Whitney, D., J. Haberman, and T. D. Sweeny. "From Textures to Crowds: Multiple Levels of Summary Statistical Perception." *The New Visual Neurosciences*. Ed. J. S. Werner and L. M. Chalupa. Cambridge: MIT, 2014. 695-710. Print.

49 From Textures to Crowds: Multiple Levels of Summary Statistical Perception

DAVID WHITNEY, JASON HABERMAN, AND TIMOTHY D. SWEENY

Billions of bits of information arrive at the retina every moment, but our awareness is limited to a small fraction of this information. There are many bottlenecks in visual processing, ranging from physiological and iconic bottlenecks (Nakayama, 1990), attentional bottlenecks in space (Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1997; Whitney & Levi, 2011) and in time (Battelli, Pascual-Leone, & Cavanagh, 2007; Franconeri, Alvarez, & Enns, 2007; Marois, Yi, & Chun, 2004; Raymond, Shapiro, & Arnell, 1992), capacity limits in attention and memory (e.g., Franconeri, in press; Luck & Vogel, 1997; Scholl & Pylyshyn, 1999), and many others. One way in which the brain surmounts some of these bottlenecks, which occur at multiple levels of visual processing, is by representing the statistical regularity of the world in the form of condensed ensemble representations. The leaves of a tree, the blades of grass, and the tiles of the floor are similar and redundant, giving rise to the percepts of "treeness," "lawnness," and "floorness," respectively. The individual components of each texture are lost in favor of a concise, summary statistical representation: an ensemble percept.

In this chapter we review summary statistical perception. Most of the work on this topic has focused on perception of average and variance in displays (e.g., the average length of blades of grass, the average size of a tree's leaves, or the average expression in a crowd of faces). Several other chapters in this volume discuss in more detail texture perception (chapter 45 by Landy), crowding (chapter 48 by Levi), and scene and gist perception (chapter 51 by Oliva). Although these are very clearly related, our focus is on the intersection between these topics, on the nature of ensemble representations, what can form an ensemble and at what levels of visual processing, and how ensembles might be represented in the brain.

The concept of summary representation has recently generated significant interest and debate within the vision science community (Alvarez, 2011; Alvarez & Oliva, 2008, 2009; Ariely, 2001, 2008; Chong & Treisman, 2003, 2005a, 2005b; de Fockert & Marchant, 2008; Haberman & Whitney, 2007, 2009; Koenderink, van Doorn, & Pont, 2004; Myczek & Simons, 2008; Simons & Myczek, 2008). Also sometimes called ensemble coding or ensemble perception, summary representation refers to the idea that the visual system naturally and directly represents an emergent quality (i.e., the gist) of a set of similar items (such as blades of grass). Such a system is intuitively appealing in terms of computational efficiency, and it has far-reaching implications for understanding awareness. For example, Chong and Treisman (2003) and, more recently, we (Haberman & Whitney, 2009) and other authors have suggested that summary representation can provide coarse information from sources across our entire field of view, driving the compelling impression that we have a complete and accurate grasp of our visual world (Haberman & Whitney, 2009). Thus, the "grand illusion" (Noe, Pessoa, & Thompson, 2000) may not be an illusion at all but rather a noisy summary representation of all that we survey. Put another way, although many of the individual details of a scene are inaccessible, ensemble coding may provide a viable algorithm to keep the gist ever present.

EARLY CONCEPTUALIZATIONS OF SUMMARY STATISTICAL PERCEPTION

The concept of ensemble representation is not a new one. Aristotle described perception as a mean of sensory inputs, which could be used to identify stimulus changes as the sense organ gathered more information. Empirical examination of this phenomenon began centuries later with investigations of Gestalt grouping (Wertheimer, 1923), although this early conceptualization was not referred to as ensemble or summary statistical perception, per se. The Gestaltists viewed emergent

8857_049.indd 695 4/18/2013 5:36:21 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

object perception as a synergy of lower-level inputs; the final percept was more than the sum of its parts. Researchers argued that the grouped object was the favored percept and that the individual features were (at worst) lost or (at best) difficult to perceive (Koffka, 1935). Although Gestaltists outlined several basic heuristics by which the visual system groups features (similarity, proximity, common fate, etc.), the underlying mechanism(s) driving this grouping, as well as the algorithm that supports it, remained elusive. It may be that Gestalt grouping amounts to a summary statistical representation, and the mechanism of ensemble coding may provide an explanation for several Gestalt phenomena.

Although Gestalt phenomenology helped to define some elemental principles of object perception, researchers in this area were not explicitly thinking in terms of ensemble perception or summary statistical representation. Some of the earliest explicit work on ensemble coding was done from a social psychology perspective. In an extensive line of research Norman Anderson outlined a simple yet flexible model called "integration theory" (Anderson, 1971). His work demonstrated that a weighted mean more precisely captured how information is integrated than a summation model. For example, subjects rated another individual more favorably when that person was described by two extremely positive terms compared to when that person was described by two extremely positive terms in addition to two moderately positive terms (Anderson, 1965). Integration theory was extended to numerous other social contexts, including group attractiveness (Anderson, Lindner, & Lopes, 1973), shopping preferences (Levin, 1974), and even the perceived "badness" of criminals (Leon, Oden, & Anderson, 1973). Thus, it appears that humans readily integrate semantic as well as social information, although the mechanism behind this process remains largely unknown. The implication is clear, however: Social perceptions and attitudes may hinge on the same sort of underlying summary computations that allow us to perceive the gist of sets of visual features like the average direction of snow blowing in a blizzard.

The modern era of summary statistical research can be divided into two stages. Psychophysical work in the 1980s and 1990s demonstrated that humans integrate low-level motion into something akin to an ensemble percept (Watamaniuk & Duchon, 1992; Watamaniuk, Sekuler, & Williams, 1989; Williams & Sekuler, 1984). These researchers proposed straightforward mechanisms for perceiving the average; local information may be pooled across a population of low-level motion detectors operating in parallel (Watamaniuk & McKee, 1998). Although these early accounts did not explicitly refer to ensembles or summary statistics, they laid the groundwork for a flood of modern work across an

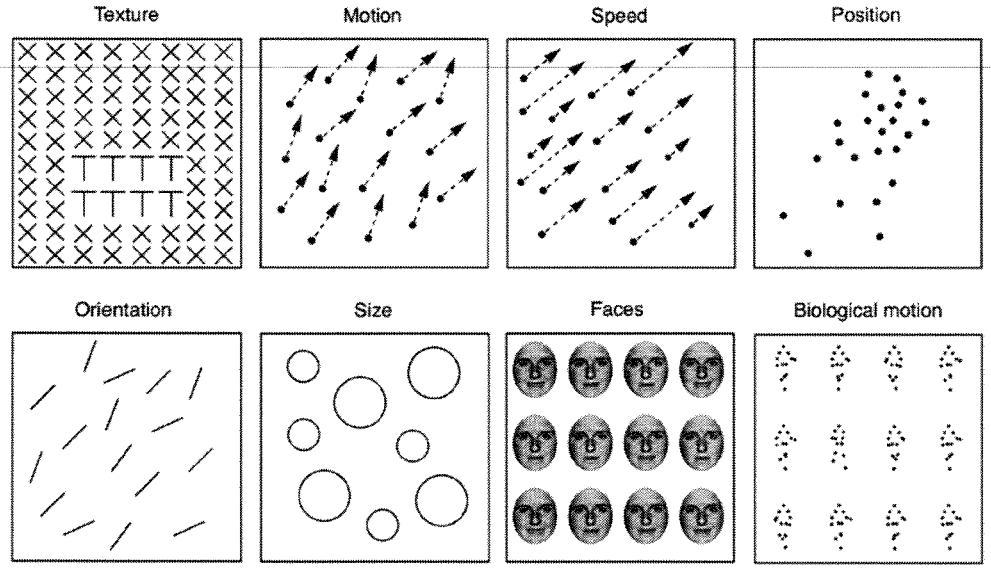

FIGURE 49.1 Summary statistical perception occurs across a wide range of stimuli. The flexibility of ensemble representation suggests that it occurs across multiple levels along the visual hierarchy.

