ELSEVIER

Progress in Neurobiology 84 (2008) 405–439

progress in Neurobiology

www.elsevier.com/locate/pneurobio

Time course of visual perception: Coarse-to-fine processing and beyond

Jay Hegdé *

Department of Psychology, 75 East River Parkway, University of Minnesota, Minneapolis, MN 55455, USA

Received 23 February 2007; received in revised form 17 July 2007; accepted 19 September 2007

Abstract

Our perception of a visual scene changes rapidly in time, even when the scene itself does not. It is increasingly clear that understanding how the visual percept changes in time is crucial to understanding how we see. We are still far from fully understanding the temporal changes in the visual percept and the neural mechanisms that underlie it. But recently, many disparate lines of evidence are beginning to converge to produce a complex but fuzzy picture of visual temporal dynamics. It is clear, largely from psychophysical studies in humans, that one can get the 'gist' of complex visual scenes within about 150 ms after the stimulus onset, even when the stimulus itself is presented as briefly as 10 ms or so. It generally takes longer processing, if not longer stimulus presentation, to identify individual objects. It may take even longer for a fuller semantic understanding, or awareness, of the scene to emerge and be encoded in short-term memory. Microelectrode recording studies in monkeys, along with neuroimaging studies mostly in humans, have elucidated many important temporal dynamic phenomena at the level of individual neurons and neuronal populations. Many of the temporal changes at the perceptual and the neural levels can be captured by the multifaceted and somewhat ambiguous concept of coarse-to-fine processing, although it is clear that not all temporal changes can be characterized this way. A more comprehensive, albeit unproven, alternative framework for understanding visual temporal dynamics is to view it as a sequential, Bayesian decision-making process. At each step, the visual system infers the likely nature visual scene by jointly evaluating the available processed image information and prior knowledge about the scene, including prior inferences. Whether the processing proceeds in a coarse-to-fine fashion depends largely on whether the underlying computations are hierarchical or not. Characterizing these inferential steps from the computational, perceptual and neural standpoints will be a key part of future work in this emerging field.

Keywords: Adaptive filtering; Awareness; Bayesian inference; Decorrelation; Fine-to-coarse; Feedback; Feed-forward; Global-to-local; Hierarchical coding; Hypothesis testing; Lateralization; Masking; Microgenesis; Natural vision; Perceptual learning; Plasticity; Priming; Recurrent/reentrant processing; Sequential decision-making

Contents

Abbreviations: BOLD, blood oxygenation level-dependent; cpd, cycles per degree; CRF, classical receptive field; DCM, dynamic causal modeling; DF, dorsal foci (human); EEG, electroencephalogram; ERMF, event-related magnetic field; ERP, event-related potential; FFA, fusiform face area (human); fMRI, functional magnetic resonance imaging; IOG, inferior occipital gyrus; IT, inferior temporal/inferotemporal visual area (macaque); ITC, inferior temporal/inferotemporal cortex (human); LGN, lateral geniculate nucleus; LOC, lateral occipital complex (human); MFC, medial frontal cortex; MT, middle temporal area; nCRF, non-classical receptive field; OFC, orbitofrontal cortex; PFC, prefrontal cortex; PHC, parahippocampal cortex; RHT, reverse hierarchy theory; TMS, transcranial magnetic stimulation; SOA, stimulus onset asynchrony; TP, temporo-occipital junction; V1, V2, V4 or V5, visual area 1, 2, 4 or 5; VEF, visually evoked potential.

* Tel.: +1 612 626 3577; fax: +1 612 626 2079.
E-mail address: hegde@umn.edu.

0301-0082/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pneurobio.2007.09.001

1. Introduction: the dynamic nature of visual perception

1.1. A simple demonstration

We all know that visual perception is not instantaneous—it takes finite time. Most of us are also aware, at least implicitly, of another temporal complexity of visual perception. It is that our perception of a given image changes depending on how long we view it, even when the stimulus itself remains unchanged. Usually, one can only get a gist of a visual scene from a fleeting glance. Perceiving the finer details of the visual scene generally requires longer scrutiny. There tends to be more to the view than

meets the eye at first; we all remember having had to do a 'double-take' of one visual scene or another.

One can easily get some basic intuitions about the temporal dynamics of visual perception, and the problems of rigorously studying these phenomena, using an informal experiment. The key is to pick a picture of an unfamiliar scene, view it as briefly as possible, look away, and describe the scene with as much rigor and detail as possible. Repeat the procedure until you feel you have seen all there is to see in the picture. You can use Fig. 1 for this purpose. Is there an animal in this picture? If so, what kind? How many looks did it take you to determine the answer to either question? How would you go about characterizing the temporal changes (or perhaps the lack thereof) in your percept?

In exercises such as these, subjects typically report that their understanding of the scene gets more detailed or specific over successive viewings. In case of Fig. 1, for instance, one initially tends to recognize broad categories of objects, such as ''brush'', ''animal'', bird''. But more specific object identification, such as ''fox'', ''kitten'', ''Blue Jay'', or ''Thuja plant'' takes longer. One's certainty about the various aspects of the scene also tends to improve over multiple viewings.

Recently, rigorous laboratory studies of such visual temporal dynamic phenomena have begun to yield some answers about how our visual percept changes in time, and what neuronal mechanisms may underlie these changes. This article will present an overview of these findings.

1.2. Scope and organization of the review

The aim of this review is to provide a broad outline, and not an exhaustive account, of visual temporal dynamics. This will entail an admittedly subjective selection of topics relevant to understanding this field. In the interest of cogency, we will also sidestep many important ancillary topics, including issues such as how the visual awareness evolves over time, a process often referred to as *microgenesis* (for reviews, see Koch and Crick, 2004; Öğmen and Breitmeyer, 2006).

Fig. 1. The degree of scene understanding varies with the duration of viewing. Is there an animal in the picture? What kind of an animal is it? How many other animals can you identify in the picture? Which aspects of the scene did you recognize instantly? Which items took some time to see?

As alluded to above, most of the psychophysical studies on this topic have been carried out in humans, and almost all of the relevant single-unit studies have been carried out in monkeys. Whole brain imaging studies of visual temporal dynamics have been mostly carried out in humans. We will treat these three sets of studies largely separately (in Sections 2–4, respectively), both because each body of research stands on its own, and because it is often unclear how they relate to each other.

Most studies of temporal changes in the visual percept focus on shape perception, *i.e.*, recognition of objects and scenes. The temporal dynamics of space perception and visually guided action have been studied less, presumably because these processes must take place in 'real time', and are believed to be mediated by a different neural pathway than that for object recognition (see Figs. 3 and 6). While this presumption itself is debatable (see Section 3.9), the corresponding bias in the literature is naturally reflected in this review.

1.2.1. Trying to make sense of temporal dynamics: frustrations and rewards

At present, we seem to know just enough about the various temporal dynamic phenomena to discern that there is no single unifying explanation yet for all of them, but not enough to know what an eventual explanation will look like. Indeed, there is no particular reason to expect that there is a single explanation for all the disparate temporal dynamic phenomena, any more than there is to expect that there is one for vision at large. Thus, the field of visual temporal dynamics currently amounts to a bewildering collection of results that are loosely interconnected as best, and this review will inevitably reflect this. However, while this uncertainty may make this review a challenging reading, it is also precisely what makes this field so exciting.

A recurring theme in visual temporal dynamic literature is that visual processing progresses in a global-to-local, or coarse-to-fine, fashion. While the scientific usefulness of this concept is unclear at best (see Section 6), it is nonetheless serves as a useful point of departure, whereby temporal dynamic phenomena can be understood as variations of, or deviations from, this theme.

1.2.2. Temporal dynamics as serial decision-making

As noted in Abstract, a more promising approach to understanding temporal dynamics is to view it as reflecting a series of probabilistic, Bayesian inferences about the nature of the visual scene and about possible courses of action. Although the underlying concepts are mathematically rigorous (which is one of the strengths of this framework), they can be intuitively understood (for more rigorous accounts, see Kersten et al., 2004; Doya et al., 2006; Ma et al., 2006; Yuille and Kersten, 2006). Briefly, this framework posits that the visual system infers the likely nature of the image and optimal course of action by jointly evaluating available information about the image and prior information about the visual world in a probabilistic fashion (using the Bayes' law of conditional probability). Thus, in this framework, vision is an inferential process, and visual temporal dynamics is simply the temporal dynamics of the inferential process—the visual system tests the

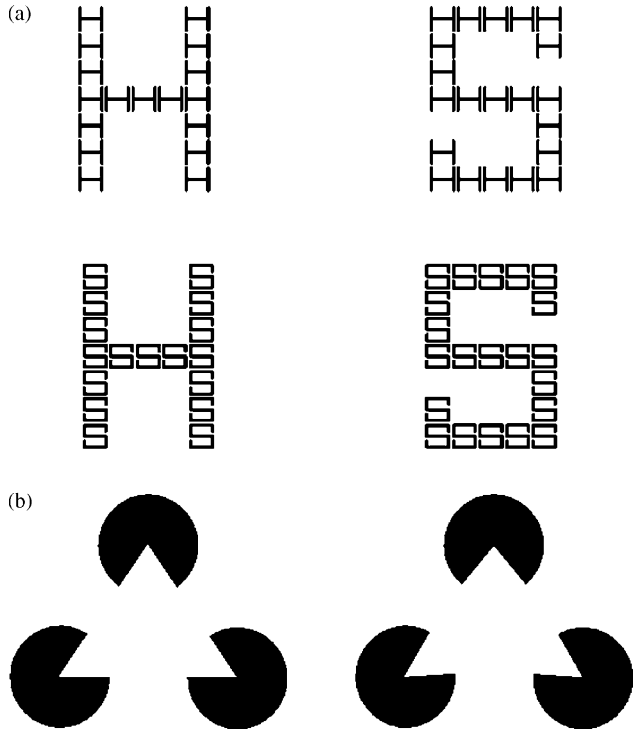


Fig. 2. (a) Stimuli used by Navon (1977). (b) Stimuli used by Reynolds (1981). See text for details.

various hypotheses as they arise and as information becomes available against which to test them.

The effectiveness of the Bayesian framework in explaining temporal dynamic phenomena remains largely untested and unproven, although there have been some promising starts (see Ma et al., 2006, and the references therein). In the realm of vision in general, however, the effectiveness of the framework has been well tested and well established (see Kersten et al., 2004; Doya et al., 2006).

Note that this framework subsumes and extends the coarse-to-fine processing framework: to the extent that the underlying computations are hierarchical (*e.g.*, the object of interest in Fig. 1 is an animal, a fox, and a particular fox named Biel), the processing may indeed proceed in a coarse-to-fine manner. To the extent that the underlying computations are non-hierarchical, such as those involving parameter estimation (*e.g.*, judging the time of day in the scene shown in Fig. 1), the inferences may not proceed in that manner. Thus, understanding the temporal dynamics of a given visual phenomenon becomes a matter of understanding the nature of the underlying decisions, how the information that supports these decisions accumulates (or just changes) over time, and how the brain arrives at a decision given the information available at the time.

Of course, when it comes to temporal dynamics, the above Bayesian framework is only a pedagogical tool as yet, since the relevant inferential steps have been elucidated for none of temporal dynamic phenomena. We will examine these issues in greater detail in Section 6. But in the meantime, it will be helpful to the reader to use this framework to help conceptualize various results described below.

1.3. Some complexities of natural vision

Under natural conditions, the retinal image is often extraordinarily complex, and tends to change in complex ways over time. Among other things, this is because the eyes, the head, the observer, the objects in the visual scene, and the source/s of illumination all can move relative to each other. In addition to these stimulus-driven, or 'bottom–up' factors, many cognitive, 'top–down' factors also tend to change dynamically over time. Quite understandably, most temporal dynamic studies so far have avoided these complexities, instead using simple, typically static, stimuli viewed without eye movements (*i.e.*, with the eyes fixating a specified target).

Moreover, most temporal dynamic studies focus on the changes that occur over a relatively brief time period—typically a few hundred milliseconds. The choice of this time range is not only practical, but also happens to be neurobiologically meaningful. Under natural viewing conditions, humans move their eyes once every 300 ms on average (mode, 230 ms; range, $<$50 ms to $>$1000 ms; Henderson and Hollingworth, 1998). These parameters are largely similar for monkeys (Wilson and Goldman-Rakic, 1994). In between a pair of eye movements, *i.e.*, during each fixation, the retinal image is largely stable (but not absolutely so, see Henderson and Hollingworth, 1998).

Thus, visual processing in laboratory experiments using a stimulus duration of a few hundred milliseconds is roughly comparable to the visual processing during a single fixation episode during natural viewing (see, *e.g.*, DiCarlo and Maunsell, 2000). However, whether and to what extent these short-term changes in visual processing are related to the adaptive changes that occur over much longer periods of time such as hours, days and even years is unclear (see Fahle and Poggio, 2002; Hochstein and Ahissar, 2002; Sharpee et al., 2006). We will therefore sidestep this larger question (see Box 4), and focus more narrowly on the short-term changes.

2. Characterizing the temporal changes at the perceptual level

Although the dynamic nature of visual perception seems obvious with the benefit of hindsight, it has historically received scant attention. Gestalt psychologists of the early 20th Century seem to have thought that, as Navon (1977) put it, the visual system is a "perfectly elastic device that can swallow and digest all visual information at once, no matter how rich it is" (p. 353). Many other contemporary psychologists did, however, note that the more one looks at an image, the more one gets from it (*e.g.*, Helson and Fehrer, 1932; Bridgen, 1933). But much of what we know substantively about the temporal dynamics of visual perception has been learned in the last three decades or so.

2.1. Insights from early psychophysical studies: seeing the forest before trees

One of the earliest studies to clearly illustrate the temporal changes in visual perception was that by Navon (1977). He

explored whether the global or the local structure of a visual stimulus has greater precedence in visual perception. One of his stimuli was a large H character built using many smaller H characters as building blocks. He similarly created a large H with smaller S's, and large S's with smaller H's or S's (Fig. 2a). He found that subjects perceived a given larger letter much more readily than its smaller building blocks. The identity of the smaller letters had no effect on the recognition of the larger ones. On the other hand, the identity of the large letter did substantially affect the recognition of the smaller letter. For instance, smaller letter H was easier to recognize if it was part of a large H rather than of a large S. In a related experiment, Navon asked subjects to report whether similar composite stimuli created using geometric shapes (*e.g.*, triangles and squares) presented in pairs were the same or different. The subjects performed much better when the global shapes were the same, regardless of the local shapes, than when the local shapes were.

These findings indicate that global object features take precedence over local features in visual processing ('forest before trees'). The reasons why are not entirely clear. It may be that this phenomenon simply reflects global-to-local processing, whereby the visual system processes the global image feature before the local ones. Or it may be that it takes longer to deploy attention (or, equivalently, to make saccades) to local features, or because zooming attention *vs.* shifting it have different costs and benefits (Stoffer, 1993; Torralba et al., 2006). It is also worth noting that whether 'global' or 'local' shape takes precedence depends on the relative size of the relevant image elements. For instance, it is easy to imagine that as the overall retinal size of the stimulus increases, it becomes easier to recognize the local shapes than the global shape.

Reynolds (1981) used a somewhat more complex stimulus that consisted of three pacmen that induced the percept of the classic Kanizsa illusory triangle. Depending on the condition, the pacmen were such that the resulting illusory triangle appeared straight or curved (Fig. 2b). The pacmen were each presented for 50 ms. After a stimulus onset asynchrony (SOA, or the delay relative to the onset of the pacmen) that varied systematically from 50 to 150 ms, a mask was presented so as to essentially prevent further processing of the pacmen stimulus (for more on masking, see Section 2.4).

Reynolds found that if the mask immediately followed the pacmen (*i.e.*, had an SOA of 50 ms), a majority of the subjects saw the pacmen, but failed to see the triangle. A few subjects that did report seeing a triangle often mistook a straight triangle for a curved one, or vice versa. When the mask had an SOA of 100–125 ms, all observers reported seeing a triangle (and, of course, the pacmen), and were generally more accurate about the curvature of the triangle.

These and additional experiments with more complex stimuli illustrated two related characteristics of visual temporal dynamics: first, even when the stimulus does not change, its perception grows more elaborate over time. Second, the subjects' inferences about the scene also grow more accurate over time. With striking prescience, Reynolds cast his results in terms of a sequential inference model, in which the visual system tests hypotheses of increasing complexity against accumulating image information. With longer processing time, the amount of the processed image information increases, against which progressively more complex hypotheses about the nature of the image can be tested, and conclusions can be drawn with greater certainty. This explains why the subjects' percepts got both more elaborate and more accurate over time.

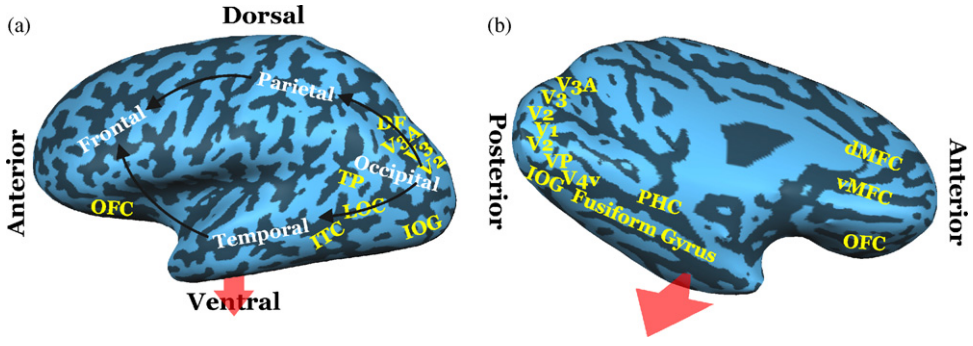


Fig. 3. The two major visual pathways in the human brain. The cerebral hemisphere is digitally inflated to reveal the entire cortical surface, including the sulci (folds), shown here are dark regions. The blue regions are the gyri (cortical bulges). Subcortical visual structures, including the eyes and the LGN, are not shown. (a) Lateral view of the inflated hemisphere. (b) Ventro-medial view. In either panel, the *translucent red arrow* lies in the sagittal plane and points in the due ventral direction. The *black arrows* denote the main two pathways though which feed-forward visual information travels from the occipital lobe to the frontal lobe. Feed-forward visual processing refers to computations where outputs from a given stage of processing travel to next stage/s of processing, progressively further away from the eyes. In the ventral (or temporal, 'what?') pathway, which is believed to mediate shape perception, the feed-forward information travels from the occipital lobe to the frontal lobe via the temporal lobe. In the dorsal (or parietal, 'where?') pathway, which is believed to mediate space perception, the feed-forward information travels from the occipital lobe to the frontal lobe via the parietal lobe. Feedback and lateral connections, which collectively mediate recurrent processing, are not shown for either pathway. Approximate locations of some key brain regions referred to in the text are denoted in *yellow*. Visual cortical areas other than V1 (visual area V1 or striate cortex) are collectively known as the extrastriate visual areas (not labeled). Prefrontal cortex (or PFC; not labeled) is the anterior part of the frontal lobe. The functional anatomy of human visual pathways is similar to that of the macaque brain (Fig. 6). The dorsal and ventral pathways may be much less functionally distinct than generally thought (see Section 3.9; also see Merigan and Maunsell, 1993). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Thus, Reynolds' model captured the qualitative essence of the more quantitative present-day framework of sequential inference, and illustrated the usefulness of such models as an explanatory and research tool. Unfortunately, as will be apparent from the remainder this review, few subsequent studies have utilized this framework (also see Box 4).

The aforementioned studies also serve to highlight three subtler, but larger, issues in the field of visual temporal dynamics. The first is the fact that the *processing duration* (usually specified as SOA) tends to be operationally far more important than the *stimulus duration*. In general, the visual stimulus need only be presented briefly (for a few tens of milliseconds or less) as long as the system is allowed more time to process the retinal input before the processing is interrupted, *e.g.*, by a mask (Di Lollo and Wilson, 1978; Coltheart, 1980). For this reason, temporal dynamic studies typically explore the temporal changes in the percept as a function of the processing duration, while holding the stimulus duration constant.

The second issue has to do with the distinction between how the reported percept varies as a function of the processing duration *vs.* how the percept itself varies over the course of a given processing duration. For instance, in the pacmen experiment outlined above, subjects report perceiving only the pacmen but not the illusory triangle for processing durations of up to 100 ms, but report perceiving both for processing durations of 150 ms or longer. Does this mean that for processing durations of $\geq$150 ms, the subjects first perceive just the pacmen and then both the pacmen and the illusory triangle? In other words, does change in the percept as a function of the processing duration also describe the change in the percept as a function of time? The answer, which hinges on whether the temporal changes in the visual percept follows the same deterministic path, and whether intermediate steps in processing are consciously perceived, is largely unknown. Nonetheless, these questions serve to highlight the potential distinction between the two scenarios (Efron, 1967; Koch and Crick, 2004).

The third issue is that the temporal changes in the visual percept are intimately associated with many other behavioral parameters, including task performance, accuracy, and the certainty in the percept. These parameters may also change over time even when the percept itself does not. Out of practical necessity, most studies typically measure only one of these outcomes at a time, but this necessarily underemphasizes the complexities (*e.g.*, non-linearities) in visual perception. Also, since the likelihood of a given behavioral outcome is rarely the same from one subject to the next or one session to the next, it makes sense to express the metric in explicitly probabilistic terms, as implied by the aforementioned probabilistic framework. But few temporal dynamic studies currently do this. This failure to explicitly account for the multivariate, probabilistic nature of the response metrics is among the reasons why although our current understanding of visual temporal dynamics has considerable explanatory power, it has little predictive power.

Altogether, the importance of the above early studies is that they brought into sharp relief many of the key features of visual temporal dynamics. Subsequent research has focused largely on characterizing the various temporal dynamic phenomena. Collectively, these studies appear to support the notion that, to a first approximation, visual perception proceeds in a coarse-to-fine manner: *Vision at a glance*, or visual perception on an ultra-rapid time scale, operates differently than visual perception on a longer time scale, or *vision with scrutiny* (Hochstein and Ahissar, 2002; also see Section 2.6). While this is a useful pedagogical distinction for our purposes, it is important to remember that it may or may not reflect two distinct types of underlying temporal dynamic phenomena.

2.2. Vision at a glance: ultra-rapid categorization

In a landmark study, Thorpe et al. (1996) measured brain activity using electroencephalogram (EEG) while human subjects reported whether or not a given photograph of a novel natural scene, presented for just 20 ms, contained an animal (see Section 4.2 for additional details). EEG signals, often referred to as event-related potentials (ERPs), specific to correct categorization of images emerged within about 150 ms after the stimulus onset, indicating that human beings can rapidly perceive the 'gist' of a complex natural scene at a glance. Subjects took a few hundred milliseconds longer to press a button to indicate their response (reaction time range, 382–567 ms), presumably reflecting the additional time it takes to plan and execute the motor response. Measuring ERPs bypassed this motor delay, establishing that complex natural scenes can be processed on an ultra-rapid time scale. Even more importantly, this result suggests that the gist of a complex image can be perceived solely through feed-forward processing of the visual signal, since a processing duration of 150 ms would leave little time for feedback processing (Thorpe et al., 1996, p. 522). Many subsequent studies have essentially confirmed this finding (Johnson and Olshausen, 2003, 2005a; Fabre-Thorpe, 2003; Rousselet et al., 2004).

