16 Neural Mechanisms Underlying Temporal Aspects of Conscious Visual Perception

Wei Ji Ma, Fred Hamker, and Christof Koch

In this chapter, we will examine dynamical aspects of conscious visual perception related to briefly presented stimuli, and their possible neural underpinnings. The time course of the contents of conscious perception usually reflects the time course of events in the world, but this correspondence is not absolute. As is illustrated in many examples in this book, it can break down in various ways: Physical events can get wiped out, stretched out in time, temporally blended, or modified. The last of these occurs in feature inheritance (Herzog & Koch, 2001), which we will study in some detail here. That these distortions do not seem omnipresent in daily life is because most of them occur at very short time scales of tens and hundreds of milliseconds. Their effects can often only be made apparent by presenting very specific, tightly controlled stimulus sequences. For a neuroscientist, these are exactly the interesting time scales. Because of the short duration of the stimuli involved, the stimuli can interfere with ongoing neural processing and hopefully reveal mechanisms used by the brain in more generality.

In studying temporal effects in perception, several quantities are of relevance. It is worthwhile to distinguish them clearly.

Perceived duration. How long do observers perceive an event of a given physical duration to be? While this question is hardly experimentally tractable, there is extensive literature on its active version, that is, how well humans estimate the passage of time (temporal cognition, interval timing). In the realm of millisecond stimuli, there is evidence for the existence of a minimal perceptual moment (Efron, 1970; Koch, 2004). For durations of seconds to minutes, our judgment of the subjective duration of a stimulus and its actual, physical duration can differ widely. Many studies show that the distribution of subjective durations (timed intervals) is invariant under scaling of the physical duration, a property that has implications for associative learning (Gibbon et al., 1997; Matell & Meck, 2000).

Temporal positioning of a percept. How long does it take before we see something, and do we think the event happens at the moment when we become aware of it, or

earlier? Recent years have witnessed widespread interest in the temporal positioning of a percept. This has been fueled by research on the flash-lag effect, in which a flashed stimulus, when presented spatially aligned with a moving object, is perceived as lagging spatially behind the object (MacKay, 1958; Nijhawan, 1994). It has been postulated (Eagleman & Sejnowski, 2000a) that the percept that the brain associates with the flash depends on events following it, but this view has been challenged (Patel et al., 2000; Nijhawan, 2002). A similar proposal has been made in the context of filling in the perceptual gap during saccadic suppression (Yarrow et al., 2001).

Content of conscious perception. What do we see? In this chapter, we will mostly be concerned with this topic. We will focus on two broad classes of perceptual distortions occurring in sequences of short stimuli. The first is *temporal integration*; the second is *backward masking*.

16.1 Perceptual Illusions for Short Stimuli

When a small green disk is presented on a screen for 10 ms, immediately followed by a red disk at the same location for 10 ms, observers perceive a yellow disk with a slight red hue (Efron, 1967, 1973). When either stimulus is presented by itself, it is perceived properly, that is, as either green or red. This shows that for very brief stimuli, the brain integrates stimuli over time to create a percept. This can be modeled by a convolution with a temporal filter. When the green–red sequence is presented for a sufficiently long period of time (for instance, both components for 500 ms), green is seen, followed by red. It is nontrivial to note that the sequence 10 ms green + 10 ms red, which by itself would be perceived as yellow, is now part of the stimulus sequence but nonetheless does *not* give rise to the *percept of yellow* in between the green and the red. This shows that the contents of perception are not always determined by a simple convolution; any linear filter would predict the intermediate percept of yellow. Related to this experiment is the everyday observation that a movie, recorded at 24 frames per second (i.e., each frame is 42 ms), is effortlessly perceived as a continuum.

A similar effect occurs for the rapid presentation of dot patterns (DiLollo, 1980). When a matrix of 5×5 dots is presented with one dot left out, it is easy to detect the gap. Now the same 24-dot display is split into two complementary 12-dot displays that follow each other in time, with a 10-ms blank interstimulus interval separating them. The first set of 12 dots is chosen randomly. The task of detecting which location was empty in both displays can become much harder, depending on the temporal parameters. If both halves are shown for 10 ms, they become perceptually blended and the observer will make few errors in detecting the missing dot. When

the duration of the first display is increased, the percentage of errors increases, most steeply between 80 and 160 ms. Again, we conclude that integration occurs for very brief stimuli and also that in order to explain conscious perception, an additional mechanism is needed. The reason is that we know that the integration period can bridge the 10-ms interstimulus interval. Thus, in all conditions, right after the onset of the second display, both the first and the second display would contribute to the integration, and their superposition would be visible. The idea has been invoked of a threshold value that the integrated activity due to one stimulus has to exceed for a certain amount of time in order for this stimulus to be perceived (Herzog et al., 2003c; Koch, 2004; Dehaene et al., 2003). In this view, a long duration for the first display would cause it to be perceived; after that, some form of reset would occur, and the first and the second display would not become superimposed.

That linear temporal integration cannot be the whole story is also clear from the phenomenon of backward masking. In backward masking, the visibility of a stimulus is destroyed or reduced by another stimulus following it; thus, there must be some mechanism at work besides that of mere temporal integration. Masking studies support the concept of the formation of an "object" as central to visual perception, as reported by Enns in this volume and in earlier work with DiLollo and Rensink (DiLollo et al., 2000). The key idea of their *object substitution theory* is that sensory input is processed in two stages. First, a feedforward sweep originating in the retina subsequently causes activation in the lateral geniculate nucleus (LGN), primary visual cortex, after which it moves toward higher and higher areas in the visual hierarchy. In the areas sequentially activated in this propagation, the receptive fields of cells are larger, and more and more complex visual features are encoded. Then, a feedback sweep acts to compare the generalized pattern activation generated at a high level, a "hypothesis," with the ongoing, high-resolution activity at a (nonspecified) lower level (see also Lee et al., 1998; Ullman, 2000; Lee & Mumford, 2003). This would serve the purpose of resolving ambiguities within a pattern hypothesis and of binding patterns to specific locations. Only after confirmation of the perceptual hypothesis are its contents, an "object," perceived. In this framework, a mask can interfere with the feedback sweep and reset the entire process, after which only the mask is perceived.

