweiler and Rauschecker (1993) found that the onset latency (contrast 82%) ranges between 30 and 170 msec (median 65 msec). The peak latency is between 43 and 245 msec (median 85 msec). This indicates a large scatter of response latencies.

Best et al. (1986) measured response latencies in the different layers of cat area 17. For the whole of area 17, the latency distribution also showed a large scatter. Part of this scatter reflects different latencies in different cortical layers. Indeed, the average latency for the supragranular layers was longer by 20 msec than the average latency in layer 4. This large difference could reflect different
pathways for the transmission of X and Y inputs, or a long integration time for polysynaptic responses, or both.

The visual latencies in the LGN and different cortical areas (17, 18, 19, PMLS, and area 7) of cat cortex have been determined by Dinse and Krüger (1994). The interesting aspect of that study is that the latencies in different areas were determined under similar experimental conditions. This gives the possibility to determine the interarcual transmission delays with visual stimulation. As one progresses from the LGN to area 7, the mean latency is found to increase. In parallel to an increase in mean latency, there is also an increase in latency scatter. However, it is in area 18 that the shortest response latencies were obtained (40.4 msec on average). Those in area 17 were longer by 12 msec (52.7 msec). This latency difference is quite close to the latency difference reported between X (the main input to area 17) and Y (the main input to area 18) cells in cat retina and LGN (see Section 3.1.1).

It has been proposed that cat cortical areas can be organized in a hierarchy according to the pattern of reciprocal connections between them (Felleman and Van Essen, 1991). The hierarchy proposed by Felleman and Van Essen suggests that the order of activation should be 17-18-19-PMLS. The order revealed by the mean latencies obtained by Dinse and Krüger (1994) is clearly different: as already mentioned, the earliest latencies are observed in area 18, followed by area 17, then by area PMLS and area 19. Therefore, if visual cortical areas of the cat processed information in a serial manner, as suggested by the hierarchical arrangement, the functional order would be 18-17-PMLS-19. This suggests that the rank of a given area in the anatomical hierarchy does not necessarily correspond to that of the order of activation. Other examples will be found among the cortical areas of the primate visual system in Section 3.3.

3.3. Visual Latencies in Primate Cortex

3.3.1. Latencies in Primate Areas V1 and V2

Visual response latencies have been reported for area V1 in a number of studies in anesthetized and in awake, behaving monkeys. These studies also differed with respect to the visual stimuli activating the neurons and to the methods used to determine latency of the response. For these reasons, the mean or median response latencies can differ by as much as 40 msec from one study to another. At one of the two extreme ends, one can find the latencies of Bartlett and Doty (1974) and those of Maunsell and Gibson (1992), both using flashed visual stimuli (median for single units, 45 msec). At the other end, Raiguel et al. (1989) obtained a median latency of 85 msec with moving bar stimuli.

The latency distribution for some of the studies performed in area V1 is presented in Fig. 4 below, which summarizes the results of a number of latency studies in monkey visual cortex using flashing stimuli (with the exception of the work of Kawano et al., who used sudden onset of fast-moving stimuli). Even among studies using flashing stimuli in area V1 in awake animals (all the reports
except that of Nowak et al., 1995), there is a large variability in absolute latency. The median latency varies between 45 and 60 msec. Even larger intersubject variability is seen in the earliest and latest latencies; the 10% centile of the distribution varies between 27 and 48 msec and the 90% centile varies between 65 and 120 msec.

The latency scatter (difference between the 90% and the 10% centile) in area V1 varies between 30 and 80 msec in different reports. It is larger in anesthetized animals than in awake, behaving animals. This might be related to the influence of anesthesia, which can modify the timing of retinal to geniculate information transfer for some cells (Mukerjee and Kaplan, 1995).