The notion that ultra-rapid visual processing is not dependent on feedback signals is supported by additional studies which indicate that such processing is unaffected by top–down influences such as attention and prior knowledge, which are mediated by feedback signals. It is also known that the ultra-rapid scene categorization can take place pre-attentively, *i.e.*, without attentional deployment. Li et al. (2002) have shown that subjects can rapidly detect whether or not a given natural scene, presented at an eccentricity of about 6°, contained an animal while simultaneously performing an attentionally demanding letter discrimination task foveally (also see Rousselet et al., 2002). Moreover, training in, or expertise with, the task typically does not speed up the performance in ultra-rapid perceptual tasks. When the subjects received extensive 3-week training in the task and were tested with novel stimuli, they were no faster at the task, using familiar or novel stimuli (Fabre-Thorpe et al., 2001).

How much of the visual scene do we perceive at a glance, *i.e.*, within the first 150 ms or so of processing time? The emerging consensus seems to be that one can perform two types of visual tasks at a glance: (i) detecting the presence of an

object, such as determining whether a given natural image contains an object, and (ii) categorizing it at a basic level, classifying the given object as a face, a bird or a house (Grill-Spector and Kanwisher, 2005; also see Fabre-Thorpe, 2003; Johnson and Olshausen, 2003, 2005a; Rousselet et al., 2004). Note that the definition of a 'basic level' category is somewhat arbitrary and operational—it typically refers to the predetermined category that non-expert subjects tend to classify a given object into upon a glance (also see Section 2.3). Importantly, detection and categorization take about the same length of time (Fig. 4; also see Section 2.3), so that as soon as one detects the presence of an object, one knows what it is (Grill-Spector and Kanwisher, 2005). It may even be that the two tasks are fundamentally the same, although we can semantically distinguish between them.

Human performance in ultra-rapid categorization tasks is remarkably robust across a relatively large range of stimulus parameters and viewing conditions. Subjects can still perform the task at more than 60% when the stimuli are centered at an eccentricity of about 70°, and regardless of whether the stimulus is upright or inverted (Thorpe et al., 2001; Rousselet et al., 2003). Processing two natural images is as fast as processing one (Rousselet et al., 2002, 2004). Subjects can perform the task well above chance levels at about 10–12% of initial contrast (Macé et al., 2005b). And, as noted above, subjects can perform the categorization task pre-attentively, *i.e.*, without deploying attention.

Fabre-Thorpe et al. (1998) have shown that the performance of monkeys in this task is essentially similar to that of humans. Indeed, the reaction times, *i.e.*, the time between the stimulus onset and the response, of monkeys were shorter by 100–180 ms in the detection task compared to the fastest human subject for the same stimuli and task (also see Delorme et al., 2000; Fabre-Thorpe, 2003; Macé et al., 2005a). This is presumably because monkeys have smaller heads, so that the neuronal signal has shorter physical distances to travel.

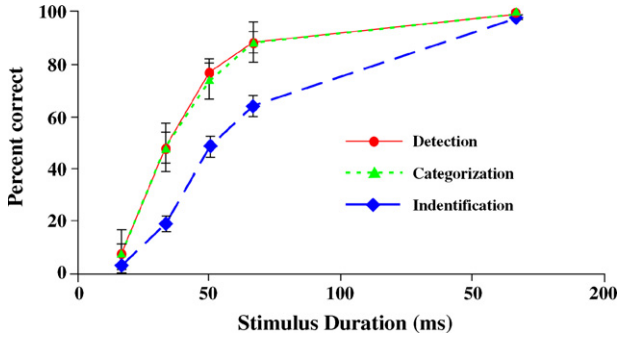


Fig. 4. The performance of human subjects in three different object recognition tasks using natural images. Stimuli were presented for various durations as indicated, and masked immediately thereafter. The performance of the subjects was measured as a function of the stimulus duration in each of the three tasks. In the detection task, subjects had to decide whether or not a given image contained an object. In the categorization task, subjects had to categorize the object in the picture at a 'basic level' (*e.g.*, animal, house, flower). The basic categories were pre-specified, *i.e.*, the subject was given the categories to classify the objects into. In the identification task, subjects had to assign the object to a subordinate-level category (*e.g.*, German shepherd), also pre-specified. Error bars indicate standard errors of the mean. Figure redrawn from Grill-Spector and Kanwisher (2005) with permission.

How much of the visual scene seen at a glance are we aware of? VanRullen and Koch (2003) showed human subjects common indoor or outdoor visual scenes, each containing 10 distinct visual objects. Each scene was presented for 250 ms, slightly longer than the typical processing duration in the aforementioned ultra-rapid categorization experiments. After viewing each given scene, subjects were able to recall on their own the names of 2.5 objects on average. Subjects were then forced to guess from a list of object names, and the subjects were able to pick out an additional 2.5 objects. Interestingly, the objects that the subjects consistently failed to recall showed a negative priming effect in later picture-word matching task, indicating that objects stayed in memory even when they were not consciously recognized or recalled.

Liu and Jiang (2005) revisited the above experiment by requiring the subjects to visually match, rather than verbally name, the objects in a visual scene. They found that subjects recalled fewer than one object per scene at glance, but were able to recall more objects after viewing the scene for several seconds. Liu and Jiang suggest that the reasons for the discrepancy between the two studies are that verbal memory and guessing contribute substantially to how much we are able to recall from a glimpse, and that the contribution of these factors becomes less prominent with longer viewing durations. The delay between stimulus presentation and testing is also important, since we rapidly forget what we see (Potter et al., 2002). Thus, how much we remember from what we see at glance depends on how the recall is measured.

2.3. Vision with scrutiny: elaboration of the visual percept

While one can categorize visual scenes rather rapidly, it is also clear that finer details of image, such as the identity of the object of interest, take longer to emerge (Johnson and Olshausen, 2003). Grill-Spector and Kanwisher (2005) measured the time it takes to perform different types of object recognition tasks by measuring the performance in a given task as a function of the stimulus duration. Subjects were presented with novel, natural scenes for 17, 33, 50, 68, or 167 ms, followed immediately by a mask. In the *object detection task*, the subjects had to report whether the given image contained an object, regardless of what the object was. In the *object categorization task*, subjects classified the object at a 'basic' level, *i.e.*, assigned the object to a 'basic', pre-specified category, such as a car, animal or flower. In the third, *within-category identification task*, subjects were asked to assign the object to a more specialized, also pre-specified, category such as a Volkswagen Beetle or German shepherd, and so forth.

The authors measured the performance as a function of the stimulus duration (same as the processing duration in this case). The rationale was that if a given task takes less time to complete, the performance in that task will be higher than the performance in a task that takes longer to complete. As mentioned above, the performance was indistinguishable for detection *vs.* categorization tasks (Fig. 4), indicating that the

two tasks require the same amount of processing time. For stimulus durations of ≤68 ms, the performance in the detection/categorization tasks was always better than the performance in the identification task. For a stimulus duration of 68 ms, the performance was at near-chance levels for the identification task, but was much better for detection/categorization tasks, indicating that 68 ms of processing time was largely enough for the latter tasks, but not for the former one. Thus, detection/categorization of the object occurs before the identification. For a stimulus duration of 167 ms, the performance in all tasks including identification was close to 100% correct.

Note that the hierarchical nature of the above three tasks naturally captures the hierarchical nature of the visual world and how we perceive it (Mervis and Rosch, 1981). In other words, in perceiving a given visual object in increasingly greater detail, we naturally place it in categories of increasing specificity, such as an object, an animal, a dog, a German Shepherd, a particular German Shepherd named Max, and so forth. The Bayesian framework, then, would explain the observed temporal dynamics of this categorization as a function of the hierarchical nature of the underlying categorical hypotheses and the time it takes for the visual system to 'develop' the necessary information and test the various hypotheses against this information.

Is object identification always slower than object detection/categorization? This seems to be the case for a given object, but not necessarily across all objects. For instance, one may be as fast or faster at identifying one's grandmother than detecting or categorizing an unfamiliar insect. Moreover, while one can conceptually distinguish between categorization *vs.* identification tasks, these and other object recognition tasks are all fundamentally similar, in that they are all classification tasks. The specificity with which one can classify a given object depends on many factors, including the object in question, the subject's familiarity with the object, and the stimulus duration. Therefore, in order to compare the time it takes to complete the various object recognition tasks, one has to be able to account for the differences in these parameters, which is extremely hard to do. Thus, while it is true that all other things being equal, identification (*i.e.*, fine categorization) generally takes longer than the corresponding coarse categorization, it is impossible to assert this *a priori* for a given pair of different objects.

2.3.1. Characterizing the temporal evolution of visual percepts

Traditional methods of measuring visual percepts are inadequate for this purpose, not only because natural images tend to be exceedingly complex, but also because our understanding of them is ultimately semantic and subjective, as you may have noticed in case of Fig. 1. Fei-Fei et al. (2007) devised a novel, and decidedly unorthodox, methodology to address this problem. Using a Google image search based on keywords suggested by experimental subjects, they collected photographs of 44 indoor scenes and 46 outdoor scenes. In the first stage of the experiment, another set of naive subjects freely viewed these stimuli for 27, 40, 53, 67, 80, 107, or 500 ms, depending on the stimulus and subject. Note that only the 500 ms stimulus duration is generally considered long enough to allow eye movements (but see Henderson and Hollingworth, 1998, 1999), although this issue is not crucial in this context.

A given subject saw a given image only once during the entire experiment, but different subjects viewed a given image for different stimulus durations, so that each image was viewed for a given duration by many subjects. After viewing a given stimulus and the ensuing mask, the subject freely recalled and wrote down, in as much detail as possible, what he/she saw in the image. In the second stage, the authors created, based on the images, a list of 105 scene descriptors belonging to six different scene categories: inanimate objects, animate objects, outdoor scenes, indoor scenes, visual/perceptual features, and event-related. Within each category, the descriptors were hierarchical. For instance, in the animate subject category, the descriptors ranged from generic (*e.g.*, 'animal' or 'people') to progressively more specific (*e.g.*, 'bird', 'penguin', and so forth). Based on this list, yet another set of naive subjects determined whether a given free-recall description by the subjects in the previous stage used one or more of the scene descriptors and, if it did, whether the descriptor was accurate given the image. Thus, this process attempts to provide an objective description of the temporal changes in the percept by assessing the shared aspects of the subjective visual experience. This is the crucial innovation of this study—objectivity by shared subjectivity.

The authors found that within a single glance (*i.e.*, stimulus durations of ≤107 ms), human subjects perceive and recall much object and scene level information (*cf*. Section 2.2). For instance, for a particular image presented for 107 ms, one subject accurately reported having seen an outdoor scene "... with a black, furry dog running toward the right of the picture. His tail is in the air and his mouth is open. Either he had a ball in his mouth or he was chasing after a ball". As expected, the percepts tended to be more detailed when the same given stimulus was presented for 500 ms. For instance, at this stimulus duration, the aforementioned image was described by a different subject as that of "... a black dog carrying a gray frisbee in the center of the photograph. The dog was walking near the ocean, with waves lapping up on the shore. It seemed to be a gray day out."

Subjects were somewhat more likely to perceive a given natural scene as an outdoor scene rather than as an indoor scene. Another intriguing asymmetry was that the subjects were much more likely to accurately perceive the overall scene context when they recognized an inanimate object in the scene *vs.* when they recognized an animate object. The reason for either asymmetry is unclear. In addition, the reporting of sensory or feature level information of a scene, such as shading, shape, etc., consistently preceded the reporting of the semantic-level information. Subjects tended to perceive shape-related information slightly sooner than semantic information. But once subjects recognized the more semantic-level components of a scene, there was little evidence of a bias toward scene level recognition (*e.g.*, outdoors) over object level recognition (*e.g.*, dog), or vice versa.

The result that visual percept gets more elaborate and more semantic, or abstract, over time is consistent with the results of

a large body of previous research using simpler, geometric stimuli and more conventional methods (Henderson and Hollingworth, 1999; Pylyshyn, 2003). Note that semantic understanding of the scene is not the same as verbal understanding. There is evidence that the former is faster than the latter. For instance, it takes slightly less time to categorize an object from its image than it takes to categorize the same object from its spoken name (Potter and Faulconer, 1975). But the nature of semantic *vs*. verbal understanding of visual objects is a matter of debate (Henderson and Hollingworth, 1999; Pylyshyn, 2003).

2.3.2. Occlusion, clutter, and other complexities of natural scenes

One important limitation of a great majority of the temporal dynamic studies is that, even when they use natural stimuli, they typically avoid many of the common complexities of natural visual scenes. In natural scenes, factors such as occlusion, visual clutter, camouflage, and variations of lighting, shadows, color texture, etc., can complicate scene segmentation, object recognition and scene understanding. For instance, it is unclear whether viewers can detect occluded or camouflaged animals as rapidly as the unoccluded animals in plain sight as in the studies of Thorpe et al. (1996). For one thing, studies using simulated occlusion (as opposed to natural images with occluded natural objects) indicate that top–down factors such as prior knowledge of the occluded object are likely needed in order to compensate for the missing information about the occluded objects (Bar, 2004; Johnson and Olshausen, 2005b; Hegdé et al., in press). It has been shown recently object-selective brain regions such as in the lateral occipital complex (LOC) and the dorsal foci (DF) contain subregions that are more responsive to occluded objects than their unoccluded counterparts (Hegdé et al., in press; also see Lerner et al., 2002). But whether such functional specializations may somehow obviate the need for top–down influences is unclear. Ultimately, understanding the effects of these complexities is a key part of understanding natural vision (Box 4).

2.4. Insights from masking studies

Masking refers to the reduced visibility of one stimulus, called the target, by another stimulus, called the mask (for reviews, see Breitmeyer, 1984; Öğmen and Breitmeyer, 2006). In a typical masking experiment, referred to as backward masking, a target is followed by the mask after a given delay. Either stimulus need be presented only briefly, for as little as few tens of milliseconds, but the duration of the SOA is much more critical. While there are many other types of masking paradigms, such as forward masking, mutual masking and so forth, backward masking is most directly useful for studying the temporal dynamics of visual perception, since it allows the experimentor to disrupt the processing of the target at various time points while keeping the stimulus duration itself unchanged, and ask how the corresponding visual percept changes. The mechanisms of masking are far from clear, but it is generally thought that it works by disrupting the reentrant processing of the target (Enns et al., 2006). Recurrent (or reentrant) processes bring the top–down, cognitive factors to bear on processing of the target, typically after the initial image-driven, feed-forward sweep (see Section 3.1 for details). For this reason, tasks that (arguably) require little or no top–down processing, such as detection, are less sensitive to masking than those that almost certainly require it, such as identification (Grill-Spector and Kanwisher, 2005; also see Fig. 4).

In a classic backward masking experiment, Bachmann and Allik (1976) used two geometric forms of equal area, without noise or background, as the target and the mask, each presented briefly (10 ms). They measured the ability of the subjects to identify the target as a function of increasing SOAs. They found that when the SOA was 0 ms, *i.e*., when the two stimuli came on simultaneously, the two stimuli tended to be perceived as a single blended object. With SOAs of 40–80 ms, the target was least visible and the identification of the target decreased to about chance level, suggesting that the recurrent processing during this period is crucial for identification. It took an SOA of about 150 ms for the target to be perceived as a stable, distinct shape, and reliable identification took an SOA of 250 ms or so, indicating the time points by which the corresponding processes are effectively complete.

Masking studies such as this serve to illustrate two larger points. First, they highlight the critical importance of recurrent processing, since the target is not perceived without it. In other words, it is a huge mistake to think of temporal dynamics, and of visual perception at large, as a primarily image-driven process (also see Hegdé and Felleman, 2007). Second, the temporal dynamics of the underlying neural processes are probably a major determinant (or bottleneck) of the temporal dynamics of visual perception. In the above experiment, for instance, the underlying stimuli and task were simple and did not vary with the SOA, so that the percept more or less directly reflected the processing time allowed. Thus, in this case, the processing bottlenecks were likely to have been a more important determinant of the temporal dynamics of the percept than the underlying perceptual hypotheses. Note that the Bayesian framework readily accounts for this, by taking the information processing bottleneck into account.

2.4.1. 'Masking' by transcranial magnetic stimulation (TMS)

In TMS, a very brief electromagnetic pulse, lasting 300 μs or so, is delivered non-invasively from a magnetic coil placed over the scalp (for reviews, see Anand and Hotson, 2002; Kammer, 2006). TMS pulse is known to produce a complex pattern of excitation and inhibition in the brain region directly underneath the coil, as well as a pattern of more indirect activity in more distant regions, and probably acts by eliciting inhibitory postsynaptic potentials in the cortex (see Moliadze et al., 2003, and the references therein).

In one of the first studies of its kind, Amassian et al. (1989) showed that a single pulse of TMS applied over the occipital cortex in human subjects disrupted visual processing, depending on when the stimulation was delivered relative to visual

stimulus. Subjects were presented with three different random letters simultaneously, and a TMS pulse was delivered after various delays. When the delay was less than about 60 ms, or more than about 140 ms, subjects were able to correctly identify the letters they saw. At delays of 80–100 ms, a blur or nothing was seen. The authors concluded that the neural activity subserving letter recognition is probably transmitted from the occipital cortex within 140 ms of the visual stimulus. Thus, the essential advantage of TMS is that, unlike visual masking, it can disrupt the processing in restricted parts of the brain. Moreover, since TMS is non-invasive, it can be easily used in human subjects who, unlike monkeys, can be readily made to perform complex tasks and report nuanced, abstract percepts. Its main disadvantage is its poor spatial resolution, especially compared to the more invasive techniques such as microstimulation (see Section 3.8).

2.5. Effects of stimulus history: priming

Priming is the phenomenon where the perception of a given stimulus, or the prime, affects the perception of a succeeding stimulus, or the target, even when the prime is presented after a long delay or is not explicitly perceived (for reviews, see Wiggs and Martin, 1998). Depending on the stimuli involved, the prime can facilitate or impair the perception of the target (*positive-* or *negative priming*, respectively). The effects of priming are usually evident as changes in the reaction time, target recognition performance, and/or error rates. Whether priming has other significant temporal dynamic effects, *e.g.*, whether it 'short-circuits' visual processing or qualitatively alters the percept, is of course much harder to address and remains unclear. While priming has been the subject of considerable research over the last two decades, its relevance to the present context is that it is one of the more easily controlled top–down factors that influence the temporal dynamics of visual processing.

In an illustrative experiment, Vorberg et al. (2003) primed the subjects with a leftward arrow for 14 ms and, after systematically varying SOAs, presented the target, also an arrow, for 140 ms. Depending on whether the trial represented a congruent or incongruent condition, the target arrow pointed in the same or the opposite direction, respectively, as the prime. The authors required subjects to report direction in which the target arrow pointed, and measured the reaction times in either condition as a function of the SOAs that ranged from 0 to 70 ms. In the congruent condition, the reaction times fell monotonically as a function of SOA, indicating that in this case, the prime speeded up the perception of the target (*positive priming*). In the incongruent condition, the prime slowed down the perception of the target (*negative priming*). In this experiment, the subjects did not perceive the brief prime (making it an instance of *subliminal-* or *masked priming*), but the results were similar when the subjects did perceive the prime (Schwarzbach and Vorberg, 2006). For intermediate differences in the orientation of the two arrows, the effects were correspondingly intermediate, indicating that the degree of priming was roughly commensurate with the similarity between the two stimuli.

In a similar experiment, Zago et al. (2005) studied the effect of varying the duration of the prime stimulus while holding the SOA constant. The facilitation by priming showed a classic 'rise-and-fall' pattern, in which the facilitation continued to increase for prime durations of up to 250 ms and decreased to lower levels thereafter and remained significantly above chance levels for up to 1900 ms. In other words, top–down influences have time windows during which they are most effective, but the optimal time windows vary depending on many factors, including the task.

How does priming work in natural scenes, where the object shapes can vary in forbiddingly complex ways? Perhaps the simplest case is *repetition priming*, when repeated looks at the same object makes it easier to see, something you may have noticed in demo experiment in Section 1.1. For images of isolated natural objects, priming (both positive and negative) has been shown to be object-specific and, to a limited degree, invariant to various transformations (Bar and Biederman, 1999). It is unclear whether this means that priming operates in similar fashion in natural *scenes* with multiple objects, although it has been reported that those objects in a natural scene that the subjects consistently fail to see tend to be those that also elicit negative priming when presented alone (VanRullen and Koch, 2003).

As noted above, the main known temporal dynamic effect of priming is that it can alter the speed and reliability of object recognition. The neuronal mechanisms of this process are largely unclear. Given the limited translational invariance of subliminal priming, Bar and Biederman (1999) proposed that areas at intermediate levels of visual processing, such as areas in the posterior temporal cortex, may mediate priming. Using functional magnetic resonance imaging (fMRI) of the temporal cortex, Zago et al. (2005) have reported evidence that suggests that there are two different temporal dynamic processes at work in priming—one earlier acting process that fine-tunes (or sharpens) the selectivity of neurons in the relevant areas, and a slightly delayed adaptation process that 'sparsens' the responses. This sparsening selects a small subpopulation of neurons and 'shuts down' the rest. Zago et al. suggest that the collective effect of these two processes is the classic 'rise-and-fall' pattern, in which the effect of priming increases for prime stimulus durations of up to 250 ms and decreases thereafter.

It is worth noting that the aforementioned sparsening is reminiscent of the sparsening at the neuronal level described in Section 3 (also see Box 3), although it remains to be seen whether the two sets of processes are fundamentally related.

2.6. Perceptual learning and familiarity

It is intuitively obvious that familiar objects, such as Einstein's face or the Eiffel Tower, can be recognized much more readily than unfamiliar objects. Conversely, the specificity with which one perceives a given object or scene tends to depend on one's expertise with the subject. A bird expert might detect a Eurasian Tree Sparrow in an image, whereas a naive subject may simply see a bird under the same experimental conditions. There is a large body of evidence that learning and

familiarity can improve object recognition performance by decreasing reaction time, increasing the specificity of classification, reducing error rates, and/or increasing the certainty of decisions (for reviews, see Fahle and Poggio, 2002; Hochstein and Ahissar, 2002; Ahissar and Hochstein, 2004). Indeed, the priming effects outlined in the previous section reflect a rapid form of perceptual learning.

Whether prior knowledge and familiarity simply expedites the recognition process or changes the process more qualitatively is a matter of some debate (see Fahle and Poggio, 2002). In any event, there is evidence that some types of object recognition, such as the aforementioned ultra-rapid categorization, cannot be speeded up by training. Fabre-Thorpe et al. (1998) trained human subjects extensively over a 3-week period to report whether a given natural scene contained an animal or not. After the training, they monitored the ERPs while the subjects performed the same task. The authors found no difference in the brain activity elicited by novel *vs.* familiar stimuli; both types of stimuli could be categorized equally fast. One explanation for this phenomenon is that the ultra-rapid categorization represents the fastest possible formation of a visual percept, and cannot be speeded up further. Another possibility is that the training did improve performance, but the task parameters (*e.g.*, stimulus duration or low-level image differences) were not difficult to distinguish between the performances with the two sets of images. This latter scenario is plausible, since the accuracy of the subjects was slightly higher for familiar stimuli than for novel stimuli (96.9% *vs.* 94.7% correct, respectively).

