Review

Representing multiple objects as an ensemble enhances visual cognition

George A. Alvarez

Vision Sciences Laboratory, Department of Psychology, Harvard University, 33 Kirkland Street, William James Hall, Room 760, Cambridge, MA 02138, USA

The visual system can only accurately represent a handful of objects at once. How do we cope with this severe capacity limitation? One possibility is to use selective attention to process only the most relevant incoming information. A complementary strategy is to represent sets of objects as a group or ensemble (e.g. represent the average size of items). Recent studies have established that the visual system computes accurate ensemble representations across a variety of feature domains and current research aims to determine how these representations are computed, why they are computed and where they are coded in the brain. Ensemble representations enhance visual cognition in many ways, making ensemble coding a crucial mechanism for coping with the limitations on visual processing.

Benefits of ensemble representation

Unlike artificial displays used in laboratory experiments, where there is no reliable pattern across individual items, the real world is highly structured and predictable [1,2]. For instance, at the object level, the visual field often consists of collections of similar objects – faces in a crowd, berries on a bush. At a more primitive feature level, natural images are highly regular in terms of their contrast and intensity distributions [3,4], color distributions [5–8], reflectance spectra [9,10] and spatial structure [2,11–14]. Where there is structure, there is redundancy, and where there is redundancy, there is an opportunity to form a compressed and efficient representation of information [15–17]. One way to capitalize on this structure and redundancy is to represent collections of objects or features at a higher level of description, describing distributions or sets of objects as an ensemble rather than as individuals.

An ensemble representation is any representation that is computed from multiple individual measurements, either by collapsing across them or by combining them across space and/or time. For instance, any summary statistic (e.g. the mean) is an ensemble representation because it collapses across individual measurements to provide a single description of the set. People are remarkably accurate at computing averages, including the mean size [18,19], brightness [20], orientation [18,21,22] and location of a collection of objects [23]; the average emotion [24], gender [24] and identity [25] of faces in a crowd; and the average number for a set of symbolically presented numbers [26,27]. These are all measures of central tendency for a collection of objects. Other statistics that describe a set, such as variance [28], skew and kurtosis, are also ensemble representations, although the ability to compute and represent these statistics has been the focus of less attention in recent research (but see [29,30] for reviews on earlier research). Finally, the concept of ensemble representations can be extended beyond first-order summary statistics, to include higher-order summary statistics [31–33].

Ensemble representations have been explored under various names in the literature, including 'global features' [32,34,35], '(w)holistic' or 'configural' features [36–38], 'sets' [18,39] and 'statistical properties' or 'statistical summaries' [19,40]. Each of these terms shares the notion that multiple measurements are combined to give rise to a higher level description. The term 'ensemble representation' is used here as an umbrella term encompassing these different ideas. Although there is, as yet, no unifying model of ensemble representation across these domains, recent research on ensemble representation is unified by a common principle: representing multiple objects as an ensemble enhances visual cognition.

The power of averaging

How can computing ensemble representations help overcome the severe capacity limitations of our visual system? The answer lies in the power of averaging: simply put, the average of multiple noisy measurements can be much more precise than the individual measurements themselves. For instance, one can measure reaction time with millisecond precision even when rounding reaction times to the nearest 100 ms (Box 1). The same principle is at play in the 'wisdom of crowds' effect, in which people guess the weight of an ox and the average response is closer to the correct answer than are the individual guesses on average [41]. These benefits arise because, when measurements are averaged, random error in one individual measurement will tend to cancel out uncorrelated random error in another measurement. Thus, the benefits of averaging depend on the extent to which the noise in individual measurements is correlated (less correlated, more benefit) and the number of individual measurements averaged (more measurements, more benefit). The benefit of averaging can be formalized mathematically, given certain assumptions regarding the noise in the individual measurements (Figure 1).

If the human visual system is capable of averaging, then observers should be able to judge the average size of a set more accurately than they can judge the individuals in the set. This is exactly what was demonstrated by Dan Ariely's

Corresponding author: Alvarez, G.A. (alvarez@wjh.harvard.edu).

