

# Cognition and Perception

How Do Psychology and Neural Science Inform Philosophy?

Athanassios Raftopoulos

A Bradford Book  
The MIT Press  
Cambridge, Massachusetts  
London, England

© 2009 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, email [specialsales@mitpress.mit.edu](mailto:specialsales@mitpress.mit.edu) or write to Special Sales Department, MIT Press, 55 Hayward Street, Cambridge, MA 02142.

Set Stone Sans and Stone Serif by SNP Best-set Typesetter Ltd., Hong Kong. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Raftopoulos, Athanassios.

Cognition and perception : how do psychology and neural science inform philosophy? / Athanassios Raftopoulos.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-262-01321-5 (hardcover : alk. paper)

1. Cognition. 2. Perception. 3. Philosophy. I. Title.

BF311.R25 2009

121'.34—dc22

2009000716

10 9 8 7 6 5 4 3 2 1

# 1 The Role of Attention in Visual Processing

As I stated in the introduction, this book aims to examine whether non-conceptual content is possible given our perceptual makeup, to delineate the nature of nonconceptual content, should such a content exist, and to propose a causal theory of reference based on the nonconceptual content of our perceptual states. The success of the endeavor depends on whether we are equipped with perceptual mechanisms that allow the retrieval of information from the environment in a purely bottom-up way (that is, in a way that is immune to top-down conceptual interference) and on whether the information thus retrieved is epistemologically interesting (that is, whether it can be used to promote epistemological issues). Should our conceptual framework intervene at all stages of perception, the contents of perceptual states would be irrevocably conceptually contaminated and any further discussion about whether perceptual content is conceptual or nonconceptual would be moot. Attention and its mechanisms are important in such a discussion. There seems to be ample evidence that spatial attention modulates perceptual processing from its very early stages, and since attention can be cognitively driven (endogenous attention) a strong argument could be made that, through attentional effects on perception, cognition and thus our conceptual schemes modulate all perceptual processing, or at least the part that delivers states with epistemologically interesting content.

In the present chapter, I discuss attention and its role in visual processing. The aims of the discussion are (1) to draw a picture of visual processing and of its stages and (2) to delineate the role of attention (and cognition) in perceptual processing. Since I do not intend to keep the reader in suspense, I will reveal the outcome of the investigation on attention and perception here: There is a part of visual processing—which I will call perception, and which corresponds, to a certain extent, to Pylyshyn's (2003, 2007) early vision—that is cognitively encapsulated and thus retrieves

information from visual scenes in a purely bottom-up manner. The qualification is needed because perception the way I construe it is a part of early vision, the other part being sensation. The difference between perception and sensation, we shall see, lies in the kind of information that is processed.

In section 1.1, I discuss various models of attention with a view to shedding light on the way it functions and on its role in visual processing. In section 1.2, I address the matter of the representations involved in visual processing and whether and to what extent they are stored in memory, allowing us to have a coherent view of the world across visual scenes. In section 1.3, I bring forth the issue of the top-down constraints in visual processing, as they are mediated by attention. I analyze these constraints and discuss the predominant role of spatial attention in realizing these constraints

## 1.1 Attention

Attention is a selection process in which some inputs are processed faster, better, or deeper than some other inputs, so that they have a better chance of producing or influencing a behavioral response, although a bodily response is not necessary; attention limits processing to items that are behaviorally relevant. Attentional mechanisms are needed because a typical visual scene contains more information than the visual system can process at any given time; in other words, the visual system can select only one or a few objects for more thorough processing, as cases of inattentive blindness that I discuss later reveal. Since the visual system does not have the capacity to process simultaneously all inputs in the retina, attention intervenes to select some inputs and to filter some others; attention favors the processing of some inputs by enhancing the responses of neurons that represent the behaviorally relevant stimuli and thereby biasing the competitive interactions among the stimuli.

Attention induces increased (Desimone and Duncan 1995) and synchronous (Fries et al. 2001) neuronal activity of the neurons processing the attended stimuli. The increased neural activity suffices to explain why the associated stimuli are processed faster and deeper. Attention may enhance the output of the salient feature detectors by lowering firing thresholds (Egeth et al. 1984; Kahneman et al. 1992). It can also increase the activity of neuronal systems that process the salient type of information (Ungerleider and Haxby 1994). Spatial attention, more specifically, results in accurate detection and discrimination of stimuli at the attended location

(LaBerge 1995); it does so by increasing the magnitude of stimulus-evoked neural activity for stimuli at the attended location. In other words, it acts like a gain-control mechanism that most likely serves to improve the signal-to-noise ratio of inputs at the attended location so that more relevant information for the task at hand can be extracted from them (Hillyard et al. 1998). The effect of attending to some stimuli on the firing rates of cells in the visual cortex is widely accepted as the neural correlate of attention (Desimone and Duncan 1995).

Stimuli can be behaviorally relevant in two senses: either they are located at behaviorally relevant locations, or they involve objects or have features that are relevant to current behavior. In either case, attention reflects a top-down expectation. The subject actively searches either for a specified feature or object, or for a specified location, depending on which information is available to the subject—that is, depending on whether the subject has information about some feature or object or about the location of the behaviorally relevant feature or object that she seeks. As will become evident below, another function of attention is to solve the binding problem and provide a coherent object representation.

Attentional selection is historically related to “spotlight of attention” models, according to which attention serves to limit processing to a single location in the visual field (Desimone 1999). When one searches for a behaviorally relevant object in a scene, one is essentially engaged in a serial process during which one shifts the spotlight of attention from one object in a scene to the next until the target object is found. Attention serves to enhance neuronal responses to a stimulus at that specific spatial location in the visual field. This response is observed at the extrastriate cortex, and there is also evidence that it is found in striate cortex. According to these models of attention, all visual attention is inherently spatial; objects are selected by attention being directed to their spatial locations (Posner 1980; Treisman 1988). Even objects defined by features (shape, color, etc.) must be found by examining the objects in a scene by this spotlight that serially scans locations.

As an example of an inherently spatial model of attention, consider Treisman’s (1993) Feature Integration Theory (FIT), which posits that objects are retrieved from scenes by means of selective spatial attention that picks out objects’ features, forms feature maps, and integrates those features that are found at the same location into forming objects. Treisman’s theory presupposes the spotlight model of attention, according to which attention acts serially to conjoin elements of a scene on the basis of their common location; the elements themselves are searched in

parallel. FIT belongs to the family of theories that hold that when one attends to an object one automatically encodes all of its features in visual working memory and has them available for further processing (Duncan and Nimmo-Smith 1996; O'Craven, Downing, and Kanwisher 1999).

However, Jonides (1980, 1983) proposed, on the basis of experiments using spatial cues, that attention might function in two distinct modes: a focal mode and a spread mode. The former, used when spatial information about a target is available, involves a serial search of locations in the visual field; thus selection operates in the space dimension. The latter can spread in parallel over the visual field and focuses on features of objects rather than on spatial locations.

In general, research on visual selective attention suggests two theoretical accounts for selection (Vecera 2000). The attention that focuses on the spatial dimension is known as "spatial-attention" or "space-based attention"; stimuli are selected on the basis of spatial location by a spotlight, a zoom lens, or a spatial gradient, depending on the specific theory. The function of spatial attention is evidenced by experiments showing that targets appearing at cued locations are processed more efficiently than targets appearing at uncued locations. The attention that focuses on features of objects or objects is known as "object-centered" or "object-based" attention.<sup>1</sup> In this case, stimuli are not selected for their location but either on the basis of some feature or as whole organized objects or shapes. Vecera and Farah (1994) and Egeth and Yantis (1997) argue that either attentional mode can be obtained, depending on the characteristics of the task at hand.

Posner and Rothbart (1992) and Posner and Raichle (1994) also argue for the existence of two separate attentional circuits. The first is a posterior attention network responsible for orienting attention. It involves the parietal lobe (for releasing attention from its current focus), the midbrain (for moving attention from its current location to the new location of a cue), and the thalamus (for selecting and enhancing the contents of the attended area). The second is an anterior attention network (the executive attention network) that mediates awareness of attended objects. The latter circuit intervenes once attention has shifted to a certain location by means of the former and the visual contents there have been transmitted forward in the visual areas of the brain. It activates the anterior cingulate gyrus in conjunction with other frontal areas, such as lateral areas of the upper prefrontal cortex (Posner and Raichle 1994). It is plausible that the distinction between the posterior network for orienting attention and the anterior network that mediates awareness of objects can be mapped onto the

distinction between the dorsal and the ventral pathways of the visual processing system. Interesting as this mapping may be, I will not pursue it further here. Alternatively, it may be that the former circuit may be responsible for spatial attention, whereas the latter is responsible for object-centered attention.

An alternative to Treisman's serial space-based spotlight theory of attention is Duncan and Humphreys's (1989, 1992) Attentional Engagement Theory (AET), according to which there is an initial pre-attentive parallel phase of perceptual segmentation and analysis that encompasses all of the visual items present in a scene. At this phase, descriptions of the objects in a visual scene are generated at a number of scales, and grouping principles organize the visual input into structural units; the outcome of this parallel phase is a multiple-spatial-scale structured representation. Selective attention intervenes after this stage to select information that will be entered into visual short-term memory. This is a serial stage that allows conscious processing of the items in the visual field. The task's requirements may be translated to a target template—an object that one expects, as a result of a preceding cue, to appear somewhere in one's visual field among other distractor objects, for instance. The visual input is matched to the internal stored templates; thus, visual items that match the task's requirements are most likely to be selected for visual short-term memory (see also Hollingworth and Henderson 2002, 2004). The target template exercises a top-down influence in visual processing, in that it activates a neuronal assembly in memory which sends top-down signals to the visual cortex that enhance the activation of the neuronal assembly that represents the target object. Duncan and Humphreys's theory posits both a parallel stage (in which items in the visual field—both the target object and distractors—activate neuronal assemblies in the brain) and a competition stage (in which these items compete until only one object is selected). AET, unlike FIT, does not require an attentional spotlight serially searching locations in the visual field and binding into objects features that are located at the same location. Furthermore, the selection need not rely exclusively on spatial information but may rely on featural information.

Humphreys (1999) elaborates further on the way objects are represented in space. He proposes the existence of two forms of spatial representations of objects, by which he means representations of objects in space: within-objects representations, in which elements are coded as parts of objects, and between-objects representations, in which elements are coded as distinct independent objects. Both kinds of representation are realized in parallel in the visual system. The former are mainly used for object

recognition, and thus are formed in the ventral system; the latter are used for navigation in the environment and action, and thus are formed in the dorsal system. Dorsal processing areas and the representations they subserve are recruited by the visual system when attention focuses from one part of an object to another, or when the spatial relations between parts are important for the identification of the object. Against Treisman's Feature Integration Theory, which posits spatial attention as a necessary condition for detection of objects, Humphreys argues that visual elements are encoded and bound together in an initial parallel phase without focal attention, and that attention serves to select among the objects that result from this initial grouping.

The coding of space devoid of objects is extremely limited, if it exists at all. The memory of locations across fixations depends on coding the relative positions of objects. Although forms of grouping depend on the proximity of the elements and although distance effects modulate selection of objects, which seems to suggest that elements are represented in terms of their position in space, Humphreys (1999) argues that the coding of distance itself is modulated by grouping between stimuli and adduces evidence that representations of space itself involve the relations between objects and are modulated by grouping between parts.

An initial parallel bottom-up phase during which all input is processed in parallel without an attentional bottleneck is also posited in "biased-competition account of visual processing" (Desimone and Duncan 1995; Reynolds and Desimone 2001). In this account, attention acts to bias the competition between neuronal populations that encode environmental stimuli. All the stimuli in a visual scene are initially processed in parallel and activate neuronal assemblies that represent them. These assemblies eventually engage in competitive interactions, either because they project onto cells in topographically organized cortical areas in which neurons have restricted receptive fields and thus cannot process all stimuli or because some behaviorally relevant feature or object must be selected among all present stimuli. Thus, in the biased-competition model of attention, multiple representations of objects (or, as we shall later see, proto-objects) are active and compete to be selected to drive a motor output (pressing a button, reaching to grasp an object, or some other motor behavior).

There are two sources of attentional control in visual searches. First, there are bottom-up influences on processing that arise from environmental stimuli; the scene in the visual field, which constitutes the visual input, provides the bottom-up information that will be searched through and

which indicates the locations of objects and the kinds of features present in each location. Second, there are top-down influences that derive from the current behavioral goals of the perceiver, as they are determined by an experimenters' instructions, by goal-oriented plans, and/or by contextual constraints. The sources of the top-down effects lie in the inferotemporal (IT) cortex. Attending to a stimulus at a particular behaviorally relevant location or with a particular behaviorally relevant feature biases the competition in favor of neurons that respond or (equivalently) are tuned to the location or the feature of the attended stimulus. As a result, the activation of the cells that represent the behaviorally relevant stimuli are enhanced, and these cells win the competition for further processing, suppressing at the same time cells representing distracting stimuli. These stimuli are thus attended. Shipp (2004, p. 227) synthesizes the essence of the bias-competition models as follows: "Activity within feature maps depends on a combination of visual input and the top-down bias signal, and the feature maps' output signals are pooled within a topographic, modality-free element labeled 'posterior parietal' (PP). The latter acts something like a salience map." The salience map is a source of top-down bias on visual search.