16.2 Feedforward or Feedback?

From the perspective of neuroscience, backward masking and temporal integration raise questions about the processes determining whether a stimulus is consciously perceived. An important and long-standing issue in this context is whether feedback

interactions are necessary for conscious awareness. Neuroanatomically, feedback connections are a dominant feature in visual cortex (Felleman & Van Essen, 1991), reaching all the way back into the LGN. Yet, their functions remain veiled. Some apparently complicated tasks, such as distinguishing animal pictures from nonanimal ones, can be performed by the brain very fast, suggesting—although most certainly not proving—that feedback is not necessary for those (Thorpe et al., 1996; VanRullen & Koch, 2003a). While classical models of backward masking (Breitmeyer, 1984; Breitmeyer & Öğmen, 2000; Öğmen, 1993) are based on local, lateral connections, object substitution theory posits feedback interactions as an essential ingredient. Several physiological studies in the macaque monkey show that in figure–ground segregation tasks, the awareness of the figure is correlated with a late component of V1 activity (Scholte et al., this volume; Lamme et al., 2000; Lamme et al., 2002; Lamme & Roelfsema, 2000). This has been taken to indicate that feedback into V1 is essential for visual awareness. However, on the basis of a computational study (Li, 2000) it has been argued that local V1 mechanisms can account for these figure–ground effects. In a study using transcranial magnetic stimulation, it was shown that the percept of a moving phosphene evoked in V1 can be masked by applying stimulation to V1 at a time that would interfere with feedback from MT (Pascual-Leone & Walsh, 2001). However, the biophysical effects of such stimulation are still poorly understood, and it should in addition be noted that rather than demonstrating the necessity of feedback, this result shows the necessary involvement of V1 at a later stage in conscious visual processing. As was pointed out in Öğmen et al. (2003), it is possible to have a graph-theoretically feedforward architecture with an anatomically descending connection, that is, from a higher to a lower area in the visual hierarchy (for instance, IT to V1). We will use such an architecture in our model ("Template" to "Object" in figure 16.1). In many studies, the distinction between anatomical and functional feedback is not properly drawn.

As a starting point to modeling the time course of visual perception, we take the above idea of *perceptual hypothesis testing*: Before an object can be consciously perceived, the brain first confirms its identity by comparing it with the sensory input at a later time. Especially when input is rapidly changing or when significant amounts of extrinsic or intrinsic noise are present, it is ecologically meaningful to ascertain whether the initial input reflects the current state of the world before engaging in a behavioral response. Models of consciousness often posit the necessity of a "coalition" of cortical areas, connected with each other through loops of feedforward and feedback connections (Koch, 2004; Baars, 2002; Baars et al., 2003; Grossberg, 1999; Dehaene et al., 2003). In these models, the main goal of a feedback circuit is presented primarily as a form of working memory rather than as a means of testing perceptual hypotheses; these two viewpoints may coincide.

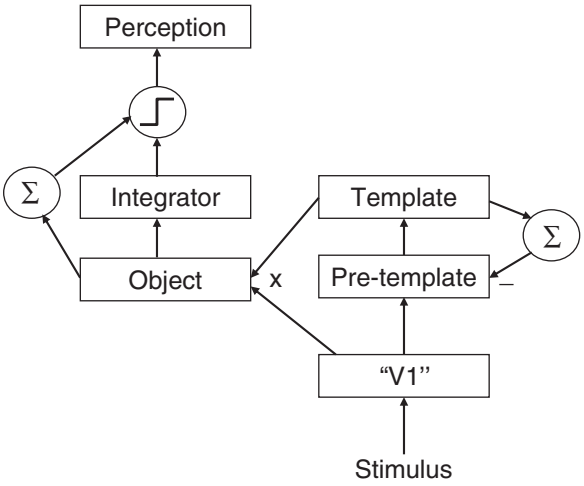

Figure 16.1
Schematic of the model. Images pass through orientation-selective filters in V1. The pretemplate–template subsystem creates an inert hypothesis about the stimulus. The hypothesis is tested against the later input into the object area. Finally, integration and thresholding determine the contents of perception.

Perceptual hypothesis testing can be mathematically modeled in terms of neural populations. One starts by describing a visual stimulus as a bundle of perceptual features, such as position, orientation, color, and shape, which are together—and not necessarily independently—encoded in neural populations. This encoding is noisy, so a sharply defined input will give rise to a broad population response. Let us concentrate on a single feature dimension, say, orientation, at one particular point in the visual field. The task of the brain is to decode the value of this quantity from the noisy population response, possibly making use of prior knowledge. Each possible orientation can be regarded as a hypothesis. Because of the numerosity of the hypotheses, the problem is one of parameter estimation, and hypothesis testing will consist of comparing neural population responses.