 1364-6613/$ – see front matter doi:10.1016/j.tics.2011.01.003 Trends in Cognitive Sciences, March 2011, Vol. 15, No. 3

Box 1. The power of averaging

Imagine you are running an experiment with an expected effect size of 20 ms, which is not uncommon in behavioral research (e.g. negative priming or simple detection tasks). Do you need to worry about the sampling rate of your keyboard? First let us consider what would happen if we simply rounded reaction times to the nearest 100 ms. By averaging multiple samples, individual errors owing to rounding will tend to cancel each other out, and it is possible to obtain millisecond precision in the estimate of the mean despite rounding. Figure Ia shows the results of a simulation with ten virtual subjects and only 30 trials per subject. The true average of the population is 600 ms, and subjects are normally distributed around this mean (i.e. each subject has their own true mean, but the average across subjects will be 600 ms). For each simulated trial, reaction time was simulated as the subject's true mean plus 15% random noise around their true mean. This is fairly typical of reaction time data, but the simulation results do not depend crucially on this value. The simulated reaction times were then rounded to the nearest 100 ms. When the true reaction times (from the simulation) are compared to the rounded reaction times, the mean and variance of the two data sets are nearly indistinguishable.

Now suppose your keyboard checks for a key once every 100 ms. This would be equivalent to rounding each reaction time up to the nearest 100 ms, which on the face of it sounds like it would add error to the estimate of the mean and variance of each condition. Indeed, it would lead to overestimates of the reaction time in each condition. However, the relative difference between conditions could be preserved. The simulation above was repeated with two conditions in which the true mean between conditions was simulated so that condition two was 20 ms slower than condition one on average. Figure Ib shows the results of the simulation, in which condition two was reliably slower than condition one for each individual subject, and the 20 ms difference is significant at $p < 0.05$ using a standard within-subject *t*-test. In general, whether the effect can be detected thus will depend on the degree of rounding, the expected size of the effect and the variability of the data.

For the present purpose, the important point is that, by averaging a relatively modest number of trials, it is possible to overcome a great deal of noise in individual estimates to obtain a precise representation of the mean (Figure Ia) and to detect a subtle difference between two conditions (Figure Ib).

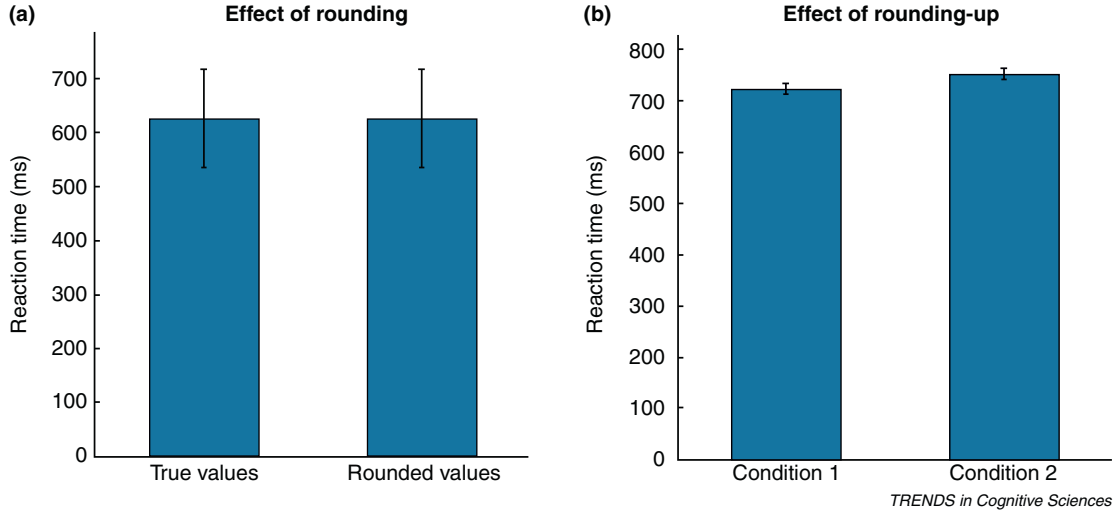

Figure I. **(a)** The effect of rounding on estimating the mean and variance in a single condition. Error bars depict the standard deviation across subjects. **(b)** The effect of rounding-up on the comparison of two conditions in which the true mean differs by 20 ms. Error bars depict the within-subject standard error of the mean.

influential research on the ability of people to perceive the mean size of a set [18], which showed that observers can estimate with high accuracy the average size of a set of objects, even when they appear unable to report the size of the individual objects in the set.