Werner—The New Visual Neurosciences
8857_049.indd 696
4/18/2013 5:36:21 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

impressive range of low- to high-level visual features (figure 49.1).

The current era in the study of ensemble statistics began with the discovery that humans also derive a summary representation for the *size* of a set of arbitrary objects (Ariely, 2001; Chong & Treisman, 2003, 2005a, 2005b) and that this summary representation is favored over a representation of the individual items composing the set. The striking aspects of this research are twofold. First, it suggested that we perceive ensembles implicitly, perhaps through parallel mechanisms. Second, it showed that summary statistical perception could occur for objects. This raises several interesting questions, including these: Are there low-level feature detectors designed to operate on object size in a manner akin to motion or orientation? If not, how does average size perception, if it is indeed parallel, bypass traditional limitations of serial attention? Does ensemble coding extend beyond low-level stimuli to complex objects important for social interactions (e.g., faces and moving bodies)?

Although open questions remain (some of which are addressed below), it is clear that ensemble coding is connected to several areas of vision science, and this, in part, explains the growing interest in summary statistical perception. In addition to providing a window on gist, ensemble perception has implications for the way we understand visual search, texture, depth, scene perception, object recognition, spatial vision, attention, and awareness. The remainder of this chapter surveys the history of this subfield, highlights in greater detail some of the more influential work, and speculates as to where future work should be directed.

MULTIPLE LEVELS OF SUMMARY STATISTICAL PERCEPTION

We begin our survey by discussing which kinds of visual features are known to form ensemble percepts and at what levels of visual processing these summary representations might be formed. The purpose of this approach is to highlight the incredible versatility of ensemble coding for perceiving various feature dimensions and to reveal the possible stages of visual processing in which ensemble codes are likely to be generated. Throughout our review we highlight evidence that ensemble percepts are formed implicitly and automatically but are nonetheless susceptible to manipulations of attention. As with the study of almost any visual process, exploring the role of attention will help to reveal the underlying mechanisms of summary statistical perception and refine our understanding of awareness.

Perceiving Average Motion and Speed

As mentioned above, low-level motion was one of the first visual features shown to produce an ensemble or gist percept. Humans precisely perceive the average direction of a group of dots moving along unique local vectors (Watamaniuk, Sekuler, & Williams, 1989; Williams & Sekuler, 1984). Similar results are found for a group of dots that vary in speed (Watamaniuk & Duchon, 1992). Rather than perceiving each moving dot individually, the dominant percept is the average direction or speed of motion. Although interesting on their own, these demonstrations also provide descriptions of pooling mechanisms that might underlie summary statistical perception with other features. For example, perceiving the average direction of motion from a set of moving dots (or blowing snow) is consistent with established physiological mechanisms of motion perception (Britten & Heuer, 1999; Britten et al., 1992; Newsome & Pare, 1988); information may be pooled across low-level motion detectors operating in parallel, potentially obviating the involvement of serial attention (Watamaniuk & McKee, 1998; but see also Bulakowski, Bressler, & Whitney, 2007). These early accounts were not referred to as ensemble perception per se, but they clearly provide some of the first evidence of summary statistical perception.

Perceiving Average Position

Several psychophysical experiments have shown that humans are sensitive to average or centroid position (Hess & Holliday, 1992; Morgan & Glennerster, 1991; Whitaker et al., 1996). Recent work shows that this sensitivity is based on a real statistical decision, much like a *t*-test. Fouriezos, Rubenfeld, and Capstick (2008) found that in an attempt to judge which of two crowds of vertically oriented bars had the greater average height (which could rely in part on a judgment of the average position of the top of the bars), performance was improved when groups with more bars were visible, but it was impaired when these groups had higher variability.

Perceiving average position can occur when awareness is compromised, such as in crowding. Greenwood, Bex, and Dakin (2009) asked observers to indicate whether a horizontal bar intersected above or below the midpoint of a peripherally located vertical bar. Similar flankers with above- or below-the-midpoint intersections accompanied the target. Position information in the flankers influenced where observers perceived the intersection in the target; observers based their decisions on the pooled position information across the

Werner—The New Visual Neurosciences
8857_049.indd 697 4/18/2013 5:36:21 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

flankers and target—*position averaging*. Complementary work by Alvarez and Oliva (2008) further suggests that selective attention may play a minimal role in ensemble position perception. Using a multiple object-tracking task (Intriligator & Cavanagh, 2001; Pylyshyn & Storm, 1988), Alvarez and Oliva (2008) found that even when observers were unable to localize individual unattended objects, they could localize the centroid of those objects, and their performance was perfectly predicted by averaging the noisy representations of the individual objects. Although Chong and Treisman (2005b; discussed below) demonstrated that distributed attention could improve an estimate of the mean, this work (Alvarez & Oliva, 2008) suggested that ensemble position might be derived even beyond the focus of attention.

Perceiving Average Orientation

Humans readily perceive average orientation. Dakin (2001) was one of the first to demonstrate this when he showed that humans pool the orientations of multiple Gabor patches to estimate the mean orientation of a texture composed of heterogeneous orientations. This pooling occurs over large areas of space, and the number of samples used to estimate the average is approximately the square root of the size of the set. This suggests that, unlike tracking multiple objects (Franconeri, in press), perceiving ensembles may not be limited by a fixed number of features. In fact, ensemble orientation perception becomes more precise when more items are in a set; Robitaille and Harris (2011) showed higher precision and reduced response times when larger sets were available to make mean orientation and size judgments. The efficiency of this pooling, however, can be compromised when attention is overloaded; Dakin and colleagues (2009) showed that an attentionally demanding task reduced the effective number of local orientations observers used to estimate the mean.

There is both psychophysical and physiological evidence suggesting that representing average orientation is a parallel process. Some of the strongest evidence for this comes from Parkes and colleagues (2001), who showed that the orientation of a Gabor patch crowded out of awareness (i.e., observers were unable to discriminate its orientation) nonetheless influenced the perceived average orientation of an entire set of surrounding patches. Even though observers could not consciously individuate or scrutinize the target Gabor patch, orientation detectors could process the set in parallel and subsequently pool the information into a summary percept. A similar conclusion was reached by Alvarez and Oliva (2009). These results with crowding suggest that an orientation averaging system is not directly dependent on mechanisms of selective attention. This is consistent with the notion that average orientation representation reflects an automatic, low-level physiological mechanism (Bosking, Crowley, & Fitzpatrick, 2002; Victor et al., 1994; Vogels, 1990). Several accounts have advocated a back-pocket model of visual texture (see chapter 45 by Landy), in which a second-stage mechanism pools the outputs of local filters, as a plausible mechanism for ensemble orientation perception.