The biased-competition model has been expanded to include object-based attention (Desimone 1999; Vecera 2000). According to Desimone (1999, p. 13), the class of theories that view attention as the result of a biased competition among neuronal assemblies suggests that any enhancement of neuronal responses in the extrastriate cortex due to attention is better understood "in the context of competitive interactions among neurons representing all of the stimuli present in the visual field." Desimone continues: "These interactions can be biased in favor of behaviorally relevant stimuli as a result of many different processes, both spatial and nonspatial and both bottom-up and top-down" (*ibid.*, p. 13). Top-down influences are derived mostly from working memory. As a result of the biased interaction, behaviorally irrelevant stimuli are suppressed. In this framework, attentional selection is better understood not so much as the enhancement of neuronal responses but more as the modulation of the competitive interaction of the stimuli in the visual field, and attention is better viewed as a dynamic property of the system than as a separate mechanism.

Notice that the biases that feed back to extrastriate cortex (where attentional effects are mostly observed) form higher neuronal assemblies in the brain (working memory circuits in the prefrontal cortex, for instance) are not limited to the cells with receptive fields at a single locus in the visual

field; that is, the bias is not necessarily spatial. The processing can be biased by object features (color, shape, etc.), in which case searching for such an object does not require serial scanning of locations, as it does in space-based models of attention in the focal mode. In fact, all stimuli present in a scene are initially processed in parallel. This leaves open the issue of search for conjunctive features, for which it was thought that serial searching of the scene was required. The binding of some features (e.g., the color and shape of an object) seems to require attention, whereas feature combinations (such as shape and location) and other feature conjunctions that lead to segregations are detected pre-attentively (Lamme 2003; Roelfsema 2005). When attention is needed, it is commonly thought that the search for such a conjunction is serial, as in the case of color and shape (Treisman 1988, 1993).

More specifically, there is quite an extensive body of evidence suggesting that searching for particular conjunctions of features does not produce steeply sloped response-time functions by set size (McLeod et al. 1988; Nakayama and Silverman 1986). In searching for the conjunction of form and motion (McLeod et al. 1988), for instance, the search seems to be effectuated in parallel. Eckstein (1998) and Palmer et al. (2000) argue that even when the evidence shows steep slopes in response times, the underlying mechanism may not be a serial search but a parallel search and the plot of response times with respect to set size can be predicted by probabilistic models based on signal-to-noise ratios. In view of the accumulated evidence that undermines the classical view of two different search modes, one serial and the other parallel, it is suggested that the different ranges of response times by set size slopes emerge from a single process, probably parallel in nature, the outcome of which is determined by the relative salience of the target and the distractors (Duncan and Humphreys 1989; Eckstein 1998; Mounts and Tomaselli 2005; Palmer et al. 2000; Spivey 2007; Wolfe 1998).

Suppose that a cue (say, a certain feature) is presented to a subject who, after a delay, is asked to perform a task involving the selection of an object (the target object) with that cued feature among other objects that do not have the specific feature (distractors). After the cue has been presented, the neuronal assemblies in the prefrontal cortex that represent that cue are activated and remain activated for the duration of the task. The description of the target provided by the demands of the task creates a “template” (Duncan and Humphreys 1989) that is stored in visual working memory for the duration of the task (otherwise put, the subject keeps the cue in her working memory to use it in the selection process). The activation of

this assembly is fed back to the extrastriate inferior temporal cortex, thereby activating only the neurons that respond to the cued feature. Thus, the features of the cue are temporarily stored in working memory, even when the stimulus has been withdrawn (Miller and Cohen 2001; Rainer et al. 1998; Schall and Hanes 1993; Super et al. 2001b). Working memory biases activity in favor of cells that select the cued feature. When the choice array is presented and the subject has to select the target object, all cells in the IT cortex that respond to any feature in the visual field are initially activated and compete to be further processed. Thus, cells representing different stimuli engage in mutually suppressive interactions, which are biased in favor of the cells that represent the cued feature. The bias is due to the top-down activation of the cells from the signals that originate in working memory. When the subject makes her choice, the activation of cells responding to nontarget stimuli has been suppressed.

A similar account of the mechanism at work when an item in a scene is selected amongst competitors is provided by Findlay and Gilchrist (2003, chapter 6), who propose that the visual display is monitored in parallel with increasing weighting for proximity to the fovea. This last assumption allows this model to take into account the fact that many visual functions show gradually declining ability as the stimuli are placed more eccentrically (that is, as the stimuli are removed from the fovea), although there are some notable exceptions, such as the monitoring for change in the environment that is considered to be mainly a function of peripheral vision. The salience map in this account is a representation in which information originating from the retinal image is represented in a two-dimensional spatial way. This map can be seen as the pattern of the activations of the units in a two-dimensional neural network in which the visual field is being mapped in a retinotopic way. The level of neural activity at each point of the neural network (and, therefore, at each point of the two-dimensional representation of the visual field) encodes the salience. "It is assumed," Findlay and Gilchrist write (2003, p. 115), "that information feeds into this salience map so that the level of activity corresponds to the level of the evidence that the target is present at any location. As a result, items sharing a feature with the target will generate a higher level of activation than items sharing no target feature. Proximity to fixation also increases an item's salience. Within this framework, the saccade is made to the location of highest activity on the salience map."

Furthermore, the saccade to the location in space where the eyes will be directed once the target has been spotted is explained by an account of what transpires at the superior colliculus (SC). The buildup cells<sup>2</sup> in the

intermediate layers in the region of the SC that corresponds to the location in space at which the target is located gradually increase their activity. At the same time, cells in the fixation center show a decrease in activity. At some point, the activity of the latter cells ceases and the burst cells start firing. At that time, the characteristic activity of saccades occurs in the midbrain reticular formation (MRF) and the paramedian pontine reticular formation (PPRF) (Findlay and Gilchrist 2003, chapter 4).

The main characteristic of models that construe attention as a “biased competition” among stimuli is that attention becomes the outcome of a competitive process among neighboring objects of which one or a few victors prevail and survive, rather than the cause of selection of one of these objects. To put it differently, it is the object or objects that win the competition that are said to be attended rather than the case being that some faculty, namely attention, determines or selects the winner. Independent evidence that decisions in simple memory and perceptual tasks (which are traditionally thought to require “focusing of attention” on some object) in fact result from biased competition among stimuli according to the outcome of which the winner is selected and affects behavior comes from neurobiological studies of single neuron firings in such tasks and from mathematical models that model both the behavior of these neurons and the behavioral data from similar psychological studies. The most successful models are the sequential-sampling models that assume that decisions are based on accumulating noisy information about the stimulus. There are two broad classes of such models: the random-walk models and the accumulator models (Smith and Ratcliff 2004). In the random-walk models, the information is accumulated as a single total and eventually one response (for instance, a certain selection out of two or more alternatives in a memory selection task) reaches a response criterion and the subject makes the corresponding selection. In these models, evidence for one response is evidence against the other alternatives. In the accumulator models, information about the two responses accumulates separately and eventually one of the two is the first to reach the response criterion and wins the competition.

When the competitive interaction is biased in favor of some stimulus that is behaviorally relevant because of its location, attention becomes spatially directed (it works on the space dimension). Spatial attention can be controlled by bottom-up signals, such as the raw visual qualities forming feature maps that determine the salience of an object in a scene, making it “pop out” (that is, draw attention to itself). The mechanisms of these signals may be implemented by colliculo-thalamic interactions along the

ventral visual pathway. Spatial attention can also be controlled by top-down signals that involve wide cortical areas, including the frontoparietal system and the IT (Shipp 2004). This way, spatial attention involves a feedback bias that modulates the interactive competition between the attended and unattended stimuli in the visual field (Clark and Hillyard 1996; Luck and Hillyard 2000).

Although Desimone and Duncan's (1995) biased-competition account of visual processing posits the existence of a parallel bottom-up stage at which information from the environment is fed toward the visual areas of the brain, it is not clear what kind of information is so processed. Their stimuli include object features, such as color and oriented lines, but they do not explicitly deal with the problem of feature binding that may occur during the parallel stage of processing; that is, they do not specify which features retrieved in the parallel mode may combine during this mode to form a more complex structure. Vecera (2000) concentrates on object-based attention and extends the biased-competition account of visual search to the segregation or segmentation of objects from backgrounds and the selection of these objects by attentional processes. Object segmentation is the set of preattentive visual processes that determine which features combine to form the shapes present in a visual scene (Driver et al. 2001; Scholl 2001; Vecera 2000). These processes segment a shape from the background and segregate it from other shapes that are similarly segmented from the background. Vecera defines object-based attention as the visual processes that select a segregated shape from among several segregated shapes.

Given that the visual system cannot process all stimuli present in multi-object scenes within the visual field, objects or regions in space compete with one another for processing in two respects. Vecera (2000, pp. 359–360) writes: "First, there is a competition within object-based segregation processes and the segregated regions formed by segregation processes; the outcome of this competition is a perceptual group or figure that is more salient than other groups or figures. Second, there is a competition within object-based attentional processes; the outcome of this competition is the selection of one perceptual figure or group over another."

The competitions are resolved by a bottom-up bias (which arises from image cues or salient information in the environment) and a top-down bias (which arises from task-relevant or goal-relevant information). These two sources of bias operate in parallel and compete or cooperate with one another. Bottom-up information (that is, information contained in the environment and retrieved by bottom-up visual processes) may define

perceptual groups and may bias some groups rendering them more easily perceived than others; it biases object segregation. Then, these perceptual groups bias the allocation of visual object-based attention in a bottom-up manner and determine a perceptually salient object. However, a scene may contain more than one perceptually salient objects or regions that are task-relevant or goal-relevant, or it may even contain an object that is perceptually more salient than a behaviorally relevant object. In this case, top-down sources of information are needed to bias the competition in favor of one behaviorally relevant object or in favor of the behaviorally relevant object over the more perceptually salient but goal-irrelevant object.

To implement this process of parallel competitive interactions in figure-ground segmentation, Vecera and colleagues (Vecera 2000; Vecera and Farah 1997; Vecera and O'Reilly 2000) rely on parallel distributed processing (PDP) connectionist models. Their model posits top-down feedback signals from object representations to an earlier process of figure-ground segregation. The latter extracts information from the stimuli in a parallel bottom-up manner (in the way of biased-competition models of visual processing), which consists in simple image features (edges) present in the visual scene. The information is stored in the boundary layers and is fed to the next layer in the network that extracts the surfaces that are "figure" as opposed to ground. The figure layer receives top-down feedback from the object-recognition layer in which familiar shapes are represented and sends feedback signals to the boundary layers. There are two kinds of competitive interactions in this network. The first of these is the overall competition between bottom-up and top-down signals in the figure layer. Thus, the selection of figure is biased both from object representations top-down signals and from bottom-up image cues. Second, there is competition among the boundary layers, as different boundaries shared by regions in the figure-ground display compete (the model presupposes that some regions or boundaries are activated selectively over others) to become foreground figure. This competition is implemented by inhibitory connections among opposing boundary units.

This model envisages a role for competition between top-down and bottom-up biases only if the bottom-up image cues are ambiguous (as in bi-stable or ambiguous figures). The input provided in the simulations by Vecera and colleagues contained no bottom-up cues that would resolve the competition between candidate boundaries. "Because the bottom-up input is insufficient to resolve this competition," Vecera argues (2000, p. 378), "there must be top-down inputs to bias the bottom-up competition. In our model, the top-down inputs came from object representations: One of the

two regions in the figure-ground display corresponded to a familiar object that was represented by one of the object units." It seems thus that the presence of unambiguous perceptual cues in a scene suffices to resolve the competition and to segment figure from ground in the initial parallel bottom-up processing step. In addition, bottom-up signals are stronger than top-down signals, the latter acting more as gain enhancers of activity that is already present than as activators of neurons that are otherwise silent (Hupé et al. 1998).

The "competition" models of attention emphasize the competition among structures that are derived from a visual scene during a first stage of perceptual processing in which information is retrieved in a parallel bottom-up way from the visual scene. Different stimuli in a visual scene will activate in parallel neuronal assemblies responding to, and thus encoding, these stimuli. If one feature or object must be selected, and/or if a local region of cortex receives input from these neuronal assemblies, there is a competitive interaction between the stimuli (the competition is the strongest when the two stimuli fall within the same receptive field of neurons in visuotopically organized areas in which neurons have restricted receptive fields). Eventually a structure wins the competition and gets the opportunity to be further processed upward in the visual stream in a second serial stage of visual processing (the stage is serial because of the attentional bottleneck that allows objects to be processed one at a time).