This type of averaging provides a potential mechanism for coping with the severe limitations on attentional processing. Attention appears to be a fluid and flexible resource: we can give full attention to a single item and represent that item with high precision, or we can divide our attention among many items but consequently represent each item with lower precision [42–44]. In general, objects outside the focus of attention are perceived with less clarity [45], lower contrast [46] and a weaker high-frequency response [47,48]. Presumably all objects in the visual field are represented with varying degrees of precision, depending on the amount of attention they receive. In some cases, objects outside the focus of attention are so poorly represented that it seems like we have no useful information about them at all. However, it turns out to be possible to combine that imprecise information to recover an accurate measure of the group [23].

Figure 2 illustrates how attention might affect the fidelity of ensemble representations. Inside the focus of attention (red beams), individual items will be represented with relatively high precision. The average of these items will be represented with even higher precision, as expected from the benefits of averaging. For items outside the focus of attention, we assume that they must be attended to some extent to be perceived at all. For instance, the results of inattentional blindness studies have shown that without attention, there is little or no consciously accessible representation of visual information [49–51]. These studies typically aim for participants to completely withdraw attention from the tested items, and in some cases observers even actively inhibit information outside of the attentional set [51]. However, when observers know they will be asked about information outside the focus of attention, it is probable that they diffusely attend to those items. Figure 2 implies a parallel system with multiple foci of

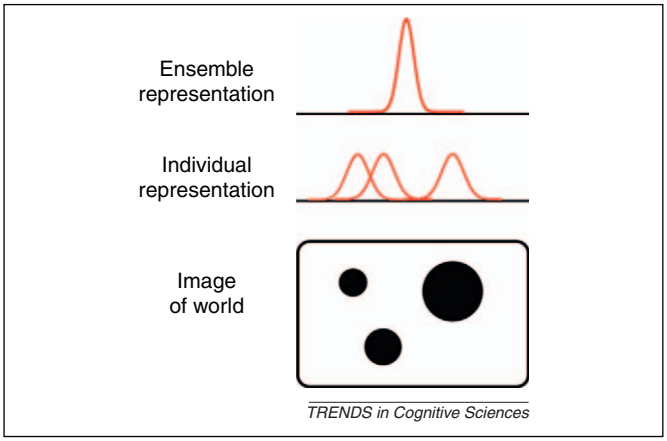

Figure 1. Gaining precision at a higher level of abstraction. By taking individual measurements and averaging them, it is possible to extract a higher-level ensemble representation. If error is independent between the individual representations, then the ensemble average will be more precisely represented than the individuals in the set. This benefit can be quantified after making certain assumptions. For instance, if each individual were represented with the same degree of independent, Gaussian noise (standard deviation = σ), then the average of these individual estimates would have less noise, with a standard deviation equal to $\sigma/\sqrt{n}$, where n is the number of individual measurements. The process is depicted for the representation of object size, but the logic holds for any feature dimension.