Although it is clear that crowding is not necessary for the extraction of ensemble information, one intriguing possibility is that it enhances the precision of the summary representation. Similar to distributed attention that improved average size representation (Chong & Treisman, 2005a), crowding (Levi, 2008; Pelli, Palomares, & Majaj, 2004; Whitney & Levi, 2011) by definition disrupts any serial attentive process (Intriligator & Cavanagh, 2001), which may force observers into an attentional strategy more conducive to summary representation. Thus, crowding might facilitate the condensation of (even consciously inaccessible) information into efficient "chunks."

Perhaps even more intriguing is the notion that people may actually *perceive* the mean orientation. For example, Morgan, Chubb, and Solomon (2008) as well as Ross and Burr (2008) showed that orientation pooling allows people to gauge the variance in a texture of orientations, and as long as the overall variability across the set is below a certain threshold, each local element takes on the appearance of the mean orientation of the group (Parkes et al., 2001). In other words, we actually perceive the mean even when it is physically absent from the set.

Perceiving Average Size

The current era in the study of ensemble statistics began when Ariely (2001) provided evidence that observers could implicitly derive the average of a set of differently sized dots. In fact, this summary representation was the favored representation. Observers viewed sets of dots for 2 s and then indicated whether a subsequently viewed test dot was a member of the set. The striking aspect of these data was not just that observers performed poorly at the member identification task. As the size of the test dot approached the average size of the array of dots, observers were much more likely to respond that the test dot was a member of the set. Even though observers were instructed to attend to the individual members, they instead represented the summary of the set constituents. When explicitly asked, observers were nearly as precise in discriminating the mean size

Werner—The New Visual Neurosciences
8857_049.indd 698 4/18/2013 5:36:21 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

of several dots as they were in discriminating the size of a single dot. As with orientation (Dakin, 2001), mean discrimination performance seemed invariant to the number of dots in the set (up to 16), possibly suggesting that serial attention mechanisms were not required.

The impact of this seminal work is probably responsible for the fact that several, if not the majority of, investigations of ensemble perception have regarded the perception of size. Accordingly, this subsection of our review is more extensive than the others. This is not to say that summarizing information about size is more important than summarizing other features (e.g., orientation or faces). It merely reflects the abundance of relevant work.

Seeing average size is not just a physical calculation. Im and Chong (2009) harnessed the Ebbinghaus illusion to create sets of circles that differed in their perceived size and physical size. The ensemble percept followed the perceived size, which is probably encoded in V1 at the earliest (Arnold, Birt, & Wallis, 2008; Murray, Boyaci, & Kersten, 2006). Choo and Franconeri (2010) provided further evidence that ensemble size is computed in early stages of visual processing. When they used masking to truncate the representation of the size of a subset of circles to lower-level stages of processing, these circles continued to influence the perception of the mean size.

As with average orientation perception, perception of average size follows statistical rules. de Gardelle and Summerfield (2011) found that observers discounted extreme values (i.e., outliers) when computing the mean shape of a set of circles and squares, suggesting that variance may be encoded in addition to the average (similar outlier exclusion has been found with faces [Haberman & Whitney, 2010]). Solomon, Morgan, and Chubb (2011) provide complementary findings in which observers were able to perceive the variance of a set of circles (or oriented Gabors [Solomon, 2010]) more efficiently than they could perceive the mean size (or orientation). The notion that we are all statisticians appears to be more than an anecdote.

Most investigations of summary statistical perception involve estimating the mean of a static array of features, but the world is dynamic, and recent research shows that the visual system accounts for this by pooling information across time. Albrecht and Scholl (2010) showed that a pooling mechanism takes multiple samples across time to precisely represent the average size of a continuously changing circle.

Average size perception is also robust and appears to occur automatically and without intention. Chong and Treisman (2003) showed that perceiving which of two sets of 12 circles had a larger mean size was immune to changes in presentation (simultaneously versus successively) and duration (even with 50-ms presentations, although see Whiting & Oriet, 2011, for evidence that 200 ms is a more appropriate lower limit). Moreover, observers' discrimination of the average size of the set was nearly as precise as their discrimination of the size of a single circle. The precision of mean representation (at least for size) is best when attention is spread over a large spatial extent (Chong & Treisman, 2005a). Nevertheless, Demeyere and colleagues found that a patient with simultanagnosia (Balint syndrome) could perceive ensemble size (and color) in an array of stimuli despite severely limited spatial attention (Demeyere et al., 2008).

Average size is even computed across multiple sets, in parallel, preceding or perhaps bypassing limitations imposed by the attentional bottleneck. Chong and Treisman (2005b) found that when observers discriminated the average size of a subset of an array of circles that was segregated from the rest of the array by color, average size perception did not depend on whether the color cue preceded or followed the array of circles and was no worse even when only a single color was presented. Ensemble size perception is even possible when attention is divided across stimulus modalities. Albrecht, Scholl, and Chun (2011) asked observers to listen to a sequence of tones while simultaneously viewing a sequence of differently sized disks. Depending on a cue, they made a subsequent judgment about the mean of one set or the other. Ensemble tone or size judgments were unaffected by whether or not the cue preceded or followed the sequences. In other words, dividing attention across the two modalities had no cost on the efficiency of perceiving the means. There is, however, a cost in ensemble precision when attention is divided between two feature dimensions (e.g., size and speed) (Emmanouil & Treisman, 2008).

Attentional manipulations may not only affect the precision or efficiency of the ensemble code but, under certain conditions, can also bias the representation in predictable ways. For example, priming observers to a dot of a particular size (either the largest or the smallest dot) biased estimates of mean size in the direction of the prime (de Fockert & Marchant, 2008). One interpretation is that observers allocated more resources to the primed dot, resulting in a biased estimate of mean size or a representation that reflected only a spatially proximal subset of the dots. A complementary result found by Brady and Alvarez (2011) is that the mean size estimate reflects ready access to one of several hierarchical means within the set. Observers were biased by the mean size of a set even though they were asked to attend to an individual member. Interestingly, this bias

Werner—The New Visual Neurosciences
8857_049.indd 699 4/18/2013 5:36:21 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

emerged only when attention was directed to a particular feature dimension (i.e., whether the member came from a red or a blue set, which were simultaneously presented). These studies suggest that observers represent multiple means simultaneously, but that those representations are predictably biased by attention.

Although the role of attention in average size representation is an ongoing debate (Ariely, 2008; Chong et al., 2008; Myczek & Simons, 2008; Simons & Myczek, 2008), these studies provide support for the existence of an automatic mechanism responsible for average size computation.

Perceiving Ensembles of Faces

For many years, the focus of research on summary statistical perception has been on low-level stimuli (motion, orientation, position, size, etc.). However, given our effortless interaction with highly complex scenes and our subjective impression of a rich and complete visual world, it is reasonable to think that the ensemble coding heuristic might operate on a processing level beyond that of orientation, size, or texture. Haberman and Whitney (2007, 2009) and Haberman, Harp, and Whitney (2009) explored the possibility that observers could extract an average representation from high-level stimuli, including faces. The authors created a series of morphs, varying the expression of faces ranging from extremely happy to extremely sad. Observers viewed sets of these emotionally varying faces and were asked whether a subsequent test face was happier or sadder than the mean expression of the previous set. Remarkably, observers could discriminate the average expression of the whole set as well as they could discriminate the expression of a single face. This phenomenon proved to be robust and flexible, operating implicitly and explicitly (Haberman & Whitney, 2009), across a variety of expressions as well as gender morphs (Haberman & Whitney, 2007), at short exposure durations (as low as 50 ms, although with reduced precision (Haberman & Whitney, 2009), and on sets containing as many as 20 faces (Haberman, Harp, & Whitney, 2009; see figure 49.2). Control experiments demonstrated that the mean discrimination of expression declined when subjects viewed sets of inverted or scrambled faces, suggesting that the visual system extracts summary statistical information about the configural or holistic properties of faces, not just about low-level visual cues such as spatial frequency (Oliva & Torralba, 2001; Torralba & Oliva, 2003) or orientation.