In any event, it is easy to conceptualize the role of perceptual learning, and knowledge resulting from it, in temporal dynamics within the Bayesian framework: prior knowledge is one of the many top–down influences that helps the visual system to better interpret the visual image by 'making up' for the ambiguities in the bottom–up information. The reason why perceptual training does not improve detection performance (to the extent it does not) may be that the results of bottom–up processing suffice for the relevant detection tasks, and top–down processes are not needed.

It is important to emphasize that, while the Bayesian framework shows that the brain *can* use prior knowledge to resolve ambiguities, it cannot yet explain *how* the brain does it. For instance, the Bayesian framework cannot yet explain how the brain selects which aspect of the vast body of prior knowledge is relevant to the task at hand (Hegdé and Felleman, 2007). But the *reverse hierarchy theory* (RHT; Hochstein and Ahissar, 2002; Ahissar and Hochstein, 2004) provides a comprehensive framework within which to relate the effects of visual perceptual learning with visual perception. RHT and the Bayesian framework are entirely consistent with each other, but RHT is a more neural-level framework.

Briefly, RHT has two interrelated components. The first, and more directly relevant to the present context, has to with the visual processing steps. RHT posits that the initial *vision at a glance* is mediated by high-level areas, and that the subsequent *vision with scrutiny* involves recruitment of progressively lower visual areas, which contain fine-grained image information, to add finer details to the percept. In other words, during the initial, feed-forward sweep of processing, the information processing proceeds up the various stages of the visual hierarchy (Felleman and Van Essen, 1991) in an automatic and implicit fashion. Our initial explicit (or conscious) percept of the scene, or vision at a glance, reflects the gist of the scene based on the activity of the topmost level of the hierarchy after this initial feed-forward sweep. Subsequent vision with scrutiny is mediated by feedback processing, which proceeds in the reverse direction down the hierarchy, focusing attention to specific, active, low-level units, and incorporating the more detailed information available there into the conscious percept (see Fig. 1 of Hochstein and Ahissar, 2002).

The second component of RHT is that perceptual learning parallels perception, in that it proceeds in a top–down fashion. Learning begins at high-level areas of the visual system, and when learning-dependent changes in these areas do not suffice, it proceeds to progressively lower levels as necessary, where the neurons convey finer-grained visual information. Thus, this theory predicts not only that training improves performance in easier tasks first, but also that these early training effects should be most evident in higher visual areas. Conversely, more difficult perceptual tasks, such as those that require finer discrimination, should show slower improvement with training, and the training-dependent effects should be more evident in lower visual areas. In general, these predictions have been empirically borne out (Hochstein and Ahissar, 2002; Ahissar and Hochstein, 2004).

The relevance of RHT to our context is that it represents a neural model of visual perception in general, and visual temporal dynamics in particular. But there are alternative models of visual perception that also appear to explain many aspects of visual perception, including its temporal dynamic aspects, as outlined in the next section.

2.7. Facilitation of scene perception by context

Visual perception is facilitated by the visual context. It is likely easier to recognize a piece of tinsel on a Christmas tree than along a wilderness trail. Thus, our expectation of what we are likely to see influences what we do see, and how quickly or easily we see it (for reviews, see Henderson and Hollingworth, 1999; Bar, 2004).

In an influential study, Biederman (1972) showed that subjects were much more accurate in identifying a single cued object in a coherent real-world scene than when the scene was jumbled. Jumbling the objects in a scene impaired identification performance even when the subjects knew where to look and what to look for. Thus, meaningfulness of an object's context appears to affect the course of perceptual recognition itself, and not just peripheral scanning or memory.

However, the mechanisms by which visual context facilitates visual perception are not altogether clear. One broad set of views is that the analyses of context and of objects proceed in parallel during the initial temporal phase of visual processing, and only later are the two sets of information brought together. The support for this view comes, among other things, from the

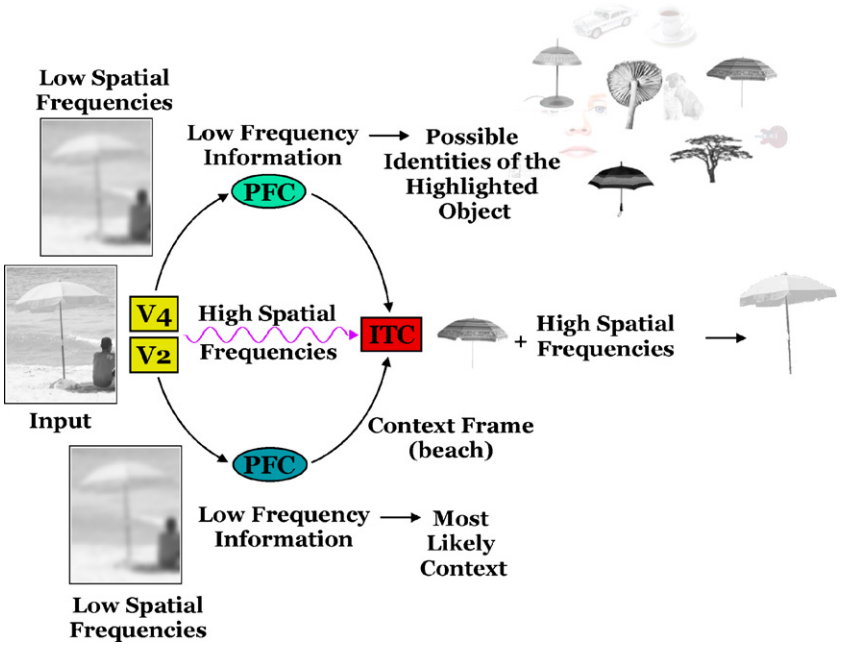


Fig. 5. The Bar model of visual perception. The model posits that a coarse-grained (or low-spatial frequency) representation of the object originates from the early visual areas (including V2 and V4), and reaches high-level brain areas (including PFC) relatively fast (*top left arrow*). Top–down signals about some of the plausible interpretations of the object project from PFC to object-selective regions in ITC (*top right arrow*). In ITC, these plausible interpretations are combined with the finer grained, high-spatial frequency-based representation of the image reaching the ITC more slowly (*squiggly arrow*) to generate a likely interpretation of the object. Contextual information is processed similarly (*bottom half of the loop*), except that the relevant plausible interpretations are thought to be generated in PHC instead of PFC. Figure redrawn from Bar (2004) with permission. Subsequent studies have added some anatomical details to the model (Bar et al., 2006).

fact that basic level object categorization can take place rapidly and in the apparent absence of contextual information (Henderson and Hollingworth, 1999; also see Section 2.2).

A second class of views holds that the visual system first extracts the context information relatively rapidly, and uses the contextual information to facilitate object recognition (Biederman, 1972; Palmer, 1975). For instance, as Bar (2003, 2004) points out, an umbrella, a mushroom and an acacia tree all have the same approximate overall shape, but each is easier to identify in its correct context.

In support of this latter set of views, Bar (2004) has proposed a model of how top–down effects such as visual context and prior object knowledge facilitate object recognition (Fig. 5; also see Bar, 2003; Bar et al., 2006). He proposes that low-spatial frequency content of the image, which carries information about the 'gist' of the scene, is processed relatively fast in the early visual areas and conveyed to specific high-level areas in the prefrontal cortex (PFC). Based on this low-resolution bottom up information, PFC areas generate some likely interpretations of the objects in the visual scene and convey this information to the object-selective regions in the inferior temporal cortex (ITC). Similarly, the parahippocampal cortex (PFC) and other brain regions use the low-frequency information to generate likely interpretations of the scene context. The high-spatial frequency content of the image, which contains finer-grained information about the object, is processed more slowly by the early visual areas and projected directly to ITC. In ITC, the high-spatial frequency-based, bottom–up information coming in from the early visual areas and the top–down information about the plausible interpretations of the object and the scene context coming in from higher visual areas are combined to arrive at a likely interpretation of the image (also see Section 4.5).

While a detailed comparison of this model and the aforementioned RHT is beyond the purview of this review, it is worth noting that the two models seek to explain essentially the same perceptual phenomena using very different styles of coarse-to-fine processing (although RHT does not explicitly use the phrase 'coarse-to-fine'). Other notable differences include the fact that RHT explicitly posits potential roles for all visual areas and for selective attention. On the other hand, the Bar model formulates visual perception explicitly as a process of inference that generates a likely interpretation given the image data and prior knowledge, in the same vein as the aforementioned Bayesian framework. The Bar model also specifically addresses the facilitative effects of visual context (for more on the Bar model, see Section 4.5).

One aspect of the effects of visual context that has received little direct attention is that visual context itself is often hierarchical, so that the processing of contexts may follow the same overall pattern as that of objects outlined above. For instance, indoor scenes can be readily distinguished from outdoor scenes, but perceiving whether a given scene is that of living room or dining room is likely to take longer.

2.8. Temporal changes in the processing of low-level image parameters

In addition to the aforementioned temporal changes in the high-level percepts, the processing of many low-level stimulus

parameters also changes rapidly in time (for a review, see Rodieck, 1998). These phenomena are worth taking a close look at, especially because it is unclear whether or how they fit into the aforementioned scheme of visual perception as coarse-grained categorization followed by finer-grained classification, such as identification.

2.8.1. Stereoscopic disparity

Most computational models of stereopsis postulate that the visual system first computes the correspondence between images from the two eyes on a coarse spatial scale, and subsequently uses this information to constrain binocular correspondence on a finer spatial scale, thus minimizing the number of false matches between the two monocular views (Howard, 2002). While there is psychophysical evidence for this view, there is also evidence for a fine-to-coarse progression of binocular fusion, in which information on a finer spatial scale can constrain the correspondence on a coarse spatial scale (Smallman, 1995; Rohaly and Wilson, 1993, 1994; also see Menz and Freeman, 2003; Ringach, 2003).

Thus, either coarse-to-fine or fine-to-coarse processing by itself fails to fully capture the complexity of the process. A more nuanced view may be that the system dynamically uses information across different spatial scales as needed to constrain estimations of binocular correspondence (also see Ringach, 2003, p. 8).

2.8.2. Luminance and luminance contrast

Human observers detect brighter flashes of light faster than they detect dimmer ones under otherwise identical conditions. This is an aspect of *Pieron's law*. Conversely, *Bloch's law* states that as the stimulus intensity I increases, it takes correspondingly shorter stimulus duration T to produce the same response R, so that $R = I \times T$. It has also been shown that reaction times decrease approximately as the power function of the product of contrast and spatial frequency (see Ludwig et al., 2004, and the references therein).

Although these results were obtained using simple geometric stimuli and it is unclear to what extent they apply to natural stimuli, they nonetheless raise the possibility that under natural viewing conditions, image elements that have higher luminance and/or contrast have greater perceptual salience, and are perceived more readily. Indeed, there is strong evidence that the high-contrast regions tend to be more effective in attracting attention, presumably by attracting saccades (Einhauser and Konig, 2003; also see Henderson and Hollingworth, 1998, 1999). Beyond this, whether the temporal dynamics of the processing of luminance and contrast plays a role in other perceptual-level changes, such as shape perception, is unknown.

2.8.3. Color sensitivity

It is known that for brief stimulus durations at a given spatial frequency, visual perception is monochromatic, *i.e.*, we perceive the stimulus as gray. At moderate presentation durations, the perception is dichromatic, *i.e.*, we fail to perceive blue-yellow contrast variations. At longer durations, the perception is trichromatic. But the color sensitivity covaries with that of space and time, so that in all three cases, human subjects are maximally sensitive to medium frequencies (Wandell, 1995). Again, it is unclear precisely how the temporal dynamics of color processing influences that of natural scene perception, although it is known that in natural scenes, colored objects are easier to recognize and remember than black-and-white objects (Gegenfurtner and Rieger, 2000).

2.8.4. Spatial frequency

We will examine this topic in some detail, because many key models of visual perception, including the aforementioned Bar model, critically depend on it (see Sections 2.7, 3.3–3.8, 4.5, 6; Fig. 5). Here, we will limit ourselves to examining the relevant psychophysical evidence.

Much of what we know about the temporal aspects of spatial frequency processing comes from the studies of contrast sensitivity as a function of temporal frequency (*i.e.*, how fast the stimulus flickers in time). But given the non-linearities in the visual system, it is hard to infer the temporal changes in spatial frequency sensitivity from the temporal frequency sensitivity data (Watson, 1986). In other words, it is hard to extrapolate from temporal frequency to processing duration.

One of the few studies that have directly addressed the spatial frequency processing as a function of processing duration was carried out by Hughes et al. (1996). They addressed whether the aforementioned global precedence effect (Section 2.1) can be explained in terms of spatial frequency processing. Recall the two hallmarks of the global precedence effect: first, global shapes are processed faster than the local shapes (global dominance). Second, when the global *vs.* local shape cues conflict as in a global T created by many local S's, it becomes harder to recognize the shape of the local shapes, but not the global shape (asymmetric interference).

In the case of Hughes et al., the global stimulus was a low-spatial frequency (1 cpd) sinusoidal grating oriented vertically or horizontally. The local stimulus was a similar grating, but of a higher spatial frequency (9 cpd), superimposed on the 1 cpd grating. They found that the subjects were faster and more accurate in reporting the orientation of the 1 cpd grating than that of the 9 cpd grating (global dominance effect). When the two gratings had different orientations, subjects performed better with the 1 cpd grating than with the 9 cpd grating (asymmetric interference). Furthermore, by systematically varying the SOAs of the low-frequency grating, Hughes et al. found that the low-frequency grating was able to interfere with the processing of the high-frequency grating for at least 100 ms after the onset of the high-frequency grating, but not the other way around. Taken together, these results indicate that the processing of the low-frequency gratings precedes that of the high-frequency gratings. Thus, the dynamics of spatial frequency processing may underlie the global precedence effect, since the former captures the essential features of the latter rather well. However, Hughes et al. caution that "low frequencies and global image attributes need not be one [and] the same thing" and that global image attributes are not attributable to low-spatial frequencies (p. 225).

Parker et al. (1992, 1997) have provided another line of evidence for spatial frequency-based coarse-to-fine processing. For each of their raw images, they created three bandpassed images that contained the low, medium or high-spatial frequency contents of the raw image. Subjects had to discriminate between a given raw image presented for 120 ms from a simultaneously shown sequence of the three corresponding bandpass images, each presented for 40 ms. They found that the subjects were significantly less able to tell the raw images from the bandpass ones if the bandpass images were presented in low-to-high sequence than if they were presented in a high-to-low sequence. In other words, the low-to-high sequence appeared similar to the raw image, suggesting that the visual system processes the image in that order.

These results have inspired many subsequent models of coarse-to-fine processing rooted in the processing of low- *vs.* high-spatial frequencies. These models have differed on whether the spatial frequencies are defined in terms of the retinal image or of the object itself. The 'retina-based' models posit that spatial frequency channels analyze the various spatial frequencies in the retinal image in parallel, with the lower spatial frequency analysis preceding the analysis of the higher frequencies of the retinal image (Parker et al., 1992, 1997; Hughes et al., 1996; Bar, 2003, 2004; Bar et al., 2006).

Another type of these models, known as 'flexible-use' or 'object-based' models, takes note of the fact that low-frequency content of the image is not always the most informative, and posits that the visual system dynamically extracts the global *vs.* local information based both on the information available in the image and many top–down factors, such as prior knowledge and the task (Schyns and Oliva, 1997). The distinction between the two types of models is not just academic, because if the retina-based models are correct, then it becomes straightforward (but not necessarily easy, see Section 3.4) to explain the temporal dynamics of image processing in terms of the known neural mechanisms of low- *vs.* high-spatial frequency processing.

A critical test that distinguishes between these two sets of models is whether the relative 'informativeness' of the various spatial frequencies scales with the retinal image. On this count, as on many others, both sides claim some supporting psychophysical evidence (see Morrison and Schyns, 2001; Sowden and Schyns, 2006).

The promise of the spatial frequency-based models notwithstanding, it is far from clear that any of these models provide a general, comprehensive explanation of visual temporal dynamics. For one thing, natural images cannot be fully explained solely in terms of its spatial frequency content (Gallant, 2004). For another, the most informative features of visual scenes are not always definable by their spatial frequency content (see, *e.g.*, Biederman, 1995; Oliva and Schyns, 1997; Torralba and Oliva, 2003; Torralba et al., 2006; Ullman, 2007).

2.9. Summary of psychophysical findings

While the results vary considerably across the various studies, two sets of findings seem fairly clear. First, one can detect or categorize objects in complex natural images on an ultra-rapid time scale. Top–down factors such as attention or learning have little effect on this process. Second, finer-grained understanding of the images takes longer, depending on a large number of bottom–up and top–down factors, including image complexity, attention, context, familiarity and expertise. Many, although not all, temporal dynamic phenomena at the perceptual-level progress in such coarse-to-fine manner. The reverse hierarchy theory and the Bar model represent two very different types of models of the coarse-to-fine perceptual phenomena. The Bayesian framework of sequential inference appears to provide a more general, albeit entirely untested, framework within which all temporal dynamic phenomena can be understood, regardless of whether or not they follow a coarse-to-fine time course.

3. Temporal dynamics at the neuronal level

3.1. Some relevant temporal dynamic properties of visual cortical neurons

Much of what we know about visual temporal dynamics at the neural level comes from single-unit studies in the macaque monkey. The functional organization of the macaque visual system (Fig. 6) is similar to that of humans, but has been understood in much greater detail.

Most neurophysiological studies, including those in this review, assume *rate coding*, which posits that neurons convey information by modulating the rate at which they fire spikes. However, other aspects of the neuronal response, such as spike timing, can also convey information (Box 1; also see Rieke et al., 1996).

Visual cortical neurons typically fire spikes at a low, 'background' level in the absence of overt visual stimulation (see, *e.g.*, Fig. 8a). When a static visual stimulus is presented in the neuron's classical receptive field (CRF), there is usually a brief delay before the response rises above background levels. This delay, or *latency*, tends to be on the order of a few tens of milliseconds in early visual cortical areas such as V1 or V2, and becomes progressively longer (lasting several tens of milliseconds) in higher visual areas, such as those in the inferotemporal cortex and the frontal cortex (Schmolesky et al., 1998; Bullier, 2001; also see Lamme and Roelfsema, 2000). In the macaque, where the latencies have been measured in detail, some neurons in all visually responsive areas, including frontal and motor cortices will have been activated by about 190 ms after the stimulus onset (Fig. 6b; also see Lamme and Roelfsema, 2000). The response latencies in the human brain are generally longer by a few tens of milliseconds, depending on the brain region (see Yoshor et al., 2006, and the references therein). This is presumably because the human brain is physically larger, so that the neuronal signals must travel farther. The feed-forward and feedback connections conduct information at a velocity of about 2–3.5 m/s (Girard et al., 2001; Bullier, 2001). The lateral connections, or connections between neurons within a given area, conduct about ten times slower, about 0.33 m/s. Note that this does not necessarily mean that lateral interactions have longer latencies,

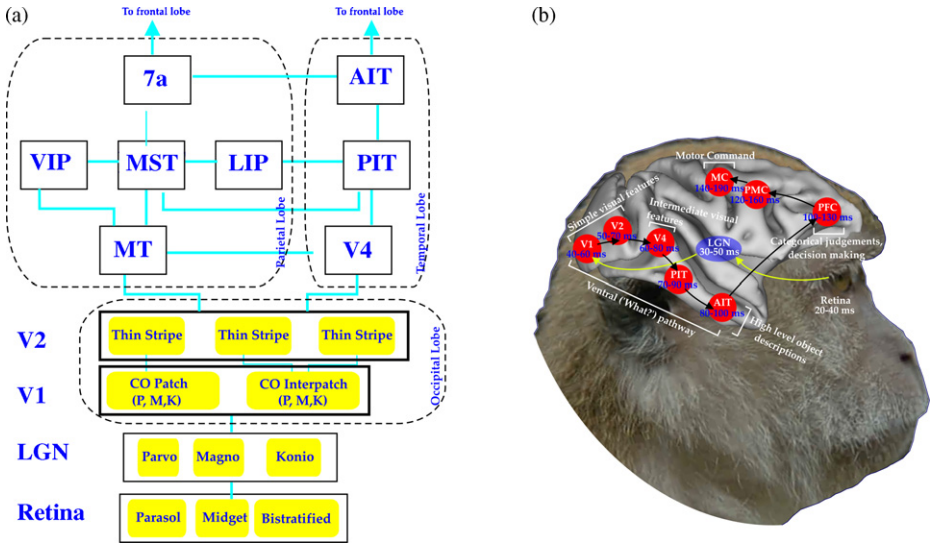


Fig. 6. Functional organization of the macaque visual system. (a) Flow chart illustrating the feed-forward flow of visual information from the retina to the frontal lobe along the dorsal pathway (*parietal lobe*, *left*) and the ventral pathway (*temporal lobe*, *right*). The boxes illustrate some key stages of visual processing. Only some of the known visual cortical areas are shown, and pathways of recurrent processing are not shown. Macaque IT contains many subdivisions, two which (PIT, AIT) are shown here. For details on functional organization of the macaque visual system, see Felleman and Van Essen (1991) and Hegdé and Felleman (2007). (b) Response latencies of key brain regions involved in a typical object recognition task, leading up to the motor command. *Yellow* and *black arrows* denote subcortical and cortical information flow, respectively, in the feed-forward direction. Feedback and lateral connections are not shown. Only the temporal pathway is shown, because it is presumed to dominate object recognition. The presumed role of key areas in object processing is indicated (*square brackets*). Figure in panel b was adapted from Thorpe and Fabre-Thorpe (2001) with permission. The brain shown is an actual macaque brain, slightly unfolded for greater clarity using the CARET toolkit (Van Essen et al., 2001). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

since lateral connections are often physically shorter (Lamme and Roelfsema, 2000; Bullier, 2001). Processing mediated by feedback and/or lateral connections is often collectively referred to as recurrent or reentrant processing. Across the visual cortex, the synaptic delays are negligible for electrical synapses, but tend to be about 5–20 ms per synapse for chemical synapses (see Azouz and Gray, 1999, and the references therein).

Altogether, response latencies in a given area vary considerably depending on the neuron, stimuli, and the experimental and analytical methods used for measuring the latency, and typically overlap substantially with those of other areas, so that a given latency is not unambiguously diagnostic of any single area. This also means that response latencies usually cannot be directly compared across studies.

When a static stimulus is presented within its CRF, a typical visual cortical neuron, after a delay, fires rapidly for a few tens of milliseconds, depending on the neuron. After this initial transient (or onset transient) response, the firing rate decays rapidly, before largely stabilizing at a lower response level over the next few hundred milliseconds. It is thought that feed-forward inputs fully account for the response transients, whereas recurrent processing plays a major role in shaping the post-transient response (see Lamme and Roelfsema, 2000; Scholte et al., 2006). When the stimulus is turned off, the response shows another, smaller, transient increase ('offset response'; see *arrowhead* in Fig. 8a), before decaying to background levels. The offset response tends to be more common in lower visual areas than in higher visual areas.