The latency scatter in V1 can be caused by several factors: the first possibility is that the scatter observed in the LGN is reproduced in the cortex. This can be determined by comparing the latency scatters in layers 4Ca and 4CB (Nowak et al., 1995), the main recipients of the M and P channels, to those observed for presumed magnocellular and parvocellular neurons of the LGN (Marrocco, 1976); both studies were performed on anesthetized monkeys. It clearly appears, then, that the latency scatter observed for the presumed magnocellular neurons (37.7 msec) is similar to the one obtained in layer 4Ca (45 msec). The latency scatter observed in layer 4CB (60 msec) appears slightly larger than that obtained for color opponent cells (43 msec) and for nonopponent, sustained cells (47.5 msec). Notice also that part of the increase in latency scatter in layer 4 could be due to some polysynaptic transmission as well: neurons within these layers establish contacts with each other and receive in addition a dense innervation from layer 6 neurons (Levitt et al., 1996).

Another factor that can account for the latency scatter is the difference in response onset for neurons driven by the magnocellular pathway compared to those driven by the parvocellular pathway. Knowing that the difference in latency between M and P neurons in the LGN, it is of interest to investigate whether such a difference persists at the cortical level. This has been done by comparing the latencies of neurons in layers 4Ca and 4CB (Nowak et al., 1995). A value of 20 msec was found, in keeping with the latency difference between M and P neurons in the LGN (see Section 3.1.2). Combining this 20-msec latency difference to the latency scatter in layer 4CB gives already a total latency scatter of 80 msec for the whole of layer 4. Another possibility is that some neurons are driven by koniocellular cells, and, if they exist in the monkey, by lagged cells of the LGN (as mentioned above, these neurons have the longest latencies in the lateral geniculate nucleus).

The third factor that may account for the latency scatter observed in area V1 is the polysynaptic nature of some of the responses. Neurons at the top of the supragranular layers, for example, receive their inputs not directly from layer 4, but after an intervening relay in the lowest part of the supragranular layers (Levitt et al., 1996). Hence, these neurons are distant by two synapses from layer 4. Assuming an integration time between 5 and 10 msec and a negligible axonal delay, we have that these neurons would be activated only 10–20 msec after neurons in layer 4. The latency scatter now would be close to 100 msec. This is in fact the latency range obtained by Nowak et al. (1995) for ON responses (for OFF responses, a few neurons with exceptionally long latencies were observed). Final-
ly, another path at the origin of polysynaptic responses in area V1, leading to long visual latency, could be one that involves feedback connections from extrastriate cortex.

Beyond the thalamorecipient layers in V1, the M and P channels appear to converge in many neurons (Malpeli et al., 1981; Nealey and Maunsell, 1994; Levitt et al., 1996) and it is therefore difficult to interpret differences in latencies of V1 neurons as reflecting different types of thalamic inputs. In area V2, it is known that the thick cytochrome oxidase bands are mainly innervated by layer 4B (the M pathway; Livingstone and Hubel, 1987). This may explain why latencies of neurons in that subdivision are among the shortest recorded in V2 (Fig. 3; Munk et al., 1995). The pale cytochrome oxidase bands receive inputs from the interblobs in area V1 (Livingstone and Hubel, 1984). Short latencies were also observed in the pale bands, in keeping with the convergence of M and P inputs in the interblobs (Nealey and Maunsell, 1994; Yoshioka et al., 1994).

Although the thin cytochrome oxidase bands in V2 are driven by the cytochrome oxidase blobs in V1, which themselves receive converging inputs from all channels (Casagrande, 1994; Hendry and Yoshioka, 1994; Nealey and Maunsell, 1994; Yoshioka et al., 1994), their latencies are substantially longer than those recorded in pale and thick bands (Fig. 3; Munk et al., 1995). One possible reason for these latencies is that the blobs, in contrast to layer 4B and to the interblobs, do not receive direct inputs from layer 4Ca. Instead, they receive their magnocellular inputs indirectly, after a relay in layer 4B (Lachica et al., 1992; Yoshioka et al., 1994). This additional relay may introduce a temporal delay of several milliseconds due to the integration time.