Perceived facial expression also rapidly integrates over time (Haberman, Harp, & Whitney, 2009). Observers viewed sequences of different faces presented at various temporal frequencies and made judgments about the mean expression of those sequences. Observers were able to accurately derive a mean expression in a sequence of 20 faces presented at 20 Hz. The ensemble required 800 ms of temporal integration. Although this integration time is higher than that for low-level motion (Burr, 1981; Nakayama, 1985; Snowden & Braddick, 1989), it compares favorably with the time it takes the visual system to perceive biological motion (Neri, Morrone, & Burr, 1998) and suggests the work of a parallel mechanism.

Although the results from many of these investigations of summary perception and countless accounts of crowding suggest that access to information about individuals (e.g., the size of a particular circle or the expression of a particular face) is lost when the summary statistic is calculated (e.g., Rosenholtz, 2011), there is also good reason to suspect that holistic object-level representations remain intact but are somehow blocked from conscious access in favor of the ensemble. Fischer and Whitney (2011) demonstrated a result that built on that from Parkes and colleagues (2001) in which the expression of a face crowded out of awareness influenced the average expression perceived in a crowd. Crucially, this pooling did not occur when the crowded face was inverted; the ensemble only incorporated holistic information from the face, indicating that a high-level object representation was intact and influencing the average but was not accessible to awareness.

High-level ensemble coding is further supported by other work showing that observers can rapidly perceive the mean identity of sets of faces (de Fockert & Wolfenstein, 2009; Yamanashi-Leib et al., 2012) as well as research showing rapid within-hemifield emotional averaging predicted by properties of neural averaging (Sweeny et al., 2009).

Perceiving Ensembles of Biological Motion

If ensemble perception occurs for high-level forms such as faces, then it is reasonable to believe that it might also occur for perception of biological motion, a high-level visual feature defined by the integration of form and motion. Sweeny, Haroz, and Whitney (2012a) used a design reminiscent of that used by Dakin (2001) to examine this possibility. Observers estimated the headings of briefly presented crowds of point-light walkers that differed in the number and headings of their members (i.e., people in differently sized crowds had identical or increasingly variable directions of walking). They found that observers rapidly pooled information from multiple walkers to precisely estimate the overall direction of the crowd. The striking aspect of these

Werner—The New Visual Neurosciences
8857_049.indd 700 4/18/2013 5:36:21 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

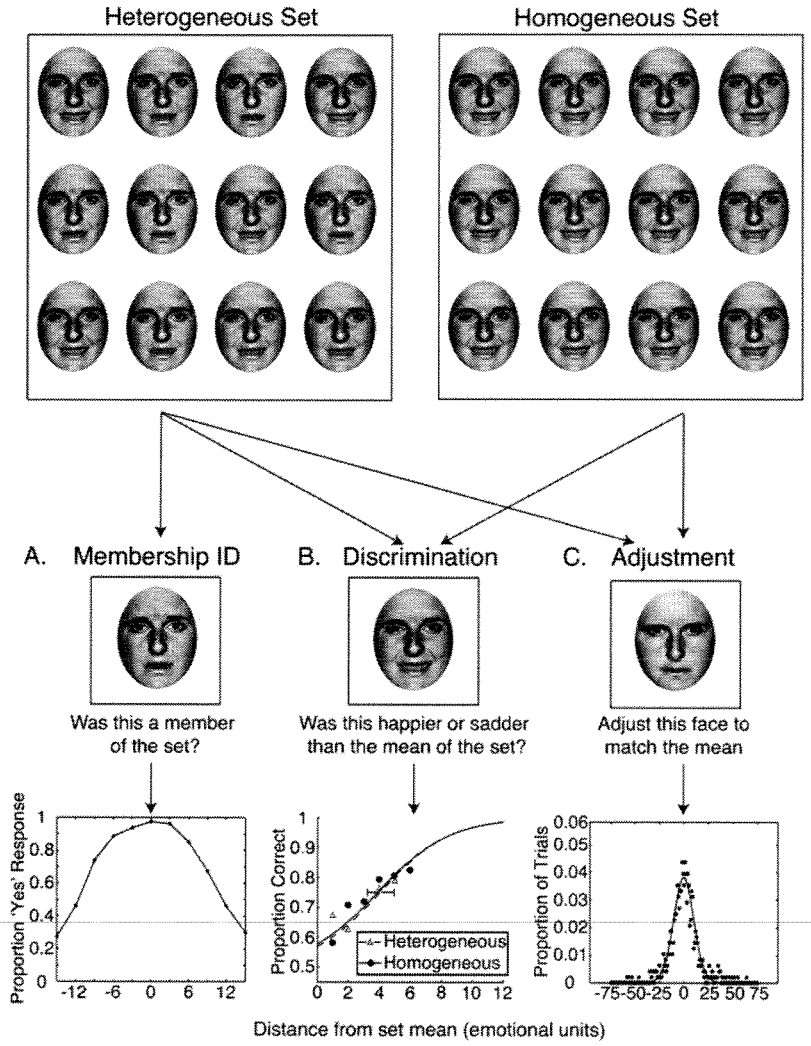

FIGURE 49.2 Several ensemble perception paradigms. Observers view sets of stimuli (e.g., faces). (A) In one experiment observers had to identify whether a test face was a member of the previously displayed set (membership ID). Observers were most likely to indicate a test face was a set member when it approached the mean expression of the set (0 indicates the mean expression). Thus, observers were unable to represent the individual set constituents but instead favored the ensemble. (B) Observers were explicitly asked about the average expression in a set (discrimination task). Surprisingly, they could discriminate the mean expression as well as they could discriminate any single face. (C) Observers used the mouse to adjust the test face to match the mean expression of the set. This provided the full error distribution of the mean representation (0 indicates the mean expression). Responses tended to cluster around the mean expression of the set. (Adapted from Haberman & Whitney, 2011b.)

results is that pooling across noisy individual walker directions allowed observers to perceive the direction of a crowd better than the direction of an individual. This precise ensemble percept required upright orientations and human configurations, suggesting that the code was formed in high-level visual areas where form and motion are integrated. Moreover, it occurred within a surprisingly brief amount of time (200 ms), showing that ensemble percepts of even the most complex visual stimuli can be formed in parallel.