The response pattern of subcortical neurons tends to be somewhat different from that of cortical neurons. Most notably, the post-transient response decay is often less pronounced in subcortical neurons (Purpura et al., 1990; Hawken et al., 1996).

The noise, or random variations in the firing rate across different presentations of the same stimulus, roughly follows the same overall temporal pattern as the firing rate (Hegdé and Van Essen, 2004, 2006). However, the temporal interplay of signal and noise is such that the signal-to-noise ratio, or related measures of information transmission, are generally maximal during the initial transient response and lower later, although they tend to remain statistically significant throughout the response (Oram and Perrett, 1992; Müller et al., 2001; Hegdé and Van Essen, 2004, 2006). Roughly 10% of the noise is *correlated noise*, in which trail-to-trial variations in the response are correlated between two (usually nearby) cells regardless of the stimulus. While correlated noise can be computationally important, most neurophysiological studies ignore this potential complexity (Averbeck et al., 2006).

It is worth emphasizing that the above description of the typical response pattern of a visual cortical neuron is necessarily simplistic, in that it describes the response resulting from presenting static stimuli with the CRF. Additional factors that come into play under natural viewing conditions – including the stimulation of the surrounding non-classical receptive field (nCRF), movements of the eye, observer and visual objects, and top–down influences – modulate the neural response in complex, often unpredictable, ways (Lamme and Roelfsema, 2000; Bullier, 2001; Gallant, 2004; Smith et al., 2006).

In Sections 3.2–3.9, we will examine some key temporal dynamic phenomena at the neuronal level. Readers should be forewarned that not all these results will fit neatly into a coherent, larger temporal dynamic picture. The reasons for

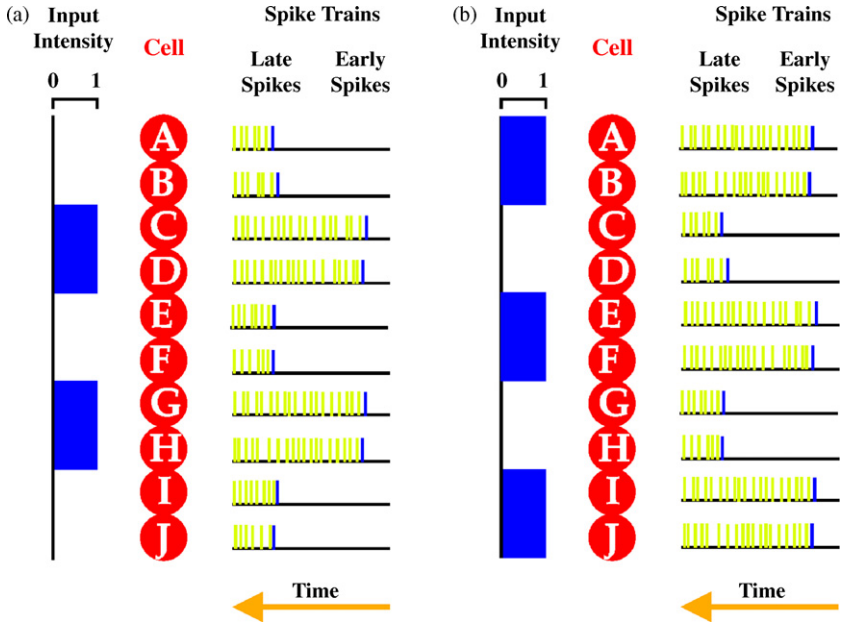
Box 1. Ultra-rapid image processing using spike timing

Most neurophysiological studies assume rate coding, *i.e.*, that neurons convey information by modulating the rate at which they fire spikes. However, many recent studies have shown that, under some circumstances, the timing of a given spike can also convey information. Such spike-time coding is based on the notion that spike latency is a function of the value of given visual feature. In the domain of luminance, for instance, this means that cells respond sooner to a brighter stimulus, and with a greater delay for a darker stimulus. Thus, when stimulated with a black-and-white pattern (*blank* and *blue* regions, respectively; *far left*, panels a and b), the cells corresponding to the brighter parts fire their first spikes (*blue vertical line* in the corresponding spike train) earlier than the cells with darker parts of the image in their receptive fields. Thus, earlier spikes correspond to brighter parts of the image. Note that time progresses from right to left in this figure, so that the earlier spikes are to the right of the later spikes.

An extreme version of spike-time coding is the rank order (or 'recruitment' order) coding, in which the temporal order, rather than the precise timing, of the first spikes convey the image intensity information. Note that in all of these schemes, spikes that occur after the first spike in a given neuron (*yellow vertical lines*) convey no additional information about the stimulus. Thus, rank order coding is essentially first-spike coding. It is roughly analogous to deciding election results based on the very first vote cast across the various precincts.

Neurophysiological studies of visual, olfactory, somatosensory and auditory systems indicate that such coding is a plausible mechanism for ultra-rapid, coarse-grained representation of the sensory stimulus (VanRullen et al., 2005). These studies also serve to illustrate the larger point that sensory representations can be achieved using coding schemes other than rate coding. Rate coding can be no faster than spike time coding because, by definition, rate coding must sample spikes that occur over a stretch of time.

A major problem with spike-time coding is that in order to interpret this code, the system needs to know the time of stimulus onset, so that a given spike thereafter can be designated as the first spike from the given neuron to the given stimulus (VanRullen et al., 2005). In more general terms, the system needs to have a reference time point against which a given spike is the first spike. Under natural viewing conditions, designating such reference time points can be problematic, since the retinal image changes dynamically and complex ways with no easily definable onset.



presenting these results in their disparate complexity is that they represent the current state of the field, and that an understanding of the response dynamics of different visual areas is critical to an eventual fuller understanding of visual temporal dynamics.

3.2. Redundancy reduction and adaptive filtering in early visual processing

Dan et al. (1996) compared the responses of lateral geniculate (LGN) cells in the anesthetized cat to 20–60 min stretches of natural movies (one of which happened to be *Casablanca*) and of white noise control stimuli. As expected, there was substantial temporal correlation, or redundancy, in the natural movies, with more power at the lower temporal frequencies, so that the power spectrum was 'pink'. But the LGN responses to the movies had a 'whitened' temporal power spectrum, with roughly comparable power over a relatively wide range of temporal frequencies (3–15 Hz), including the frequencies that were redundant in the movies. Such whitening makes the neural responses more efficient by reducing the redundancy in the information, just like removing duplicate pages from a book would. Interestingly, the responses of

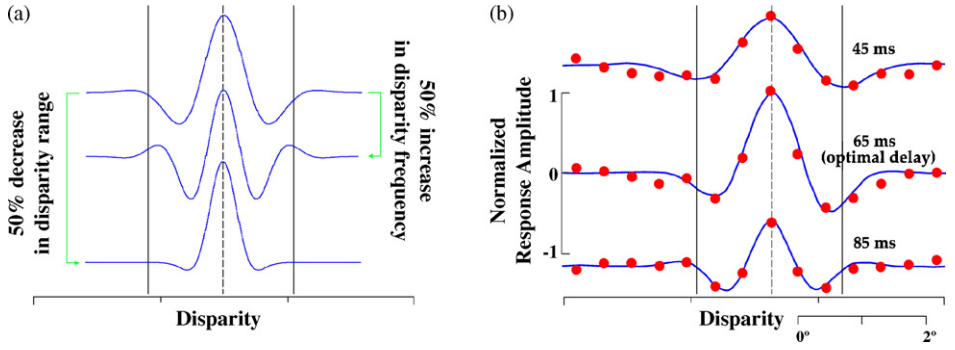


Fig. 7. Coarse-to-fine changes in disparity tuning. (a) Schematic illustration of changes in two different measures of disparity tuning. Either a decrease in the disparity range (*arrow on left*) or an increase in disparity frequency (*arrow on right*) denotes a coarse-to-fine change in disparity tuning. The dashed vertical line denotes the preferred disparity. (b) Coarse-to-fine changes in the disparity tuning of an actual complex cell in cat area 17 over a 40 ms period. The responses to individual disparity stimuli (*dots*) were fitted with a disparity tuning curve (*blue lines*). The optimal delay for this cell was 65 ms (*middle*), when the modulation of the tuning curve was most pronounced. The cell showed a 25% decrease in disparity range and a 71% increase in disparity frequency 20 ms after this time point relative to 20 ms before. Panel (b) was redrawn from Menz and Freeman (2003) with permission. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

the same cells to white noise did not show this whitening effect, suggesting that the LGN cells are optimized to convey efficient information about natural scenes.

Importantly, this effect was evident in the responses collected over several tens of minutes, and not evident in the shorter subsets of response. As such, it is unknown whether comparable whitening effects can take place over a few hundred milliseconds (and a correspondingly scaled temporal frequency range).

It is also unclear whether LGN cells are hard-wired to whiten natural stimuli to begin with, or whether this property somehow adaptively evolves over the course of the movie or the experiment. However, there is evidence that cells in the cat primary visual area (striate cortex or area 17, homologous to monkey V1) dynamically adapt to visual stimuli. Sharpee et al. (2006) have reported that the responses of many cells in the striate cortex of the anesthetized cat adaptively change to a given set of stimuli, over the course of 40 s to many minutes, so as to maximize the information conveyed about the stimuli. The main effect of this adaptive change is to increase the sensitivity of the cells to underrepresented spatial frequencies. In other words, these cells act like a set of filters that adapt to optimally process the visual input.

3.3. Disparity tuning in the primary visual cortex

As noted in Section 2.8, many computational and psychophysical studies indicate that the extraction of disparity proceeds in a coarse-to-fine fashion, although there is also some evidence for the opposite, fine-to-coarse, scenario. Recent notable studies by Menz and Freeman (2003, 2004) elucidate the underlying neural mechanisms. Menz and Freeman estimated the disparity tuning curves of individual cells in the cat striate cortex using the reverse correlation technique (for a review, see Ringach and Shapley, 2004). This method yields one tuning curve for each given delay between the stimulus train and the spike train it elicits. Menz and Freeman calculated the tuning curve for each delay between 0 and 200 ms in 5 ms increments, and designated the tuning curve with the most pronounced modulation (specifically, the largest root mean squared signal strength) as the curve at the optimal delay (Fig. 7). The relationship between the optimal delays on the one hand and the firing rate on the other (*e.g.*, whether the optimal delay occurred before, during or after the onset transients) is unclear, although this does not weaken the analysis itself.

For each given neuron, Menz and Freeman then examined tuning curves 20 ms before and after the optimal delay, and measured the changes in the disparity frequency (a measure of disparity resolution) and in the disparity range (a measure of the range of disparities that the cell responds to) as illustrated in Fig. 7a. They found that, over the 40 ms time span centered on the optimal delay, the disparity frequency narrowed for nearly 80% of the cells, and broadened for none. On the other hand, the picture was somewhat more equivocal by the disparity range measure: for about 38% of the cells, the disparity range decreased, indicating a coarse-to-fine change. But for about 7% of the cells, disparity range underwent an opposite, or fine-to-coarse, change. Furthermore, some cells underwent a coarse-to-fine change by one measure and a change in the opposite direction by the other measure (see Fig. 3b of Menz and Freeman, 2003), illustrating the complexity of the temporal dynamic changes and the attendant challenges of characterizing these changes. These trends were also generally true for the other analyses of this data set carried out by Menz and Freeman (2003, 2004); also see Ringach (2003).

3.4. Processing of other low-level image characteristics

3.4.1. Luminance and luminance contrast

Individual cells in the cat LGN can adjust their integration time and gain to the local luminance and local contrast of the stimulus (Mante et al., 2005). These adaptive changes in LGN occur extremely rapidly, within a few tens of milliseconds, unlike those described in Section 3.2, which take closer to minutes or hours. Importantly, the gain control mechanisms for luminance and for contrast operate largely independently of each other. Mante et al. show that the two parameters are also largely independent in natural stimuli, and therefore the gain

control mechanisms in LGN are well suited for processing natural stimuli.

Albrecht et al. (2002) have reported that in the primary visual cortex of both monkeys and cats, the contrast response function of individual cells generally scales and shifts over time. In fact, such scaling and shifting accounts for about 95% of the response variance of a given cell. They also found that key non-linearities that help adapt the neuronal responses to changing stimulus conditions during natural vision, such as contrast gain control and response expansion, are evident within about 10 ms after the cell starts responding to the stimulus (*i.e.*, within a few tens of milliseconds after the stimulus onset).

3.4.2. Orientation

There is some debate about whether orientation tuning in the primary visual cortex follows a coarse-to-fine tuning pattern. Ringach et al. (1997) studied the temporal dynamics of orientation tuning of V1 cells in anesthetized monkeys using the reverse correlation technique. They found that orientation tuning first develops 30–45 ms after the stimulus onset, and persists until 40–85 ms thereafter. The tuning generally becomes sharper over time, and this coarse-to-fine change is more pronounced for output layers (layers 2, 3, 4B, 5 or 6) than for the layers that receive direct input from LGN (4Cα and 4Cβ). Importantly, the observed temporal dynamic patterns could not be accounted for by simple feed-forward inputs, but can be readily accounted for by feedback circuits.

Many recent studies, using somewhat different experimental paradigms and analytical techniques, have reported similar findings (Ringach et al., 1997; Volgushev et al., 1995; Shapley et al., 2003; Xing et al., 2005; also see Chen et al., 2005). However, some other studies have reported that orientation tuning remains largely unchanged over time (Celebrini et al., 1993; Gillespie et al., 2001; Müller et al., 2001; Mazer et al., 2002; Sharon and Grinvald, 2002). While the reasons for the discrepancies between the two sets of studies is beyond the scope of this review, it is worth noting that this debate, often exemplary in its technical rigor, serves to highlight the complexities of the temporal dynamic phenomena and the problems of studying them.

3.4.3. Spatial frequency

The spatial frequency tuning of macaque V1 cells tends to be generally broad at the start of response, and become shaper, *i.e.*, more selective for specific spatial frequencies, over the course of the next 100 ms or so (Bredfeldt and Ringach, 2002; Mazer et al., 2002; Frazor et al., 2004). The preferred spatial frequency tends to shift from low to higher frequencies over the course of about a 100 ms so after the stimulus onset. In addition, the response to low-spatial frequencies tends to be suppressed below the background levels to a greater extent later in the response. Spatial frequency tuning in the cat area 17 and area 18 shows a similar coarse-to-fine change (Frazor et al., 2004; Nishimoto et al., 2005; Allen and Freeman, 2006).

Malone et al. (2007) have recently shown that this coarse-to-fine temporal dynamic pattern could arise by temporal changes in the receptive field size of V1 simple cells. In turn, such patterns in V1 can arise from the observed patterns of center-surround delay of individual LGN neurons and from convergent input from multiple LGN cells with different receptive field sizes and response latencies (Allen and Freeman, 2006).

The fact that the early *vs.* late responses in V1 are dominated, respectively, by the low- *vs.* high-frequency contents of the image does lend support to the spatial frequency-based models of coarse-to-fine processing outlined earlier (Section 2.7). However, many of the more specific issues remain unresolved, including whether the time-course and the magnitude of the changes in V1 can fully explain the dynamics at the perceptual level (see, *e.g.*, Malone et al., 2007).

3.5. Processing of shape characteristics

Hegdé and Van Essen (2004) examined the temporal dynamics of shape coding in visual area V2 of awake, fixating macaques using static 2D shape stimuli presented in the cell's CRF. The stimuli consisted of not only the conventional oriented bars and sinusoids, but also more complex line stimuli and non-Cartesian gratings that represent potential shape and texture primitives (Fig. 8c–e; see Hegdé and Van Essen, 2004, for details). They found that, during the response transients, individual V2 cells are responsive to most shape stimuli (see, *e.g.*, Fig. 8c), so that the cell's shape selectivity is rather broad, or the sharpness of its shape 'tuning' is low, during this period (Fig. 8b). After the transients, the response becomes sparser, in that the cell responds well only to a few stimuli, and the responses to the remaining stimuli decay to near-background levels, so that the cell's shape selectivity sharpens (Fig. 8d and e). That is, sparsening effectively makes given cell's responses more explicitly selective to specific shapes. In addition to such sparsening of the cell's responses across the stimuli, responses can also sparsen over time along other response dimensions, including across the population (Box 3; also see Vinje and Gallant, 2000, 2002). Thus, response sparsening appears to be a widespread temporal dynamic phenomenon in the visual cortex (also see Section 3.6, Box 3).

The above analysis solely considers the cell's shape *signal*, or the variation in the cell's response from one stimulus to the next. In this case, the cell's signal varies in a coarse-to-fine fashion over time. But a more complex picture emerges when one takes into account the response *noise*, or the random variation in the cell's response to the same given stimulus from one trial to the next. Hegdé and Van Essen found that the signal-to-noise ratio, which is a measure of the amount of shape information conveyed by a given cell, follows a roughly opposite pattern as the aforementioned tuning sharpness measure, in that the ratio is maximal during the transients and decays thereafter. Importantly, while the signal-to-noise ratio decreases substantially after the transients, it nonetheless remains statistically significant throughout the remainder of the response, up to at least 300 ms after the stimulus onset.

The fact that initial transients are the most informative phase of a cell's response is somewhat surprising, since neurons tend to respond well to most stimuli during this period (see, *e.g.*,

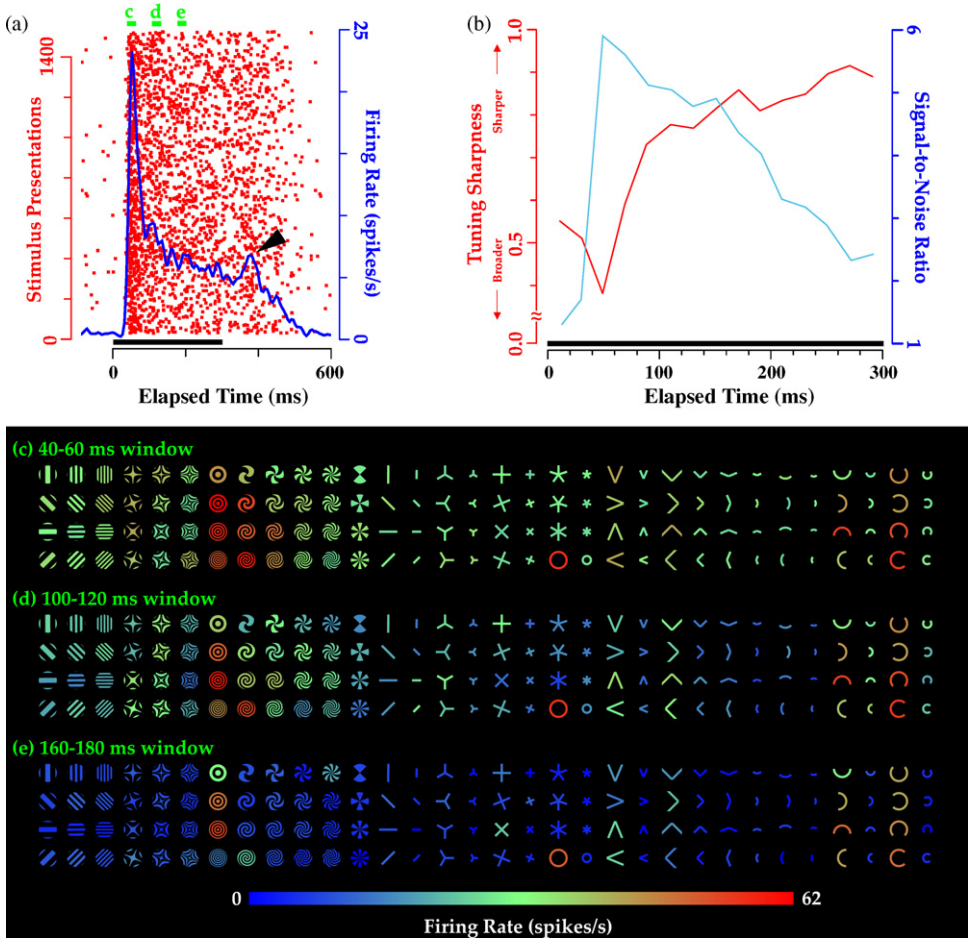


Fig. 8. Temporal changes in 2D shape selectivity of a neuron in macaque V2. (a) Temporal dynamics of the firing rate. Static 2D texture and shape primitives (shown in panels (c–e)) were presented within the cell's CRF for 300 ms (*thick black horizontal line* in panel (a)). Each line of dots in panel a represents the spikes fired by the neuron during a 600 ms interval spanning the given presentation of a given stimulus. The *blue line* represents the peristimulus time histogram. As is typical for visual cortical cells presented with a static stimulus, the cell initially responds with brief period of brisk firing ('onset transient') followed by a more sustained, but lower, firing rate. The *arrowhead* denotes the off-response. (b) Temporal dynamics of two different metrics of the cell's shape selectivity. Note that the Tuning Sharpness (*left axis*) shows a coarse-to-fine temporal pattern, whereas the signal-to-noise ratio (*right axis*) shows a roughly opposite pattern over this time range. Measures of sparseness (not shown) follow the same general temporal pattern as Tuning Sharpness. (c–e) Responses of the cell to the various shape stimuli during the transient (panel (c)) and at two different time bins after the transient (panels (d and e)). The responses are color-coded in a heat map format according to the *color scale at bottom*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Fig. 8c). Since the information conveyed by a given cell depends on the extent to which its responses vary from one stimulus to the next, it would therefore appear that these unselective responses are *less* informative during the transients. But, as first pointed out by Müller et al. (2001), the reason why the information conveyed by the cell is maximal during the transients is that the noise is much lower relative to the signal. While both the signal and the noise decay after the transients, the signal decays somewhat faster than the noise, thus decreasing the signal-to-noise ratio (Hegdé and Van Essen, 2004, 2006). In other words, because of this complex interplay between the signal and noise, the apparently contradictory pictures of shape selectivity emerge depending on whether one considers the signal alone, or both the signal and the noise.