The long latencies of neurons in thin cytochrome oxidase bands of V2 may also be due to the predominance of nonoriented color-selective cells in that subdivision (although this is a matter of debate; see Salin and Bullier, 1995, for review). Since cytochrome oxidase blobs in V1 also contain a majority of nonoriented color-coded cells (Livingstone and Hubel, 1984; Yoshioka and Dow, 1996), it is possible that neurons in that subdivision are mainly under the influence of the parvocellular system, hence responding with longer visual latencies. Such long latencies would also be consistent with the projection of the later activated K channel in the cytochrome oxidase blobs (Casagrande, 1994, Hendry and Yoshioka, 1994). A third explanation is that the blobs—thin bands channel may have response latencies delayed by strong inhibitory blockade from the M channel (L. G. Nowak, M. H. J. Munk, P. Girard, and J. Bullier, unpublished; see also Sato et al., 1994).

The thick cytochrome oxidase bands in V2 project to area MT (Shipp and Zeki, 1985). The short latencies in this subdivision suggest that MT neurons are activated rapidly, a conclusion in keeping with published work (see Section 3.3.3) and with the very fast conducting axons between V1 and MT (see Section 2.1.4c). Area V4, the other major output of area V2, receives some short latency activity through the pale bands and long latency activity through the thick bands. Thin and pale bands might process different aspects (color in thin bands and orientation of borders in pale bands) of the visual scene. There is evidence that a segregation between color and orientation modules also exists in V4 (Tanaka et al., 1986; Yoshioka and Dow, 1996). This suggests that in V4, some modules are
activated earlier than others. A possible interpretation of such a segregation will be given in Section 4.3.

3.3.2. Transmission of Visual Signals May Rely on One Neuronal Layer per Area

In several studies the visual latencies have been determined with the same method in two different areas. Under these conditions it is possible to determine the interarea information transfer time. The latency difference that has been reported between areas V1 and V2 is 11 msec in Raiguel et al. (1989) and 10 msec for single-unit recordings in Nowak et al. (1995). The latency difference that has been reported between areas V1 and MT, which are reciprocally and mono-synaptically connected, is 9 msec in the study by Raiguel et al. (1989) and 11 msec in the results of Maunsell (1987). Hence, for cortical areas that are directly connected, the latency difference appears to be close to 10 msec (see also Thorpe and Imbert, 1989).

Maunsell (1987) also reported a 22-msec latency difference between the transient cells in area V1 and neurons in V4. Areas V1 and V4 are connected through a relay in V2. The 22-msec difference between the latencies in V1 and V4 corresponds approximately to twice the interarea transfer time of 10 msec. Similar reasoning can be applied to the results obtained by Vogels and Orban (1990, 1994), who reported visual latencies in areas V1 and TE of macaque monkeys with the same experimental paradigm (in one monkey, recording was done in both areas). The median latency in TE was found to be larger than the median latency in V1 by 40 msec. This 40 msec corresponds exactly to what can be expected if the latency difference between two directly connected areas is 10 msec, because there are four relays between V1 and TE (V2, V4, TEO, TE).

As already mentioned, an interareal delay of 10 msec corresponds roughly to the sum of the axonal conduction delay between two cortical areas and the integration time. To explain a 22-msec latency difference between areas V1 and V4, or a 40-msec latency difference between areas V1 and TE, it must be assumed that there is only one synaptic layer per cortical area and only one action potential per neuron (Thorpe and Imbert, 1989). This is unexpected since neurons in layer 4 that receive information from a lower stage usually do not possess axons projecting in other cortical areas. However, the basal dendrites of neurons in lower layer 3 can enter in layer 4. In addition, except in monkey area V1, neurons of the infragranular layer could be contacted on their apical dendrites as they pass through layer 4. Furthermore, it is known that feedforward axons often terminate not only in layer 4, but also in the lower part of layer 3 and in the infragranular layers. In other words, the feedforward inputs can directly contact output neurons. For example, terminals of feedforward inputs from area V1 onto area MT are found in layer 4 and the lower portion of layer 3 (Rockland, 1989). It is also in the lower portion of layer 3 that neurons projecting from area MT to the frontal eye field are concentrated (Schall et al., 1995).