K2

Werner—The New Visual Neurosciences
8857_049.indd 701
4/18/2013 5:36:22 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

Multiple Levels and Multiple Pathways of Ensemble Coding

This brief survey is necessarily incomplete, but it provides a glimpse at the history of ensemble perception. The robust summary of statistical representations found across domains suggests that ensembles are calculated at multiple levels in both the dorsal and ventral streams. Some ensembles, such as average brightness, color, and orientation, may be generated at the earliest cortical (and possibly even subcortical) stages. Others, such as motion and position, may be generated along the dorsal stream. High-level shape and face ensembles are likely generated along the ventral, object-processing stream. Finally, biological motion ensembles are likely generated after the convergence of the dorsal and ventral pathways. Because ensemble percepts can emerge at independent levels of analysis, for example on holistic representations of faces independent of the ensemble brightness, orientation, or facial features in a scene, no single visual or cortical area is likely to be responsible for ensemble perception. Consequently, although there are several physiologically inspired models that might generate ensemble representations at single levels of visual processing (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Rosenholtz et al., 2012), they are not prepared to capture the repetitive and independent nature of ensemble representations at multiple stages along the visual hierarchy. This holds especially true for high-level objects such as crowds of walking humans or faces.

Despite the distinct object properties processed at each level, the uniting commonality is that any set may be represented by a single ensemble percept. This percept is created and maintained for conscious access, while the individual constituents are lost (via limitations of visual working memory, crowding, etc.). Because the visual system creates a representation of many of the items within a set, (conscious) loss of the individual is inconsequential. Many unanswered questions remain, such as how many concurrent ensemble percepts can be maintained, whether there is interference between different levels of ensemble analysis (e.g., average facial expression, brightness, and orientation), and whether the ensembles bypass the limited capacity of attention and visual short-term memory or instead simply act as "chunks" of information, increasing processing efficiency while still drawing on the finite resources of attention and memory.

It is easy to imagine how summary statistics explain texture appearance—the graniteness, stucconess, and so on of surfaces. Although textures have been extensively studied (Beck, 1983; Landy & Graham, 2004; Malik & Rosenholtz, 1997; Nothdurft, 1991), and summary statistical representation of low-level features holds for typical textures, the finding that groups of faces or walking people are perceived as an ensemble—as a texture—suggests that textures can occur at multiple, distinct levels of the visual processing hierarchy.

Is Ensemble Perception Just a Prototype?

The demonstration of summary statistical representation for faces may raise the concern that the results are simply due to a prototype effect (Solso & McCarthy, 1981). Indeed, there has been significant research providing evidence that observers implicitly develop statistical sensitivities to arbitrary patterns over time (Fiser & Aslin, 2001; Posner & Keele, 1968). However, unlike the prototype effect, ensemble coding requires no learning; summary statistical representation is a perceptual process, and observers are sensitive to it after only a single trial. Prototypes suggest that observers falsely recognize an average face due to predominant exposure to specific facial features over an extended period (Solso & McCarthy, 1981). The average face (or size, orientation, etc.) in ensemble coding, though, changes on a trial-by-trial basis and is immediately recognizable. Ensemble perception is therefore a much more flexible pooling of important information into computationally palatable chunks. Observers never actually see the average face of a set, and yet they favor the ensemble percept over the individuals.

ENSEMBLES AS AN EXPLICIT CODE

Given the explosion of work convincingly showing that the visual system is sensitive to summary statistics, and the ease with which they are represented, there has been surprisingly little work exploring the supporting mechanisms. Although there have been a few studies that have addressed this directly, as described below, there are many fundamental questions left unanswered. For example, how are ensembles computed in the brain? Are there multiple levels of representation corresponding to the level at which each exemplar is analyzed (e.g., is average orientation computed in early visual cortex, and average emotion computed in face selective regions)?

Recent work exploring how ensembles are represented demonstrated an aftereffect specific to the average size. Aftereffects occur as a result of neural adaptation and reveal that a given feature is directly represented in the brain (e.g., Suzuki, 2005). In other words the existence of an aftereffect demonstrates dedicated neural coding for a given feature dimension. In

Werner—The New Visual Neurosciences
8857_049.indd 702 4/18/2013 5:36:22 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

their study Corbett and colleagues (2012) had observers adapt to sets of dots varying in size and then judge which of two test circles was larger. The test dots were perceived as smaller after adaptation to a set with a large mean size, and vice versa, suggesting that average size is an explicitly represented feature dimension. However, one concern with the study is that the authors did not distinguish between local adaptation to the individual elements (which must occur) and adaptation to the average. To clearly establish that these aftereffects indeed reflected adaptation to the mean size, the results must demonstrate adaptation independent of what is predicted by local adaptation effects. Nevertheless, this finding makes a clear prediction for future research; it should be possible to identify and characterize an ensemble representation in the brain.

So while evidence is accumulating to support the notion that summary statistics are directly represented in the brain, the algorithm for computing this summary value is not entirely clear. A linear pooling mechanism is probably the most plausible and popular candidate (see figure 49.3). This type of mechanism is fairly straightforward (representations of individual features are integrated, and the average is determined), and modeling illustrates that it can provide a reasonable approximation for perception of average orientation (Parkes et al., 2001) and average biological motion (Sweeny, Haroz, & Whitney, 2012b). These pooling

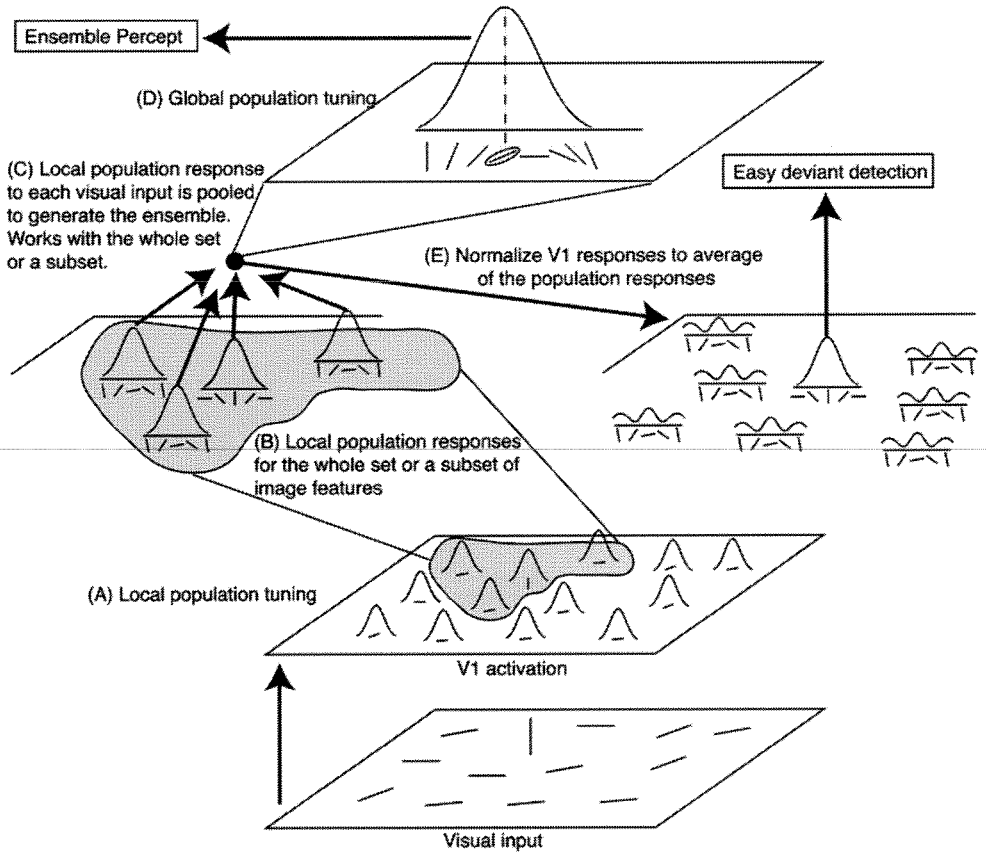

FIGURE 49.3 One possible physiological mechanism driving pop-out. (A, B) Orientation-selective cells (possibly in V1) fire in response to visual input. (C, D) The activity from some or all of the orientation-selective cells is combined to create the ensemble. (E) Via feedback or horizontal connections, the activity from orientation-selective cells is normalized to the population response (i.e., ensemble). Any cell activity remaining will correspond to the deviant. One of the strengths of this model is that it can operate in parallel, negating the computationally inefficient method of comparing each item with every other one. (Adapted from Haberman & Whitney, 2011b.)