At the population level, responses in V2 show a different type of coarse-to-fine change over time. During the transient response, the V2 population response is largely correlated or redundant, in that three different groups of stimuli (color-coded as *red*, *green* or *blue* stimulus clusters in Fig. 9) elicit different population responses. That is, during this period, most V2 cells tend to respond similarly to all the stimuli within each stimulus cluster, but differently from one cluster to the next. Thus, the population response during the transients is better able to categorize the stimuli into broad groups, rather than distinguish among the individual stimuli. After the transients, the population response gradually decorrelates, so that the cell-to-cell variation in the response pattern increases. Since different cells tend to have different response patterns at this stage, the population as a whole is better able to distinguish between individual stimuli, including within the grating stimulus group or either line stimulus group. Thus, the post-transient population response carries finer-grained information about individual stimuli. Although there are many plausible mechanisms by which the population response may decorrelate, the actual mechanism appears to be a differential decrease in, or a sparsening of, the responses across different

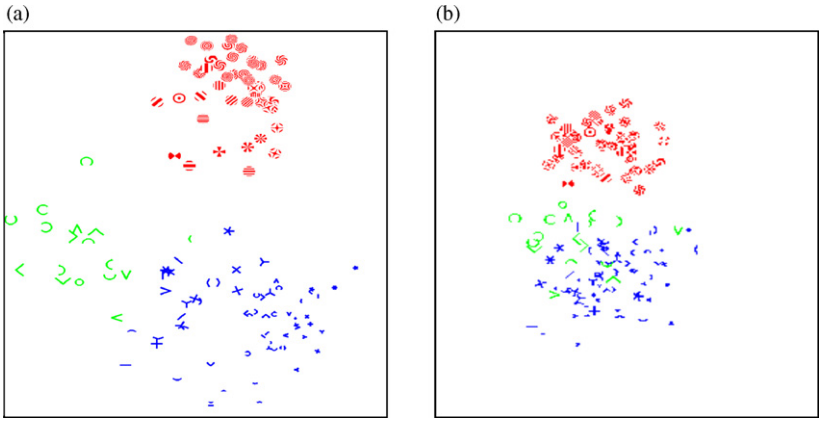


Fig. 9. Coarse-to-fine changes in the V2 population response. Panels (a) and (b) denote the population response during and after the onset transients, respectively, as visualized by multi-dimensional scaling (MDS). MDS plots the stimuli so that the stimuli that elicit similar responses from one V2 cell to the next are clustered together, and the stimuli that elicit dissimilar responses from different cells are dispersed correspondingly father apart. During the transient responses in V2 (40–60 ms time window, panel (a)), three such response patterns are evident, as denoted by the corresponding stimulus clusters, color-coded in this figure in *red*, *green* and *blue*. The clustering in panel a means that most V2 cells responded similarly to the stimuli within each given cluster, and the responses of the cells tended to change collectively from one cluster to the next. This means that the population response was effective in distinguishing among the stimuli that belonged to different clusters, but less effective in distinguishing among the stimuli that belonged to the same cluster. This is a form of efficient coding, the visual system only need sample the responses of very few cells to extract most of the information. After the transients (280–300 ms time window, panel (b)), the response patterns diverged (or decorrelated) from one cell to the next, so that the within-cluster *vs.* across-cluster distinction became blurred. This means that the population response was better able to distinguish among the various stimuli, regardless of whether they belonged to the same or different clusters. Reproduced from Hegdé and Van Essen (2004) with permission. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

cells and stimuli (see Box 3; see Hegdé and Van Essen, 2004, 2006, for details).

This overall response pattern of V2 cells to 2D shape stimuli, both at the level of individual cells and of the population, has also been reported for 2D shape stimuli in macaque V4 and V1, and for 3D stereo stimuli in V4 (Hegdé and Van Essen, 2006). In addition, comparable response patterns have been reported for face stimuli in the macaque IT (Sugase et al., 1999; Matsumoto et al., 2005; also see Section 3.7). Therefore, the aforementioned distinction between the responses during *vs.* after the transients appears to be widespread in the visual cortex, both at the level of individual cells and of the population.

Brincat and Connor (2006) studied the temporal dynamics of shape selectivity in macaque posterior IT. They used relatively simple straight and curved line stimuli (such as arcs, angles, etc., at various curvatures and orientations, etc.), and various multipart complex stimuli created by appropriately combining the simple stimuli, such as two arcs overlaid on each other to make a more complex shape. They explored whether or not the response to the complex shapes could be explained as simple weighted sum of the corresponding simple stimuli, *i.e.*, whether response to a given complex stimulus was a linear function of the responses to its building blocks or not. The former scenario represents a linear response, and the latter represents a non-linear response. Across IT cells they studied, the linear response strength reached 90% of maximum at about 120 ms after the stimulus onset, whereas the non-linear response evolved later, reaching 90% of maximum at about 184 ms after the stimulus onset. Using computer simulations, they showed the gradual linear-to-non-linear transformation can be explained by a recurrent input from neurons within IT with dissimilar selectivity patterns. Putative feedback connections from higher areas alone were not sufficient to produce this pattern; putative lateral (or local) connections within IT had to be taken into account. This suggests that the response of IT neurons to a given complex shape stimulus evolves over time by means of a recurrent network that in effect compares the response of many individual IT cells to the building blocks of the given complex shape.

3.6. Temporal dynamics of center-surround modulation

A large majority of temporal dynamic studies at the neuronal level have focused on the responses to stimuli presented within the CRF. But understanding the temporal dynamics of the nCRF is important because, after all, it too receives visual stimulation and contributes to the percept under natural viewing conditions. The temporal dynamics studies of nCRF have focused largely on early- to intermediate-level visual areas such as V1, V2, and V4, where nCRFs are manageable in size. Particularly noteworthy among these studies are those that deal with the temporal dynamics of scene segmentation and response sparsening.

Border ownership is a prototypical scene segmentation effect. When an oriented edge is presented in the CRF of a V4 neuron, in many cases the neuron is more responsive to the edge if the edge is a part of a larger figure, such as a closed rectangular surface, with the other three edges located in the nCRF (Zhou et al., 2000). The cell typically responds less if the edges do not form a closed rectangle, even when the edge located in the CRF is the same. Thus, the nCRF can convey information about whether the given edge (or border) belongs to the figure or the background. Such 'border ownership' effects are also evident, although progressively less pronounced, in V2 and V1.

This phenomenon has a revealing temporal component: the distinction between the figure *vs.* ground is not evident during the initial response transients, but emerges afterwards. That is, the figure-ground distinction is not apparent during the initial, feed-forward sweep of cortical processing, but emerges with successive passes mediated by recurrent processing (Zhou et al., 2000; Lamme and Roelfsema, 2000; also see Roelfsema et al., 2004; Roelfsema, 2006).

Related scene segmentation effects mediated by nCRF, such as relative depth, da Vinci stereopsis, and texture segmentation, also have comparable temporal dynamics (for reviews, see Roelfsema et al., 2004; Lee and Yuilee, 2006; Roelfsema, 2006; Scholte et al., 2006). In case of texture segmentation in V1, there is also some evidence that figure *vs.* ground differential activity of individual neurons reflects the figure-ground percept itself, and not just the corresponding physical configuration of the stimulus. Moreover, this activity appears to be mediated by feedback connections from other areas, rather than by lateral connections with neighboring cells from within V1 (Supèr et al., 2001; also see Scholte et al., 2006).

Many of the above studies do not explicitly take into account response noise (and, in some cases, cell-to-cell response variation). Therefore, it remains possible that the figure *vs.* ground differential effects will be apparent even during the response transients when the noise is taken into account (*cf*. Section 3.5). This possibility notwithstanding, there is little doubt that recurrent processing mediated by nCRF plays a key role in scene segmentation.

Another nCRF-mediated phenomenon involving recurrent processing is that of response sparsening. Vinje and Gallant (2000, 2002) have shown that when natural stimuli are presented to both the CRF and the nCRF of V1 cells, the responses become sparser (*i.e.*, more selective for certain stimuli) as the total size of the stimulus gets larger, *i.e.*, as more of the nCRF is stimulated. This sparsening tends to become more pronounced over a period of a few seconds. Many different information theoretic measures of the cell's response, including entropy, show this overall trend. Vinje and Gallant also show that entropy increases over the course of stimulation due to a differential increase in total entropy relative to noise entropy. The precise nature of the interplay between feed-forward and recurrent processing that results in response sparsening, however, is largely unclear (but see Chen et al., 2005; Smith et al., 2006).

Barlow (1994) has proposed that a key function of the center-surround modulation is to spatially decorrelate the neuronal responses. Indeed, Barlow argues that a chief task of the neocortex is to remove the correlations in the sensory input "that has already been identified through past experience" (p. 20), and that neocortical areas, specifically the center-surround interactions therein, have evolved to accomplish this. Of course, the latter hypothesis is far harder to test experimentally, and neither hypothesis seems to have been tested so far.

3.7. Face processing in the macaque IT

Responses of face-selective neurons in the macaque IT also show a coarse-to-fine temporal progression at the level of individual cells and of the population. Sugase et al. (1999) examined the time course of face representation by individual IT cells in awake, fixating macaques. Their stimulus set consisted of 38 stimuli, made up of 3 different human faces with 4 different expressions each, 4 different monkey faces with 4 different expressions each, and 10 different geometric shapes. A representative subset of these stimuli is shown in Fig. 10 (inset). The three stimulus types (human faces, monkey faces, and geometric stimuli) constituted the global categories, and the four fine categories comprised of human identity, human facial expression, monkey identity, and monkey facial expression. The authors presented the stimuli for 350 ms, and measured the information transmission rate for the global- and fine categories during each given 50 ms sliding window between 50 and 500 ms. They found that the responses during the transients conveyed significant information about the global categories, whereas the information about the fine categories evolved an average of 51 ms later, during the post-transient sustained response. Thus, the post-transient responses conveyed significant information both about global and fine categories, whereas the transient responses conveyed significant information only about the global categories.

This global-to-local change was evident in a relatively small subset of IT neurons. Of the 1874 IT neurons studied by the authors, only 158 (8%) were selective for face stimuli, of which 86 neurons were further studied. Of these 86 neurons, 32 (37%, or 1.7% of the IT neurons examined) showed the global-to-local effect. This is notable for two reasons. First, comparable global-to-local effects in shape processing in the lower visual areas such as V1, V2 and V4 have been found at the population level, but these effects are not evident at the level of single cells, at least as measured by the signal-to-noise ratio (Hegdé and Van Essen, 2004, 2006). But these studies examined all the cells encountered, and did not select cells on the basis of stimulus selectivity, as Sugase et al. (1999) did. Thus, it remains possible that similar effects exist in V1, V2 and/or V4 for at least some shape stimuli. Second, if the global-to-local effect is relatively rare in IT, is it enough for the brain to use it in face recognition tasks?

A recent study by Hung et al. (2005) suggests that it is. They recorded the responses of macaque IT cells to 77 objects that belonged to one of eight categories (human and monkey faces, foods, toys, etc.). They then devised a linear classification algorithm that classifies statistical patterns in the responses, and trained it to learn the object categories and identities from the neuronal responses. They found that the classifier could reliably classify and identify the objects using the responses from as few as ~100 randomly selected IT neurons and using response durations as brief as 12.5 ms. Thus, information about both object category and object identity is widespread in IT but, in any event, only a tiny fraction of the IT cells need be sampled in principle for a small time interval for either task (also see Keysers et al., 2001; Matsumoto et al., 2005). The results were similar regardless of whether the classifier used multi-unit activity, single-unit activity, or local field potentials for the tasks.

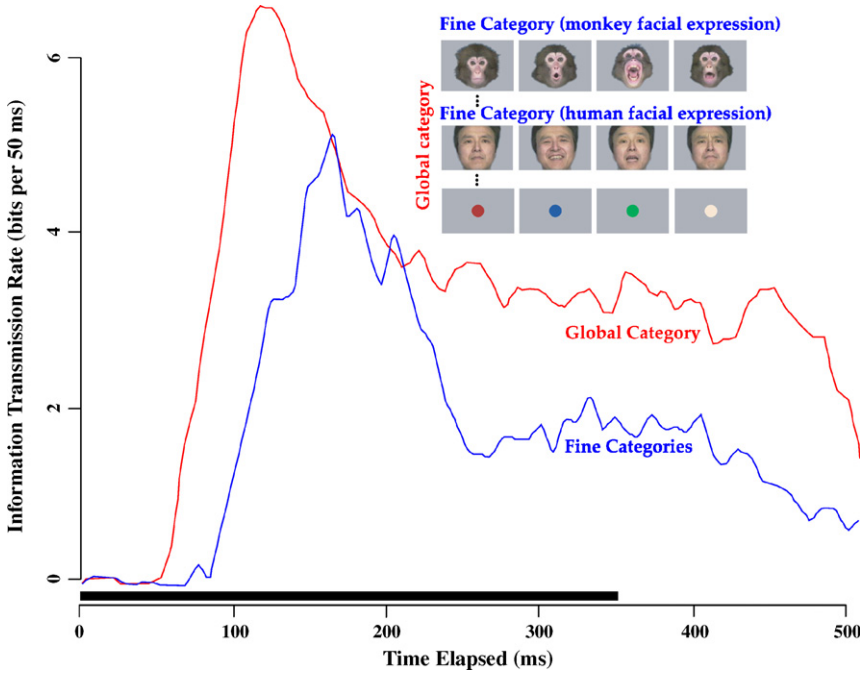


Fig. 10. Coarse-to-fine tuning of shape categories in the macaque IT. The stimulus set consisted of 38 stimuli, a few of which are shown in the *inset*. The global shape categories consisted of human faces, monkey faces, and geometric shapes (*inset*, *vertical axis*). The fine categories (*inset*, *horizontal axis*) consisted of the various facial identities and expressions. Four fine categories were defined: human identity, human expression, monkey identity, and monkey expression. Fine categories were not defined for geometric shapes (*inset*, *bottom row*). The plots show the cumulative information transmission rate of 32 IT neurons that conveyed information about both the global and fine categories (*red* and *blue*, respectively). The *thick horizontal line* denotes the stimulus duration. Adapted from Sugase et al. (1999) with permission. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Three additional results from this study (Hung et al., 2005) serve to highlight the complexities of studying coarse-to-fine representation at the level of single cells. First, neuronal responses as early as 125 ms after the stimulus onset contained enough information to perform both the categorization and the identification tasks. This may be because the algorithm of Hung et al. took into account both the signal and the noise. As noted earlier (Sections 3.5 and 3.6), analyses that take noise into account can produce results different from those that discount noise. Second, the information available for the categorization *vs.* identification tasks had very similar time courses. Thus, across many cells, there was no evidence that the category (or global) information developed earlier than local (identification) information. Third, during the 12.5 ms time bin starting at 125 ms, IT cells fire 0–2 spikes per second on average, or 0.18 ± 0.26 (mean $\pm$ S.D.) spikes/bin. But the classifier performs at about 70% correct using this bin, indicating that individual spikes can be sufficient to carry meaningful object information even in higher visual areas (also see Box 1).

3.8. Face processing: insights from microstimulation

While the above results collectively suggest that the global-to-local temporal dynamic variation could, in principle, mediate the corresponding changes at the perceptual level, they do not by themselves establish such a causal link. But a recent study by Afraz et al. (2006) provides precisely such a link (also see DiCarlo, 2006).

In essence, Afraz et al. show that altering the activity of small clusters of macaque IT cells using microstimulation during *vs.* after the response transients has the corresponding expected effects at the perceptual level. As its name implies, microstimulation alters the activity of a localized cluster of neurons (typically in a region several hundred microns across) by passing a brief pulse (50 ms long in case of Afraz et al., 2006) of external current using a microelectrode (see Tehovnik et al., 2006, for a review). Depending on the parameters, microstimulation can activate or disrupt the collective activity of the target clusters. Afraz et al. activated small, selected clusters of IT neurons while the monkey categorized noisy visual stimuli (presented for 54 ms) as faces or non-faces. They found that monkeys tended to classify the stimuli as faces when face-selective neuronal clusters were stimulated, but not when non-face-selective clusters were. The magnitude of this biasing effect depended on the degree of face selectivity of the stimulation site (*i.e.*, neuronal cluster), and the size of the stimulated cluster. More importantly in the present context, the timing of the stimulation made a big difference. Microstimulation 0–50 ms after the stimulus onset, which is largely before the onset of IT responses (Bruce et al., 1981), made little difference in the behavior of the monkeys. But when the face-selective neurons were microstimulated 50–100 ms after the stimulus onset (*i.e.*, approximately during the response transients), during which face-selective IT neurons mostly convey information about global categories (see above; also see Bruce et al., 1981; Sugase et al., 1999; Kiani et al., 2005), the monkeys' behavioral responses were significantly biased

toward categorizing the stimuli as faces. On the other hand, stimulating the same neurons 100–150 ms after the stimulus onset, during which the neurons convey both global and local information about faces (Sugase et al., 1999), resulted in a face response bias in one monkey but not the other.

It is important to note that Afraz et al. only addressed the global level categorization task (*i.e.*, face *vs.* non-faces) and not a finer level task, such as recognizing facial identity or expression. In this sense, this study does not address the coarse-to-fine processing of face information, but only the temporal dynamics of the coarse-scale task of face detection. Nonetheless, this study is notable both because it is one of the clearest accounts of the temporal changes at the perceptual level in non-human species, and because it compellingly demonstrates the usefulness of the microstimulation technique in establishing a causal link between the temporal dynamics of the behavioral response and of the underlying neural activity (*cf*. Section 2.4).

3.9. Temporal dynamic changes in the dorsal pathway

The dorsal visual pathway (see Figs. 3 and 6) is sometimes referred to as the 'where?' pathway or the action pathway, because it is thought to specialize in processing information about the spatial location of visual objects and mediating visually guided action (Ungerleider and Pasternak, 2004; Goodale and Westwood, 2004). Since visually guided action must happen in real time, information about visually guided action may not evolve over time in the dorsal pathway (Goodale and Milner, 1992; Goodale and Westwood, 2004) as shape information clearly does in the ventral pathway. However, it is increasingly clear that the processing in the dorsal pathway is not limited to spatial processing. Areas in the dorsal pathway represent shape information and visual category information, although perhaps at a much more coarse-grained level (Freedman and Assad, 2006; Lehky and Sereno, 2007; also see Schoenfeld et al., 2003; Goodale and Westwood, 2004).

While not much is known about the temporal dynamics of shape information in the dorsal pathway, there are hints of important similarities and differences with the temporal dynamics of the ventral pathway. Pack and Born (2001) studied the response of macaque MT (middle temporal area) cells using a pattern of line segments drifting coherently such that the direction of the local motion components differed by 45° from the direction of global motion. That is, the parts collectively moved in a different direction than the whole. The neuronal responses were initially strongly biased by the component directions, but evolved over the course of the next 80 ms or so to represent the direction of global motion. Note, incidentally, that one can reasonably consider this an instance of local-to-global (or fine-to-coarse) change in direction selectivity—an issue we will revisit below in the context examining whether coarse-to-fine processing is a common theme of visual processing (Section 6).

Smith et al. (2005) used moving plaid stimuli in which directions of component- and global pattern motion were readily separable. They reported that for many MT cells, the responses initially represent the direction of component (local) motion, but evolve over 100–150 ms after the stimulus onset to represent the direction of the global pattern motion.

Palanca and DeAngelis (2003) measured the disparity selectivity of macaque MT neurons using static and moving disparity stimuli and compared the time course of disparity selectivity in the two cases. For static stimuli, they found that the disparity selectivity peaks around the time of initial response transient, and decreases to a lower asymptotic level afterward, much like the disparity selectivity in V4 (Hegdé and Van Essen, 2006) and the selectivity for many other static shape stimuli as described above (Section 3.5). But Palanca and DeAngelis found that for moving stimuli, disparity selectivity in MT does not show the post-transient decay, but instead remains relatively high throughout the stimulus presentation. The lack of response decay for motion stimuli may simply reflect steady motion energy flux throughout the stimulus duration (Hegdé and Van Essen, 2006).

The extent to which the dichotomy of response patterns for static *vs.* moving stimuli is widespread in the dorsal pathway – or, for that matter, in the ventral pathway – remains to be determined. Nonetheless, the scenario where tuning property of a given neuron changes for static stimuli and not for moving stimuli has a potentially important computational consequence. Many perceptual judgments about moving stimuli are explained well by models in which the visual system straightforwardly integrates the sensory information until a decision threshold is reached (see, *e.g.*, Palmer et al., 2005). The value of the decision threshold depends on the tradeoff between the speed *vs.* accuracy of decisions. However, if the amount and the nature of information conveyed by the neurons itself changes, decision-making will have to account for these factors as well. How the system accomplishes this is unclear.

3.9.1. Neuronal temporal dynamics in the Bayesian framework

Little is empirically known about the neural mechanisms of vision as Bayesian inference and even less about visual temporal dynamics in a Bayesian framework. But it is worth noting that computational and modeling studies show that Bayesian inference can be straightforwardly implemented in terms of the previously known (or postulated) mechanisms of neural coding (Glimcher, 2005; Doya et al., 2006). Essentially, this means that probability distributions on which the visual system must base its inference can be coded rather faithfully at the level of the neuronal population. In this sense, the plausible Bayesian explanations are similar to those at the perceptual level outlined above—the neural population essentially acts as the observer. The whole brain imaging results described in Section 4 below can also be conceptualized in the same fashion using the Bayesian framework.

3.10. Summary of single-unit studies

Many response characteristics of visual neurons tend to be different during *vs.* after the onset transients. Many, but by no means all, response characteristics of individual neurons and

neuronal populations tend to change in a coarse-to-fine manner over the course of the response. In some cases, whether or not a given response characteristic appears to be processed in a coarse-to-fine fashion depends on the analytical method, *e.g.*, on whether response noise was taken into account. It is unclear whether the many apparent coarse-to-fine processing phenomena are superficially similar or are fundamentally related. Nonetheless, coarse-to-fine processing may be a prevalent, although not universal, temporal dynamic theme in many visual areas. Microstimulation studies show that, at least in case of face processing in IT, the temporal dynamics of the neural responses may have a causal relationship with the corresponding changes at the perceptual level. Response sparsening, likely due to recurrent processing, appears to be an important and widespread mechanism of sharpening the response selectivity both of individual neurons and of neuronal populations.

4. Spatial aspects of temporal dynamics: insights from whole brain imaging studies

4.1. A brief overview of whole brain imaging techniques

While single-unit studies offer unmatched spatial and temporal resolution, they, by practical necessity, tend to focus on one area at time and typically on one neuron at a time. Thus, it can be hard to see the forest for trees from these studies.

The advantage of whole brain imaging techniques is that they offer a comprehensive view of the forest, are typically non-invasive, and can be carried out using human subjects that can readily perform complicated tasks and report semantically nuanced percepts. On the other hand, whole brain imaging techniques have their own sobering limitations. EEG, the least expensive of the whole brain imaging methods, offers millisecond-level temporal resolution, but the spatial resolution of the ERPs is poor for signal sources on the surface of the brain, and progressively poorer at deeper levels (see Andreassi, 2006; Luck, 2005, 2006). Magnetoencephalography (MEG) signals are referred to as event-related magnetic fields (ERMFs) or, in case of visual stimuli, visually evoked fields (VEFs). ERMFs are comparable to ERPs in terms of temporal resolution but have a decidedly greater spatial resolution (for reviews, see Hämäläinen et al., 1993; Andreassi, 2006). Moreover, MEG is sensitive to only a subset of the neural activity that can be detected by EEG (Hämäläinen et al., 1993; Andreassi, 2006). The spatial resolution of the fMRI signal, or the blood oxygenation level-dependent (BOLD) response, tends to be on the order of a millimeter or so, and is better than ERPs or ERMFs in this respect. But with a typical temporal resolution of a few seconds (Jasanoff, 2005), BOLD signal would seem almost ideally unsuited for studying millisecond-level temporal dynamic phenomena. Nonetheless, as we will see below, much has been learned by judiciously combining the various whole brain imaging techniques with each other and relating the results with psychophysical and single-unit studies (see Logothetis et al., 2001; Bar et al., 2006; Schmid et al., 2006).