This 10 msec per area rule is of course valid only on average for the short-latency neurons. For the activation of long-latency neurons, there may be some paths which involve several synaptic stages in a given area, combined with information transferred through feedback connections.
3.3.3. The Fast and the Slow Brain

Figure 4 presents a summary of a number of reports of latencies to flashed visual stimuli recorded in different cortical areas of awake monkeys. Note that there are three exceptions: the studies of Nowak et al. and Munk et al. were done in anesthetized animals and the study by Kawano et al. (1994) was done with stimuli consisting of a large array of dots suddenly set in motion at high speed.

What is immediately evident in Fig. 4 is the large scatter of latencies. This scatter is such that latencies in adjacent cortical areas overlap substantially. There are two major groups of areas in Fig. 4: areas with median latencies between 40 and 80 msec, which constitute what we call the fast brain, and areas with median latencies between 100 and 130 msec, which make up what we call the slow brain.

The fast brain corresponds in Fig. 4 to areas V1, V2, MST, and the frontal eye field (FEF). In addition, we saw earlier that areas MT and V4 can also be included in this group because their latencies are displaced with respect to V1 by only 10 or 20 msec. Short latencies of MT and MST neurons to rapidly moving stimuli have also been reported by Orban's group (Lagae et al., 1994). Given the short latencies in areas MT and MST, it is somewhat surprising that latencies in
areas LIP and 7a are not shorter. In the case of LIP (Barash et al., 1991), the methods used to measure latencies tend to overestimate them by 20 msec.

The fast brain is dedicated to an early processing of visual information in the case of V1 and occipital prestriate cortex and to a processing of information related to visuomotor integration in the case of the parietal cortex and frontal eye field. An early activation of parietal cortex is in keeping with its major input from the M channel in the parietal cortex via area MT. Most cortical areas involved in this pathway are heavily myelinated (areas MT, MST, VIP), suggesting that speed of transfer is an important parameter in this cortical region. We also know that, at least in the case of the V1 to MT pathway, axons are rapidly conducting. It is likely that rapidly conducting axons also link areas MT, MST, VIP, and FEF. The presence of rapidly conducting axons between MT and FEF would explain the early activation of FEF neurons by visual stimuli (Bushnell et al., 1991; Bichot et al., 1996; Thompson et al., 1996), despite its long distance from area MT, which probably constitutes the major relay of visual information toward frontal cortex.

Another potential explanation for the early latencies in areas of the fast brain is the large number of bypass connections in the dorsal occipitoparietal stream. Examples of such bypasses are found in the direct projections from V1 to V3 and MT, which bypass a relay through V2 (Felleman and Van Essen, 1991). Another example is given by the direct projection from area V1 to area PO, which also receives indirect inputs through area MT (Colby et al., 1988).

The slow brain groups most areas of the temporal lobe and some areas of the frontal cortex located rostral to the frontal eye field. The median latencies are in the 100–150-msec range and many neurons respond as late as 200 msec after the presentation of the stimulus. It is unclear at the moment which factors are most important in determining the long latencies of neurons of the slow brain. One could be that axons in the temporal cortex are slowly conducting. In addition, the presence of bypass connections appears to be less frequent in the ventral occipitotemporal stream, which seems to be organized in a more serial fashion. For example, to reach TE from V1, information has to relay in areas V2, V4, and TEO (although a few bypass connections also exist; Nakamura et al., 1993).

3.3.4. Latencies and the Hierarchical Model of Cortical Organization

As was the case for the cat visual cortex, latencies to visual stimulation in monkey are not ordered as expected from the hierarchical organization based on anatomical grounds (Felleman and Van Essen, 1991). The FEF region, which is at a position close to the top of the hierarchy, is activated simultaneously with neurons of areas V1, which belong to the first level (Bichot et al., 1996; Thompson et al., 1996). The latency difference between neurons in these two areas does not leave much time for whatever processing is supposed to occur at each stage of the hierarchy. Another example that was cited earlier (Bullier and Nowak, 1995) is that of areas V2, MT, and V4. The reasoning goes like this: The median latency is longer by 10 msec in area V2 than in V1 when the latencies are determined under similar experimental conditions. A 10-msec latency differ-
ence has also been reported between area V1 and area MT (Maunsell, 1987; Raiguel et al., 1989). In other words, if one uses latency to order cortical areas, then area MT should leave the fifth level it occupies in the anatomical hierarchy, to join area V2 at the second level (Raiguel et al., 1989). In the anatomical hierarchy, areas MT and V4 occupy the same level. However, the latency difference between areas V1 and V4 is 12 msec longer than that between area V1 and area MT (Maunsell, 1987; Raiguel et al., 1989). In terms of latency, therefore, it looks as if areas MT and V4 occupy two different hierarchical levels.