The competition may be biased by various top-down influences that reflect the expectations and goals of the perceiver. To repeat a point made before: Attention is seen as the dynamic cause of the competitive interaction between bottom-up and top-down processes. Attention acts to bias the competition between neuronal populations that encode environmental stimuli. The biases may be either spatial or featural, depending on the task and the kinds of cues (Deco et al. 2002; Vecera and Farah 1994; Usher and Niebur 1996)—that is, they are not limited to cells with receptive fields at a single locus in the visual field; they include biases in favor of behaviorally relevant features. If the biases are spatial, then the selection of a target object in a visual scene requires a serial scanning of locations (of the type posited in FIT, for instance). If the bias is featural (as it is, for example, when the cue concerns a feature of the target object), the selection of a target does not require a serial scanning of the scene; it takes place in a parallel stage in which all stimuli present in the scene are processed until the target object is found.

Usher and Niebur (1996) offer a model of object-centered attention. It consists of a parallel mechanism for selective object-based attention that

is complemented under certain conditions with a serial search in the focal mode—that is, with a serial spatial attention. Usher and Niebur study the behavior of cells in the IT cortex. These cells, with their large receptive fields, respond to complex structures, such as specific shapes or faces, irrespective of their position in the visual field. Research (Chelazzi et al. 1993) with tasks in which monkeys have to search for target objects that are characterized by some feature suggests that the response of IT neurons to displays that include two objects has two phases. The first of these is a parallel phase in which the activation of neurons responding to both objects is enhanced (each group of neurons responds to its preferred stimulus). The response does not depend on whether the preferred object is a target, and thus the activation is task independent. In other words, there is a parallel phase in which information is retrieved bottom-up from a scene irrespective of task demands and independent of any top-down expectation-driven feedback. This stage can be construed as “preattentive,” in that no selection is involved. In the second phase, the response of the neurons shows expectation-driven top-down modulation of processing. Activation is enhanced only for the neuronal assembly that represents the target, and is suppressed for the other neuronal assembly. The second stage underlies selective object-based attention.

The Usher-Niebur model is based on a neural network with distributed representation, so that cell assemblies that represent similar objects share some cells. The network has three layers of units: the input layer (whose units simulate the neurons in V1 primary visual cortex), the visual sensory memory level (whose units simulate the neuronal assemblies in the IT cortex), and the working memory layer (whose units simulate neurons in prefrontal cortex). The first layer sends input to the second layer, and the third layer has feedback projections to the second layer. To make the system sensitive to input, there are excitatory connections between the cells in each cell assembly and inhibitory connections among assemblies. The excitatory connections between cells in each neuronal assembly in the visual sensory memory are strong enough to generate competition between the objects in the input but not strong enough to make the activation of the cells in the visual sensory memory layer independent of the input.

Each cell assembly in working memory corresponds to a cell assembly in visual sensory memory. The activation of a cell assembly in the working memory layer is stronger than the activation in the corresponding cell assembly in the visual sensory memory, so the response of the former persists even during the delay between the presentation of a cue and the presentation of the test display (that is, during the absence of sensory

stimuli). Each working-memory assembly sends a feedback signal to its corresponding cell assembly in the visual sensory memory, strengthening its activation. This means that a cell assembly that responds to an object in the visual field which is already stored in working memory (because it has been designated as a target by a preceding cue) has, eventually, a stronger activation than a cell assembly responding to an object in the visual field but not in working memory (in the way described in Desimone 1999). This is also what it makes the search for a target an expectation-driven search. To simulate the findings of Chelazzi et al. (1993) that suggest an initial parallel stage of bottom-up target-independent extraction of information, the feedback input is weak, so that initially both objects in the visual primary cortex are activated. Eventually, the assembly that receives additional input from working memory wins the competition and suppresses the activation of the other assembly.

The model was designed to simulate the behavioral characteristics of IT cells in delayed match-to-sample tasks and to shed light on the mechanism underlying expectation-driven selective attention for object features. “The function of the described mechanism,” Usher and Niebur write (1996, p. 317), “is to ‘look for’ an expected stimulus among distractors in the display (the expected stimulus being the one that had been shown previously, i.e., the ‘cued’ one). When such a target stimulus is found, the corresponding assembly is selected for activation, while otherwise (in the absence of a target) no assembly achieves full domination of the system.” This is a parallel mechanism in that it does not compare the stored target with the distractors through a serial search on locations in the display until it finds the target; instead, it selects the most likely stimuli to be the target in an expectation-driven process.

The Usher-Niebur model is not limited to a serial processing, as FIT models are. In FIT models, the selective attention that is at work is attention in its focal mode—that is, spatial attention. Spatial attention can search for one spatial position at a time. This means that the distractors present at different locations in the display are serially searched and compared with the mnemonic trace of the target before they are suppressed. In the Usher-Niebur model, on the other hand, the selective mechanism that is driven by top-down expectations searches for one object at a time. In the preattentive stage, neurons responding to the different objects in a scene (whether they be targets or distractors) are activated in parallel. In the selective-attention stage, the neuronal assembly whose activation is enhanced by the top-down feedback input from working memory eventually wins the competition and suppresses the activations of the other

assemblies. This way, attention selects the target object for further processing; it is serial with respect to the number of targets, but not with respect to the number of distractors.

Spivey (2007) performed a series of simulations purporting to test the efficiency of the biased-competition account of attention in accounting for the amassed data on visual search. Spivey used a normalized recurrence localist network to study whether a model of visual search in which several objects in a scene are all represented in parallel could explain the various experimental results on searching for a target amidst distractors, on target-distractor similarity, and on distractor-distractor similarity. Recall that, according to the standard account for visual search, when an object is searched amidst distractors that share only one feature then search is not affected by the number of the distractors, presumably because the search is parallel and the target item pops out. However, when the search is conjunctive (that is, when the search involves looking for the conjunction of features), then, given the standard assumption that searching for conjunctive features involves attention and serial matching of the objects in the scene with the target's template in memory until a match is found, the search slope increases steeply as a function of the size of the set. Against this model, Spivey's simulation provide an existence proof that a biased-competition model of visual search in which several objects in the scene are all represented by being partially activated, and by being processed in parallel while competing simultaneously to dominate processing, can explain the steep search slopes that are thought to support a serial search pattern without positing a serial search; all that need be posited is parallel biased competition among the representations of objects that are all processed in parallel. The parallel competitive architecture mimics the linearly increasing search function without positing serial matching guided by attention. This shows, in the words of Spivey (2007, p. 228), that "with the normalized competition algorithm, conjunction searches (as well as feature searches) are capable of a wide range of response time by set size slopes. . . . The past results and current simulations suggest that visual search phenomena are best described via a continuum of search efficiency (Duncan and Humphreys 1989), rather than via a discrete distinction between parallel (sensory) and serial (attentional) processing (Treisman 1988)."

A parallel account of selective attention in search tasks leaves open the issue of the role, if any, of spatial or focal serial attention in such tasks. Usher and Niebur (1996) argue that focal serial attention is needed in two search situations: when spatial information that could help locate the

target is available (when, for example, there is a spatial cue as to the location of the target) and when the parallel selection mechanism reaches its limits (as when the target vs. distractor discrimination is difficult because they share features, or when the search task consists in searching for conjunctive stimuli). In these cases, the decision has to be made on the basis of a serial scan of all objects. I would like to add a third reason why spatial attention may be necessary even for a much more extended range of tasks, including the mundane but fundamental for vision task of dealing with the environment by representing the features and objects in it. In a nutshell: since visual processing does not result in detailed iconic representations being stored in memory, when such detailed information for an object in a scene is needed, one must orient one's eyes to the location of the object to gather the required information. This is the task of spatial attention, which, to succeed, presupposes that the spatial layout of the scene and specific positions of objects are retained in memory. I will elaborate on that in the next section.

## 1.2 Visual Representations

It has traditionally been assumed that upon viewing a scene we construct rich sensory representations of all, or most, of the objects in the scene, which we then store in a visual buffer (Neisser 1976). This buffer is supposed to integrate the contents of individual eye fixations so that a complete coherent representation of a scene can emerge (Feldman 1985; Trehub 1994). Attention is thought to be the mechanism that integrates visual features into long-lasting representations of objects (Kanwisher and Driver 1992; Treisman 1993). However, either a visual buffer that integrates representations from successive views of a scene does not exist, as is suggested by the work of Rensink and colleagues, or, if it does exist, it does not store rich point-to-point iconic sensory representations of objects (Hollingworth et al. 2001; Hollingworth and Henderson 2002, 2003).

According to Rensink's work, representations of objects in a visual scene do not accumulate as the eyes move from parts of the scene to other parts. Furthermore, objects are represented in detail only for so long as attention is focused on them. If representations of objects are not stored in a visual buffer, and if they persist only so long as attention is focused on them, then, to interact successfully with the environment, one must be able to shift the focus of processing activities effectively and quickly from one location to another and to select to represent briefly those objects that are indispensable to achieving one's goals. In this case, object representations

do not accumulate, and are not stored; rather, they are assembled as needed. To know where to look for an object in order to focus on it and gather the required representation, one must retain in memory, among other things, a map of relative locations. Spatial attention then orients the eyes to that location. I will now discuss the relevant evidence and the attentional mechanism required to ensure that capability, concentrating on Rensink's (2000a,b) work.

Rensink discusses research on change blindness (CB) and inattention blindness (IB). "Change blindness" refers to the phenomenon in which changes in an image of a real-world scene are not detected or become difficult to detect when made during a flicker, a blink, a saccade, a movie cut, or some other short interruption (Rensink et al. 1997, 2000; Simons and Levin 1997). "Inattention blindness" refers to the phenomenon in which observers attending to a particular object fail to report the appearance of irrelevant or unexpected items (Mack and Rock 1998). Both phenomena, owing to the fact that they can be induced in a large number of ways and to the fact that they occur with real-world scenes, are central to the way we represent the world around us. The explanation of CB is that under normal conditions, changes in the world are accompanied by motion signals in the input that attract attention to their location and render the changes visible. Spatial analysis of shape and spatial relations among objects, and detection of motion, are the most relevant processes on physical properties that form the basis for focused attention (Egeth et al. 1984; McCleod et al. 1991). When these motion signals coincide with and thus are masked by other transients (flickers, saccades, etc.), they cannot draw attention, and CB is induced. The explanation of IB is that the ability to report an event requires attention. If the attention is focused elsewhere, the event may go unnoticed.

The problem CB and IB pose for the view that we store rich and stable representations of visual scenes is that, if we did store such representations, then a simple comparison of the actual scene after the brief interruption with the contents of the buffer (the scene before the interruption), or a simple detection of anomalous structures formed by superimposing the scene before and the scene after the interruption, would suffice to render the changes diaphanous to consciousness and thus reportable. Thus, these phenomena suggest that there is no visual buffer that accumulates and integrates the contents of individual eye fixations, undermining the idea that a detailed representation of a scene is carried across saccades so that the visual system construct a composite perceptual system (Henderson and Hollingworth 1999). If such a buffer does not exist and as a result we do

not build complete representations of visual scenes, then the contents of successive representations due to individual eye fixation cannot be compared and any change in a scene would be difficult to notice.

Rensink (2000a,b) offers a theory of vision that purports to explain visual mechanisms in a way that is consistent with the empirical findings related to CB and IB. His account consists of two parts: a mechanism for vision and attention (the coherence theory), which deals with the “scrutinizing” aspect of vision, and an account of the nature of representations resulting from this mechanism (virtual representations), which deals with the “seeing” aspect (2000b).

The coherence field theory of attention posits three stages of visual processing of a scene.

First, there is an early preattentive stage consisting of three substages during which properties of stimuli are bottom-up retrieved rapidly (within a few hundred milliseconds) and in parallel from a visual scene. This stage is referred to as *low-level vision*. In the first substage (the transduction stage), the photometric properties of stimuli are retrieved. In the second substage (primary processing stage), image properties are measured by means of various filters. The two first substages perform “quick and clean” measurements at the expense of complexity. The third substage (secondary processing stage) performs “quick and dirty” interpretations at the expense of reliability. The interpretations form structures by binding together features retrieved at the previous substages. These structures are the proto-objects that provide local descriptions of scene structure (e.g., three-dimensional orientation, grouping of related edge fragments) and thus correspond to localized structures in the world.

Though proto-objects can be complex structures, they are coherent only over a small region and over a limited amount of time; they have limited spatial and temporal coherence (where “coherence” means that the structures refer to parts of the same system in space and time; in other words, that their representations in different locations and over different times refer to the same object). Proto-objects are very volatile, in that they are either overwritten by subsequent stimuli or else fade away within a few hundred milliseconds; the volatile representations last about half a second (Rensink 2000b). Each time the eyes move and new light enters them, the older proto-objects fade away and new proto-objects are generated.