4.2. Visualizing visual processing

The aforementioned ERP study of ultra-rapid visual processing by Thorpe et al. (1996) is notable for another reason. This study represented one of the clearest early demonstrations of the utility of methods like EEG in visualizing the spatiotemporal pattern of visual processing. It provided what in effect are time-lapse pictures of the ERPs sweeping through the brain over a 400 ms period, somewhat like a 'standing wave' in a sports arena (see Fig. 2a and b of Thorpe et al., 1996). Of course, such ERP 'maps' cannot be directly interpreted as brain activity maps, given the uncertainties involved in assigning the ERPs at a given surface location to a given source location in the brain (Andreassi, 2006; Luck, 2005, 2006). But even with this caveat, the original study by Thorpe et al. (1996), and many subsequent ones like it (*e.g.*, Johnson and Olshausen, 2003, 2005a), provide revealing spatiotemporal views of visual processing.

4.3. Spatiotemporal patterns of response persistence

As the retinal image changes, how does the activity in different parts of the brain change with it? Mukamel et al. (2004) presented human subjects with stimuli of various visual objects, such as animals, houses, faces, etc., at either 4 Hz or 1 Hz, and studied the time course of the BOLD response in various visual areas at the two presentation rates. They found that the four-fold increase in the presentation rate (from 1 to 4 Hz) results in a two-fold (200%) increase in the BOLD response in the early visual areas and the motion sensitive area MT/V5, but only resulted in a 25% response increase in higher visual areas, such as those in the fusiform gyrus. Mukamel et al. suggest that a likely explanation for this phenomenon is that the stimulation effects persist longer in the higher visual areas and decay rapidly in lower visual areas. That is, temporal dynamics of the response is more image-dependent in the lower visual areas and less so in the higher areas. While other studies support this notion, they also find that the temporal dynamic patterns of the BOLD response vary a good deal depending on the stimulus, imaging technique, attentional control, task and other behavioral parameters (Kourtzi and Huberle, 2005; Carlson et al., 2006; Philiastides et al., 2006).

Because relationship between the neuronal responses and the BOLD responses may, in principle, vary depending on the area, these results cannot necessarily be taken as proof that the stimulus-driven effects persist longer in higher visual areas than in lower, retinotopic areas. Nonetheless, it is worth noting that these results are consistent with results obtained in other whole brain imaging and assorted monkey neurophysiological studies (Rolls and Tovee, 1994; Rolls, 2004; Keysers et al., 2001; Kourtzi and Huberle, 2005; Carlson et al., 2006). Thus, it seems reasonable to conclude that the stimulus-driven effects are more transient in early visual areas than in higher visual areas. This is important, among other things because the fact that the visual information persists longer in higher visual areas provides a potential explanation for why processing duration is more important than the stimulus duration in backward masking

experiments (see Section 2.1). In other words, the persistent activations might serve as a short-term (iconic) memory mechanism for preserving a trace of the stimulus even in its absence (Mukamel et al., 2004).

4.3.1. Responses of object-selective regions during object recognition

Grill-Spector et al. (2000) have shown that the BOLD responses of the object-selective human brain regions are correlated with the performance of the subjects in correctly categorizing objects at a basic level (*e.g.*, 'flower', 'boat'). As expected, subjects recognized all objects correctly at a processing duration of 500 ms. Subjects performed nearly as well when the duration was 120 ms. But when the duration was reduced to 40 ms, the subjects performed at an average of 18% correct, and when the duration was reduced 20 ms, the performance was near zero. Thus, the performance was a highly non-linear function of the processing duration. Notably, the BOLD response in high-level object-selective regions, including LOC and DF, paralleled the performance of the subjects. By contrast, the BOLD response in V1 was much less sensitive to the performance or the processing duration. Moreover, after the subjects received training resulting in an enhanced performance, the LOC responses increased accordingly. In addition, LOC responses were larger for trained than for novel stimuli. Thus, the response patterns of LOC and other high-level regions, but not the response patterns of lower retinotopic regions such as V1, reflected object recognition.

4.4. Spatiotemporal dynamics of face processing

Liu et al. (2002) used MEG to examine the spatiotemporal patterns of face processing in human subjects. They found bilateral regions in the occipitotemporal cortex that were significantly more responsive to faces than to houses 100 ms after the stimulus onset (M100). They also found a different bilateral ERMF pattern that peaked 170 ms after the stimulus onset (M170). The regions of activation of the M100 *vs*. the M170 had distinct, albeit overlapping, spatial locations. Moreover, the M100 response was associated with correct categorization of stimuli as faces, but not with successful identification of individual faces. M170 was associated with both. Taken together, these MEG results reaffirm the notion that face processing proceeds in a coarse-to-fine fashion: initial face categorization, followed by face identification.

It is important to note that although these results are manifestly similar to those in macaque IT (Section 3.7), the two sets of results are not directly comparable, for two reasons. First, it is unclear whether the region of activation for either M100 or M170 in humans is homologous (in the phylogenetic sense), or even functionally analogous, to macaque IT. Second, the fact that categorization responses in both cases have a latency of 100 ms does not mean that the two brain regions are functionally analogous, since cortical response latencies are generally shorter in macaques than in humans (Section 3.1). Similarly, although the M170 ERMF and the N170 ERP in humans both have the same latencies and are both selective for face identity (but see Rousselet et al., 2004), there is no *a priori* reason to believe that they represent the same underlying neuronal response (Liu et al., 2002).

4.5. Imaging studies of spatial frequency-based coarse-to-fine processing

As noted in Section 2.7, many models of coarse-to-fine processing postulate that the visual system first extracts a low-spatial frequency-based 'gist' of the scene, and is followed more slowly by the higher spatial frequency content, which provides finer-grained spatial information. The coarse grained information is used to constrain the interpretation of the information on a finer spatial scale (Fig. 5; Bar, 2003, 2004; Peyrin et al., 2005). Previous fMRI studies have shown that the object-selective regions in the temporal cortex and, somewhat unexpectedly, the regions in the orbitofrontal cortex (OFC) are preferentially activated when a given object is recognized compared to when the same object is not recognized (Bar et al., 2001).

In a more recent study, Bar et al. (2006) used a combination of fMRI and MEG to compare the time courses of the MEG responses in the OFC and the fusiform gyrus during trials in which subjects reported recognizing *vs*. not recognizing a given set of line drawing objects. They found that differential responses to recognized *vs.* unrecognized stimuli peaked in the left OFC 130 ms after the stimulus onset, 50 ms before it peaked in the fusiform gyrus. But is the response in OFC correlated with the low-spatial frequency content of the image, as the Bar (2004) model predicts?

To test this, Bar et al. compared the response *magnitude* for low-pass *vs.* high-pass images of real-world objects since, for technical reasons, it is hard to directly compare response *polarity* (*i.e.*, activation *vs.* deactivation) using MEG (Bar et al., 2006, p. 452). They found that the responses to both sets of stimuli were suppressed below the baseline levels (*i.e.*, were negative) in OFC, but the response to the high-pass stimuli were more suppressed. Moreover, the suppression was comparable for low-pass *vs.* intact, unfiltered images. The significance of this response polarity (*i.e.*, suppression) is unclear, but Bar et al. (2006) point out that the differences in the response magnitudes does mean that the ''early OFC activity is differentially sensitive to spatial frequencies'' (p. 452). They also show that MEG responses covaried among OFC, fusiform gyrus and the early visual areas, and this covariance, or phase synchrony, was stronger for low-pass images, suggesting that the connectivity between the three sets of brain regions is stronger for low-pass images.

While these results are consistent with the Bar model, it is important to note that many key features of the model remain untested. These include, but are not limited to, the issue of whether or not what OFC generates are image interpretations *per se*, whether the function of the fusiform gyrus depends on the input from the OFC, and whether the apparent rapid extraction of the low-frequency information is mediated by the dorsal pathway as the Bar model posits. Moreover, it is also unclear the extent to which these effects depend on the stimuli

used, to the spatial frequency contents of the stimuli, and to the methods used to determine response latencies.

The overall strength of the Bar model is that it highlights the importance of recurrent processing in constraining the interpretation of the bottom–up information. But the model also has its weaknesses, two of which must be noted here. First, it is unclear whether the low-frequency images used by Bar et al. (2006), for instance had the same overall luminance energy as the high-frequency images. Roughly speaking, this means that the image pixels may have been brighter (or, less likely, darker) on average in low-frequency images than in their high-frequency counterparts. If true, this potentially serious confound may have contributed to the differential responses to the two sets of images. Second, the model posits that the ''dorsal magnocellular pathway'' mediates the rapid arrival of the low-frequency information to the prefrontal cortex (see, *e.g.*, Bar et al., 2006, p. 449). This is erroneous on many counts. (i) The notion that the dorsal pathway is a predominantly magnocellular pathway may be widespread, but it is untrue (Callaway, 2005; Sincich and Horton, 2005). (ii) The notion that the magnocellular pathway selectively carries low-spatial frequency information is also untrue (Kaplan, 2004, p. 484). (iii) Although magnocellular responses are somewhat faster than parvocellular responses in some respects, it is far from clear that magnocellular neurons respond faster than parvocellular neurons to lower spatial frequencies (Merigan and Maunsell, 1993, pp. 372–374; Kaplan, 2004). Together, these considerations call into question whether and to what extent the observed faster prefrontal responses are attributable to low-frequency information.

4.5.1. Evidence for lateralization of spatial frequency processing

The aforementioned preferential responses in the left OFC to low-pass images is one of the many findings that suggest that there is some hemispheric asymmetry in the coarse-to-fine processing of spatial frequency. Peyrin et al. (2005) investigated this by measuring the BOLD responses using sequences of bandpass images that progressed systematically from low-pass to high-pass, or vice versa. They found preferential responsiveness to the high-pass to low-pass sequence in the left occipitotemporal cortex, and to the opposite sequence (*i.e.*, low-pass to high-pass) in the right occipitotemporal cortex. They interpret this to mean that the left and the right hemispheres are preferentially engaged in fine-to-coarse and coarse-to-fine processing, respectively, of the image (*cf*. Iidaka et al., 2004). More recently, Peyrin et al. (2006) have shown that while this pattern holds for short stimulus durations (30 ms), for longer durations (150 ms), the right hemisphere tends to show greater activation regardless of the spatial frequency content of the stimulus, suggesting that the pattern of hemispheric lateralization changes over the course of about a 100 ms.

While resolving the lateralization issue is important to understanding the functional organization of spatial frequency processing, it is unclear whether the lateralization *per se* has any computational consequences. For instance, it remains possible that the differential brain responses to low- *vs*. high-spatial frequencies are simply a reflection of where the brain processes spatial frequency information, and do not have significant computational consequences. There are other sobering instances of structure without function in the visual system (Horton and Adams, 2005).

4.6. Analyses of functional and effective connectivities

One of the noteworthy insights to emerge from recent imaging studies is that different brain regions show similar temporal patterns of activity during a given task. Briefly, two brain regions are said to be *functionally connected* when the responses in the two regions are mutually correlated in time (for reviews, see Friston, 1994; Penny et al., 2004; B. Horwitz et al., 2005). Of course, if the two regions are selected in the first place because their responses match a third, pre-specified temporal pattern (*e.g.*, a box car function), then the activity in the two regions can be expected to be correlated. But this does not necessarily mean that the two regions are causally related, *i.e.*, the response of one region causes, or is caused by, the response of the other. Two regions are said to be *effectively connected* when their responses are causally related. This means that there is a path connecting the two regions. The connectivity analyses can be carried out using any time-varying neural response data, regardless of the method by which the data are collected, although connectivity studies commonly use BOLD response data.

A recent study by Summerfield et al. (2006) illustrates the current state of the technique. During each given block of the scan, subjects were shown low-contrast, ambiguous pictures of faces, houses and cars in random order, and were required to classify the stimuli in one of two ways depending on the block. During 'face block', the subjects had to classify each given stimulus as a face or non-face. During the 'house block', subjects had to classify similar stimuli as houses or non-houses. Car stimuli, which did not have a block of their own, served as control stimuli. Regions in inferior occipital gyrus (IOG), fusiform face area (FFA), temporo-parietal junction (TP), and amygdala were more responsive to the face stimuli than to the house stimuli, regardless of the block. The responses were larger during the face block than during the house block in dorsal and ventral medial frontal cortex (dMFC and vMFC, respectively), regardless of the physical stimuli. That is, the activity in MFC reflected the predicted percept (depending on whether it was a face- or a house block).

To test whether MFC instructs the face-selective regions what stimulus type to expect depending on the block, Summerfield et al. used a type of connectivity analysis called dynamic causal modeling (DCM). DCM estimates how much of the hemodynamic activity in a given brain region can be accounted for by the stimulus input, experimental parameters (*e.g.*, face stimulus and face block), and the known anatomical interconnectivity between the relevant regions. The analysis found that upon the presentation of face stimuli, feed-forward connections from IOG to FFA and amygdala were selectively enhanced. During the face blocks, feedback connection from vMFC to FFA and amygdala were

selectively enhanced. Therefore, the DCM analysis indicates that the top–down signals from the MFC help modulate the processing of the bottom–up signals, thus adding a dynamic dimension to the brain activations identified by the stimuli and conditions.

Functional connectivity can be also analyzed using analytical methods quite distinct from DCM, such as various types of autoregressive methods (see Valdés-Sosa et al., 2005; Ioannides, 2007, and the references therein). This means that one analytical technique can be used to complement the weaknesses of the other, using the same underlying data set.

4.7. Summary of whole brain studies

Many previous whole brain imaging studies have helped elucidate the spatiotemporal pattern of responses during many key temporal dynamic processes. Techniques from this evolving field will continue to be indispensable in temporal dynamic research, including in delineating the spatiotemporal patterns of brain activity and patterns of functional and effective connectivity.

5. Attention and eye movements: a thumbnail sketch

Under natural viewing conditions, attentional shifts are intricately related with eye movements. However, much of what we know about temporal dynamic effects of attention comes from experiments with eye position controls, typically fixation. Some results from these experiments that help illustrate a few key temporal dynamic effects of attention are briefly outlined below.

5.1. Image-driven vs. goal-driven attention

In general, attention can be image-driven, so that it is allocated to a given part of the scene based either on which part of the image is salient (*e.g.*, comes on suddenly, or has a distinctive color, shape, etc.), or it can be goal-driven, allocated on the basis of which object or image characteristic the viewer is looking for (for reviews, see Yantis and Serences, 2003; Maunsell and Treue, 2006; also see Peters et al., 2005; Zhaoping and May, 2007). Time courses of such image-driven *vs.* goal-driven mechanisms tend to be very different, as do the time courses of goal-driven attention in different tasks. In our context, this is important because it means that the dynamics of processing of the same scene can be different depending on the attentional condition.

Neurophysiological studies of image-driven attention typically use a visual search for an unknown target in an array of relatively simple, geometric shapes. For instance, the subject may be instructed to look for a unique 'odd-man-out' target among a field of various colored letters. ERP studies using such paradigms show that if the array contains the target, neural activity is increased beginning 120–150 ms after the stimulus onset, and spatial attention is directed to it about 25–50 ms later. For various technical reasons, the precise cortical origin of this ERP is unclear. However, comparable effects have been recently found in macaque V4 (Mazer and Gallant, 2003; Ogawa and Komatsu, 2004; Bichot et al., 2005; also see Maunsell and Treue, 2006). ERP and macaque single-unit studies also indicate that, when the stimulus array does not contain a target, attention tends to shift from one object to another in a more or less serial fashion once every 100 ms or so (Luck, 1999, 2006; Bichot et al., 2005; also see Maunsell and Treue, 2006).

Goal-driven attention, also referred to as task-directed attention, can have different time courses depending on whether attention is directed to an image location or to an image feature (spatial and feature-based attention, respectively; see Hayden and Gallant, 2005; Maunsell and Treue, 2006).

5.2. Goal-driven, spatial attention

In these experiments, the subject typically attends to a given spatial location, and visual stimuli are subsequently presented at the attended location and at another, comparable unattended location, and the responses are typically compared for stimuli at the attended *vs.* unattended location. In ERP studies, the first clear-cut differences generally occur in the extrastriate cortex beginning at about 60–100 ms after the stimulus onset, wherein the responses to the attended stimulus are enhanced relative to the response to the unattended stimulus. No such differential effect is discernible in the striate cortex during this time. This is generally interpreted to mean that directed spatial attention selectively enhances the feed-forward transmission of visual information in the extrastriate cortex beginning at about 60–100 ms after the stimulus onset (Luck, 1999, 2006). These findings from whole brain imaging studies are broadly consistent with those from monkey single-unit studies (Luck, 1999, 2006; also see Maunsell and Treue, 2006).

5.3. Goal-driven, feature-based attention

There is some evidence that space-based attention has different dynamics than feature-based attention. For instance, space-based attention can be transiently disrupted by the abrupt appearance of visual stimuli (Nakayama and Mackeben, 1989; Egeth and Yantis, 1997), but it is unclear whether feature-based attention is disrupted in this manner (Lamy and Tsal, 2001).

Hayden and Gallant (2005) compared the responses of macaque V4 neurons during spatial *vs.* feature-based attention. They found that neuronal responses increased when spatial attention was directed toward the CRF of the cell, and were modulated by the identity of the target of feature-based attention. Modulation by spatial attention was weaker during the initial transients, and stronger afterwards. In contrast, the modulation by feature-based attention was relatively constant throughout the response. Hayden and Gallant suggest that stimulus onset transients disrupt spatial attention, but not feature attention, and that spatial attention reflects a combination of stimulus-driven and goal-driven processes, whereas feature-based attention is purely goal driven.

5.4. Attention and eye movements in natural vision

One of the main difficulties in characterizing the temporal dynamic effects of shifting eye position and attention is that these shifts are largely unpredictable during natural viewing. Although we know that salient image locations tend to elicit eye movements to them, the temporal sequence of eye movements is largely unpredictable (Henderson and Hollingworth, 1998; Torralba et al., 2006). There is some evidence that this may not be a problem for characterizing the response dynamics in at least some visual areas. For instance, in macaque IT, the responses are virtually unaltered by free viewing (DiCarlo and Maunsell, 2000; but see Sheinberg and Logothetis, 2001). This raises the prospect that, in the future, it may be possible to characterize the temporal dynamics of natural vision largely in terms of a series of static views of the appropriate parts of the scene. However, it is not known the extent to which this is true for other high-level visual areas.

At the perceptual level, the sequence of eye movements clearly has considerable effect on the dynamics of scene understanding (Henderson and Hollingworth, 1998; Pylyshyn, 2003; also see Najemnik and Geisler, 2005), although it is much less clear whether it matters to the eventual outcome of the scene understanding process. This is one of the key questions in temporal dynamic research, but it is also one of the more difficult ones to address, since doing so would require rigorously comparing the scene understanding resulting from different sequences of a given set of fixations (Box 4).

It is important to note that it is currently altogether unclear how to incorporate attentional effects within the Bayesian framework (Hegdé and Felleman, 2007). For instance, image-driven attention amounts to estimating the image probabilities using selected part/s of the image, as opposed to the whole image as the Bayesian models currently do. But there is every reason to believe that this is an eminently addressable problem. Indeed, models that estimate the saliency of different parts of the image to construct a saliency 'map' of the image (Peters et al., 2005; Zhaoping and May, 2007) are a promising way of modeling the effects of attention, including its temporal dynamic effects, in the Bayesian framework.

6. Coarse-to-fine processing *vs*. Bayesian inference as a framework for understanding temporal dynamic phenomena

This issue is worth delving into because a wide variety of visual temporal dynamic phenomena have been described as either 'coarse-to-fine' or 'global-to-local'. Many studies have used the two terms synonymously (see, *e.g*., Iidaka et al., 2004; Peyrin et al., 2005, 2006), although some others have sought to distinguish between the two concepts (see, *e.g*., Hughes et al., 1996). This article has used the two terms interchangeably, if only because the distinction between them seems arbitrary and not widely accepted. But the more substantive issue is whether coarse-to-fine processing is a useful framework for understanding visual temporal dynamic phenomena. This section will argue that it is not, and that the Bayesian framework is a better alternative.

6.1. The computational appeal of coarse-to-fine processing

Coarse-to-fine processing is a potentially useful information processing strategy from two interrelated computational perspectives. The first is that since coarse-to-fine processing is well-suited for representing the visual world, because the visual world, or more precisely our understanding of it, is hierarchical. As an illustration, consider the following modified version of the parlor game 'Twenty Questions', where the player has to infer, by asking no more than twenty questions, the visual object in a given image without actually seeing the image. What is the best strategy for deciding which types of question to ask first, and which ones later? Clearly, it is efficient to ask about global object categories first ('Is it a person, animal, plant, or an inanimate object?'), and more fine-tuned questions later, based on the answers to earlier questions. This hierarchical strategy is useful, ultimately because our understanding of the visual world is hierarchical. When this is not the case, *e.g*., when the 'objects' consist of a dozen alphanumeric characters scattered on a page, a coarse-to-fine strategy is decidedly non-optimal for guessing them. By contrast, it can be rigorously proven that a Bayesian inference is the *ideal* strategy in *both* cases (see Kersten et al., 2004; Doya et al., 2006). In other words, Bayesian inference is a better strategy in this case because it subsumes, and extends, the coarse-to-fine framework.

Many large scale theories incorporate global-to-local hierarchical representation in a more rigorous fashion (Wersing and Korner, 2003; Amit et al., 2004; Serre et al., 2007; Ullman, 2007). While these studies generally do not deal with the temporal dimension explicitly, the temporal dimension is implicit in them. That is, these models formulate visual processing as a set of serial, typically hierarchical, Bayesian decisions, so that earlier decisions address image data on a coarser (or more global) scale, and the latter decisions deal with finer scale data.

A second sense in which coarse-to-fine processing is computationally useful is that it allows the system to balance speed *vs.* accuracy. When the processing hardware is finite (as it is in the brain), and the input is sufficiently complex (as typical nature scenes are), the amount of information available for a given decision depends on the processing duration, which means that faster decisions tend to be more error prone and making more accurate decisions takes longer. Thus, optimal decision-making requires finding an optimal speed *vs.* accuracy tradeoff. Imagine, in our parlor game example, the task is to guess what is in the picture with the fewest questions possible or, in an equivalent but more explicitly temporal sense, to guess it as quickly as possible. Again, Bayesian inference is the ideal strategy for optimizing speed *vs.* accuracy (Doya et al., 2006).

It should be noted that the above two scenarios are not mutually exclusive. The former is about finding a likely interpretation of the scene, and the latter is about finding it efficiently. Under natural conditions, where visual perception

primarily subserves action, the two are different aspects of the same process. It is easy to appreciate this by imagining that you are playing the above game against another person, and the first person to guess it correctly wins.

6.2. *Caveats about the concept of coarse-to-fine processing*

The potential computational advantages of coarse-to-fine processing aside, it is empirically clear that coarse-to-fine processing, in one form or another, is widespread in the visual system. Indeed, Allen and Freeman (2006, p. 11773) have proposed, in the context of coding at the level of individual neurons, that coarse-to-fine processing is a fundamental coding strategy in the central nervous system. But obviously, a large variety of temporal changes at the population and perceptual levels can also be described as coarse-to-fine. Indeed, this concept is currently vague and flexible enough that it seems to fit many different phenomena with minimal suspension of disbelief. For instance, the changes in the perceptual level (vision-at-glance followed by vision-with-scrutiny) seem to fit this definition, but so does orientation tuning in some cases. If both are valid instances of this process, how are the two related? If they are unrelated, what is to be gained by referring to them by the same label? If one of them better fits the definition, which one and why?