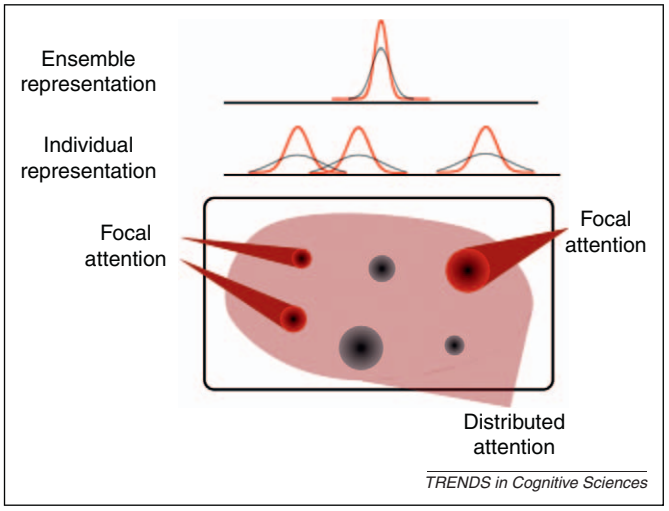

Figure 2. Effect of attention on the fidelity of ensemble representations. Two sets of items are depicted: one set inside the focus of attention (red beams) and one set diffusely attended outside the focus of attention (pink region). For illustrative purposes, both sets are composed of identical individuals, and thus both sets have the same individual and mean representations. For items inside the focus of attention, individual representations will be relatively precise (red curves). The ensemble representation of the items inside the focus of attention will be even more precise, owing to the benefits of averaging. For items outside the focus of attention which are diffusely attended, the individual representations will be very imprecise (gray curves). However, the benefits of averaging are so great that the ensemble representation will be fairly precise, even when a relatively small number of individual representations are averaged (just three in this example).

attention, plus diffuse attention spread over items outside the foci of attention. However, a similar result could be modeled with a single spotlight of attention that spends more time in some locations than others. Either way this diffuse attention results in extremely imprecise representations of the individual items, and yet averaging even just three imprecise measurements results in a fairly precise representation of the ensemble. If a large enough sample of items is averaged together, then the ensemble representation for items outside the focus of attention can be nearly as accurate as the ensemble representation for items inside the focus of attention.

The mechanisms of averaging

Although there is general agreement that human observers can accurately represent ensemble features, many questions remain regarding 'how' these ensemble representations are computed, including: (i) Are individual representations computed and then combined to form an ensemble representation, or are ensemble representations somehow computed without computing individuals? (ii) If individual representations are computed, are they discarded once the ensemble has been computed? (iii) How many individual items are sampled and included in the calculation of the mean? Is it just a few or could it be all of them? (iv) Do all items contribute to the mean equally?

Are ensembles built up from representations of individuals?

Ariely [18] proposed that the visual system performs a type of compression, by creating an ensemble representation and then discarding individual representations. Some have interpreted this proposal to mean that the ensemble representation is computed without first directly computing individual measurements. For instance, it is possible that there is a 'total activation map' and a 'number map' and that mean size is computed by taking the total activation and dividing it by the number of items [52]. However, Ariely's use of the term 'discard' suggests that his intended meaning was that the individual properties are computed, combined and then discarded. This type of averaging model has been supported by research on the computation of mean orientation [21]. Addressing this question empirically is a challenge because it is possible to compute accurate ensemble representations even from very imprecise individual measurements. Consequently, a poor representation of individual items cannot be used as evidence for mean computation without computing individuals – unless the mean can be shown to be represented more accurately than expected based on the number and fidelity of individual items represented.

Are individual representations discarded?

How do we explain such poor performance when observers are required to report the properties of individual members of a set? One possibility is that these properties are computed and then discarded. An important alternative possibility is that the individual representations are not discarded, but are simply so noisy and inaccurate that observers cannot consistently identify individuals from the set owing to this high level of noise. Alvarez and Oliva found support for this possibility by modeling their results [23], consistently finding that the accuracy of ensemble judgments is perfectly predicted from the accuracy of individual judgments – even when individuals appear to be judged with near chance accuracy. This alternative possibility fits with a framework in which the representation of an image is hierarchical, retaining information at multiple levels of abstraction [35,53].

How many items are sampled?

A great deal of enthusiasm surrounding studies on ensemble representations stems from the possibility that there are specialized ensemble processing mechanisms which are separate from the mechanisms employed to represent individual objects. However, this idea has spurred some controversy in the area of research on mean size perception, where modeling study has shown that it is possible to accurately estimate the mean by sampling a small subset of items [54]. In some cases, the average of the set could be accurately estimated by strategically sampling as few as one or two items, and estimating the average of those items alone [54]. Consistent with this subset sampling hypothesis, the accuracy of the mean estimate is typically constant as the number of items in the set increases beyond four items [18,55,56], whereas the benefits of averaging should accrue as more items are averaged together. This would be expected if observers were sampling just a subset of the items.