The second stage involves attention, which “acts as a hand that grasps a small number of proto-objects from this constantly-regenerating flux. While held, they form a coherence field representing an individuated object with a high degree of coherence over time and space.” (Rensink

2000b, p. 1473) This is mid-level vision. Attention has access only to proto-objects, the output of the secondary processing stage, and does not modulate processing in the first two substages. Thus, proto-objects are both the lowest-level operands upon which selective attention can act and the highest-level outputs of low-level vision (the preattentive parallel bottom-up stage of visual processing). Focused attention provides structures that are coherent over an extended region of space and time, and thus it is inextricably involved in object perception—that is, in the perception of objects as they are experienced through our senses. Notice that attention, through the continuity in space and time it provides to objects, is indispensable in perceiving changes in visual scenes, for it allows a new stimulus to be treated as the transformation of an existing object rather than as the appearance of a new object (as we shall see later on, this is accomplished by means of the object-files that the visual system opens for the objects it parses in a scene).

For the third stage, Rensink posits a “nexus,” the place in which the interaction between attention and proto-objects takes place. Since attention combines proto-objects, the nexus consists in a single structure representing the attended object (for example, its shape, size, color, and orientation). It becomes clear that Rensink’s proto-objects may be object parts that, when combined together, constitute a single object or distinct objects that attention combines to form a complex object. Not all properties of the proto-objects can be represented in the nexus; thus, only some among the properties of objects can be represented and perceived at any one time. This is important for perception of change too, since only if the change concerns one of the represented aspects of the object will the change be seen; otherwise, CB results.

When proto-objects are attended, links are established between them (in order to form the attended object) and the nexus. The links are bi-directional, allowing two-way transmission between the nexus and the proto-objects. Bottom-up information allows the nexus to represent selected properties of the proto-objects and allows mapping between the constantly changing retinotopic coordinates of proto-objects and the more stable viewer-centered or object-centered coordinates of the nexus. Top-down information provides stability and coherence to the attended proto-objects. It is the recurrent flow of information between the nexus and the proto-objects made possible by the links between the nexus and the proto-objects that establishes the circuit known as a *coherence field*.

After attention is released, the object dissolves back into its constituent proto-objects, losing its coherence over an extended region of space-time.

There is no attentional aftereffect of the representation of an object; once attention has been withdrawn, the object ceases to be represented (see also Wolfe 1999). This means two things. First, it is wrong to assume that when attention captures objects, they enter a visual working memory where they are stored even when attention is withdrawn. Second, as a corollary of the first point, it is wrong that visual working memory could be identified with the attentional hold, in that attending an object is both a necessary and a sufficient condition for the object to be in visual working memory. Rensink (2000a, p. 26) does not deny that there exists a working memory (which he calls *short-term memory*, abbreviated STM) for objects that have been previously attended, in which traces of object types are stored. But this is different from the standard visual short-term memory (VSTM) or visual buffer, which is supposed to be purely a visual memory that stores representations of object tokens.

Rensink argues that focused attention provides the short-term coherence and stability of one object at a time, although more than one object could be attended at one time if they form a whole new object. The attended object is represented in a short-term buffer as a working representation, but the representation contains a limited amount of information, mainly information about size, shape, color, orientation, location, and motion (Kahneman et al. 1992). This counters earlier claims that attention binds features into a complete representation of an object (Treisman 1993). Once attention is withdrawn, the only memory of the scene that persists is its gross spatial layout and its gist.

Though attention is a necessary condition for seeing change, it is not a sufficient one, since changes to objects may go unnoticed even when the objects are attended to, especially if the changes are unexpected (Simons and Rensink 2005). Simons and Rensink's finding suggests that even when coherent representations of objects are formed through the action of attention, the contents of these representations are usually limited to those that suit the task at hand, and that is why changes in behaviorally irrelevant or unexpected features go undetected even when the object is being attended to. If one considers that information on color, size, motion, location, and orientation is in most cases important to almost any task, one is led to the conclusion that Kahneman et al. 1992 reach: that the information contained in a coherent and stable representation of an object is usually limited to such information.

Accepting that working representations of objects in a scene do exist, Rensink rejects (see the detailed discussion in Simons and Rensink 2005) the strong conclusions occasionally drawn from studies on CB and IB that

these two phenomena suggest that the brain forms sparse or no representations at all (O'Regan 1992; O'Regan and Noë 2001). The working representations are fragile and easily overwritten, but they survive long enough to allow successful recognition performance (Mitroff et al. 2004). In fact, not only do we form representations, but we form multiple representations that can be used in multiple tasks. As we have seen, though attention does not bind features into a complete representation of an object, still the working representations are about a distinct object that persists in space and time and contain information about its size, color, shape, orientation, motion, and location.

The role of attention in vision can explain the findings on change blindness and inattention blindness. If seeing and reporting the change presupposes that attention is being allocated to the points in space and time at which motion change signals occur, the lack of attention renders intelligible why one could fail to report a change in a scene in change blindness or even the appearance of a new object in the scene in IB. Notice that the emphasis is on seeing and reporting a change, not on detecting or perceiving the change. The reason is that, despite the fact that without attention one cannot see and report a change (i.e., that without attention one cannot be made consciously aware of that change), that does not mean that the change is undetected or not perceived; on the contrary, it is detected and may influence the behavior of the perceiver.

This is a clear case of perception without attention and awareness, which Rensink calls "implicit perception." Indeed, studies of change blindness and inattention blindness and studies of perception in the absence of attention (Driver et al. 2001; Humphreys 1999; Kanwisher 2001; Mack and Rock 1998; Merikle et al. 2001; Moore and Egeth 1998; Treisman and Kanwisher 1998) suggest that even when there is no awareness of stimuli when either space-based or object-based attention is diverted elsewhere, stimuli are nevertheless perceived, and grouping of features into some form of objects takes place along the visual system. The claim that some form of object representations can be constructed in the absence of focal (i.e. serial) attention is further reinforced by studies suggesting that there is a parallel selection of parts within objects, and that thus, focal attention need not apply to each visual element separately.

Priming studies show, for instance, that shape can be implicitly registered (Treisman and Kanwisher 1998). Evans et al. (2000), Han et al. (2000), Heinze et al. (1998), Koivisto and Revonsuo (2004), and Paquet and Merikle (1988) present evidence from priming studies and argue that global and local stimuli are processed in parallel at preconscious processing stages

(that is, as early as 100 ms after stimulus onset, which, as we shall see, is roughly the threshold at which some form of awareness enters the picture and preconscious processing ceases) even in the absence of attention, although global stimuli are analyzed at that stage more than local stimuli (which explains in part the global precedence hypothesis of Navon (1997)). Moreover, there is evidence that semantic processing of stimuli takes place when the stimuli are not attended to, and under conditions that preclude awareness of the stimuli (Dehaene et al. 1998; Ladavas et al. 1993; Merikle et al. 2001). Thus, one should not hurry to assert that semantic top-down processing is necessarily accompanied by awareness, a finding that reinforces Kanwisher's (2001) choice to treat perception as independent of any form of awareness and to discuss awareness after perceptual processes have been analyzed.

According to Rensink's coherence field theory, there is no aftereffect of attending an object. Once attention has been withdrawn, there persists no representation of that object token. In view of this, it is natural to ask how could we interact successfully with our environment, given that such interaction presupposes reliable representation of object features in the world. Rensink's answer to that is the idea of a *virtual representation*. Granted that we need to represent various aspects of our environment, it is also true that we never need a detailed representation of all objects and object features in a scene; we need represent only those aspects that are relevant to a task at any time. Thus, instead of forming a detailed representation of all objects in a visual scene, we represent only the object and features needed for the task at hand. The objects in a scene are virtually represented, in that they could be brought forth into the focus of attention and thus represented whenever needed. This presupposes that there is a reliable mechanism that can coordinate attention so that a coherent and detailed representation of an object in a scene could be assembled by picking up relevant information from the environment whenever it is needed. In this case, "the representation of a scene will appear to higher levels as if it is real, that is, as if all objects are represented in great detail simultaneously" (Rensink 2000a, p. 28).

Successful use of virtual representation requires attentional shifts to the appropriate object at the appropriate time. This gives rise to two problems. First, to direct attention to the appropriate object presupposes that the location of the object is known before the allocation of attention. How could this be? Second, and related to the first, supposing that such knowledge is possible, knowing the location of an object in a scene requires a memory of the objects in the scene; but how can there be a memory of

the scene, given that attention has no aftereffects and representations are assembled *in situ*? These are problems related to one of the three different aspects of vision: “seeing.”

To account for a mechanism that coordinates attention, Rensink (2000a,b) proposes a “triadic architecture” of three largely independent systems. The triadic architecture is based on the notion of “pointers,” aspects of structures in the world that are extracted from a scene and serve as pointers to entities in the world. A pointer is a characteristic example of a deictic or indexical representation of objects in the world.

When one views a scene, one might represent the objects in it by forming and storing rich representations, a fact that is disputed by the evidence coming from studies of change blindness and inattention blindness. Alternatively, one could store in one’s memory the location or some characteristic feature of the objects, which, upon needing some information about these objects, would allow one to retrieve the required information. The stored location or feature of an object is a representation of that object in that it comes from that object; it acts as a pointer that has indexed the object and allows one to track it and find more about it. This representation is called “indexical” or “deictic” representation for two reasons. As is true of all indexicals (such as ‘this’ or ‘that’), its meaning consists mainly in the fact that it denotes or points to a certain object in the environment. The representation hardly has an internal structure, although it points to structures in the world, which could constitute its meaning; some philosophers go as far as to claim that an indexical does not have a meaning outside its external relation to the object it denotes. Compare the representation of an object by means of a pointer to its location, which if articulated would consist in a simple “there,” versus its representation by means of a detailed description of its properties, which if articulated would consist in a set of sentences implementing this description. In the latter case, the representation has a meaning on its own right by virtue of its having an internal structure.<sup>3</sup>

Rensink’s triadic architecture consists of the architecture of the coherence field theory that I discussed above in which a third element is added, to wit, a limited-capacity preattentive system that provides a setting that guides the allocation of attention to different parts of a scene and which ensures that attentional shifts be made to the appropriate object location at the appropriate time. The setting system can guide the allocation of attention to the appropriate locations of a scene by representing at least three aspects of a scene’s structure.

First, the gist of the scene (whether the scene is a city, a market, or something else) allows attention to be directed to objects that are important in the context of the scene. At first sight, talk about the gist of a scene perceived during early vision seems problematic, insofar as “getting the gist” seems to rely on a mental abstraction and thus seems to require high-level semantic processing. However, the gist seems to be determined within 100–120 ms after stimulus onset, before the onset of attention (Koch 2004; Potter 1993; Rousselet et al. 2002). It precedes the identification of objects in a scene and thus is extracted by means of simple measures of the image or other properties of the proto-objects—that is, during low-level vision. Of course, for the context to influence perception, the context must be available early on if it is to exert any influence on the perceptual processes. The evidence suggests that this is indeed the case. Similarly, Rensink (2000a,b) and Henderson and Hollingworth (1999) review evidence suggesting that the type of the scene can be identified as soon as 45–135 ms after stimulus onset.

Findlay and Gilchrist (2003, p. 138) point to the fact that, since objects not fixated closer than  $2^\circ$  or  $3^\circ$  are not recognized, eye movements are necessary for the identification of objects within scenes. It is known that the eyes perform three or four saccades in a second. However, the gist of the scene can be extracted from a single glimpse of the scene, which means that the scene’s abstract schema can be evoked with very little delay. It seems that “a initial rapid pass through the visual hierarchy provides the global framework and gist of the scene and primes competing identities though the features that are detected” (Treisman 2004, p. 541).

In chapter 5, I discuss Petitot’s (1995) notion of the “positional content of a scene.” The main point of that discussion is that positional content is nonconceptual and conveys information about nonvisual properties, such as causal relations (e.g.,  $X$  “transfers” something to  $Y$ ). Suppose that one witnesses a scene in which  $X$  gives  $Z$  to  $Y$ . The semantics of the scene consists of two parts: (i) the semantic lexical content of  $X$ ,  $Z$ ,  $Y$  and ‘give’ as a specific action and (ii) the purely positional local content. The latter is the image scheme of the “transfer” type.  $X$ ,  $Y$ , and  $Z$  occupy a specific location in the space occupied by the scene (just as they are the arguments in the three-place predicate ‘to give’). In the image scheme,  $X$ ,  $Y$ ,  $Z$  are thus reduced to featureless objects that occupy specific relative locations, and in that sense they can be viewed as pure abstract places.  $X$ ,  $Y$ , and  $Z$ , which in a linguistic description of the scene are the semantic roles, “are reduced to pure abstract places, locations that must be filled by ‘true’ participants.” A structured set of such relations constitutes the positional

content of a more complex scene, such as “being in a market.” It seems, thus, that the nonvisual properties in which the gist of a scene consists could be retrieved directly from a scene in a bottom-up way. This suggests that ‘gist’ should not be construed as involving an act of mental abstraction, but a perceptual act of retrieving the positional content of the scene, which is abstract in the sense explicated above.