But even the more narrow view that coarse-to-fine processing is a fundamental coding strategy at the neural level is problematic, for many reasons. First, as noted above, not all cells show a clear-cut coarse-to-fine pattern even for given visual parameter and within a given area; some even clearly show the opposite, *i.e.*, fine-to-coarse, pattern (Menz and Freeman, 2003, 2004; also see Gillespie et al., 2001; Mazer et al., 2002). As noted in Section 3.9, direction selectivity evolves a local-to-global fashion for many MT cells. Second, in some cases, such as orientation selectivity in the primary visual cortex, some studies have found a coarse-to-fine change but others have not (Section 3.4). Third, some neural temporal dynamic patterns are too complex to be captured adequately as coarse-to-fine. For instance, in macaque V1, the temporal dynamics of color selectivity in macaqueV1, for which the LGN input mechanisms are reasonably well understood, cannot be meaningfully described as coarse-to-fine (Cottaris and De Valois, 1998; G.D. Horwitz et al., 2005; also see Shapley, 1998). Fourth, as noted in Section 3.5, while some temporal changes do appear to follow a bona fide coarse-to-fine pattern when the response noise is ignored, the picture gets considerably more complex when response noise is taken into account. Finally, as noted in Box 2, whether a given temporal dynamic change can be classified as coarse-to-fine, fine-to-coarse, or something else often depends on which time points are being compared. Thus, unless the relevant parameters of comparison are all specified, coarse-to-fine processing can amount to the proverbial blind man's view of the elephant.

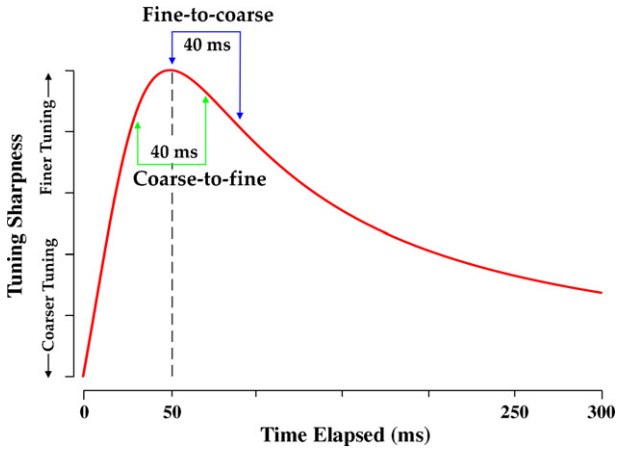
It is also unclear whether coarse-to-fine processing at the level of individual cell level necessarily leads to coarse-to-fine changes at the population level. And, regardless of whether the response readout at the individual cell or at the population level underlies the percept in a given case, the extent to which coarse-to-fine changes in the neuronal response can explain the coarse-to-fine changes in the percept is unclear. In other words, although both perceptual and neuronal changes can be described as coarse-to-fine, we do not yet know whether the various relevant processes are fundamentally related, or whether the similarities are merely superficial.

Box 2. Coarse-to-fine or fine-to-coarse?

Many neurophysiological studies have reported that, by some key measures, the representation of the stimulus changes in a coarse-to-fine fashion in time. The many complexities of this notion are outlined in Section 6. This box will illustrate an additional analytical complexity, namely that the same underlying temporal pattern might appear to represent a coarse-to-fine or fine-to-coarse change, depending on which two time points are compared.

The *red line* denotes the temporal changes in the sharpness of tuning (*i.e.*, inverse of tuning width) of a hypothetical neuron. Note that it shows the same broad pattern of fast rise followed by a slower fall as many actual neurons do (see, *e.g.*, Fig. 8a). The *dashed vertical line* denotes the time point (50 ms) at which the tuning is sharpest (*i.e.*, narrowest or finest). When the change in the tuning sharpness is measured symmetrically about this time point (*e.g.*, *green double arrow*) the tuning width will appear to have changed in a coarse-to-fine fashion. On the other hand, when tuning width change is measured relative to the time of sharpest tuning and a later time point, the tuning width will appear to have changed in a fine-to-coarse fashion. Note that in the instance shown, the tuning width changes are measured over the same time range (40 ms).

The two types of analyses, both of which are principled, produce such discrepant results whenever the underlying measure of tuning width (*red line* in figure) changes non-monotonically in time, which is presumably more common than not. Note also that it is not possible to determine whether either or neither of the two temporal changes is computationally meaningful without knowing the time point/s at which the brain samples the neuron's responses.



Altogether, while the concept of coarse-to-fine processing has undeniable pedagogical utility, it has very little explanatory or predictive value. Thus, at least as yet, both the strength and weakness of this concept may be that it simplifies complex temporal dynamic realities.

6.3. Bayesian framework in perspective

If there is a sense in which the coarse-to-fine is better than the Bayesian framework, it remains to be established. This is because the Bayesian framework subsumes the coarse-to-fine framework, as noted in various contexts above. This is not to say that the Bayesian framework has no weaknesses of its own, much less to say that it is an answer to everything. For one thing, as noted above, Bayesian framework cannot currently incorporate attentional selection or selection of relevant prior knowledge (see Sections 2.6 and 5). For another, it is far from clear what a Bayesian explanation for the potential temporal dynamic effects of affective factors, such as emotion, may even look like.

Nonetheless, it is fair to say that the Bayesian framework holds enormous promise as a framework of understanding and studying how sensory and motor systems work (Doya et al., 2006; Yuille and Kersten, 2006). In the context of temporal dynamics, its biggest weakness is simply that it is as yet untested.

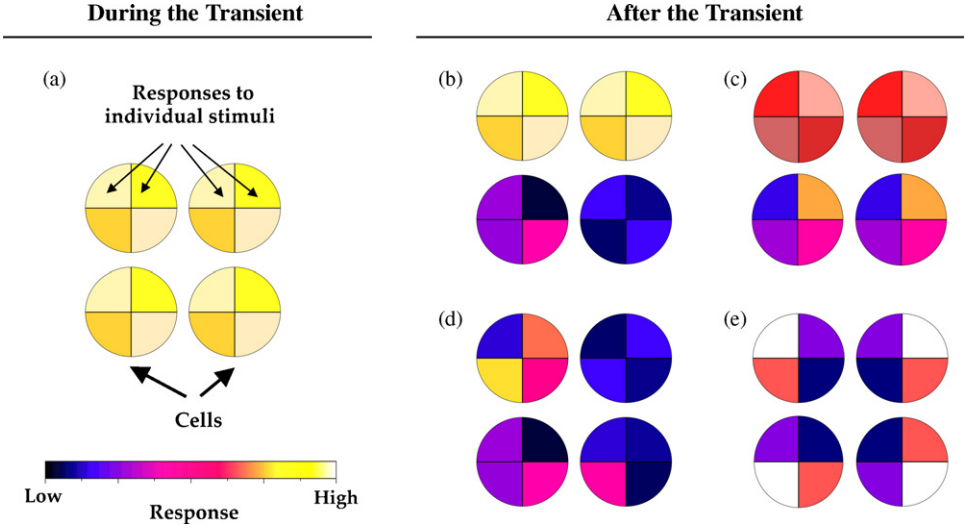
7. Future directions

Studying visual temporal dynamics for its own sake, while worthwhile, is far less interesting than studying it to understand how we see. In this regard, one of the major challenges for the future research is to rigorously characterize the various temporal dynamic changes at the neural and perceptual levels

Box 3. Correlation, decorrelation, and sparsening of the population response

The figure schematically illustrates some scenarios of response correlation and decorrelation at the population level. Each panel shows a highly idealized 'population' consisting of four cells (*circles*). Each *quadrant* of a given circle denotes the response of the cell to a given stimulus, color-coded according to the color scale at *bottom left*.

When the response pattern to a given set of stimuli is similar from one cell to the next, the population response is said to be correlated. In many visual areas, population response during the onset transient tends to be correlated not only across cells, but across stimuli as well (*panel* (a)). Following the transients, the population response decorrelates. Recall from Section 3.1 that the average response of a given cell generally decreases after the transient. This means that, in principle, the population response can decorrelate when the responses decrease differentially from one cell to the next, one stimulus to the next, or both. *Panels* (b–e) illustrate four such scenarios. (b) Responses decorrelate only in a restricted subpopulation of cells (*bottom row*), whereas the response of the remaining cells remain largely unchanged (*top row*). (c) All cells decorrelate with respect to the population response during the transient, but different subpopulations of cells (*rows*) decorrelate similarly, so that the responses remain correlated within the given subpopulation. (d) Response sparsening, whereby a given cell is responsive to only a subset of the stimuli, and only a subset of the cells in the population is responsive to a given stimulus. Thus, in the instance shown, the responses are sparsened both within and across cells. Note also that while this does result in some limited decorrelation of the population response, sparsened responses are inherently correlated, in that most of the responses are at or near zero. This type of limited decorrelation by sparsening appears to be widespread in the visual system (Section 3.5; Hegdé and Van Essen, 2004, 2006; also see Vinje and Gallant, 2000, 2002; Brincat and Connor, 2006). (e) Maximal decorrelation, where responses are as different as possible from one stimulus to the next and from one cell to the next. This type of decorrelation appears to underlie the dynamic representation of odor in the olfactory bulb (Laurent, 2002).



(see Box 4). At least in the near term, much of this research will necessarily be open-ended and exploratory in nature, since we do not yet know enough about visual temporal dynamics to construct large scale hypotheses about it.

Future research in visual temporal dynamics is likely to benefit greatly from three broad trends of current research. The first is our improving ability to study the responses of a large number of neurons in many different brain areas simultaneously, often using a combination of single-unit recoding and/or whole brain imaging techniques (see, *e.g.*, Logothetis et al., 2001; Bar et al., 2006; Schmid et al., 2006).

The second is our ability to deal with the complexity of natural images and the dynamic nature of the brain's response to them. Broadly speaking, a major drawback of the conventional neurophysiological and psychometric methods is that they, however implicitly, tend to treat vision as a static system. In such a system, the changes in the output are due solely to the changes in the input, and the system itself remains unchanged (or static) in the process. For example, the orientation tuning curve of a neuron is a static measure of its response, because it does not recognize that the orientation tuning can and does change in time. Most of the studies outlined above characterize the (temporal) dynamics of the system by measuring its static responses across several time points. While this method of extending the static methods to study dynamics is perfectly principled (see, *e.g.*, Watson, 1986), it also tends to make the analysis fundamentally descriptive. It is much like trying to understand climate change by comparing the daily weather reports from a large number of locations over a long time—it is simple and straightforward, but it is laborious, misses much, and predicts little. However, much progress is being made in recent years in developing rigorous, quantitative tools for analyzing the dynamics of the visual system (Rust et al., 2005; Wu et al., 2006; Ioannides, 2007).

Finally, as noted throughout this article, it is increasingly clear that studying vision as a probabilistic, Bayesian inference provides a powerful and unifying framework for understanding visual perception and visually guided action (Doya et al., 2006; Yuille and Kersten, 2006). But this framework currently does not explicitly account for dynamic changes in the underlying probabilistic factors. Thus, understanding vision as a *dynamic* inferential process is likely to be a fruitful avenue of future research.

Box 4. Key questions for future research

- How do various probabilistic parameters that underlie a given visual inference change over time? How do these changes affect the inference?
- How does the perception of natural visual scenes change under natural viewing conditions? How does this differ from what we have learned from using geometric stimuli with fixation controls?
- How do the response properties of individual neurons and neuronal populations in various visual cortical areas change over time? What type of feedforward and recurrent mechanisms bring about these changes? To what extent are these changes adaptive, *i.e.*, to what extent does the visual system adapt in response to the stimulus and various other task requirements?
- How are the temporal dynamic patterns in various brain areas related to each other and to those at the perceptual level?
- Is it feasible to understand the temporal dynamics of natural visual perception as a series of fixation epochs?
- Is it feasible to quantitatively model the temporal dynamics with a relatively small number of spatio-temporal parameters in a given area or set of areas? How predictive are such models?
- How do the complexities of natural scenes such as occlusion, visual clutter, camouflage, and variations of lighting, shadows, color, texture, etc., affect the dynamics of vision?
- How does the dynamics of visual perception relate to the dynamics of perceptual learning? (*cf.* Hochstein and Ahissar, 2002; Ahissar and Hochstein, 2004).

Acknowledgments

The preparation of this article was supported by ONR grant N00014-05-1-0124 to my advisor, Dr. Daniel Kersten. I am also grateful to Dr. Kersten for providing the picture in Fig. 1, and for his insights about visual processing in general. I thank Dr. David Van Essen for his permission to use the hitherto unpublished data collected in his laboratory shown in Fig. 8. I am grateful to Dr. Fei-Fei Li and Pietro Perona for sharing their data prior to publication, and Drs. Rufin VanRullen and Yasuko Sugase-Miyamoto for providing original figures from their work, and Drs. Sheng He and Gordon Legge for stimulating discussions. Many colleagues, most notably Dr. Sugase-Miyamoto and two anonymous reviewers, offered many valuable suggestions for improving the manuscript.

References

Afraz, S.R., Kiani, R., Esteky, H., 2006. Microstimulation of inferotemporal cortex influences face categorization. Nature 442, 692–695.

Ahissar, M., Hochstein, S., 2004. The reverse hierarchy theory of visual perceptual learning. Trends Cogn. Sci. 8, 457–464.

Albrecht, D.G., Geisler, W.S., Frazor, R.A., Crane, A.M., 2002. Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. J. Neurophysiol. 88, 888–913.

Allen, E.A., Freeman, R.D., 2006. Dynamic spatial processing originates in early visual pathways. J. Neurosci. 26, 11763–11774.

Amassian, V.E., Cracco, R.Q., Maccabee, P.J., Cracco, J.B., Rudell, A., Eberle, L., 1989. Suppression of visual perception by magnetic coil stimulation of human occipital cortex. Electroencephalogr. Clin. Neurophysiol. 4, 458–462.

Amit, Y., Geman, D., Fan, X., 2004. A coarse-to-fine strategy for multiclass shape detection. IEEE Trans. Pattern Anal. Mach. Intell. 26, 1606–1621.

Anand, S., Hotson, J., 2002. Transcranial magnetic stimulation: neurophysiological applications and safety. Brain Cogn. 50, 366–386.

Andreassi, J.L., 2006. Psychophysiology: Human Behavior and Physiological Response, fifth ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Averbeck, B.B., Latham, P.E., Pouget, A., 2006. Neural correlations, population coding and computation. Nat. Rev. Neurosci. 7, 358–366.

Azouz, R., Gray, C.M., 1999. Cellular mechanisms contributing to response variability of cortical neurons in vivo. J. Neurosci. 19, 2209–2223.

Bachmann, T., Allik, J., 1976. Integration and interruption in the masking of form by form. Perception 5, 79–97.

Bar, M., 2003. A cortical mechanism for triggering top–down facilitation in visual object recognition. J. Cogn. Neurosci. 15, 600–609.

Bar, M., 2004. Visual objects in context. Nat. Rev. Neurosci. 5, 617–629.

Bar, M., Biederman, I., 1999. Localizing the cortical region mediating visual awareness of object identity. Proc. Natl. Acad. Sci. U. S. A. 96, 1790–1793.

Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M., Hamalainen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., Halgren, E., 2006. Top–down facilitation of visual recognition. Proc. Natl. Acad. Sci. U. S. A. 103, 449–454.

Bar, M., Tootell, R.B., Schacter, D.L., Greve, D.N., Fischl, B., Mendola, J.D., Rosen, B.R., Dale, A.M., 2001. Cortical mechanisms specific to explicit visual object recognition. Neuron 29, 529–535 (Erratum in: Neuron 30, 299 (2001)).

Barlow, H., 1994. What is the computational goal of the neocortex? In: Koch, C., Davis, J.L. (Eds.), Large-scale Neuronal Theories of the Brain. MIT Press, Cambridge, MA, pp. 1–22.

Bichot, N.P., Rossi, A.F., Desimone, R., 2005. Parallel and serial neural mechanisms for visual search in macaque area V4. Science 308, 529–534.

Biederman, I., 1972. Perceiving real-world scenes. Science 177, 77–80.

Biederman, I., 1995. Visual object recognition. In: Kosslyn, M., Osherson, D.N. (Eds.), An Invitation to Cognitive Science: Visual Cognition, second ed., vol. 2. MIT Press, Cambridge, MA, pp. 121–165.

Bredfeldt, C.E., Ringach, D.L., 2002. Dynamics of spatial frequency tuning in macaque V1. J. Neurosci. 22, 1976–1984.

Breitmeyer, B.G., 1984. Visual Masking. Oxford University Press, New York, NY.

Bridgen, R.F., 1933. A tachistoscopic study of the differentiation of perception. Psychol. Monogr. 44, 153–166.

Brincat, S.L., Connor, C.E., 2006. Dynamic shape synthesis in posterior inferotemporal cortex. Neuron 49, 17–24.

Bruce, C., Desimone, R., Gross, C.G., 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol. 46, 369–384.

Bullier, J., 2001. Integrated model of visual processing. Brain Res. Brain Res. Rev. 36, 96–107.

Callaway, E.M., 2005. Neural substrates within primary visual cortex for interactions between parallel visual pathways. Prog. Brain Res. 149, 59–64.

Carlson, T., Grol, M.J., Verstraten, F.A., 2006. Dynamics of visual recognition revealed by fMRI. Neuroimage 32, 892–905.

Celebrini, S., Thorpe, S., Trotter, Y., Imbert, M., 1993. Dynamics of orientation coding in area V1 of the awake primate. Vis. Neurosci. 10, 811–825.

Chen, G., Dan, Y., Li, C.Y., 2005. Stimulation of non-classical receptive field enhances orientation selectivity in the cat. J. Physiol. 564, 233–243.

Coltheart, M., 1980. The persistences of vision. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 290, 57–69.

Cottaris, N.P., De Valois, R.L., 1998. Temporal dynamics of chromatic tuning in macaque primary visual cortex. Nature 395, 896–900.

Dan, Y., Atick, J.J., Reid, R.C., 1996. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. J. Neurosci. 6, 3351–3362.

Delorme, A., Richard, G., Fabre-Thorpe, M., 2000. Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. Vision Res. 40, 2187–2200.

DiCarlo, J.J., 2006. Neuroscience: making faces in the brain. Nature 442, 644.

DiCarlo, J.J., Maunsell, J.H., 2000. Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat. Neurosci. 3, 814–821.

Di Lollo, V., Wilson, A.E., 1978. Iconic persistence and perceptual moment as determinants of temporal integration in vision. Vision Res. 18, 1607–1610.

Doya, K., Ishii, S., Pouget, A., Rao, R.P.N., 2006. Bayesian Brain: Probabilistic Approaches to Neural Coding. MIT Press, Cambridge, MA.

Efron, R., 1967. The duration of the present. Proc. N. Y. Acad. Sci. 138, 713–729.

Egeth, H.E., Yantis, S., 1997. Visual attention: control, representation, and time course. Annu. Rev. Psychol. 48, 269–297.

Einhauser, W., Konig, P., 2003. Does luminance-contrast contribute to a saliency map for overt visual attention? Eur. J. Neurosci. 17, 1089–1097.

Enns, J.T., Lleras, A., Di Lollo, V., 2006. A reentrant view of visual masking, object substitution and response priming. In: Öğmen, H., Breitmeyer, B.G. (Eds.), The First Half Second. MIT Press, Cambridge, MA, pp. 127–147.

Fabre-Thorpe, M., 2003. Visual categorization: accessing abstraction in non-human primates. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 358, 1215–1223.

Fabre-Thorpe, M., Delorme, A., Marlot, C., Thorpe, S., 2001. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. J. Cogn. Neurosci. 13, 171–180.

Fabre-Thorpe, M., Richard, G., Thorpe, S.J., 1998. Rapid categorization of natural images by rhesus monkeys. Neuroreport 9, 303–308.

Fahle, M., Poggio, T., 2002. Perceptual Learning. MIT Press, Cambridge, MA.

Fei-Fei, L., Iyer, A., Koch, C., Perona, P., 2007. What do we perceive in a glance of a real-world scene? J. Vis. 7, 1–29.

Felleman, D.J., Van Essen, D.C., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47.

Frazor, R.A., Albrecht, D.G., Geisler, W.S., Crane, A.M., 2004. Visual cortex neurons of monkeys and cats: temporal dynamics of the spatial frequency response function. J. Neurophysiol. 91, 2607–2627.

Freedman, D.J., Assad, J.A., 2006. Experience-dependent representation of visual categories in parietal cortex. Nature 443, 85–88.

Friston, K.J., 1994. Functional and effective connectivity in neuroimaging: a synthesis. Hum. Brain Mapp. 2, 56–78.

Gallant, J.L., 2004. Neural mechanisms of natural scene perception. In: Chalupa, L.M., Werner, J.M. (Eds.), The Visual Neurosciences. MIT Press, Cambridge, MA, pp. 1590–1602.

Gegenfurtner, K.R., Rieger, J., 2000. Sensory and cognitive contributions of color to the recognition of natural scenes. Curr. Biol. 10, 805–808.

Gillespie, D.C., Lampl, I., Anderson, J.S., Ferster, D., 2001. Dynamics of the orientation-tuned membrane potential response in cat primary visual cortex. Nat. Neurosci. 4, 1014–1019.

Girard, P., Hupe, J.M., Bullier, J., 2001. Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. J. Neurophysiol. 85, 1328–1331.

Glimcher, P.W., 2005. Indeterminacy in brain and behavior. Annu. Rev. Psychol. 56, 25–56.

Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends Neurosci. 15, 20–25.

Goodale, M.A., Westwood, D.A., 2004. An evolving view of duplex vision: separate but interacting cortical pathways for perception and action. Curr. Opin. Neurobiol. 14, 203–211.

Grill-Spector, K., Kanwisher, N., 2005. Visual recognition: as soon as you know it is there, you know what it is. Psychol. Sci. 6, 152–160.

Grill-Spector, K., Kushnir, T., Hendler, T., Malach, R., 2000. Dynamics of object-selective activation correlate with recognition performance in humans. Nat. Neurosci. 3, 837–843.

Hämäläinen, M., Hari, R., Ilmoniemi, R., Knuutila, J., Lounasmaa, O.V., 1993. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of signal processing in the human brain. Rev. Mod. Phys. 65, 413–497.

Hawken, M.J., Shapley, R.M., Grosof, D.H., 1996. Temporal-frequency selectivity in monkey visual cortex. Vis. Neurosci. 13, 477–492.

Hayden, B.Y., Gallant, J.L., 2005. Time course of attention reveals different mechanisms for spatial and feature-based attention in area V4. Neuron 47, 637–643.

Hegdé, J., Felleman, D.J., 2007. Reappraising the functional implications of the primate visual anatomical hierarchy. Neuroscientist 13, 416–421.

Hegdé, J., Van Essen, D.C., 2004. Temporal dynamics of shape analysis in macaque visual area V2. J. Neurophysiol. 92, 3030–3042.

Hegdé, J., Van Essen, D.C., 2006. Temporal dynamics of 2-D and 3-D shape representation in macaque visual area V4. Vis. Neurosci. 23, 749–763.