16.3 Parameter Estimation

Parameter estimation based on population codes is an example of Bayesian inference, with population activity patterns representing uncertainty about stimuli in the form of probability distributions. Bayesian inference has been postulated to be a general operating principle of the brain (Pouget et al., 2003; Rao et al., 2002; Yuille & Bülthoff, 1996). The idea here is that the brain tries to infer the identity of a source Z of an event in the world by optimally making use of the noisy (probabilistic) information contained in the sensory input S. It does this using Bayes's rule:

$$P(Z|S) = P(S|Z)\frac{P(Z)}{P(S)}. \tag{16.1}$$

$P(S|Z)$ is called the *likelihood function*, and $P(Z)$ is called the *prior distribution*. The priors can contain many factors, including genetically specified biases, previous experiences, and response biases.

Psychophysical evidence that the brain performs Bayesian inference partly comes from studies of how the brain combines noisy cues from two sensory modalities, or two cues from the same sensory modality, into a single percept (for instance, see Knill & Richards, 1996; Ernst & Banks, 2002). In these combinations, the best estimate is obtained by multiplying the probability distributions obtained from the two different cues and of the priors, if any, and locating the peak of the product distribution. If we denote the two cues by S_1 and S_2, this is captured by

$$P(Z|S_1, S_2) = P(S_1|Z)P(S_2|Z)\frac{P(Z)}{P(S_1, S_2)}. \tag{16.2}$$

Our problem concerns two cues appearing at different times, and Z is the orientation to be inferred. We assume that the likelihood function corresponding to the cues is encoded in the population activity pattern, where each neuron signals the probability of the stimulus given the neuron's preferred orientation.[1] In the absence of priors—that is, if $P(Z)$ is a flat distribution—the best estimate based on the two cues is then obtained by multiplication of the activities at each cell in the population.

In our model, one can also regard multiplication of population activities as an implementation of Bayes's rule of equation 16.1, where the one population encodes the likelihood function and the other one contains the priors. This is valid if the bottom-up input is compared with an expectation that has been generated not from directly preceding stimuli but from previous experience.

There are several examples of neural circuits performing multiplication, varying from spatial receptive fields in the barn owl (Pena & Konishi, 2001), to gain fields in the monkey's posterior parietal cortex (Andersen et al., 1997), to motion-sensitive neurons in the blowfly (Egelhaaf et al., 1989). There is evidence for single-cell mechanisms able to execute multiplicative computations (Gabbiani et al., 2002). Salinas and Abbott (1996) showed that multiplicative responses can also arise in a network model through population effects.

Our model is based on the idea of optimal cue combination, but it is, strictly speaking, not Bayesian. The reason is that we use the magnitude of the product response across the population as information about the similarity of the two population activities that are multiplied: The higher the product activity, the more similar the two activity patterns were. The key idea is then to use the product activ-

ity in some higher visual area as representing an object. If earlier and later activity patterns match, the multiplication will produce a high outcome and the stimulus will be perceived. If the match is not good, because the first stimulus has been replaced by a dissimilar one, either masking or pure temporal integration occurs. The latter gives rise to a percept in which stimuli physically present at different times are perceived simultaneously, in a sort of superposition. It can also happen that the match is not perfect and the first stimulus is not perceived but still modulates the perception of the second stimulus. We claim that this is what happens in feature inheritance.

16.4 Feature Inheritance

In the orientation paradigm of feature inheritance (Herzog & Koch, 2001), detailed in chapter 15 (Herzog, this volume), a bar slightly tilted with respect to the vertical is presented for 30 ms (the target), immediately followed by a grating of 3–5 vertical bars, presented for 300 ms (the "mask"). The percept is a slightly tilted grating. The perceived tilt is much smaller than the actual tilt of the first stimulus, but always in the same direction as the (perceptually invisible) target. This illusion is surprising in several respects: first, the invisibility of the target; second, the spread of the tilt over the entire grating[2]; and finally, the long duration of 300 ms over which the effect extends.

We now consider what happens to feature inheritance when the temporal parameters are varied. When the duration of target presentation is increased to several hundred milliseconds, the target and the mask are perceived veridically, that is, as a sequence of two stimuli. When the target is flashed by itself, it is clearly perceived. When the mask is of very long duration (e.g., 1,000 ms), one first observes the illusory, feature inheritance percept, which subsequently changes into the veridical percept of the original mask. When the target and the mask are very different in orientation, both the mask and the target are visible (so-called shine through, although this term was originally coined to describe the case in which the masking grating has many lines). We would like to simulate each of these cases.

Some of the spatial characteristics of the shine-through effect that can give rise to the spread of the feature over the grating have been explained with a feedforward, two-layer neural network model (Herzog et al., 2003a) and with a system of interconnected excitatory–inhibitory neuron pairs (Li, 2003). The issue of feature inheritance, that is, of how the orientation is decoded and assigned to the mask, and how the target becomes invisible, has not yet been addressed. A crucial issue is the invisibility of the target, on the one hand, if and only if a mask follows, and the perception of a tilted mask, on the other hand. A feedforward system with lateral

interactions would have difficulties in explaining how it is that the target is not perceived at all if a mask follows. This suggests that the earlier presented target influences the neural processing of the later presented mask in some kind of top-down fashion.

In our model, we focus on the temporal characteristics of visual perception. For simplicity, our only dimension is orientation space at the central location. We will see that with the mechanism sketched above, the temporal aspects of the inheritance process can be understood without any spatial interactions.

16.5 Model

The architecture of our model is shown in figure 16.1. It consists of early visual processing, a pretemplate and a template area, an object area, an integration area, and a perception area. We describe each area in terms of an analogue population activity. This is much cruder than a biophysical model, but we believe that this level of analysis will suffice for our purposes.