Second, perception of the spatial arrangement of objects in the scene without regard to other features allows attention to be directed to a particular object in a scene. The perception of the layout of the objects, in other words, allows one to know the location in the scene of the object that one seeks. Akin to the gist, the layout of a scene is extracted from the scene by means of low-level vision, and more specifically through the proto-objects at which low-level vision culminates. Recall that proto-objects are construed by Rensink as object-parts, but they can also be objects that are combined to form complex objects. Pylyshyn (2001) thinks of proto-objects as viewer-centered structural descriptions of objects. Be that as it may, the first two aspects involve working representations.

Third, an abstract scene schema that is stored in long-term memory facilitates the perception of both gist and object layout in the scene. At the same time, the gist and the layout facilitate the long-term learning of the characteristics of the particular scene. The scene schema involves the category in which a particular scene belongs and an associated collection of representations, such as an inventory of objects that are likely to be present in the scene and their relative locations.

Rensink’s account of attention, although it puts no emphasis on the competition between top-down and bottom-up information but only on the competition between bottom-up signals generating proto-objects, agrees with biased-competition accounts that there exist two main stages of visual processing, one parallel (in which proto-objects are formed) and one serial (in which some object is selected for further processing). The theory of virtual representation leads Rensink to underscore both the spatial role and the object-based role of selective attention. Attention may be focused either on locations or on objects, because, for virtual representation to be effective, objects must be attended on request by effectively allocating attention to various parts of the scene. The allocation of attention may be guided by the layout of the scene (which emphasizes likely locations of objects in a scene) or by the gist and the causal type of the event in Petitot’s (1995) sense, (which emphasize those objects that are important in the context of the scene), or perhaps both. Recall that the layout and the gist of a scene are retrieved in early vision before the onset

of attention. It is the nature of the task that determines which kind of attention (spatial or object-based) will be deployed, a finding that is reinforced by other studies (Egeth and Yantis 1997; Vecera and Farah 1994; Vecera 2000). Thus, Rensink's work provides a role for spatial attention in addition to those discussed by Usher and Niebur (1996).

As another result of the theory of virtual representation, Rensink goes beyond biased-competition models of visual search in that he posits an explicit mechanism guiding selective attention, whether it be spatial or object-based. This mechanism is lacking from the biased-competition models, since they are specifically concerned with the way objects are selected in a scene, that is, how they win the competition against other candidate objects. The biased-competition models are not concerned with how one has to search a scene to gather information about several objects that may become behaviorally relevant at different times. A reason for that may be that some of the proponents of biased-competition models subscribe to the view that through attention a complete representation of an object is stored in memory and thus one need not constantly search the scene to gather information about objects in it; this information is stored in memory and can be retrieved from there. One could, thus, argue that the biased-competition models are mainly concerned with the "static" role of attention in selecting and perceiving an object at some specific time, whereas Rensink is also interested in the "dynamic" role of attention in providing information about objects on request.

Rensink's non-attentional setting in his triadic architecture has some affinity to Cave and Wolfe's (1994) and Wolfe's (1994) *priority maps* in models of visual search. The Cave-Wolfe model posits a parallel process that guides a focal serial search. Wolfe's (1994) guided search model also posits the existence of a parallel feature-competition stage that guides a subsequent serial visual attention stage. The parallel stage generates a priority map that is used to guide the localized control of the spatial attention mechanism at a serial stage of processing. In other words, the parallel stage forms feature maps whose outputs are pooled into a salience map that guides the attentional focus on salient locations serially from item to item in space in the visual field by representing topographically the relevance of the different parts of a visual scene. In that respect, this system functions the same way Rensink's preattentive system does. The feature-competition stage calculates the order of serial inspection of the visual field, hence the priority map. The priority map is generated after the binding of features, which means that the competition mechanisms are involved after a parallel bottom-up stage in which features are retrieved

from the scene and bound to a certain extent, in accordance with the class of biased-competition models (recall that for Rensink the gist of a scene is retrieved about 120 ms after stimulus onset).

Rensink's work on change blindness and inattention blindness also sheds light on another issue. The fact that without attention one cannot see and report a change (that is, that without attention one cannot be not made consciously aware of that change) does not suggest that the change is undetected or not perceived; on the contrary, it is being detected, and it may influence the behavior of the perceiver. Indeed, there is abundant evidence (Block 2005; Desimone 1999; Desimone and Duncan 1995; Duncan and Humphreys 1989; Fernandez-Duque and Thorton 2000; Fernandez-Duque and Thorton 2003; Kanwisher 2001; Lamme 2003; Merikle et al. 2001; Merikle and Joordens 1997; Rensink 2002; Thorton and Fernandez-Duque 1999; Usher and Niebur 1996; Wolfe et al. 2002) that there is significant processing of information that can affect behavior (for example, cause priming effects, affect semantic decisions, or even enter long-term memory (LTM)) without attention and without consciousness. For this reason, Driver et al. (2001) and Wolfe (1999) think that in change blindness and inattention blindness it is not the case that change is not perceived; change is perceived but cannot be reported and that is why the phenomena should not be called "change blindness" or "inattention blindness" but rather "inattention amnesia." Rensink calls perception without attention and awareness "implicit perception," and associates it with "sensing," one of the three different aspects of vision, the other two being, as we have seen, "seeing" and "scrutinizing" (Rensink 2000b).

Rensink's thesis that once attention is withdrawn from objects the objects lose their constitutional coherence and revert to unstructured collections of object parts is hotly debated. The debate extends to another related thesis of Rensink's, namely that IB and CB phenomena suggest that there is no accumulation of visual information from a series of viewings of a scene into some visual memory buffer. There is evidence (Hollingworth et al. 2001; Hollingworth and Henderson 2002, 2004) that undermines both of Rensink's theses. This evidence suggests that when attention withdraws, information about objects that were previously attended to has been stored in VSTM and/or LTM and can be used for comparisons of scenes or for accumulation of visual information across fixations. Wolfe et al. (2000), for instance, leave open the possibility that representations of objects may be retained in memory for some time once attention has been withdrawn. This has several repercussions regarding the issue of CB and IB that open the road to alternative accounts of the

phenomena to those offered by Rensink and his colleagues, but I will not discuss these accounts here.

What is important for the aim of this book is the clarification of the role of attention in visual processing and in the formation of representations of objects in visual scenes. Recall that the “phenomenal” preattentive percept or proto-object consists of tentatively but uniquely bound features that are segmented from visual information. This representation is a short-lived, vulnerable, and not easily reportable form of visual experience. When attention is directed toward the scene, two things happen (Hollingworth et al. 2001; Hollingworth and Henderson 2002, 2004). First, the sensory and perceptual information extracted preattentively from the scene is augmented with a more abstract higher-level visual representation, which consists of abstract visual categories that are formed in more anterior areas (medial and inferior temporal cortex) and which is stored in short-term visual memory (STVM). These higher-level representations are the representations that are functional across saccades and thus contain the content that is accumulated from saccade to saccade (Henderson and Hollingworth 2003). Although the higher-level representations contain information about the visual form of the scene and preserve a great deal of visual information, they are more abstract than the sensory representations in that they do not have the iconic form of the latter—for example, they do not retain the metric of the scene as the iconic perceptual representations do. They also do not encode the perceptual content in all its detail. For example, they do not encode the exact shade or hue of a color but only the class of the color with a rough description of its brightness or the exact texture of an object. Second, the higher-level representations through the VSTM are consolidated in due time into a more stable long-term memory representation (whether the representations in long-term memory are modal or amodal is an interesting question; see Barsalou 1999 for a discussion).<sup>4</sup>

Both lines of research—Rensink’s and Henderson and Hollingworth’s—recognize that successful visual interaction with the environment would not be possible unless implicit visual memories allowed us to select and retain information across space and time. The difference between the two classes of theories lies in that in Rensink’s work what is retained is the gist of the scene and the gross spatial layout of the objects, so that the observer could know where to focus her attention to acquire the required task-related information, whereas for Hollingworth and Henderson the mnemonic traces of previously attended scenes are richer, in that they also contain visual information about the objects in the scene, although the

representations of objects stored in memory, whether it be VSTM or LTVM, are more abstract than the sensory representations in that they do not have the iconic form of the latter.

Both accounts could use, for instance, Maljkovic and Nakayama's (1994, 1996) priming of pop-out mechanism of visual memory, which uses traces of previously attended features or locations to help guide attention and eye movements toward task-relevant objects. This mechanism goes beyond Rensink's triadic architecture that ensures the successful use of virtual representation, which requires that attentional shifts be made to the appropriate object at the appropriate time, in that in addition to the spatial layout it may include mnemonic traces of object features to guide the orientation of attention. Contextual cueing (Brockmole et al. 2006; Chung and Jiang 1998, 2003; Jiang et al. 2005) may also be a mechanism of visual memory that fits within both accounts but is better suited for Henderson and Hollingworth's work. This is a mechanism that retains in memory both the spatial layout of a scene and the specific positions of objects and can, therefore, efficiently guide search within scenes (Hollingworth 2005, 2006). Notice that both mechanisms are implicit, in the sense that the observers have no awareness of the information they use to guide spatial attention toward task-relevant aspects of the scene and in the sense that explicit knowledge that a new object will appear does not alter the probability or speed of fixating that object.

The research I have discussed thus far brings forth the role of attention in constructing representations of objects and in rendering subjects conscious of these objects. One might be tempted to construe attention as a necessary and sufficient condition for visual conscious awareness—that is, to think that consciousness requires attention and that attention guarantees consciousness. Rensink is not clear on that, although his account of CB and IB seems to suggest that he thinks attention necessary for awareness, which is closely linked to the ability to report a stimulus, although his account leaves open, if it does not outright suggest, that one can have some form of awareness of the proto-objects. As we shall see next, for Lamme neither the necessity nor the sufficiency conditions hold true.

Further evidence for the role of attention in visual processing and awareness comes from Lamme's (2000, 2003, 2005), Lamme and Roelfsema's (2000), Roelfsema's et al. (2000), and Roelfsema's (2005) studies on perceptual processing and the relation between attention, perception, and awareness. Lamme starts by offering neural definitions of the psychological phenomena of perception, attention and awareness. He argues for two kinds of processing that take place in the brain: the feedforward sweep

(FFS) and recurrent processing (RP). In the FFS, the signal is transmitted only from the lower structures of the brain to the higher or more central ones. There is no feedback, and thus no signal can be transmitted from the higher centers to the lower structures during the FFS, although there is evidence that horizontal or contextual modulation occurs very early and involves even area V1 of the visual cortex. This should not be read to mean that the early visual areas at which the FFS takes place do not receive any signals from higher centers. As we shall see in the next chapter, they do. However, the top-down signals are delayed and thus do not affect the FFS processing that occurs at the same areas. Feedforward connections can extract high-level information, which is sufficient to lead to some initial categorization of the visual scene and to some behavioral responses. In RP, signals flow in both directions.

Lamme (2000, 2003, 2005) offers a detailed account of the processes along the ventral pathway, the pathway along which objects are represented for purposes of identification and categorization. When a visual scene is being presented to the eyes, the feedforward sweep (that is, the feedforward propagation of information from the periphery to the center without any recurrent processing) reaches V1 at a latency of about 40 ms. Multiple stimuli are all represented at this stage. Then this information is fed forward to the extrastriate, parietal, and temporal areas of the brain. By 80 ms after stimulus onset most visual areas are activated, and at about 120 ms activation is found in the motor cortex transmitted through the dorsal pathway. The preattentive feedforward processing culminates within 100 or 120 ms after stimulus onset. The processing in the FFS is unconscious and is capable of generating increasingly complex receptive field-tuning properties and thus extracting high-level information that could lead to categorization. Some groupings are computed at this stage; during the FFS, contours are extracted in early visual areas, and form features and patterns of motion are extracted in higher visual areas (V2 to MT). The initial pattern of neuronal activity that results from the stage of parallel extraction of features in a visual scene by specialized areas of the visual cortex has been called "base representation" (Roelfsema et al. 2000; Roelfsema 2005).