Werner—The New Visual Neurosciences
8857_049.indd 703 4/18/2013 5:36:22 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

models include Gaussian-shaped early-stage channel noise in the encoding of each individual feature and a subsequent stage in which the outputs of these channels are combined, averaged, and perturbed by late-stage Gaussian-shaped noise. For example, the outputs of several orientation-tuned channels could be integrated into a population code. The centroid of this population of noisy individual feature representations would determine the average orientation. Population coding such as this is common (Suzuki, 2005), and it is reasonable to assume that it could apply to perception of ensembles.

Pooling, especially across large subsets of features, affords the prospect of averaging out noisy estimates of local features (Dakin, 2001; Ross & Burr, 2008; Morgan, Chubb, & Solomon, 2008). Moreover, linear pooling predicts that when encoding is particularly noisy, the average should be as precise as (or even more precise than) perception of an individual. Such precision has been borne out in several investigations (Alvarez & Oliva, 2008; Ariely, 2001; Bulakowski, Bressler, & Whitney, 2007; Haberman & Whitney, 2009; Sweeny, Haroz, & Whitney, 2012a). For example, a recent study by Yamanashi-Leib et al. (2012) found that prosopagnosics, who have difficulty recognizing individual faces (i.e., have noisy face recognition), nevertheless recognize crowds surprisingly well.

Choo and Franconeri (2010) provide a compelling explanation of how pooling might be implemented in the brain. They note that the mandatory pooling that occurs when objects are nearby in space (i.e., in what are referred to as integration fields by Pelli, Palomares, & Majaj, 2004, or areas of minimal attentional resolution by Intriligator & Cavanagh, 2001) is consistent with the length of horizontal connections in V1 and the size of receptive fields in V4. They suggest that averaging is a result of integration through these connections in lower-level areas and pooling within receptive fields in high-level areas. This hypothesis is consistent with the fact that ensemble perception is better when attention is diffusely spread (Chong & Treisman, 2005a). Furthermore, Sweeny and colleagues (2009) provide direct empirical support for this speculation. In their investigation observers viewed a briefly and simultaneously presented pair of faces with different emotional expressions and were postcued to rate the emotional expression of just one face in the pair. Critically, the locations of the faces were varied such that both faces either fell within large receptive fields of high-level neurons (both within a hemifield) or in separate receptive fields (each in a separate hemifield). Perceptual averaging occurred (the expression of a given face appeared more like the average of the pair), but only when the faces fell within what would be expected to be the same receptive fields of high-level face-tuned neurons.

Although these studies hint that ensembles may be a fundamental representation of the visual system, much work remains to be done to determine whether multiple levels of representation exist and to characterize the algorithm more precisely. In our view these represent some of the most challenging and exciting avenues for future research in the field.

IMPLICATIONS OF ENSEMBLE CODING

Overall, it is clear that although ensemble percepts may not be completely independent of attention, they do occur implicitly. What are the broader implications of implicit statistical summarization of the environment? How does this knowledge inform traditional notions of awareness? We explore these questions in this section.

Bypassing the Bottleneck

The discovery that ensembles could be represented implicitly led several researchers to speculate that summary perception might drive the sense of visual completeness in spite of limited awareness (Cavanagh, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2009). Even though the visual system can explicitly represent just a few items simultaneously (e.g., Luck & Vogel, 1997; Franconeri, in press), the world beyond the focus of attention does not fade to black. In fact, the objects and scenes we are not attending to seem remarkably rich. It is reasonable to speculate that this "grand illusion" may be due, in part, to summary statistical perception. The statistics of natural scenes (e.g., Simoncelli & Olshausen, 2001) are, indeed, quite stable (Oliva, 2005; Torralba & Oliva, 2003), and the visual system efficiently exploits this natural redundancy by generating summary percepts. Several recent findings support this hypothesis by showing (1) how ensemble representations provide strikingly precise percepts in spite of noisy encoding of individual details and (2) that moment-to-moment awareness of visual scenes more closely follows ensemble representations than abrupt changes in individual features.

Sweeny, Haroz, and Whitney (2012a) demonstrated that pooling across multiple noisy features produces an ensemble percept that surpasses the precision with which we can perceive an individual. Observers viewed a crowd composed of individual people with different directions of walking. Because the crowd spanned a large spatial extent and was only visible for 200 ms, the encoding of each walker was noisy. Nevertheless, observers perceived the crowd's average direction of walking

Werner—The New Visual Neurosciences
8857_049.indd 704 4/18/2013 5:36:23 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

more precisely than they perceived a single foveally presented person's direction. This shows that even though perception of a given feature in the periphery may be poor, perception of the group (or the whole scene) truly is precise.

Similar high-resolution pooling occurs outside the focus of attention. Alvarez and Oliva (2008) found that observers were just as good at reporting the average position of a set of dots they had been tracking as with dots they had not been tracking (i.e., beyond the focus of attention). Modeling showed that, although the position representation of the individual dots beyond the focus of attention was noisy, as expected, the average of these noisy representations nonetheless accurately predicted performance on the average position task. This suggests that ensemble information was preserved in spite of limited awareness, and it supports the assertion that ensembles provide an efficient means to maintain perceptual stability (i.e., deriving a precise ensemble code even though only noisy information was available).

Moment-to-moment awareness of a scene closely follows ensemble representations even when abrupt changes in individual features go unnoticed. Alvarez and Oliva (2009) showed that while engaged in an attentionally demanding tracking task, observers were explicitly aware of changes in the background that altered the average orientation of the top and bottom halves of the screen. However, observers were oblivious to changes of the same magnitude that preserved the overall ensemble (i.e., the average orientation). Thus, information regarding global scene statistics remains available even in the face of exhausted attentional resources.

Complementary evidence that ensembles provide low-cost perceptual stability comes from a dual-task paradigm employed by Haberman and Whitney (2011a). Observers viewed two sets of 16 successively presented faces on each trial. Within a given trial, 4 of the 16 faces changed from one emotional extreme to another (e.g., four happy faces turned sad). This shift created a change in the overall mean emotion of the set. Observers were instructed to identify (1) which of the two sets was on average happier (ensemble task) and (2) the location of just one of the four changes (change localization task). In trials in which change localization failed (i.e., when the observer could not report where on the screen the change occurred), observers were nonetheless significantly above chance in identifying which set was on average happier. Although change localization reflects the limitations of explicit awareness, ensemble codes seem to bypass these limitations. Taken together, these studies point to a robust and efficient heuristic at work, one that can maintain the stability of our visual world in the face of limited information.