16.5.1 Early Visual Processing

Early visual processing, conveniently denoted as *V1*, although it is not necessarily limited to primary visual cortex, is modeled by convolving the black-and-white image of the oriented bar with a family $G_\theta(x,y)$ of two-dimensional Gabor filters, one for every integer number of degrees of angle θ. This describes the spatial receptive fields of cells with different orientation selectivities. We need a fine resolution in orientation space, because the effect we eventually want to show is rather subtle. The response in V1 to an image $S(t)$ is a 180-unit population activity pattern $(S(t) * G_\theta)(x,y)$ at each spatial location. Now we restrict ourselves to the central point ($x = y = 0$), where both the target and one bar of the masking grating are presented; this gives a population activity pattern $I(\theta,t)$. The half width at half height of this pattern is 13°. The pattern depends on time in a manner determined only by the stimulus sequence.

The temporal dynamics of V1 responses to visual stimuli is known to be affected by synaptic depression (Chance et al., 1998). This causes cells to initially respond very strongly (the *transient response*), followed by a decrease in sensitivity and a leveling off of firing rate (the *sustained response*). For our purposes, it can be modeled simply by a multiplicative factor in the input (Chance et al., 1998). The differential equation for the V1 response $A_{V1}(\theta,t)$ at angle θ and time t reads

$$\tau_{V1} \frac{\partial A_{V1}}{\partial t} = -A_{V1}(\theta,t) + \{1 - \alpha_d S(t)\} \cdot I(\theta,t), \tag{16.3a}$$

where the synaptic depression $S(t)$ is governed by

$$\tau_S \frac{dS}{dt} = 1 - S. \tag{16.3b}$$

16.5.2 Pretemplate and Template

The template is a higher level area that receives and stores the input for comparison with the later input. This area has a biophysically very long time constant (200 ms), which can be obtained, for instance, by a local positive feedback loop or self-excitation. In order to encode a hypothesis, activity in this area should not be readily overwritten by new input. This is the purpose of the pretemplate area, a gateway that receives bottom-up input from V1 and global inhibition from the template area. A new input has to compete with the existing hypothesis before it can form a new hypothesis. The activity in the pretemplate area A_{PT} is determined by

$$\tau_{PT} \frac{\partial A_{PT}}{\partial t} = -A_{PT}(\theta,t) + [\alpha_{PT} A_{V1}(\theta,t) - A^I_{PT}(t)]_+, \tag{16.4}$$

where $[..]_+$ denotes rectification (i.e., the function value is zero when the argument is negative, and it is equal to the argument otherwise) and the inhibitory activity A^1_{PT} is determined by

$$\tau^I_{PT} \frac{\partial A^I_{PT}}{\partial t} = -A^I_{PT}(t) + \alpha^I_{PT} \sum_{\theta} A_T(\theta - \varepsilon). \tag{16.5}$$

Here, ε is a delay without which the inhibition would only amount to a subtractive normalization. The template is solely driven by pretemplate input,

$$\tau_T \frac{\partial A_T}{\partial t} = -A_T(\theta,t) + \alpha_T A_{PT}(\theta,t). \tag{16.6}$$

16.5.3 Object Representation

This area is where the comparison between template and low-level activity takes place. It receives bottom-up input from V1, which is multiplied by the activity in the template area. The activity A_O is described by

$$\tau_O \frac{\partial A_O}{\partial t} = -A_O(\theta,t) + \alpha_O A_{V1}(\theta,t) \cdot A_T(\theta,t). \tag{16.7}$$

16.5.4 Integrator

The governing equation of the integration area reads

$$\tau_I \frac{\partial A_I}{\partial t} = -A_I(\theta,t) + \alpha_I A_O(\theta,t). \tag{16.8}$$

This area implements the fact that evidence has to build up over time before neural activity is sufficient for perception. Neurons in this area may be comparable to neurons in parietal and prefrontal cortex integrating sensory evidence until a decision criterion is reached (Gold & Shadlen, 2001; Freedman et al., 2002). Although this integration is in dynamics similar to the one in the template area, it serves a different goal. Here it is a mechanism for evidence accumulation, while there it guarantees the sustaining of a hypothesis.

16.5.5 Perception

The population activity pattern in the perception area determines the contents of the percept: in our case, which orientation or orientations are seen. The activity in the integration area is thresholded as follows:

$$\tau_P \frac{\partial A_P}{\partial t} = -A_P(\theta,t) + \frac{[\alpha_P A_I(\theta,t) - T]_+}{1 + \sum_\theta A_P(\theta,t)}. \tag{16.9}$$

The denominator serves as a normalization. The threshold T is given by

$$T = \max\left\{T_0, \alpha_{thr} \sum_\theta A_O(\theta,t)\right\}. \tag{16.10}$$

There is a baseline threshold T_0 and a threshold dependent on the total activity in the object area. In feature inheritance, no percept should be created before the multiplicative interaction in the object area has finished, although the stimulus is 300 ms long. On the other hand, any single 10-ms stimulus is perceived. This means that the perceptual threshold should be dependent on activity in a lower area. In this simple model, we take this activity to be the total activity in the object area. Such a dynamic threshold is equivalent to feedforward inhibition from the object area into the integration area.

The activity in the perception area is read out by registering sufficiently high local maxima, where "sufficiently high" at a certain point in time is taken to be at least 50% of the highest overall activity in the perception area until that moment. The duration of sufficiently high activity in the perception area is interpreted as the duration of the percept. Whether this is realistic is an open question. It has been noted that distinguishing the simultaneous presence of multiple distinct stimuli from the population noise intrinsic to the encoding of a single stimulus can be a problem (Sahani & Dayan, 2003), but we do not address that issue here.