Then the feedforward signal reaches area V4, where recurrent processing enters the picture with a delay of about 100–120 ms. Horizontal and recurrent processing allows interaction between the distributed information along the visual stream and, more specifically, between neurons at that area and neurons that have been activated earlier at lower levels. In other words, recurrent processing enables information that has been processed

in higher visual areas to reenter lower-level areas, such as V1 (Lamme and Roelfsema 2000). This first recurrent processing is local in that it is confined to interactions within the visual areas. Lamme (2003; 2005) argues that at this stage visual awareness emerges, and, more specifically, a form of awareness: “phenomenal awareness.” The recurrent interactions at this stage functions as the marker of phenomenal awareness (Block 2005, Super 2001). It seems, thus, that when one perceives, in the sense that one is aware of some perceptual content no matter what this content might be, there is not a terminal perceptual area in which when the signals enter awareness emerges. Instead, awareness is the result of the activation of a whole nexus of brain areas through recurrent processing, the same brain areas that are involved in the processing of the visual stimulus in the first place. Studies by Moutoussis and Zeki (2002) suggest that the difference between mere visual processing and awareness in the nexus of brain areas that are involved in both is marked by an elevated level of activity of neurons when awareness occurs.

At this stage an initial coherent perceptual interpretation of the scene is provided, since features bind further. The base representation resulting from the stage of parallel feedforward extraction of visual features leaves many problems unresolved. Owing to the distributed nature of the neuronal representations formed in the base representation, many features of objects (e.g., location and shape) are represented at different visual areas. When there are multiple objects in a scene, the distributed representation creates the binding problem—that is, the problem of which features belong to some object rather than to another. Horizontal and local recurrent processing addresses this problem by providing “base groupings” (Roelfsema et al. 2000)—that is, by creating conjunction detectors that bind features together (Roelfsema 2005).

Thus, feedback and horizontal connections are involved in the integration of information about different parts of the visual field (Hupé et al. 1998). This contributes to solving the object-segmentation problem for those objects that are not segmented in the first FFS pass. Cortical feedback from V5 (a small area of the superior temporal sulcus) and from MT, for instance, improves discrimination between figure and ground by neurons in either striate areas (V1) or extrastriate areas (V2, and V3) by amplifying their responses and focusing their activity, although the effect is more pronounced for the extrastriate areas (Hupé et al. 1998).

Contextual or horizontal modulation and its effects are well documented (Felleman et al. 1997; Gilbert et al. 2000; Kapadia et al. 1999; Lamme 1995; Lamme and Spekreijse 2000; Lamme and Roelfsema 2000; Lamme et al.

2000; Zipser et al. 1996) in most visual areas. The contextual influences are implemented by horizontal connections that link cells with widely separated receptive fields. In V1, horizontal connections connect mainly neurons that have closely spaced receptive fields and whose preferred stimulus is collinear line elements, that is, cells that have similar orientation tuning (Gilbert and Wiesel 1989; Gilbert et al. 2000; Roelfsema et al. 1998; Schall and Bichot 1998).

There is ample evidence (Cavanagh 1988; Livingstone and Hubel 1987; Moutoussis and Zeki 1997; Zeki 1981, 1993) that the early vision module consists of a set of interconnected processes for orientation, shape, color, motion, stereo, and luminance that cooperate within it. These are functionally independent, they process stimuli in parallel, and they provide input both to each other and to other visual areas that bind incoming information and segregate figures from ground. Studies (Moutoussis and Zeki 1997; Zeki 1981, 1993) also show that color, form, and motion are perceived separately and at different times. Color is perceived first, then shape, then motion. It seems, thus, that the brain consists not only of separate processing systems, but of separate perceptual systems that form a perceptual temporal hierarchy in vision. The same studies suggest that when two areas with different specializations (say, color and shape) project to a third area in the brain, information integration or binding does not occur by direct convergence in that third area (that is, the inputs are not integrated in that third area by a converging process that takes place there), but is brought about by the action of neural connections interlinking the two separate areas—in other words, by means of the contextual or horizontal modulation.

Thus, figure-ground segregation could be based on differences in orientation, disparity, color, or luminance; contextual modulation in V1 provides such a segregation for all these cues (Lamme 1995; Zipser et al. 1996). Contextual effects create the “nonclassical” receptive field, that is, the areas to which the cell is responsive through inhibitory and excitatory connections with other cells. Roelfsema et al. (1998) and Gilbert et al. (2000) claim that horizontal connections and the contextual effects they create may be involved in the propagation of attentional modulation in early visual areas.

The feedback and horizontal modulations of cell activity in early vision suggest that the activity of neurons in early visual areas does not depend solely on feedforward inputs but depends also on the activity of other neurons in the same area with which they are linked through horizontal connections, and on the activity of other neurons in higher-order visual

areas, is the latter being fed back to earlier areas through the feedback projections. In this sense, the activities of V1 cells, for instance, express different aspects of visual processing depending on the latency. At short latencies they represent local field features, whereas at longer latencies, owing to horizontal and/or feedback connections, they may represent various aspects of perceptual organization. In tasks in which many figures segregated from background were shown (Lamme et al. 1999), it was found that at 55 ms after stimulus onset V1 cells are selective for orientation, at 80 ms they are selective for the figure-ground boundary, and at 100 ms they are selective for the figure-ground relationship of the surface features that cover the receptive field of the cell. Top-down attentional effects are registered at a latency longer than 200 ms, a time by which the enhancement and inhibition of cells correlates with the number of objects present in the scene (Lamme et al. 2000).

Feedback connections can enhance or inhibit activation of neurons in lower processing areas only if the latter are already active. Otherwise the feedback projections cannot by themselves activate or silence neurons; only feedforward connections can play this role (Hupé et al. 1998). Note also that these horizontal and feedback interactions mostly occur within latencies shorter than 100 ms and thus belong at a preattentive stage of visual processing (Lamme 2000); as we shall see, attention is effectuated when full or global RP becomes possible at a latency of about 200 ms, since it involves working memory and learning. Even when the horizontal connections are involved in the propagation of attentional modulation in early visual areas, as Roelfsema et al (1998) argue, this modulation, as we shall see in detail in chapter 2, occurs with such a long delay that places the modulatory effects after the termination of the FFS and local RP.

In its first stages, recurrent processing is limited (by attentional suppression) to early visual areas and there is no feedback from cognitive areas. At these levels there is already some competition between multiple stimuli, especially between close-by stimuli. Not all stimuli can be processed in full by the receptive fields that get larger and larger going upstream in the visual cortical hierarchy. This results in crowding phenomena that render close-by stimuli difficult to process separately. However, the competition at this level is limited, and this renders possible the phenomenal awareness of many perceptual groups. Lamme suggests that visual recurrent processing may be the neural correlate of binding or perceptual organization. Whether at this preattentive stage the binding problem has been solved is not clear. The binding of some features of a particular object (say, its color and its shape) requires attention, while other feature combinations

(such as shape and location) and segregations are detected preattentively either in the base representation (i.e., the representation formed during the FFS sweep; see Roelfsema 2005) or by means of binding operators when horizontal and local RP occurs.

At a latency of about 200 ms, recurrent processing may gradually involve higher cognitive centers (e.g., frontal and prefrontal cortex and the mnemonic circuits) in the brain and output areas. In that sense, this kind of RP is global, as opposed to local RP involving only visual areas. Suppose, for instance, that an abstract cue has been presented to a subject in a selection task or in some task that requires that the subject report her experience. In order for the cue to affect selection or allow the subject to report her experience, “parts of the brain that extract the meaning of the cue, and that able to relate this to current needs and goals, must preactivate or otherwise facilitate the appropriate sensory pathways, mostly via corticocortical feedback or subcortical routes” (Lamme 2003, p. 15). The transformation of visual information into motor activity enables a report (or, in general, a behavioral response), which is based on the content of phenomenal experiences.

At this level there is considerable competition, since not many stimuli can interact with the higher levels. Further selection becomes necessary when stimuli reach the brain but only one response is possible; then one stimulus must be selected and processed further so that a response is possible. Attentional selection intervenes at this stage to resolve this competition. The selection results from the combination of the information processing with short-term and long-term memory, which recover the meaning of input and relate it to the subject’s current goals.

The work of Lamme and colleagues shows that access awareness arises about 270 ms after stimulus onset, a result that is supported by studies of attentional blink in which the P3 component of event-related potentials (ERPs) is elicited at latencies of about 270–300 ms (Hopfinger et al. 2004) and by studies correlating behavioral visibility ratings and recordings of ERPs (Sergent et al. 2005). The P3 component is generally taken to index explicit detection (of, say, changes in objects in a scene) and, thus, conscious awareness (Evans et al. 2000; Niedeggen et al. 2001). This means that the subject has access to the content of her perceptual states and can therefore report them; hence the use of the term *access awareness* or *report awareness* for this kind of awareness. Access awareness is the characteristic mark of conscious visual experience.

Block (2005a,b) raises the possibility that a “doubter” might argue that the contents formed up to the local RP level are not really objects, not

even in the sense of proto-objects, and thus they are not contents but proto-contents; they become contents when and if they are grasped by attention and enter the global workspace that constitutes the basis of consciousness. Against that objection, Block cites work of Super et al. (2001b) which shows that the marker of phenomenal consciousness (that is, recurrent processing) is present in monkeys trained to saccade to a target, independently of whether the monkey accesses the target—that is, independently of whether attention focuses on the target. This shows that phenomenal awareness is activated even in the absence of a subsequent phase at which attention intervenes and transfers contents to the realm of access awareness.

The proposal by Lamme et al. that access awareness or conscious visual experience results from global recurrent processing involving higher areas in the brain, such as frontal and prefrontal areas and mnemonic circuits, finds support in recent work (Dehaene et al. 1998; Dehaene et al. 2003; Dehaene and Changeux 2005; Sergent et al. 2005) that puts forth the hypothesis of the “global neuronal workspace model.” According to this hypothesis, the step to conscious visual experience (which Dehaene and colleagues call “conscious perception”) consists in the entry of processed visual stimuli into a global brain state that relates distant areas, including parietal, prefrontal, frontal areas, and anterior cingulate nodes. These higher cortical association areas have neurons that are interconnected through long-distance connections and which send top-down signals to sensory areas, mobilizing them in a top-down manner. As a result of this combined bottom-up and top-down activity, a dynamic neural system emerges that is characterized by recurrent processing involving both visual areas and higher cortical areas. The entry of perceptual information into such a dynamic loop renders it reportable by multiple means, and stimuli gain access to consciousness by mobilizing a global workspace. For a stimulus to enter into the global workspace, its duration, and therefore the strength of the bottom-up signal along thalamocortical connections that the stimulus elicits, must exceed a threshold. Thus, there exists a critical stimulus duration beyond which the early visual areas can receive top-down signals from the workspace neurons reverberating their activation.

The same work also supports the thesis that visual processing consists of two stages of stimulus processing: an early stage of parallel unconscious processing that comprises Lamme’s FFS and local RR<sup>5</sup> and a later stage, of limited capacity, that imposes a bottleneck on visual processing, and which is responsible for access awareness or conscious visual experience. Dehaene and Changeux write: “Because of its long-distance brain-scale connectivity,

the global workspace establishes a central processing bottleneck such that, in the presence of two competing stimuli, processing of the first temporarily blocks high-level processing of the second." (2005, p. 0002) It is at the later stage that stimuli enter the global neuronal workspace. It is important to notice a dimension of this work, not covered much by Lamme, about the fate of those stimuli of the first stage that somehow do not pass the bottleneck of the second stage and thus do not make it into consciousness. The processing of these stimuli is not stopped when they fail to pass to the next stage; it can continue for a long time within the left temporal lobe (Sergent et al. 2005). As a result, conscious and unconscious processing proceed along partially distinct and parallel anatomical pathways, and they may overlap in time (see the discussion on the processing of unconscious stimuli in this chapter).

Lamme and Roelfsema discuss the nature of information that has achieved only local recurrent embedding and therefore has not reached the level of access awareness, and one can only be phenomenally aware of it. The information is situated between feedforward (unconscious or pre-conscious) and globally recurrent (access conscious) processing. The local recurrent interactions relate features to other features (recall that some binding and the formation of complex stimuli also take place during the FFS sweep), allowing binding and segregation to occur; this results in some form of perceptual organization at that stage, which produces "phenomenal awareness." Marr's 2½D sketches are formed at that stage (thus, we are phenomenally aware of them), and in general this stage delivers a structural representation of an object. In addition, motion and size can be retrieved in the preattentive stage, since these are represented in cortical areas in which the FFS and local RP take place. Color is included in the list of features that can be preattentively extracted from a scene, although binding shape and size seems to require attention (Lamme 2003).