Visual Search

The possible connection between ensemble coding and visual search is appealing. Despite the rich literature on the properties of visual search (Treisman, 1982; Verghese, 2001; Wolfe, Cave, & Franzel, 1989), a physiologically plausible mechanism (e.g., an algorithm or neural implementation; Marr, 1982) that generates popout is still debated (Eckstein, 1998; Itti & Koch, 2000; Wolfe, 2003). Summary statistical representations—ensemble coding—may serve as a computationally efficient means of calculating deviance. Several models have made similar suggestions (e.g., Callaghan, 1984; Duncan & Humphreys, 1989). Often these models suggest that similarity influences popout (Duncan & Humphreys, 1989). However, what counts as "similar" or "dissimilar" is unclear. Summary statistical representations, per se, could provide the underlying metric of similarity, one that affords deviance detection (Rosenholtz et al., 2012). Recent accounts of visual search also acknowledge the possibility that much of the periphery may be represented as an ensemble and processed preattentively and that this nonselective ensemble pathway generates a gist impression that guides a selective pathway, leading to more efficient real-world search (Wolfe et al., 2011).

How might a very simple, physiologically plausible, population-coding algorithm extract ensemble information and generate popout? Figure 49.3A shows an example of an array of oriented lines that might stimulate many local populations of orientation-selective cells (e.g., in V1). If a subset of *locally* tuned receptive fields is sampled (figure 49.3B), and its output is pooled (figure 49.3C), a *global* population tuning curve is represented [only a subset of the items needs to be sampled]; cf. Dakin & Watt, 1997; Morgan, Chubb, & Solomon, 2008; Myczek & Simons, 2008). This global population curve is the average of local tuning curves and ultimately produces an ensemble percept (figure 49.3D). Note that the impact of any deviant orientation is mitigated in the global population curve, as most of the inputs are of similar orientations. The global population response then normalizes the local tuning (via feedback or horizontal connections; figure 49.3E). Most of the local population responses are reduced to near 0, and what is left is activity corresponding to the deviant orientation. Although low-level normalization or contextually dependent procedures have been implemented in other models (e.g., Itti, Koch, & Niebur,

Werner—The New Visual Neurosciences
8857_049.indd 705 4/18/2013 5:36:23 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

1998; Li, 1999), this model implicates ensemble coding and the generation of ensemble percepts as the basis for popout. A particular strength of this model is that the normalization operation may be carried out in parallel, without repetitive comparisons across local population responses.

CONCLUSION

Despite many bottlenecks in visual processing and the limited nature of awareness, humans rapidly extract an enormous amount of information from scenes (Oliva & Torralba, 2001; Potter, 1976; Thorpe, Fize, & Marlot, 1996; Torralba & Oliva, 2003). It is becoming clear that much of this information may take the form of condensed summary statistics—computationally efficient ensemble representations of similar features and objects in scenes. Ensembles are encoded from the lowest levels of feature processing to the highest levels of object and face perception. Ensemble perception occurs quickly, automatically, and outside the focus of attention, although it is also modulated by attention. More broadly, ensemble perception may underlie much of our impression of perceiving a complete and rich visual world.

REFERENCES

Albrecht, A. R., & Scholl, B. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science, 21,* 560–567.

Albrecht, A. R., Scholl, B., & Chun, M. M. (2011). Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli. *Journal of Vision, 11*(11), 1210. doi:10.1167/11.11.1210.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Neurosciences, 15,* 122–131. doi:10.1016/j.tics.2011.01.003.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19,* 392–398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 7345–7350.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression-formation. *Journal of Experimental Psychology, 70*(4), 394–400.

Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review, 78,* 171–206.

Anderson, N. H., Lindner, R., & Lopes, L. L. (1973). Integration theory applied to judgments of group attractiveness. *Journal of Personality and Social Psychology, 26,* 400–408.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12,* 157–162.

Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics, 70,* 1325–1326.

Arnold, D. H., Birt, A., & Wallis, T. S. A. (2008). Perceived size and spatial coding. *Journal of Neuroscience, 28,* 5954–5958.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12), 1–18. doi:10.1167/9.12.13.

Battelli, L., Pascual-Leone, A., & Cavanagh, P. (2007). The "when" pathway of the right parietal lobe. *Trends in Cognitive Sciences, 11,* 204–210.

Beck, J. (1983). Textural segmentation, 2nd-order statistics, and textural elements. *Biological Cybernetics, 48,* 125–130.

Bosking, W. H., Crowley, J. C., & Fitzpatrick, D. (2002). Spatial coding of position and orientation in primary visual cortex. *Nature Neuroscience, 5,* 874–882.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble bias memory for individual items. *Psychological Science, 22,* 384–392.

Britten, K. H., & Heuer, H. W. (1999). Spatial summation in the receptive fields of MT neurons. *Journal of Neuroscience, 19,* 5074–5084.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual-motion—a comparison of neuronal and psychophysical performance. *Journal of Neuroscience, 12,* 4745–4765.

Bulakowski, P. F., Bressler, D. W., & Whitney, D. (2007). Shared attentional resources for global and local motion processing. *Journal of Vision, 7*(10), 810–817. doi:10.1167/7.10.10.

Burr, D. C. (1981). Temporal summation of moving images by the human visual-system. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 211,* 321–339.

Callaghan, T. C. (1984). Dimensional interaction of hue and brightness in preattentive field segregation. *Perception & Psychophysics, 36*(1), 25–34.

Cavanagh, P. (2001). Seeing the forest but not the trees. *Nature Neuroscience, 4,* 673–674.

Chong, S. C., Joo, S. J., Emmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics, 70,* 1327–1334.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393–404.

Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics, 67,* 1–13.

Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45,* 891–900.

Choo, H., & Franconeri, S. L. (2010). Objects with reduced visibility still contribute to size averaging. *Attention, Perception & Psychophysics, 72,* 86–99.

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20,* 211–231.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 18,* 1016–1026.

Dakin, S. C., Bex, P. J., Cass, J. R., & Watt, R. J. (2009). Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision, 9*(11), 1–16. doi:10.1167/9.11.28.

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research, 37,* 3181–3192.

de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics, 70,* 789–794.

Werner—The New Visual Neurosciences

8857_049.indd 706 4/18/2013 5:36:23 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology, 62,* 1716–1722.

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 13341–13346.

Demeyere, N., Rzeskiewicz, A., Humphreys, K. A., & Humphreys, G. W. (2008). Automatic statistical processing of visual properties in simultanagnosia. *Neuropsychologia, 46,* 2861–2864.

Duncan, J., & Humphreys, G. W. (1989). Visual-search and stimulus similarity. *Psychological Review, 96,* 433–458.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science, 9,* 111–118.

Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Perception & Psychophysics, 70,* 946–954.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology, 106,* 1389–1398.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504.

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics, 70,* 456–464.

Franconeri, S. L. (in press). The nature and status of visual resources. In D. Resiberg (Ed.), *Oxford handbook of cognitive psychology.* Oxford: Oxford University Press.

Franconeri, S. L., Alvarez, G. A., & Enns, J. T. (2007). How many locations can be selected at once? *Journal of Experimental Psychology. Human Perception and Performance, 33,* 1003–1012.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195–1201.

Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 13130–13135.

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9*(11), 1–13. doi:10.1167/9.11.1.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17,* R751–R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology. Human Perception and Performance, 35*(3), 718–734.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception & Psychophysics, 72,* 1825–1838.

Haberman, J., & Whitney, D. (2011a). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review, 18,* 955–959.