16.6 Results

We presented this network with an oriented bar at an angle of 80° with respect to the horizontal for 30 ms, followed by a vertical bar for 300 ms. We tuned the parameters in the model such that it produces the desired phenomenology. Figure 16.2 shows the activity in each of the areas in the model when the target is oriented at 80°. The model produces a single peak at an orientation of 87°, corresponding to the percept in feature inheritance. The target is rendered invisible. Figure 16.3 shows the multiplication of V1 with template activity in the object area at $t = 200$ ms.

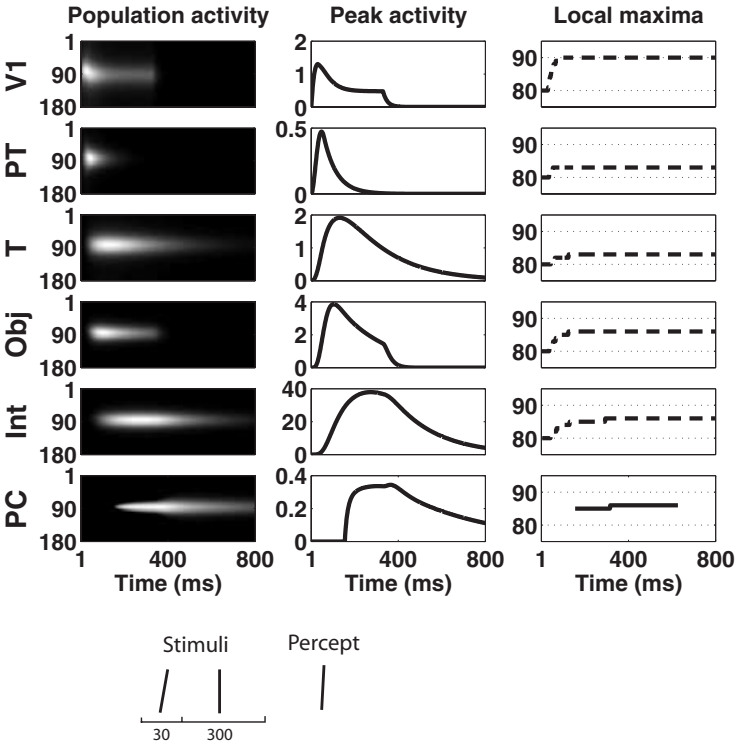

Figure 16.2
Feature inheritance. In all figures, the left column shows plots of the activity in each area as a function of time (*x*-axis) and orientation (*y*-axis). The second column shows the maximum activity in each area as a function of time. The third column shows the location of the local maxima in orientation space. Only for the perception area (PC), the third column shows all sufficiently high local maxima. Activity units are arbitrary but are consistent throughout the simulations. The diagram at the bottom shows the time course of stimulation (in milliseconds) and the resulting percept. Int, integrator; Obj, object; T, template; PT, pretemplate.

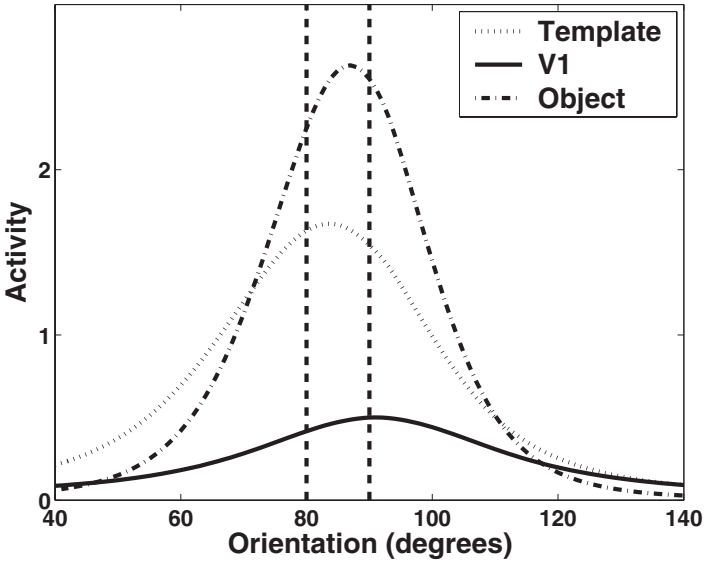

Figure 16.3
The multiplicative interaction occurring in the object area, taken at time $t = 200$ from figure 16.2. The dotted vertical lines represent the target and mask orientations.

A single short stimulus (10 ms) is perceived even though the activity it produces in the integration area is much lower (figure 16.4). Because of the integrator, the perceived duration is much longer than the stimulus duration. Figure 16.4b shows the effect of varying physical duration on perceived duration. The graph has a plateau at a value of about 250 ms. Because of the many parameters in our model and the ad hoc readout rule, we cannot reliably use this as a quantitative test, but it is qualitatively in accordance with the concept of a "minimal perceptual moment" (Efron, 1970).

A sequence of long-duration stimuli gets perceived veridically because the match between template and bottom-up activity is good (figure 16.5a). We compared a target at 80° (figure 16.5a) with one at 40° (figure 16.5b). The first one produces a smooth transition between the percept of target and mask, because the orientations are close enough together that the template activity representing the target can interact with the mask activity. This can be interpreted as apparent rotation. (In the real experiment [Herzog, this volume], observers were presented with an entire grating rather than with a single bar; the grating could serve as a perceptual cue against apparent rotation.) In the second case, we find a sudden transition. We do not know of any psychophysical studies of the strength of apparent motion as a function of spatial and temporal distance for rotations, although there have been such studies of linear motion (Korte's laws, and Burt & Sperling, 1981).