However, there are several reasons to believe that observers are not strategically subsampling when they compute the mean. In the case of crowded items, observers simply cannot sample individual items, thus it is unlikely that judgments for crowded displays [21] reflect a sampling strategy. When items are not crowded, it has been shown that intermixing conditions that would require different sampling strategies does not impair performance on mean size estimation [57], suggesting that subjects either are not using a strategic sampling strategy or can instantly deploy a new strategy based on some property of the display. This latter possibility is unlikely, given that the displays in [57] were only presented for 200 ms. One study on perceiving the average facial expression has shown that observers discount outliers when computing the average, but a sampling strategy would show a large effect of outliers [58]. Moreover, the accuracy of centroid estimates suggests that 'all' of the items must be averaged to compute the centroid with the level of precision observed, requiring the representation of a minimum of eight individual items [23].

If observers are not strategically subsampling, the fact that the precision of mean size estimation is constant with the number of items beyond four presents a bit of a mystery. One possibility is that the benefits of averaging accrue quickly, and that one would predict a steep improvement in the precision of mean estimation from one to four items, with a leveling off beyond four items [58]. Another possibility is that the precision with which each individual item is represented decreases as the number of items increases, because each item receives less attention [42,44] and/or because items are more crowded and appear further in the periphery on average. If this were the case, then the benefits from averaging additional items would be offset by the decrease in precision with which the individual items are represented, as illustrated in Figure 3. This account predicts that the slope of the function relating the precision of mean judgments to the number of items would depend on the degree to which the noise in individual items increases with the number of items. In practice, this slope is often fairly shallow or even flat [18,55,56]. This raises the intriguing possibility that averaging perfectly offsets the increase in noise that occurs as the number of items increases.

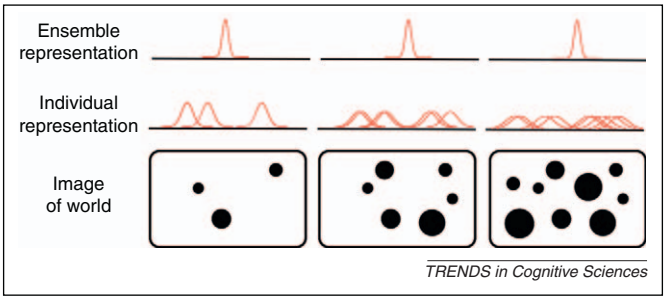

Figure 3. Effect of set size on the fidelity of individual and ensemble representations. The ensemble average should become more precise as the number of individual items increases, because the benefits of averaging accrue with each additional item averaged (with diminishing returns, of course). However, if the precision with which individual items can be represented decreases with set size, as depicted here, it is possible for this decrease to perfectly offset the benefits of averaging so that the precision of the average remains constant with set size.

Do all items contribute to the mean equally?

There is already some evidence that not all items contribute equally to the mean [58]. Intuitively, if some measures are very unreliable, and other measures are very reliable, we should give the more reliable measures more weight when combining these measurements. In general, computing a weighted average in which more reliable estimates are given greater weight will minimize the error in estimates of the mean. To illustrate this point, Figure 4 shows the results of a simulation in which the mean size of eight items was estimated. Half of the individual item sizes were estimated with high precision (low variance), whereas the other half were estimated with low precision (high variance). The individual measurements were then averaged using the standard equal-weight average or using a precision-weighted average in which each individual measurement was weighted proportional to its precision. A total of 1000 trials were simulated, and for each trial error was measured as the difference between the actual mean size and the estimated mean size. The error distributions show that error was lower for the precision-weighted average than for the standard, equal-weighted average.

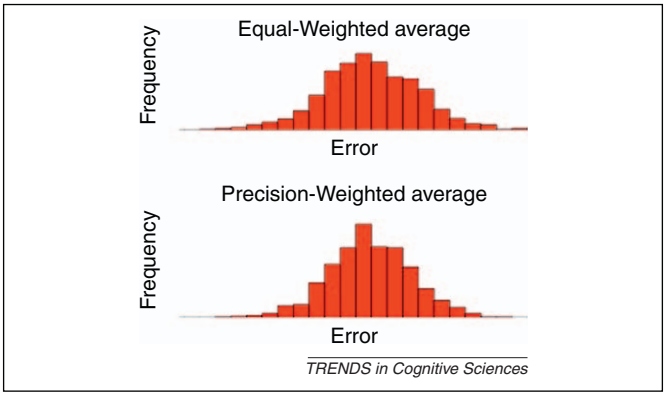

Figure 4. Benefits of precision-weighted averaging. A standard equal-weighted average will be less precise on average than a precision-weighted average in which more reliable individual measurements are given more weight in the average. Thus, if the precision of individual measurements is known, the optimal strategy for computing the average is to combine individual measurements with more weight given to more reliable individual measurements.