Thus, if one assumes that latencies to visual stimuli provide a reasonable estimate of the order of activation of cortical areas, it appears that this order does not follow the one suggested by the anatomical hierarchy of cortical areas. The major factor that influences the speed of activation of cortical areas appears to be whether or not a cortical area belongs to the dorsal stream that is almost exclusively activated by the M channel.

4. Conclusion: How Does Cortex Process Information

So Rapidly?

4.1. What Connects Together Fires Together

Latencies in adjacent areas differ by 10 msec on the average. The latency distributions in different areas show dispersions much larger than 10 msec (Fig. 4). In addition, neurons of the visual cortex give responses lasting usually more than 100 msec even to the shortest stimuli (Rolls and Tovee, 1994; Kovacs et al., 1995). Because of the large latency scatter and the sustained responses of visual neurons, neurons in adjacent cortical areas emit spikes during the same post-stimulus period. In the case of areas V1 and V2, this overlap of latency leads one to conclude that much of the processing in these two areas is done in parallel and not in a sequential fashion, as suggested by the results of reversible inactivation of V1 (Schiller and Malpeli, 1977a; Girard and Bullier, 1989). Given the simultaneity of firing of neurons, it is no surprise that many pairs of neurons in adjacent areas give synchronized responses (J. I. Nelson et al., 1992; Nowak et al., 1994). Neighboring areas are also often strongly linked with reciprocal connections (Felleman and Van Essen, 1991; Young, 1992). It is tempting to suggest that the synchronization of firing of neurons in adjacent areas is due to the dense network of reciprocal connections between them.

Beyond a certain distance on the cortical surface, connections become rare or absent. This is particularly the case in the temporal lobe, in which connections appear to link a small set of neighboring areas (Desimone et al., 1980). It is interesting to remark that neurons in V1 and V2 which are not directly connected to areas of the temporal cortex also display much shorter latencies. On the contrary, despite their large separation on the cortical surface, neurons of areas MT/MST and FEF are interconnected and display similar latencies to visual stimulation. In other words, paraphrasing an earlier statement (Löwel and Singer, 1992), “what connects together fires together.”
4.2. How Rapidly Is Response Selectivity Established?

Given the rapid activation of neurons in most cortical areas, an interesting question, discussed earlier by Thorpe and Imbert (1989), is to determine whether the early part of the response is already selective to different stimuli or whether the response is first broadly tuned and selectivity is acquired progressively as the response develops.

Face-selective cells in the STS have visual latencies as short as 80 msec (Perrett et al., 1982). Response selectivity for different views of the face and head is already present only 5 msec after the onset of the response, even in the case of rapidly responding cells (Oram and Perrett, 1992). A fast acquisition of response selectivity has also been observed by Celebrini et al. (1993), who found that orientation selectivity in area V1 is already present at the onset of the response of even the most rapidly responding neurons. In their intracellular study of V1 neurons, Volgushev et al. (1995) showed that the subthreshold synaptic response is initially broadly tuned to orientation, but acquire its definitive tuning before action potentials are fired.

The very fast acquisition of response selectivity has an important consequence for the neurobiological basis of receptive field construction: at least for the rapidly responding neurons, the receptive field properties must be determined by the pattern of feedforward inputs (Thorpe and Imbert, 1989). If, in addition, inhibition is required in the sculpturing of response selectivity (Sillito et al., 1980; Douglas and Martin, 1991), this implies that inhibitory neurons must have very short visual latencies and a very rapid integration time. Whether this is truly the case remains to be established for the visual cortex, but results consistent with this view have been obtained in the barrel field of rat somatosensory cortex (Welker et al., 1993).