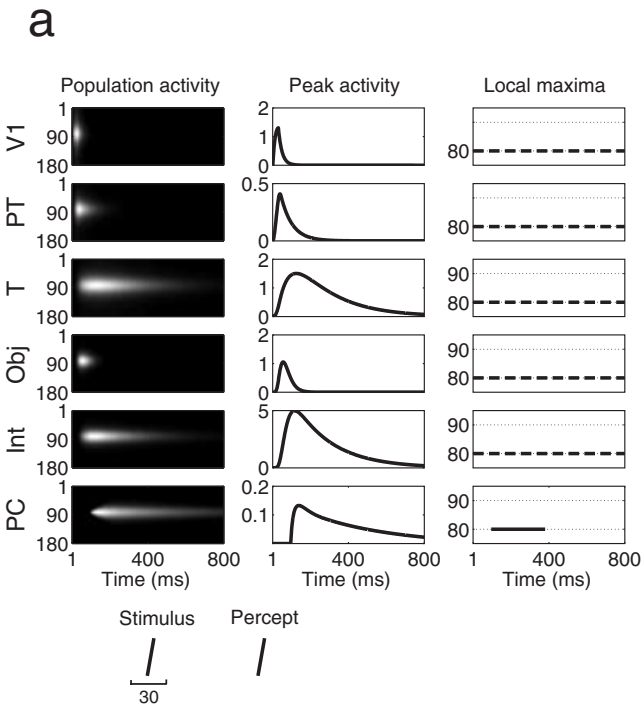

Figure 16.4
(*a*) A single 30-ms stimulus is perceived. PC, perception area; Int, integrator; Obj, object; T, template; PT, pretemplate. (*b*) The model qualitatively reproduces the "minimal perceptual moment": up to a certain physical duration, perceived duration has a constant value. This value is higher than the physical duration and is between 200 and 300 ms.

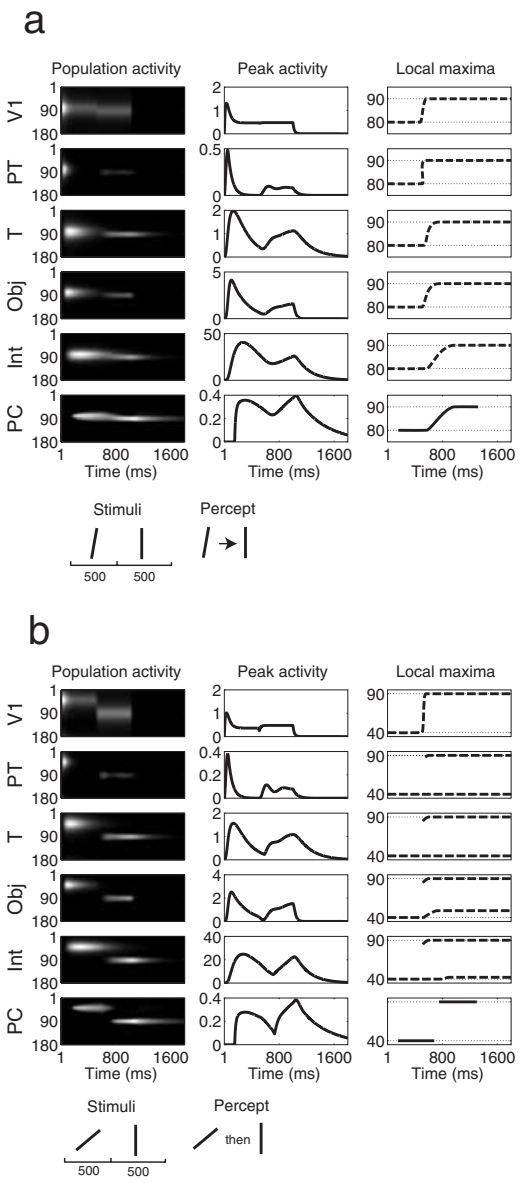

Figure 16.5
(*a*) Target and mask are presented sequentially, each for 500 ms, with a small orientation difference. They are perceived sequentially, with apparent rotation. (*b*) If two orientations with a large orientation difference (target, 40°; mask, 90°) are presented for the same durations (500 and 500 ms), they are observed sequentially, without any apparent motion. PC, perception area; Int, integrator; Obj, object; T, template; PT, pretemplate.

For two brief stimuli (both 20 ms), the percept also depends on the orientation difference. The model shows that when the difference is small, a single line with the average orientation is perceived (figure 16.6a). This is different from feature inheritance in that it does not involve any hypothesis testing; the line is also predicted to be at a slightly different orientation. When the difference is large, two lines superimposed on each other will be perceived (temporal integration; see figure 16.6b). This percept is essentially different from the one in figure 16.5b, while, physically speaking, the 20 + 20-ms stimulus sequence is contained within the 500 + 500-ms sequence. A condition that the model can also deal with is that of a target at 40° for 30 ms, followed by a masking grating for 300 ms, that is, the temporal parameters at which feature inheritance would occur were the target at 80°. In this case, a sort of superposition is observed (see figure 16.7). In the cases of figures 16.6 and 16.7, we can again not interpret the onsets and durations as quantitative predictions. In the superposition percepts, different perceptual durations for the components do not necessarily mean that an observer will see them superimposed only part of the time.

Table 16.1 shows the impact of leaving out certain aspects of the model on the reproduction of the phenomenology in three conditions, on the existence of a minimal perceptual moment, and on consistency with object substitution theory.