Haberman, J., & Whitney, D. (2011b). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), *A festschrift in honor of Anne Treisman.* Oxford: Oxford University Press.

Hess, R. F., & Holliday, I. E. (1992). The coding of spatial position by the human visual-system—effects of spatial scale and contrast. *Vision Research, 32,* 1085–1097.

Im, H. Y., & Chong, S. C. (2009). Computation of mean size is based on perceived size. *Attention, Perception & Psychophysics, 71,* 375–384.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology, 43,* 171–216.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1254–1259.

Koenderink, J. J., van Doorn, A. J., & Pont, S. C. (2004). Light direction from shad(ow)ed random Gaussian surfaces. *Perception, 33,* 1405–1420.

Koffka, K. (1935). *The principles of Gestalt psychology.* London: Routledge and Kegan Paul.

Landy, M., & Graham, N. (2004). Visual perception of texture. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (Vol. 2, pp. 1106–1118). Cambridge, MA: MIT Press.

Leon, M., Oden, G. C., & Anderson, N. H. (1973). Functional measurement of social values. *Journal of Personality and Social Psychology, 27,* 301–310.

Levi, D. M. (2008). Crowding—an essential bottleneck for object recognition: A mini-review. *Vision Research, 48,* 635–654.

Levin, I. P. (1974). Averaging processes in ratings and choices based on numerical information. *Memory & Cognition, 2,* 786–790.

Li, Z. (1999). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proceedings of the National Academy of Sciences of the United States of America, 96,* 10530–10535.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281.

Malik, J., & Rosenholtz, R. (1997). Computing local surface orientation and shape from texture for curved surfaces. *International Journal of Computer Vision, 23,* 149–168.

Marois, R., Yi, D. J., & Chun, M. M. (2004). The neural fate of consciously perceived and missed events in the attentional blink. *Neuron, 41,* 465–472.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: W. H. Freeman.

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A "dipper" function for texture discrimination based on orientation variance. *Journal of Vision, 8*(11), 9. doi:10.1167/8.11.9.

Morgan, M. J., & Glennerster, A. (1991). Efficiency of locating centres of dot-clusters by human observers. *Vision Research, 31,* 2075–2083.

Murray, S. O., Boyaci, H., & Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. *Nature Neuroscience, 9,* 429.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70,* 772–788.

Nakayama, K. (1985). Biological image motion processing—a review. *Vision Research, 25,* 625–660.

Nakayama, K. (1990). The iconic bottleneck and the tenuous link between early visual processing

Werner—The New Visual Neurosciences

8857_049.indd 707 4/18/2013 5:36:23 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

and perception. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 411–422). Cambridge: Cambridge University Press.

Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. *Nature, 395*, 894–896.

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience, 8*, 2201–2211.

Noe, A., Pessoa, L., & Thompson, E. (2000). Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition, 7*, 93–106.

Nothdurft, H. C. (1991). Texture segmentation and pop-out from orientation contrast. *Vision Research, 31*, 1073–1078.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*(12), 1136–1169. doi:10.1167/4.12.12.

Posner, M. I., & Keele, S. W. (1968). On genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology. Human Learning and Memory, 2*, 509–522.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179–197.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology. Human Perception and Performance, 18*, 849–860.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 368–373.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12), 1–8. doi:10.1167/11.12.18.

Rosenholtz, R. (2011). What your visual system sees where you are not looking. In B. E. Rogowitz & T. N. Pappas (Eds.), *Proceedings of the SPIE 7865, Human Vision and Electronic Imaging, XVI*, 786510, San Francisco, CA.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Llie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision, 12*(4), 1–17. doi:10.1167/12.4.14.

Ross, J., & Burr, D. (2008). The knowing visual self. *Trends in Cognitive Sciences, 12*, 363–364. doi:10.1016/j.tics.2008.06.007.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology, 38*, 259–290.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*, 261–267. doi:10.1016/S1364-6613(97)01080-2.

Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics, 70*(7), 1335–1336.

Snowden, R. J., & Braddick, O. J. (1989). The combination of motion signals over time. *Vision Research, 29*, 1621–1630.

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision, 10*(14), 1–16. doi:10.1167/10.14.19.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for statistics of size discrimination. *Journal of Vision, 12*(12), 1–11. doi:10.1167/11.12.13.

Solso, R. L., & McCarthy, J. E. (1981). Prototype formation of faces—a case of pseudo-memory. *British Journal of Psychology, 72*, 499–503.

Suzuki, S. (2005). High-level pattern coding revealed by brief shape aftereffects. In C. Clifford & G. Rhodes (Eds.), *Advances in visual cognition: Vol 2. Fitting the mind to the world: Adaptation and aftereffects in high-level vision* (pp. 135–172). New York: Oxford University Press.

Sweeny, T. D., Grabowecky, M., Paller, K., & Suzuki, S. (2009). Within-hemifield perceptual averaging of facial expressions predicted by neural averaging. *Journal of Vision, 9*(3), 1–11. doi:10.1167/9.3.2.

Sweeny, T. D., Haroz, S., & Whitney, D. (2012a). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, Advance online publication. doi:10.1037/a0028712.

Sweeny, T. D., Haroz, S., & Whitney, D. (2012b). Reference repulsion in the categorical perception of biological motion. *Vision Research, 64*, 26–34. doi:10.1016/j.visres.2012.05.008.

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

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network (Bristol, England), 14*(3), 391–412.

Treisman, A. (1982). Perceptual grouping and attention in visual-search for features and for objects. *Journal of Experimental Psychology. Human Perception and Performance, 8*(2), 194–214.

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron, 31*, 523–535.

Victor, J. D., Purpura, K., Katz, E., & Mao, B. Q. (1994). Population encoding of spatial-frequency, orientation, and color in macaque V1. *Journal of Neurophysiology, 72*(5), 2151–2166.

Vogels, R. (1990). Population coding of stimulus orientation by striate cortical-cells. *Biological Cybernetics, 64*, 25–31.

Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual-system averages speed information. *Vision Research, 32*, 931–941.

Watamaniuk, S. N. J., & McKee, S. P. (1998). Simultaneous encoding of direction at a local and global scale. *Perception & Psychophysics, 60*, 191–200.

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays—the integration of direction information. *Vision Research, 29*, 47–59.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung, 4*, 301–350.

Werner—The New Visual Neurosciences

8857_049.indd 708 4/18/2013 5:36:23 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

Whitaker, D., McGraw, P. V., Pacey, I., & Barrett, B. T. (1996). Centroid analysis predicts visual localization of first- and second-order stimuli. *Vision Research, 36,* 2957–2970.

Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review, 18,* 484–489.

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences, 15,* 160–168. doi:10.1016/j.tics.2011.02.005.

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *Vision Research, 24,* 55–62.

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences, 7,* 70–76. doi:10.1016/S1364-6613(02)00024-4.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search—an alternative to the feature integration model for visual-search. *Journal of Experimental Psychology. Human Perception and Performance, 15,* 419–433.

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences, 15,* 77–84.

Yamanashi-Leib, A., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia, 50,* 1698–1707. doi:10.1016/j.neuropsychologia.2012.03.026.

8857_049.indd 709 4/18/2013 5:36:23 PM

PROPERTY OF MIT PRESS: FOR PROOFREADING AND INDEXING PURPOSES ONLY

8857_049.indd 710 4/18/2013 5:36:23 PM