Exactly how to implement precision-weighted averaging depends on how the problem is formulated. When faced with a group of samples to average, we could either assume that each individual item is a sample drawn from a single distribution or that each individual item is a sample drawn from a separate distribution. If we assume that individual measurements are separate samples from a single distribution, and the goal is to estimate the central tendency of the underlying distribution, then each measurement i should weighted by $1/\sigma_i^2$ (where σ_i^2 is the variance for item i). For instance, if one of the items has infinite variance, it will be completely ignored. This type of weighted average has been used extensively in the cue integration literature to define the optimal strategy for combining cues that have different degrees of reliability [59]. Alternatively, if the items are considered samples from separate distributions, and the goal is to estimate the mean of the sample, then items should never be given zero weight in the average. One strategy would be to compute the mean and variance of the samples, and to adjust the mean towards more reliable measures in proportion to their variance. In this case, an item with infinite variance would be included in the initial estimate of the mean, but there would be no additional updating of the mean towards this item. This strategy was employed in the simulations shown in Figure 4.

For ensemble averaging mechanisms to employ this type of precision-weighted averaging, the visual system would either have to know the degree of reliability with which items are represented or have a heuristic to calculate it. Both of these routes are plausible. Some models of visual perception model representations of individual items as probabilistic [59–61], in which knowledge is stored as a probability distribution that explicitly contains a representation of the reliability/variance of the representation. Alternatively, certain heuristics could be employed for estimating reliability, such as giving peripheral items less weight because visual resolution is known to drop off with eccentricity. Similarly, items inside the focus of attention might be weighted more than items outside the focus of attention because the precision with which items are represented is proportional to the amount of attention we give them. These heuristics would not be explicit representations of reliability, but they are cues that are tightly correlated with reliability, and thus they could be used to weigh individual items as a proxy for reliability.

It has been suggested that attended items are given more weight in the averaging of crowded orientation signals [62]. One study has shown that when attention is drawn to a particular item in the set, the mean judgment is biased towards that item [63]. One possible interpretation of this finding is that attention enhances the resolution with which the attended item is represented [42–44,48], and that items are weighed by their precision or reliability when computing the mean [40]. This possibility is speculative and has not been directly tested in uncrowded displays.

Beyond spatial averaging

Recent research on ensemble representation has gone beyond assessing the ability of observers to average visual features across space, including: (i) the ability to average features across time; (ii) the ability to represent other ensemble properties, such as the number of items in a set; (iii) the ability to represent spatial patterns; (iv) the relationship between ensemble representation and crowding; and (v) the neural correlates of ensemble representation.

Computing ensemble representations across time

In addition to spatial structure, there is a great deal of temporal structure and redundancy in the input to the visual system, and thus it would be advantageous to be able to also compute ensemble representations across time. Recent research has shown that observers can judge the mean size of a dynamically changing item or groups of items [40], or the mean expression of a dynamically changing face [56]. These findings demonstrate that perceptual averaging can operate over continuous and dynamic input, and that averaging across time can be as precise as averaging across space. Whether temporal averaging mechanisms constantly accumulate information or sample from high information points, such as salient transitions or discontinuities in the input stream, remains an open question. However, there is some evidence that certain information in a temporal sequence will be given more weight in the average than other information, possibly related to the amount of attention allocated to different points in the temporal sequence [40].

Number as an ensemble representation

Perhaps the most basic summary description for a collection of items is the number of items in the set. Without verbally counting, observers are able to estimate the approximate number of items in a set [64–66]. Similar to the perception of mean properties, the ability to enumerate items in a set occurs rapidly. It is also possible to extract the number of items across multiple sets in parallel [39]. Surprisingly, there is even evidence that number is directly perceived in the same way as other primary visual attributes [67]. Burr and Ross [67] demonstrated that it is possible to adapt to number in the same way that it is possible to adapt to visual properties such as color, orientation or motion. Number literally seems to be a 'perceived property' of sets. The relationship between the mechanisms underlying number representation and perceptual averaging is an important topic for future research.