16.7 Discussion

Modeling the contents of perception based on neuronal population activity usually requires many simplifications. With our approach, we merely hope to outline a mechanism that can simultaneously explain the temporal characteristics of feature inheritance, backward masking, and temporal integration. The key properties of our network, which—so we claim—give rise to a visual percept, are as follows:

- a Bayes-motivated comparison interaction between a dynamic template and bottom-up input, leading to the formation of an "object"
- linear temporal integration, followed by a threshold
- feedforward inhibition proportional to the strength of an "object"

Speculations about the existence of both a threshold and an integration period for perception are supported by neurosurgical experiments (Libet, 1966, 1973, 1993; for an overview, see Koch, 2004).

If we examine the architecture of the model, we see that there is no loop between the model areas V1, template, and object; thus, there is no feedback in a graph-theoretical sense. However, it is possible that the template area is higher in the cortical hierarchy than the object area. Thus, a physiologist may characterize this interaction as a feedback interaction, although there is no recursion involved (cf.

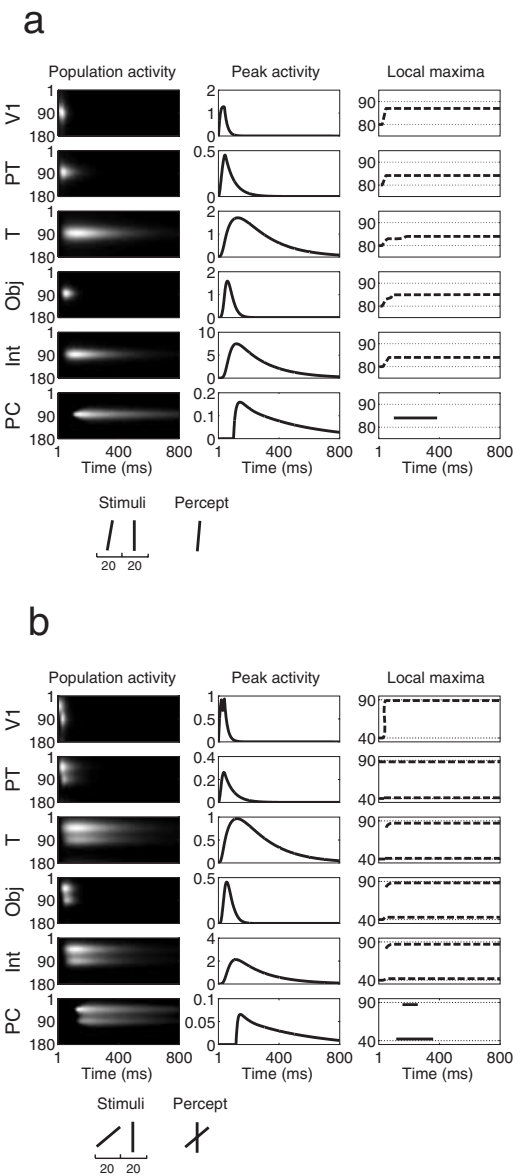

Figure 16.6
Perception of two 20-ms stimuli immediately following each other is determined by temporal integration. (*a*) When the orientation difference is small, a single line at the average orientation is perceived. This is an effect different from feature inheritance. (*b*) When the orientation difference is large, the two lines are perceived simultaneously and separately. PC, perception area; Int, integrator; Obj, object; T, template; PT, pretemplate.

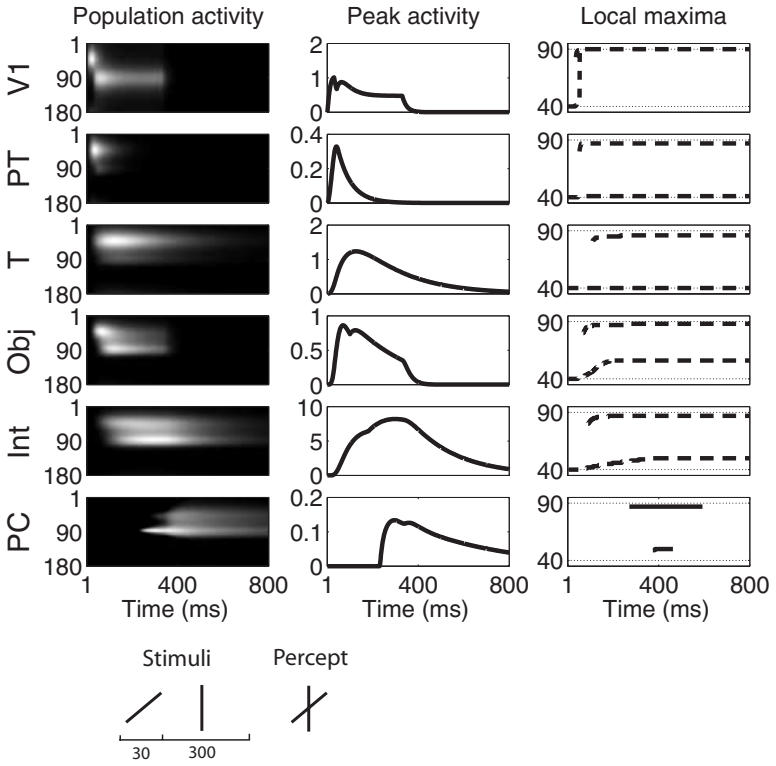

Figure 16.7
With a large orientation difference and the same timing as in feature inheritance, a superposition is observed. PC, perception area; Int, integrator; Obj, object; T, template; PT, pretemplate.