Hegdé, J., Fang, F., Murray, S.O., Kersten, D. Functional specialization for occluded objects in the human extrastriate cortex. J. Vis., in press.

Helson, H., Fehrer, E.V., 1932. The role of form in perception. Am. J. Psychol. 44, 79–102.

Henderson, J.M., Hollingworth, A., 1998. Eye movements during scene viewing: an overview. In: Underwood, G. (Ed.), Eye Guidance in Reading and Scene Perception. Elsevier, New York, NY, pp. 269–283.

Henderson, J.M., Hollingworth, A., 1999. High-level scene perception. Annu. Rev. Psychol. 50, 243–271.

Hochstein, S., Ahissar, M., 2002. View from the top: hierarchies and reverse hierarchies in the visual system. Neuron 36, 791–804.

Horwitz, B., Warner, B., Fitzer, J., Tagamets, M.A., Husain, F.T., Long, T.W., 2005. Investigating the neural basis for functional and effective connectivity. Application to fMRI. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360, 1093–1108.

Horwitz, G.D., Chichilnisky, E.J., Albright, T.D., 2005. Blue-yellow signals are enhanced by spatiotemporal luminance contrast in macaque V1. J. Neurophysiol. 93, 2263–2278.

Horton, J.C., Adams, D.L., 2005. The cortical column: a structure without a function. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360, 837–862.

Howard, I.P., 2002. Seeing in Depth. Basic Mechanisms, vol. 1, Porteous, Toronto, Canada.

Hughes, H.C., Nozawa, G., Kitterle, F., 1996. Global precedence, spatial frequency channels, and the statistics of natural images. J. Cogn. Neurosci. 8, 197–230.

Hung, C.P., Kreiman, G., Poggio, T., DiCarlo, J.J., 2005. Fast readout of object identity from macaque inferior temporal cortex. Science 310, 863–866.

Iidaka, T., Yamashita, K., Kashikura, K., Yonekura, Y., 2004. Spatial frequency of visual image modulates neural responses in the temporo-occipital lobe. An investigation with event-related fMRI. Cogn. Brain Res. 18, 196–204.

Ioannides, A.A., 2007. Dynamic functional connectivity. Curr. Opin. Neurobiol. 17, 161–170.

Jasanoff, A., 2005. Functional MRI using molecular imaging agents. Trends Neurosci. 28, 120–126.

Johnson, J.S., Olshausen, B.A., 2003. Timecourse of neural signatures of object recognition. J. Vis. 3, 499–512.

Johnson, J.S., Olshausen, B.A., 2005a. The earliest EEG signatures of object recognition in a cued-target task are postsensory. J. Vis. 5, 299–312.

Johnson, J.S., Olshausen, B.A., 2005b. The recognition of partially visible natural objects in the presence and absence of their occluders. Vision Res. 45, 3262–3276.

Kammer, T., 2006. Masking visual stimuli by transcranial magnetic stimulation. Psychol. Res. [Epub ahead of print].

Kaplan, E., 2004. The M, P and K pathways of the primate visual system. In: Chalupa, L.M., Werner, J.M. (Eds.), The Visual Neurosciences. MIT Press, Cambridge, MA, pp. 481–493.

Kersten, D., Mamassian, P., Yuille, A., 2004. Object perception as Bayesian inference. Annu. Rev. Psychol. 55, 271–304.

Keysers, C., Xiao, D.K., Foldiak, P., Perrett, D.I., 2001. The speed of sight. J. Cogn. Neurosci. 13, 90–101.

Kiani, R., Esteky, H., Tanaka, K., 2005. Differences in onset latency of macaque inferotemporal neural responses to primate and non-primate faces. J. Neurophysiol. 94, 1587–1596.

Koch, C., Crick, F., 2004. The neuronal basis of visual consciousness. In: Chalupa, L.M., Werner, J.M. (Eds.), The Visual Neurosciences. MIT Press, Cambridge, MA, pp. 1682–1694.

Kourtzi, Z., Huberle, E., 2005. Spatiotemporal characteristics of form analysis in the human visual cortex revealed by rapid event-related fMRI adaptation. Neuroimage 28, 440–452.

Lamme, V.A., Roelfsema, P.R., 2000. The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci. 23, 571–579.

Lamy, D., Tsal, Y., 2001. On the status of location in visual attention. Eur. J. Cogn. Psychol. 13, 305–342.

Laurent, G., 2002. Olfactory network dynamics and the coding of multidimensional signals. Nat. Rev. Neurosci. 3, 884–895.

Lee, T.S., Yuilee, A., 2006. Efficient coding of visual scenes by grouping and segmentation. In: Doya, K., Ishii, S., Pouget, A., Rao, R.P.N. (Eds.), Bayesian Brain: Probabilistic Approaches to Neural Coding. MIT Press, Cambridge, MA, pp. 145–188.

Lehky, S.R., Sereno, A.B., 2007. Comparison of shape encoding in primate dorsal and ventral visual pathways. J. Neurophysiol. 97, 307–319.

Lerner, Y., Hendler, T., Malach, R., 2002. Object-completion effects in the human lateral occipital complex. Cereb. Cortex 12, 163–177.

Li, F.F., VanRullen, R., Koch, C., Perona, P., 2002. Rapid natural scene categorization in the near absence of attention. Proc. Natl. Acad. Sci. U. S. A. 99, 9596–9601.

Liu, J., Harris, A., Kanwisher, N., 2002. Stages of processing in face perception: an MEG study. Nat. Neurosci. 5, 910–916.

Liu, K., Jiang, Y., 2005. Visual working memory for briefly presented scenes. J. Vis. 5, 650–658.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157.

Luck, S.J., 1999. Direct and indirect integration of event-related potentials, functional magnetic resonance images, and single-unit recordings. Hum. Brain Mapp. 8, 115–201.

Luck, S.J., 2005. An Introduction to the Event-related Potential Technique. MIT Press, Cambridge, MA.

Luck, S.J., 2006. The operation of attention – millisecond by millisecond – over the first half second. In: Öğmen, H., Breitmeyer, B.G. (Eds.), The First Half Second. MIT Press, Cambridge, MA, pp. 187–206.

Ludwig, C.J.H., Gilchrist, I.D., McSorley, E., 2004. The influence of spatial frequency and contrast on saccade latencies. Vision Res. 44, 2597–2604.

Ma, W.J., Hamker, F., Koch, C., 2006. Neural mechanisms underlying temporal aspects of conscious visual perception. In: Öğmen, H., Breitmeyer, B.G. (Eds.), The First Half Second. MIT Press, Cambridge, MA, pp. 275–294.

Macé, M.J.-M., Richard, G., Delorme, A., Fabre-Thorpe, M., 2005a. Rapid categorization of natural scenes in monkeys: target predictability and processing speed. Neuroreport 16, 349–354.

Macé, M.J.-M., Thorpe, S.J., Fabre-Thorpe, M., 2005b. Rapid categorization of achromatic natural scenes: how robust at very low contrasts? Eur. J. Neurosci. 21, 2007–2018.

Malone, B.J., Kumar, V., Ringach, D.L., 2007. Dynamics of receptive field size in primary visual cortex. J. Neurophysiol. 97, 407–414.

Mante, V., Frazor, R.A., Bonin, V., Geisler, W.S., Carandini, M., 2005. Independence of luminance and contrast in natural scenes and in the early visual system. Nat. Neurosci. 8, 1690–1697.

Matsumoto, N., Okada, M., Sugase-Miyamoto, Y., Yamane, S., Kawano, K., 2005. Population dynamics of face-responsive neurons in the inferior temporal cortex. Cereb. Cortex 15, 1103–1112.

Maunsell, J.H., Treue, S., 2006. Feature-based attention in visual cortex. Trends Neurosci. 29, 317–322.

Mazer, J.A., Vinje, W.E., McDermott, J., Schiller, P.H., Gallant, J.L., 2002. Spatial frequency and orientation tuning dynamics in area V1. Proc. Natl. Acad. Sci. U. S. A. 99, 1645–1650.

Mazer, J.A., Gallant, J.L., 2003. Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. Neuron 40, 1241–1250.

Menz, M.D., Freeman, R.D., 2003. Stereoscopic depth processing in the visual cortex: a coarse-to-fine mechanism. Nat. Neurosci. 6, 59–65.

Menz, M.D., Freeman, R.D., 2004. Temporal dynamics of binocular disparity processing in the central visual pathway. J. Neurophysiol. 91, 1782–1793.

Merigan, W.H., Maunsell, J.H., 1993. How parallel are the primate visual pathways? Annu. Rev. Neurosci. 16, 369–402.

Mervis, C.B., Rosch, E., 1981. Categorization of natural objects. Annu. Rev. Psychol. 32, 89–115.

Moliadze, V., Zhao, Y., Eysel, U., Funke, K., 2003. Effect of transcranial magnetic stimulation on single-unit activity in the cat primary visual cortex. J. Physiol. 553, 665–679.

Morrison, D.J., Schyns, P.G., 2001. Usage of spatial scales for the categorization of faces, objects, and scenes. Psychon. Bull. Rev. 8, 454–469.

Mukamel, R., Harel, M., Hendler, T., Malach, R., 2004. Enhanced temporal non-linearities in human object-related occipito-temporal cortex. Cereb. Cortex 14, 575–585.

Müller, J.R., Metha, A.B., Krauskopf, J., Lennie, P., 2001. Information conveyed by onset transients in responses of striate cortical neurons. J. Neurosci. 21, 6978–6990.

Najemnik, J., Geisler, W.S., 2005. Optimal eye movement strategies in visual search. Nature 434, 387–391.

Nakayama, K., Mackeben, M., 1989. Sustained and transient components of focal visual attention. Vision Res. 29, 1631–1647.

Navon, D., 1977. Forest before trees: the precedence of global features in visual perception. Cogn. Psychol. 9, 353–383.

Nishimoto, S., Arai, M., Ohzawa, I., 2005. Accuracy of subspace mapping of spatiotemporal frequency domain visual receptive fields. J. Neurophysiol. 93, 3524–3536.

Ogawa, T., Komatsu, H., 2004. Target selection in area V4 during a multidimensional visual search task. J. Neurosci. 24, 6371–6382.

Öğmen, H., Breitmeyer, B.G., 2006. The First Half Second. MIT Press, Cambridge, MA.

Oliva, A., Schyns, P.G., 1997. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Cogn. Psychol. 34, 72–107.

Oram, M.W., Perrett, D.I., 1992. Time course of neural responses discriminating different views of the face and head. J. Neurophysiol. 68, 70–84.

Pack, C.C., Born, R.T., 2001. Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. Nature 409, 1040–1042.

Palanca, B.J., DeAngelis, G.C., 2003. Macaque middle temporal neurons signal depth in the absence of motion. J. Neurosci. 23, 7647–7658.

Palmer, J., Huk, A.C., Shadlen, M.N., 2005. The effect of stimulus strength on the speed and accuracy of a perceptual decision. J. Vis. 5, 376–404.

Palmer, S.E., 1975. The effects of contextual scenes on the identification of objects. Mem. Cogn. 3, 519–526.

Parker, D.M., Lishman, J.R., Hughes, J., 1992. Temporal integration of spatially filtered visual images. Perception 21, 147–160.

Parker, D.M., Lishman, J.R., Hughes, J., 1997. Evidence for the view that temporospatial integration in vision is temporally anisotropic. Perception 26, 1169–1180.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Modelling functional integration: a comparison of structural equation and dynamic causal models. Neuroimage 23 (Suppl. 1), S264–S274.

Peters, R.J., Iyer, A., Itti, L., Koch, C., 2005. Components of bottom-up gaze allocation in natural images. Vision Res. 45, 2397–2416.

Peyrin, C., Mermillod, M., Chokron, S., Marendaz, C., 2006. Effect of temporal constraints on hemispheric asymmetries during spatial frequency processing. Brain Cogn. 62 pp. 214–120.

Peyrin, C., Schwartz, S., Seghier, M., Michel, C., Landis, T., Vuilleumier, P., 2005. Hemispheric specialization of human inferior temporal cortex during coarse-to-fine and fine-to-coarse analysis of natural visual scenes. Neuroimage 28, 464–473.

Philiastides, M.G., Ratcliff, R., Sajda, P., 2006. Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. J. Neurosci. 26, 8965–8975.

Potter, M.C., Faulconer, B.A., 1975. Time to understand pictures and words. Nature 253, 437–438.

Potter, M.C., Staub, A., Rado, J., O'Connor, D.H., 2002. Recognition memory for briefly presented pictures: the time course of rapid forgetting. J. Exp. Psychol. Hum. Percept. Perform. 28, 1163–1175.

Purpura, K., Tranchina, D., Kaplan, E., Shapley, R.M., 1990. Light adaptation in the primate retina: analysis of changes in gain and dynamics of monkey retinal ganglion cells. Vis. Neurosci. 4, 75–93.

Pylyshyn, Z.W., 2003. Seeing and Visualizing. MIT Press, Cambridge MA.

Reynolds, R.I., 1981. Perception of an illusory contour as a function of processing time. Perception 10, 107–115.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., Bialek, W., 1996. Spikes. MIT Press, Cambridge, MA.

Ringach, D.L., Hawken, M.J., Shapley, R., 1997. Dynamics of orientation tuning in macaque primary visual cortex. Nature 387, 281–284.

Ringach, D.L., 2003. Look at the big picture (details will follow). Nat. Neurosci. 6, 7–8.

Ringach, D., Shapley, R., 2004. Reverse correlation in neurophysiology. Cogn. Sci.: Multidisciplinary J. 28, 147–166.

Rodieck, R.W., 1998. The First Steps in Seeing. Sinauer Associates, Sunderland, MA.

Roelfsema, P.R., 2006. Cortical algorithms for perceptual grouping. Annu. Rev. Neurosci. 29, 203–227.

Roelfsema, P.R., Lamme, V.A., Spekreijse, H., 2004. Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. Nat. Neurosci. 7, 982–991.

Rohaly, A.M., Wilson, H.R., 1993. Nature of coarse-to-fine constraints on binocular fusion. J. Opt. Soc. Am. A. Opt. Image Sci. Vis. 10, 2433–2441.

Rohaly, A.M., Wilson, H.R., 1994. Disparity averaging across spatial scales. Vision Res. 34, 1315–1325.

Rolls, E.T., 2004. Consciousness absent and present: a neurophysiological exploration. Prog. Brain Res. 144, 95–106.

Rolls, E.T., Tovee, M.J., 1994. Processing speed in the cerebral cortex and the neurophysiology of visual masking. Proc. Biol. Sci. 257, 9–15.

Rousselet, G.A., Fabre-Thorpe, M., Thorpe, S.J., 2002. Parallel processing in high-level categorization of natural images. Nat. Neurosci. 5, 629–630.

Rousselet, G.A., Macé, M.J., Fabre-Thorpe, M., 2003. Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. J. Vis. 3, 440–455.

Rousselet, G.A., Thorpe, S.J., Fabre-Thorpe, M., 2004. How parallel is visual processing in the ventral pathway? Trends Cogn. Sci. 8, 363–370.

Rust, N.C., Schwartz, O., Movshon, J.A., Simoncelli, E.P., 2005. Spatiotemporal elements of macaque V1 receptive fields. Neuron 46, 945–956.

Sharon, D., Grinvald, A., 2002. Dynamics and constancy in cortical spatiotemporal patterns of orientation processing. Science 295, 512–515.

Sheinberg, D.L., Logothetis, N.K., 2001. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J. Neurosci. 21, 1340–1350.

Schmid, M.C., Oeltermann, A., Juchem, C., Logothetis, N.K., Smirnakis, S.M., 2006. Simultaneous EEG and fMRI in the macaque monkey at 4.7 Tesla. Magn. Reson. Imaging 24, 335–342.

Schmolesky, M.T., Wang, Y., Hanes, D.P., Thompson, K.G., Leutgeb, S., Schall, J.D., Leventhal, A.G., 1998. Signal timing across the macaque visual system. J. Neurophysiol. 79, 3272–3278.

Schoenfeld, M.A., Woldorff, M., Duzel, E., Scheich, H., Heinze, H.J., Mangun, G.R., 2003. Form-from-motion: MEG evidence for time course and processing sequence. J. Cogn. Neurosci. 15, 157–172.

Scholte, H.S., Jolij, J., Lamme, V.A.F., 2006. Edge detection and scene segmentation. In: Öğmen, H., Breitmeyer, B.G. (Eds.), The First Half Second. MIT Press, Cambridge, MA, pp. 73–87.

Schwarzbach, J., Vorberg, D., 2006. A reentrant view of visual masking, object substitution and response priming. In: Öğmen, H., Breitmeyer, B.G. (Eds.), The First Half Second. MIT Press, Cambridge, MA, pp. 297–314.

Schyns, P.G., Oliva, A., 1997. Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. Perception 26, 1027–1038.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T., 2007. Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. 29, 411–426.

Shapley, R., 1998. In the mind's eye of the beholder. Nature 395, 845–846.

Shapley, R., Hawken, M., Ringach, D.L., 2003. Dynamics of orientation selectivity in the primary visual cortex and the importance of cortical inhibition. Neuron 38, 689–699.

Sharpee, T.O., Sugihara, H., Kurgansky, A.V., Rebrik, S.P., Stryker, M.P., Miller, K.D., 2006. Adaptive filtering enhances information transmission in visual cortex. Nature 439, 936–942.

Smallman, H.S., 1995. Fine-to-coarse scale disambiguation in stereopsis. Vision Res. 35, 1047–1060.

Smith, M.A., Bair, W., Movshon, J.A., 2006. Dynamics of suppression in macaque primary visual cortex. J. Neurosci. 26, 4826–4834.

Smith, M.A., Majaj, N.J., Movshon, J.A., 2005. Dynamics of motion signaling by neurons in macaque area MT. Nat. Neurosci. 8, 220–228.

Sowden, P.T., Schyns, P.G., 2006. Channel surfing in the visual brain. Trends Cogn. Sci. 10, 538–545.

Stoffer, T.H., 1993. The time course of attentional zooming: a comparison of voluntary and involuntary allocation of attention to the levels of compound stimuli. Psychol. Res. 56, 14–25.

Sincich, L.C., Horton, J.C., 2005. The circuitry of V1 and V2: integration of color, form, and motion. Annu. Rev. Neurosci. 28, 303–326.

Sugase, Y., Yamane, S., Ueno, S., Kawano, K., 1999. Global and fine information coded by single neurons in the temporal visual cortex. Nature 400, 869–873.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., Hirsch, J., 2006. Predictive codes for forthcoming perception in the frontal cortex. Science 314, 1311–1314.

Supèr, H., Spekreijse, H., Lamme, V.A., 2001. Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). Nat. Neurosci. 4, 304–310.

Tehovnik, E.J., Tolias, A.S., Sultan, F., Slocum, W.M., Logothetis, N.K., 2006. Direct and indirect activation of cortical neurons by electrical microstimulation. J. Neurophysiol. 96, 512–521.

Thorpe, S.J., Fabre-Thorpe, M., 2001. Seeking categories in the brain. Science 291, 260–263.

Thorpe, S.J., Gegenfurtner, K.R., Fabre-Thorpe, M., Bulthoff, H.H., 2001. Detection of animals in natural images using far peripheral vision. Eur. J. Neurosci. 14, 869–876.

Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522.

Torralba, A., Oliva, A., 2003. Statistics of natural image categories. Network 4, 391–412.

Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M., 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. 113, 766–786.

Ullman, S., 2007. Object recognition and segmentation by a fragment-based hierarchy. Trends Cogn. Sci. 11, 58–64.

Ungerleider, L., Pasternak, T., 2004. Ventral and dorsal cortical processing streams. In: Chalupa, L.M., Werner, J.M. (Eds.), The Visual Neurosciences. MIT Press, Cambridge, MA, pp. 541–562.

Valdés-Sosa, P.A., Sánchez-Bornot, J.M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., Canales-Rodríguez, E., 2005. Estimating brain functional connectivity with sparse multivariate autoregression. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360, 969–981.

Van Essen, D.C., Drury, H.A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C.H., 2001. An integrated software suite for surface-based analyses of cerebral cortex. J. Am. Med. Inf. Assoc. 8, 443–459.

VanRullen, R., Guyonneau, R., Thorpe, S., 2005. Spike times make sense. Trends Neurosci. 28, 1–4.

VanRullen, R., Koch, C., 2003. Competition and selection during visual processing of natural scenes and objects. J. Vis. 3, 75–85.

Vinje, W.E., Gallant, J.L., 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287, 1273–1276.

Vinje, W.E., Gallant, J.L., 2002. Natural stimulation of the nonclassical receptive field increases information transmission efficiency in V1. J. Neurosci. 22, 2904–2915.

Volgushev, M., Vidyasagar, T.R., Pei, X., 1995. Dynamics of the orientation tuning of postsynaptic potentials in the cat visual cortex. Vis. Neurosci. 12, 621–628.

Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., Schwarzbach, J., 2003. Different time courses for visual perception and action priming. Proc. Natl. Acad. Sci. U. S. A. 100, 6275–6280.

Wandell, B.A., 1995. Foundations of Vision. Sinauer Associates, Sunderland, MA.

Watson, A.B., 1986. Temporal sensitivity. In: Boff, K., Kaufman, L., Thomas, J. (Eds.), Handbook of Perception and Human Performance. Wiley, New York, NY, pp. 6-1–6-43.

Wersing, H., Korner, E., 2003. Learning optimized features for hierarchical models of invariant object recognition. Neural Comput. 15, 1559–1588.

Wiggs, C.L., Martin, A., 1998. Properties and mechanisms of perceptual priming. Curr. Opin. Neurobiol. 8, 227–233.

Wilson, F.A., Goldman-Rakic, P.S., 1994. Viewing preferences of rhesus monkeys related to memory for complex pictures, colours and faces. Behav. Brain Res. 60, 79–89.

Wu, M.C., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory neurons by system identification. Annu. Rev. Neurosci. 29, 477–505.

Xing, D., Shapley, R.M., Hawken, M.J., Ringach, D.L., 2005. Effect of stimulus size on the dynamics of orientation selectivity in Macaque V1. J. Neurophysiol. 94, 799–812.

Yantis, S., Serences, J.T., 2003. Cortical mechanisms of space-based and object-based attentional control. Curr. Opin. Neurobiol. 13, 187–193.

Yoshor, D., Bosking, W.H., Ghose, G.M., Maunsell, J.H., 2006. Receptive fields in human visual cortex mapped with surface electrodes. Cereb. Cortex [Epub ahead of print].

Yuille, A., Kersten, D., 2006. Vision as Bayesian inference: analysis by synthesis? Trends Cogn. Sci. 10, 301–308.

Zago, L., Fenske, M.J., Aminoff, E., Bar, M., 2005. The rise and fall of priming: how visual exposure shapes cortical representations of objects. Cereb. Cortex 15, 1655–1665.

Zhaoping, L., May, K.A., 2007. Psychophysical tests of the hypothesis of a bottom–up saliency map in primary visual cortex. PLoS Comput. Biol. 3, e62.

Zhou, H., Friedman, H.S., von der Heydt, R., 2000. Coding of border ownership in monkey visual cortex. J. Neurosci. 20, 6594–6611.