The "phenomenal" preattentive percept consists of tentatively but uniquely bound features that are segmented from visual information and are candidate objects for further processing (Driver et al. 2001; Lamme 2003, 2005; Wolfe et al. 2002). This information is a short-lived, vulnerable, and not easily reportable form of visual experience (Lamme 2003, p. 15). Pylyshyn (2001) calls this kind of object representation (that is, the representation output by early vision) a "proto-object." According to Pylyshyn (*ibid.*, p. 361), the "proto-objects" are classes provided by early vision that are, at a first approximation, classes of viewer-centered shapes expressible in the vocabulary of geometry. Early vision, Pylyshyn argues (1999), delivers a small set of alternative proto-hypotheses, in the form of

shape-based perceptual options, from among which focal attention selects an option for further processing. Thus, Pylyshyn's proto-objects are related to those of Rensink (2000a,b), although for Pylyshyn the proto-objects are structural descriptions of objects, whereas for Rensink they should more appropriately be viewed as object parts, or objects that are bound together to form more complex objects.

However, these two ways of construing proto-objects need not be taken as contradictory. Recall that according to Humphreys (1999) there are two ways of representing objects in space, both of which exist in parallel in the visual system and which may serve different purposes: within-objects representations (in which elements are coded as parts of objects) and the between-objects representations (in which elements are coded as distinct independent objects). Thus, it may be that Rensink's work emphasizes the former form of representation, whereas Pylyshyn's work focuses on the latter. This is further supported by the fact that Rensink is mainly interested in the way attention selects among the proto-objects to identify objects in a visual scene; between-objects representations are formed in the ventral system, which is implicated in object identification. Hence, Rensink construes proto-objects as object parts that are glued together by attention to form the stable percepts that lead to object identification. Pylyshyn's work, on the other hand, stemmed from his Multiple Object Tracking experiments, in which subjects had to select points on a screen and mentally group them together, and then follow their motions in space amidst other identical points that served as distractors. This task involves apprehension of relative spatial relations among the points, motion in space, and navigation in the environment, tasks that are closely related to the dorsal system, in which the between-objects representations are constructed. Hence Pylyshyn construes proto-objects as individual objects in space. Of course, in most cases these streams interact in vision and thus both representations are formed in parallel and interact with each other (Glover 2004; Goodale and Milner 1992, 2004).

The formation of perceptual content through local RP is essential for the stimuli to reach some form or other of awareness, meaning phenomenal or report awareness. Walsh and Cowey (1998) review work with Transcranial Magnetic Stimulation (TMS), in which magnetic stimulation in the form of a single magnetic pulse or a train of pulses is applied for about 1 millisecond to the scalp. The magnetic stimulation produces functional disruption in the area affected for a short time (10–30 ms), and thus can be used as a lesion technique. TMS was initially applied to the occipital cortex of subjects performing a letter-identification task (Amassian et al.

1989). When the pulse was applied 0–40 ms after stimulus onset, there was no effect on performance. Application of the pulse 60–140 ms after stimulus onset did impair performance. A more pronounced effect, such that the subjects were incapable of detecting any of the letters, was found for applications of the pulse 80–100 ms after stimulus onset. In the framework of Lamme's model of visual processing, these results can be easily explained, and this provides additional support to the model. More specifically, early application of the pulse affects the FFS sweep but because of the short time of the effect enough information reaches the areas in which local RP and later on global RP takes place, and this allows the subjects to identify and report the letters. When the disruptive effect is applied with latencies of 80–100 ms, it coincides with the critical timing at which local RP since it disrupts the local RP that occurs at about 100 ms after stimulus onset. This prevents the representation of the letter form, and thus the performance deteriorates completely.

Attention determines the passage from phenomenal awareness to access awareness. It does so because a conscious report is a motor output and thus involves the motor cortex, and a selection or a decision in some task is situated between the early RP stage and the motor output. Whenever either a motor output or a decision or selection takes place, the meaning of stimuli must be recovered and related to the subject's goals and needs, and this involves higher cognitive areas in the brain. This, in turn, as we have seen, requires selective attention.

To recapitulate: Lamme, first, defines attention as a selection process in which some input is processed faster, better, and/or deeper than other input. Thus, it has a better chance of producing or influencing a behavioral response or of being memorized. Attentional selection modifies sensorimotor processes. Attention selects, according to behavioral needs, some among the proto-objects that are processed by the visual system in parallel and of which we are only phenomenally aware. Those that are thus selected will enter the realm of report awareness or access awareness. However, there are other forms of selection that are non-attentional. These include the processes that prevent many stimuli from reaching awareness, even when attended to. Such stimuli are the high temporal and spatial frequencies, anti-correlated disparity, physical wavelength (instead of color), crowded or masked stimuli, and so forth. Lamme, second, defines awareness as the occurrence of recurrent processing. Without RP there is no awareness whatsoever. The processes in the FFS are necessarily unconscious. When RP occurs, awareness arises. Initially, when RP is limited (e.g. by attentional suppression) to early areas, there is only phenomenal experience, and thus

this form of awareness is called “phenomenal awareness.” When RP also involves mnemonic and output areas, there is “access awareness.” Since the content of this awareness can be typically reported, this form of awareness is also called “report awareness.” It is at this point that selective attention intervenes to resolve competition that is caused by the processing bottleneck, allowing only some of the information available at that point to be further processed. Thus, the evolution from phenomenal awareness to access awareness depends on attentional selection mechanisms; only the information available at the local RP that is selected by attention will enter the realm of report awareness.

However, whether neurons engage in recurrent interactions, and thus whether one goes from unconscious to conscious processing, is determined by neural mechanisms independent of attention. The type of information of which one is phenomenally aware is situated between feedforward (unconscious) and globally recurrent (access-conscious) processing. As I have said, the information of which one is phenomenally aware is a short-lived, vulnerable, and not easily reportable form of visual experience (Lamme 2003, p. 3). In contrast, the “access awareness” (that is, the awareness that accompanies our normal experience) is more stable and is easily reportable.

Lamme’s account of awareness and attention renders the former independent of the latter. One can have phenomenal awareness without attention and one can have attention without any form of awareness; recall that many of the stimuli never reach beyond the FFS even though they are being attended to. Recurrent processing is all it takes to ensure some form of awareness. Attention, on the other hand, is a selection process that arises for independent reasons and is implemented by different neural circuits; thus, at the neural level attention and awareness can be defined as two fully separate mechanisms. However, attention is related to awareness, since it is when attention intervenes that one goes from phenomenal awareness to access awareness. Attention is the competition between neural inputs for output space. Awareness in general—that is, the passage from unconscious to conscious processing—is the result of recurrent processing, independent of this competition, but its extent and its type depend on this competition.

Kanwisher (2001) reaches more or less the same conclusion. Kanwisher distinguishes perception from awareness: one can be in perceptual states while being unaware that one is in such states. Philosophers call these states “subpersonal,” since they are not available to the person’s awareness. Perceptual awareness “involves not only activation of the relevant percep-

tual properties, but the further construction of an organized representation in which these visual properties are attributed to their source in external objects and events" (ibid., p. 90). Thus, perceptual awareness presupposes binding of activated features with a representation that specifies a specific token, as opposed to type, object. This binding requires (ibid., p. 108) "visual attention," which thus becomes crucial for visual awareness, although attention is not, strictly speaking, necessary for awareness, since we can become aware of things that we do not attend to; one can be conscious of an unattended voice although not of the spoken words (Fernandez-Duque et al. 2003; Treisman 2004; Treisman and Kanwisher 1998), since, as we shall see next, there are many ways one can be aware of a stimulus. Kanwisher's "perceptual awareness," insofar as it requires attention, is equivalent to Lamme's and Block's report awareness. However, for Kanwisher various forms of binding occur before the subject becomes aware of a percept.<sup>6</sup>

There are many ways one can be aware of a stimulus. One may be aware of a mere presence (as opposed to absence), or of certain of the features of the stimulus, or of the category of the object present, or of the gist of a complex scene (Kanwisher 2001, p. 97). Thus, one can be visually aware of features or of the presence of an object and one can be also aware of the content of an organized representation. These are different kinds of awareness and only awareness of an organized representation requires visual attention. It is reasonable to argue, therefore, that the distinction between phenomenal awareness and access awareness or report awareness is latent in Kanwisher's work. The distinction between the awareness of the mere presence of an object with some properties and the awareness of the content of an organized representation is Kanwisher's distinction between the perceptual awareness of a type and the perceptual awareness of a token. Notice that the fact that without attention one can be aware only of a type instead of a token does not mean that one is not aware of a specific object. It means, rather, that without attention one cannot be aware of the fact that one has a representation with a specific content; one only perceives the content of a representational state that one has, but one is not aware of the fact that one has an organized representation of an object. To put it differently: One does not relate the features to a source in space and time, in the sense of having a representation of these features as features of that specific object token, although one perceives the features bound to an object. Representing a specific object-token involves a separate mental act that requires attention. Treisman and Kanwisher (1998) remark that although object recognition can occur within 100–200 ms after

stimulus onset, it takes another 100 ms for the required subsequent processes to bring this information into awareness so that the perceiver is aware of the presence of a token object, which means that the perceiver has access awareness to the contents of her cognitive states only after the object has been identified. This fits well with work by Rensink and Lamme, since in their accounts awareness requires attention, and it is attention that allows object identification. It is also consistent with Lamme's finding (which I will discuss in chapter 2) that access awareness or report awareness arises much later than the onset of spatial attention and that in that sense awareness is a late development and has a broad dorso-parietal distribution.

Kanwisher argues further that the *contents* of awareness are being represented in the domain-specific areas of the ventral system, where the activation of a large region, the lateral occipital complex (LOC), shows strong correlation with awareness (2001, p. 98). Kanwisher speculates that the identification of an object takes place in the ventral system, a hypothesis that is also advanced by studies on the distinction between ventral and dorsal systems (which I will review in chapter 3), although, as she notes, there are known exceptions to her hypothesis. But the *content-independent aspects* of perceptual awareness—perhaps what it feels like to be aware of something—depend on the interaction between the attentional network in dorso-frontal-parietal areas with the dorsal pathway. Kanwisher's claim that the content-independent aspects of perceptual awareness may be correlated with activity along the dorsal system, whereas the contents of awareness being represented along the ventral system conforms with Matthen's (2005) view that dorsal processing gives rise to the feeling of presence of objects in a visual scene, although awareness of the features of these objects occurs in the ventral system.

Kanwisher's claim that representations of objects that can be formed without attention (and thus without report awareness) correspond to representations of types of objects rather than to representations of tokens of objects fits well with both Lamme's account of vision and Rensink's. Recall that for Kanwisher representation of a type rather than of a token of an object does not mean that one does not perceive a specific object. One is aware of the existence of an object; and notoriously we are not aware of types, only of specific object tokens. For Kanwisher, representation of a type rather than of a token means, rather, that one is not aware of the fact that one has a representation with an organized content; one perceives the content of one's representation that is formed before the onset of attention. This, in turn, implies that the preattentive information about an

object is restricted to what information is retrieved directly from the environment in early vision and cannot benefit from information stored in memory or from information constructed in later stages of vision. As we have seen, in Lamme's and Rensink's accounts, in the absence of attention, it is possible to construct only some fleeting unstable representations of objects that contain sparse information and lack specific details about objects. Thus, these representations correspond to types rather than tokens of objects, types being at a more general level of abstraction and thus containing less information than tokens. This remark will help me elucidate (in chapters 4 and 5) some problems besetting the philosophical discussion pertaining to nonconceptual content, and more specifically the issue of the fine-grained as opposed to coarse-grained nature of the non-conceptual content of experience.

### 1.3 Top-Down Attentional Influences on Perception

As we have seen, in theories that construe vision as a biased-competition account attention is viewed less as the enhancement of the processing of the attended information than as the result of the biased competition between neuronal populations that encode environmental stimuli. The bias can be either bottom-up or top-down, and usually it is both. Closure and common region, for instance, are properties of the stimulus that bias the competition toward segmenting and selecting objects that have the properties of forming closed or common regions. The goals of the observer, on the other hand, affect these same processes in a top-down way; obviously, we tend to select in a scene those objects that are behaviorally relevant.

According to Vecera (2000) there are at least three sources of top-down information that can bias either object segregation and/or object attention: object-recognition processes, perceptual set processes, and spatial attention processes. It is known that familiar objects have an advantage over unfamiliar objects in object figure-ground segregation (Peterson 1994; Vecera and O'Reilly 2000) and in object attention (Vecera and Farah 1997). Using Desimone's (1999) account of biased competition, one can explain these findings by arguing that familiar objects, which are stored in visual long-term memory, in the appropriate context (that is, when they become task relevant), activate cells in visual working memory that represent the familiar objects' features, thus providing top-down feedback that enhances the activation of neurons in the visual cortex that respond to these objects, giving them an edge in their competition against neuronal assemblies that respond to unfamiliar objects.