Table 16.1
Impact of leaving out selected aspects of the model on the reproduction of phenomenology

	30/300 orientations close, figure 16.2	30/300 orientations far apart, figure 16.7	20/20 orientations far apart, figure 16.6b	Minimal perceptual moment? figure 16.4b	Consistent with object substitution theory?
Full model	X	X	X	X	X
Template constant (linear feedforward)			X	X	
Short τ_T				X	X
No pretemplate			X	X	X
No integrator	X				X
Fixed threshold		X	X		X

Note. X = yes.

figure 7 in Öğmen, 2003). V1 is essentially involved in later interactions, but there is no feedback into V1.

The integration area can be thought of as a form of *iconic memory*, a high-capacity, rapidly decaying form of storage, lasting for at least a few hundred milliseconds (Coltheart, 1983, 1999). That does not mean that we think of iconic memory as instantiated in a single area. This role of the integration area is consistent with the claim that iconic memory is essential for visual awareness (Koch, 2004): It is as if enough evidence has to be collected before one can become aware of a stimulus.

Because we only consider orientation space at one location in physical space, our model is rather limited in scope. A next step would be to combine it with a detailed model of lateral interactions in space (e.g., Herzog et al., 2003a, or Li, 2004) in order to arrive at a more complete explanation of feature inheritance. It might then be possible to explain the spreading of the feature over the whole grating. Moreover, feature inheritance is a form of incomplete backward masking. In describing the phenomenology of backward masking, the multiplicative interaction would implement hypothesis testing in the sense of DiLollo et al. (2000). Both the transition from temporal integration to masking and the transition from temporal integration to normal vision can be seen in our model. The former is determined by the duration of the mask, the latter by the duration of the target.

Another shortcoming of our model is the absence of stochastic noise in the cells. Including it ought to explain variability of human performance in feature inheritance and thus allow a fit of the psychometric curves of Herzog and Koch (2001). Such a model would likely also allow the application of a more fundamental decision rule, such as a maximum-likelihood one (instead of "sufficiently high local maxima"). Noise is also the reason that for very short stimulus durations, seeing becomes a signal-detection task.

A general issue that requires further study is the implementation of Bayesian mechanisms in neural circuitry. While many instances of multiplication have been found in the brain, Gabbiani et al. (2002) remains the only study of the local biophysical mechanisms involved. One proposal to perform Bayesian inference without coding multiplication in neurons has been put forward by Rao (2004). By representing probabilities in the logarithmic domain, multiplication is turned into addition, which is more easily implemented in the neural circuitry. He has shown that feedforward and recurrent connections perform Bayesian inference for arbitrary hidden Markov models; however, some strong mathematical assumptions are made in the process.

In conclusion, we have implemented the simple idea that the contents of perception are the result of a continuous comparison of sensory input with a template updated with some inertia, an integration stage, and a dynamic threshold.

16.8 Appendix: Model Details

All areas are modeled as populations of analogue, noiseless neurons. Orientation filtering in "V1" occurs through convolutions with Gabors. The horizontally oriented one is

$$G(x,y) = \frac{2\cos \alpha y}{\sqrt{\sigma_x \sigma_y}} e^{-\left(\frac{x^2}{2\sigma_x} + \frac{y^2}{2\sigma_y}\right)},$$

with $\alpha = 0.5$, $\sigma_x = 5$, and $\sigma_y = 30$. The time constants of the different areas are as follows: $\tau_{v1} = 20\,\text{ms}$, $\tau_S = 40\,\text{ms}$, $\tau_{PT} = 60\,\text{ms}$, $\tau_{PT}^1 = 10\,\text{ms}$, $\tau_T = 200\,\text{ms}$, $\tau_O = 20\,\text{ms}$, $\tau_I = 200\,\text{ms}$, and $\tau_P = 20\,\text{ms}$. The time constants of some of these areas are very long, but they can be effective time constants due to a positive feedback loop or attractor dynamics. The synaptic delay ε between pretemplate and template areas is taken to be 3 ms; between other areas, synaptic delay is irrelevant for our model. The pretemplate–template subsystem with global feedback inhibition (see equations 16.4–16.6) implements the fact that a candidate hypothesis, sent from V1 into the pretemplate area, has to compete with the stored hypothesis. A threshold proportional to the total activity in the template area is equivalent to an inhibitory term in the pretemplate activity.

Weights are as follows: $\alpha_{V1} = 1$, $\alpha_{PT} = 1$, $\alpha_{PT}^1 = 4/180$, $\alpha_T = 15$, $\alpha_T^1 = 3$, $\alpha_O = 0.5$, $\alpha_O^1 = 20$, $\alpha_1 = 20$, $\alpha_P = 1$, and $\alpha_{thr} = 0.26$. Synaptic depression in V1 has weight $\alpha_d = 0.8$. These have been tuned to reproduce the experimental results; this tuning is not necessarily unique. The baseline threshold is $T_0 = 0.05$. The minimum value that a local maximum in the perception area has to exceed at a certain time in order to "be perceived" is taken to be 50% of the overall maximum until that time.

The differential equations were integrated using the Euler method in MATLAB (MathWorks, Inc.). All initial activities were zero.

Acknowledgments

We thank Michael Herzog for countless valuable discussions, for comments on the manuscript, and for conducting many pilot experiments for us. We also thank Haluk Öğmen for many pleasant and useful discussions. Wei Ji Ma is supported by the Netherlands Organisation for Scientific Research and the Swartz Foundation. We are grateful to both. We thank the organizers and the other participants of the "First Half Second" workshop.

Notes

1. An alternative approach, making use of the noise distribution of neuronal responses, can be found in Pouget et al. (2003).

2. In the offset paradigm (i.e., when the target is a pair of offset lines, a vernier), the offset is also bequeathed to the entire grating, but the observer's attention is always on one of its edges (Herzog & Koch, 2001). It is unknown whether this holds for the orientation paradigm as well.