What happens to selectivity beyond the early response of the first activated neurons? Using an analysis based on information theory, Tovee et al. (1993) observed that the quantity of information reaches a high value very early after the onset of the response. A fast rise in information content of neuronal response has also been reported for areas V1, V2, and V4 by McClurkin and Optican (1996). However, the information content has also been found to increase during later parts of the response (Richmond and Optican, 1990; Heller et al., 1995; McClurkin and Optican, 1996). Similarly, if the results obtained by Oram and Perrett (1992) showed a very fast acquisition of response selectivity, it is only 80 msec after response onset that the strongest difference between responses is reached. Therefore, it looks as if there is a very rapid acquisition of receptive field properties, followed by a further refinement of the selectivity. Other evidence that there are properties that develop only after the response onset can be found in the time course of responses of V1 neurons to pop-out stimuli (Knierim and Van Essen, 1992; Lamme, 1995).

The arguments presented above indicate that the rapid acquisition of response selectivity relies only on the feedforward transmission of information across cortical areas. However, the further refinement that follows that initial response could be related to the overlap of visual response latencies between different cortical areas that could allow bidirectional interactions involving the feedback connections.
4.3. A Functional Model of the Slow Brain

Seeing is not just detecting the orientation of a bar stimulus. There are many situations in which ambiguities must be resolved rapidly in order to identify the different elements of the visual scene. We present here a model of the part of the visual system involved in pattern recognition, the slow brain. The model contains several of the features that were stressed above: (1) The wave of incoming information reaching the cortex is not synchronous; it is composed of at least three and possibly four (M, P, K, and lagged inputs) waves that are activated up to 200 msec after the stimulus onset. (2) Neurons function mainly as coincidence detectors and signal the coincidence of properly timed waves of activity. (3) Feedback connections are used to carry different hypotheses concerning the elements of the visual scene to be tested against the incoming waves of afferent activity (Ullman, 1995).

The model is an elaboration of that presented by Ullman (1995) to explain pattern recognition with a bidirectional flow of information. The idea is that the M channel acts as an ignition device to activate representations stored in the temporal cortex. The poor temporal resolution and the lack of color selectivity of the M channel leads to a sketchy view of the scene. Although limited in precision, this is the first information that reaches the temporal cortex. Different potential interpretations of an object lead to activation of different neurons that send activity down through feedback connections. As proposed in the model of Ullman (1995), these waves of descending activity meet the subsequent waves of feedforward activity (P, K, lagged), giving rise either to a reinforcement or to an extinction by a mechanism of coincidence detection between properly timed feedforward and feedback activity waves. The proper interpretation of the image is the one that leads to reinforced activity at all levels of the visual cortex.

If we integrate such a model of the slow brain with the existence of the fast brain, we arrive at the conclusion that analysis of a visual scene proceeds at different speeds depending on which aspects are privileged. A very fast analysis system appears to be devoted to the processing of localization and visuomotor interactions. It corresponds to the fast brain, is under the influence of the M system, and may function mostly in the feedforward mode. A relatively fast processing system, mostly under the influence of the M system in the slow brain, achieves a coarse analysis of broad features in the feedforward mode. Finally, a slow processing system deals with the fine detail analysis and the precise recognition using P and K systems and feedback connections.

5. References


Bullier, J., McCourt, M. E., and Henry, G. H., 1988, Physiological studies on the feedback connection to the striate cortex from cortical areas 18 and 19 of the cat, Exp. Brain Res. 70:90–98.


CHAPTER 5


Nowak, L. G., and Bullier, J., 1978b, Axons, but not cell bodies, are activated by electrical stimulation in cortical gray matter: II. Evidence from selective inactivation of cell bodies and axon initial segments, Exp Brain Res., in press.


Nowak, L. G., James, A. C., and Bullier, J., 1997, Cortico-cortical connections between visual areas 17 and 18a of the rat studied in vitro: Spatial and temporal organization of functional synaptic responses, Exp Brain Res. 117:283–305.