The second top-down source of bias on object segmentation and attention is *perceptual set*, which refers to the expectancies or goals held by the observer and which, in the context of an experimental set-up, are usually determined by the experimenter's instructions (Vecera 2000). Neisser and Becklen (1975) and Baylis (1994) have shown, indeed, that perceptual set can influence object-based attention. Folk et al. (1992) uses the term *attentional control settings* to denote all those factors that guide perceptual acts (that is, the perceptual goals held by the observer during the task, such as visual search). Such goals may include either the experimenter's instructions (search for a certain object, or focus at that location) or the subjects' plan of action (search for the book in the library). When it comes to object-based attention, perceptual set corresponds to Duncan and Humphreys's (1989) *template*, that is, the description of a target object, as defined by the experimenter, that the observer must keep active in her working memory for the duration of a task. Being active, this description enhances and thus biases the activations of neuronal assemblies that represent the target object and allows the selection of the target object. In this way, perceptual set influences object-based attention.

The activation of a template or a perceptual set may explain the process by which perceptual set operates with bi-stable stimuli (stimuli that support two perceptual interpretations, such as the Necker cube or the duck/rabbit figure), in which only one stimulus is present and there is not a target object that must be selected amidst other distractors; this is a case of figure-ground segmentation. In this case, the template facilitates one interpretation over another. Peterson and Hochberg's (1983) work with bi-stable stimuli sheds light on the mechanisms that underlie the way perceptual set biases object segmentation, that is, on the visual processes involved in top-down biases. Their findings show that the intention of the observer (i.e., that she is looking for a certain figure) does not affect by itself the organization of the stimulus. Some crucial points of fixation influence the organization of the stimulus, that is, by stimulus bottom-up information. The way a bi-stable stimulus can be perceptually interpreted depends on where the observer fixates her attention, because there are in the figure crucial points fixation on which determines the perceptual interpretation. This means that the mechanism underlying the bias of perceptual set in figure-ground segmentation involves the voluntary control of spatial attention: the instructions of the experimenter, or the attentional setting in general, induces the observers to allocate their attention to a specific region in the stimulus (Peterson and Gibson 1994). This means that the mechanism underlying the bias of perceptual set in figure-ground segmentation

involves the voluntary control of spatial attention and not directly the modulation of early perceptual processing. Further evidence that the perception of bi-stable figures is not determined during early visual processing by suppression of monocular cells but in higher visual areas (such as V4 and MT) in which shape is encoded is adduced by Leopold and Logothetis (1996) and Logothetis and Schall (1989). I discuss this evidence in chapter 2.

The third top-down source of bias on object segmentation and attention is spatial attention. In fact, it seems that spatial attention may be the mechanism that underlies perceptual set, in the sense that the effects of perceptual set are mediated by the control of spatial attention. In other words, the cognitive states of the observer induced by perceptual set drive the observer to allocate her attention to a region in space that is behaviorally salient. There is evidence that spatial attention affects figure-ground segregation (Driver and Baylis 1996). Subjects performed a contour-matching task with ambiguous figure-ground displays in which they had to match the contour of one of the regions of the ambiguous display. Before the display, a spatial “pre-cue” appeared that either predicted or did not predict the region that the subjects would have to match. The subjects performed faster only when the cue was predictive of the region to be matched. Since only the pre-cue influenced performance, the study shows that spatial attention influenced the figure-ground matching process and thus object segregation.

As in the studies by Peterson and Gibson (1994), Peterson and Hochberg (1983), and Driver and Baylis (1996), spatial attention seems to be the mechanism that implements the effects of perceptual set by guiding the observer to focus attention on critical points at locations that bias the competition and determine the outcome of the visual process. Now, we have seen that object segmentation takes place at various levels of visual processing both early and late. If the effects of spatial location can be registered with latencies that are within the time course of early perceptual processing, then this seems to be clear evidence for modulation of early perceptual processing by cognition through the effects of endogenous (i.e., cognitively driven) spatial attention. As we shall see in the next chapter, the spatial effects are indeed registered at short latencies (about 70 ms after stimulus onset), and thus there is a *prima facie* case for the cognitive penetrability of perception. However, things will turn out differently, owing to what I will call the indirect character of the manner in which spatial attention influences perception.

This section has shed some light on the role of top-down constraints in object segmentation and identification. A theme emerged clearly from the

discussion, namely the predominant role of attention in general and spatial attention in particular in mediating the top-down cognitive influences on perception. Thus, attention becomes crucial in any discussion of the interface between perception and cognition, and therefore in any discussion of the cognitive penetrability of perception.

#### 1.4 Concluding Comments

In the next chapter I will discuss the effects of attention on perceptual processing, emphasizing the role of spatial attention. But here I would like to note some things and recapitulate the main findings of my first chapter.

First, the reader will have noticed that Lamme's phenomenal content is very similar to Pylyshyn's (1999, 2003) proto-hypotheses or proto-objects and to Vecera's and Rensink's proto-objects, in that phenomenal content is retrieved in a bottom-up preattentive stage from the visual scene, and has a shaky existence (in the sense that it has limited spatial and temporal coherence) that differs from the stable percept of which one has access awareness or report awareness and which is delivered at the final stage of visual processing. Furthermore, Lamme and Rensink agree that it is attention that makes the formation of the stable percept possible. The proto-objects are not cognitively accessible and thus one does not have access awareness or report awareness of them, and they are in competition for further processing. However, they are within the realm of phenomenal awareness—Block's (1995) phenomenal consciousness. The proto-objects and their features constitute the content of Lamme's and Block's phenomenal awareness and of Raftopoulos and Muller's (2006a,b) nonconceptual contents of experience. Thus, the locally recurrent processing is the neural correlate of phenomenal experience *per se*, or phenomenal awareness (Block 1995). Lamme's work suggests that even though the output of the ventral system is the content of our ordinary experience of which we are aware, a substantial part of the processing in the ventral system is unconscious; it also suggests that there is information processed in this system that never reaches awareness even when attended to, and information of which we are only phenomenally aware and which we can report only with difficulty, if at all.

Second, note that all the aforementioned models differ from traditional models of selective attention (such as FIT) that also posit that visual processing consists of two stages. FIT distinguishes two stages of visual processing: a preattentive parallel stage at which all information across the visual field is processed and which extracts primitive features from the scene

without integrating them, and a serial attentive stage at which only some information is selected for further processing and is integrated to form shapes and eventually objects (Treisman and Gelade 1980). The difference is due to the fact that, in the class of models of attention and of visual processing discussed here, the parallel stage of processing delivers structures that integrate to some extent features in the stimuli. There is, indeed, extensive evidence that in some cases complex information can be extracted in parallel across the visual display (Enns and Rensink 1991; Gilchrist et al. 1997). Some researchers (e.g. Wolfe et al. 2002) argue that the object-segmenting stage (that is, the separation of object-structures from the background) is preattentive, rather than being performed by serial scanning for each likely object location. These structures are the proto-objects (Driver 2001; Pylyshyn 2001; Rensink 2000a,b; Scholl 2001; Vecera 2000).

Thus, the preattentive parallel stage does not consist merely in the extraction of primitive features that are not integrated. As Spekrijse (2000, p. 1179) remarks, “pre-attentive mechanisms transform the visual input rapidly and in parallel, and parse the image into coherent parts. One of these parts may pop out and trigger a behavioral response. However, in many cases pre-attentive mechanisms are not sufficient, and visual attention needs to be invoked.” If one substitutes “proto-objects” for “parts” and reads “coherent” bearing in mind the role of attention in the formation of coherent representations out of the more volatile proto-objects, then the above statement fully expresses the thesis defended in the present chapter.

Feature integration, as a process that binds parts of a scene into units, and thus object segmentation or segregation, takes place at many levels of visual processing, some early and some late (Driver et al. 2001; Scholl. 2001). Lamme’s (2000; 2003), Lamme and Roelfsema (2000), and Roelfsema’s (2005) model of visual processing emphasizes, among other things, that feature binding takes place at many levels of visual processing and that top-down and lateral recurrent interactions between cortical areas are important in feature binding. Though the preattentive stage is naturally related to the feedforward sweep and the recurrent processing is naturally related to attentive grouping, this does not mean that there is no grouping without attention. As we have seen, Lamme argues that grouping and recovery of structure take place both during the FFS sweep and during the stage of local RP—that is, before the onset of attention.

To sum up: The evidence examined in this chapter suggests that during early vision there is a bottom-up (in the sense of a process that is guided

only by the stimuli and not by cognitive influences), preliminary segregation of the sensory data into separate candidate objects, or, rather, proto-objects. Top-down effects, including familiarity with objects or scenes or some form of attentional setting, may override this initial segregation in favor of some other parsing of the scene into objects. The top-down effects also resolve ambiguities when the bottom-up processes do not suffice to segment a scene into its objects (Treisman and Kanwisher 1998). However, these top-down effects occur after early vision has performed its first pass into parsing the scene into separate objects. Feature integration and object segregation, thus, is better seen not as a separate stage of visual processing higher in the brain, but as “an emergent phenomenon due to interactive activation among the cortical areas” (Deco et al. 2002, p. 2939).

It seems, therefore, that there are various stages within visual processing that can be summarized in the following distinction among three stages of visual processing, to wit, sensation, perception, and observation<sup>7</sup>: All processes that apply to the information contained in the retinal image fall within the scope of *sensation*. Thus, we have processes that compute information on light intensity. Sensation includes parts of early vision, such as those processes that compute changes in light intensity by locating and coding individual intensity changes. Sensation includes Marr’s *raw primal sketch* that provides information about zero crossings, bars, blobs, boundaries, edge segments, and so on.<sup>8</sup> The idea is that much of the information about surfaces is encoded in changes in the intensity of reflected light on the retina. Thus, the task of the very early visual system is to decode this information by locating, representing, and interpreting both the intensity changes and the ways in which the intensities are reorganized at various spatial scales by more abstract properties, such as the alignment of termination. Sharp changes in intensity, for instance, are interpreted as surface boundaries. Since, however, this is not a world of uniformly illuminated smooth flat surfaces, the visual system must also represent and interpret gradual changes in intensity. The properties of stimuli recorded at this level (high temporal and spatial frequencies, anti-correlated disparity, etc.) never reach awareness. There are non-attentional selection mechanisms involved here that filter out information (Lamme 2003, 2004). In neuroscientific terms, sensation consists in those processes that belong to Lamme’s feedforward sweep. It may be that they occur before the binding of features extracted from the retinal image. The “image” resulting from sensation, which initially is cognitively useless, is gradually transformed along the visual pathways in increasingly structured representations, via *perception*.

The processes that transform sensation to a representation that can be processed by cognition constitute *perception*. The output of these processes is a cognitively impenetrable content that is retrieved from a visual scene in a bottom-up way. A subset of this output—that which can be brought to a kind of awareness called “phenomenal awareness”—is the “phenomenal content.” In Lamme’s theory, phenomenal awareness requires local recurrent processing. It follows that only content that is formed by means of local RP can be “phenomenal content.” Another subset is the content of subpersonal information-processing states. As an example of perception, consider Marr’s various grouping procedures applied to the edge fragments formed in the *raw primal sketch*. They yield the *full primal sketch*, in which larger structures with boundaries and regions are recovered. Through the *primal sketch*, contours and textures in an image are captured in a purely bottom-up way, although processing at that level involves lateral and local top-down flow of information, which, however, being within early vision, does not threaten the bottom-up character of the relevant processes. Perception comprises the intermediate-level vision that includes processes (such as the extraction of shape and of spatial relations) that cannot be purely bottom-up but which do not require information from higher cognitive states, since they rely on lateral and local top-down flow of information (Hildreth and Ullmann 1995). Note that since the extraction of shape and of spatial relations require local RP, they are within the scope of phenomenal awareness. Thus, being nonconceptual, perceptual processes are not affected by our knowledge about specific objects and events. In Marr’s model, the  $2\frac{1}{2}$ D sketch is the final product of perception. As we have seen, spatial relations, position, orientation, motion, size, viewer-centered shape, surface properties, and color are all bottom-retrievable by low-level vision processes. It may be that, in Lamme’s framework, perception consists in those stages of the FFS that bind features in the image, such as edge fragments, and thus result in states that have a rudimentary structure, and in the stage of vision that involves local recurrent processing. Both sensation and perception constitute Pylyshyn’s (2001, 2003, 2007) early vision.

All subsequent visual processes fall within *cognition*, and include both the post-sensory/semantic interface at which the object-recognition units intervene, as well as purely conceptual processes that lead to the identification and recognition of the array (high-level vision). At this level, we have *observation*. In Marr’s theory, the culmination of visual processes is the *3-dimensional model* of an object. The recovery of the objects cannot be purely data-driven, since what is regarded as an object depends on the subsequent usage of the information, and thus is cognitively penetrable.

Several theories of vision hold that object identification is based on part decomposition, which is the first stage in forming a structural description of an object and which seems to depend on knowledge of specific objects. Other theorists, including Edelman (1999), propose that objects are identified by template-matching processes. Object recognition requires matching between the internal representation of an object stored in memory and the representation of an object generated from the image. Similarly, template matching relies on knowledge of specific objects and is, consequently, cognitively driven, since the templates result from previous encounters with objects that have been stored in memory.