Representing spatial patterns

Statistical summary representations, such as the mean or number of items in a set, are extremely compact representations, collapsing the description of a set down to a single number. However, images often consist of spatially distributed patterns of information, also referred to as spatial regularities or spatial layout statistics. For example, natural images consist of regular distributions of orientation and spatial frequency information [34,68]. In one study, Oliva and Torralba [34] measured orientation energy at different spatial scales over thousands of images and conducted a principal components analysis on these measurements. This analysis revealed that there are regularities in the structure of natural images, with certain patterns of

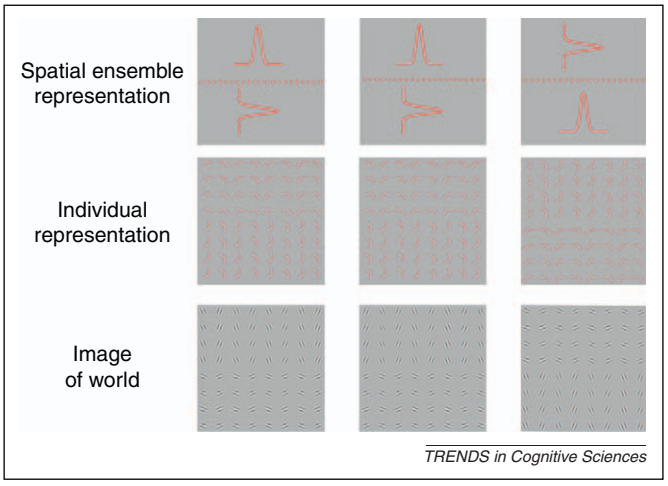

Figure 5. Spatial ensemble representations. Individual orientation measurements can be combined to represent patterns of orientation information. For each pattern, local orientation measurements are made (depicted as Gaussian curves centered around the true orientation), but each individual measure has a high degree of noise or uncertainty. Similar orientation signals are then pooled together to characterize regions with similar orientation signals using the average orientation. In the first column, the top half of the image has a mean orientation of vertical, whereas the bottom half has the a mean orientation of horizontal. The same is true for the image in the middle column. However, the pattern is flipped for the third column, here the top half has a mean of horizontal and the bottom half has a mean of vertical. Crucially, at the level of individual representations, the left and middle columns are just as different from each other as the left and right columns. However, at the ensemble level, the left and middle columns are more similar to each other than the left and right columns.

spatial frequency and orientation more likely to occur than other patterns. A schematic of a common pattern is shown in Figure 5, in which orientation signals tend to be more similar to each other within the top and bottom halves of the image than they are across the top and bottom halves. It would be efficient for the visual system to capitalize on the redundancy in natural images by using visual mechanisms that are tuned to the statistics of the natural world [11,69]. Indeed, a great deal of research has suggested that low-level sensory mechanisms are tuned to real-world statistical regularities [17,70–72].

The representation of such spatial ensemble statistics is robust to the withdrawal of attention, as would be expected if these ensemble representations are computed by pooling together local measurements [31]. For example, while attending to a set of moving objects in the foreground, changes to the background were only noticed when they altered the ensemble structure of the display, not when the ensemble structure remained the same, even though these changes were perfectly matched in terms of the magnitude of local change [31]. This suggests that the visual system maintains an accurate representation of the spatial ensemble statistics of a scene, even when attention is focused on a subset of items in the visual field.

Ensemble representation and crowding

Items in the visual field are often spaced too closely for each individual item to be resolved. For instance, it is unlikely that one can perceive the individual letters three sentences above or below this one. Yet, one can tell that there are letters present, that these letters are grouped into several words and so on. What is the nature of our perceptual representation when looking at a crowded collection of objects? There is a growing body of evidence suggesting that one perceives the higher-order summary statistics of information within the crowded region [21,73]. For a crowded set of oriented items, one perceives the average orientation [21]. For more complex patterns, such as a set of letters, the perceived pattern appears to result from a more complex statistical representation [73]. Balas and colleagues generated stimuli using a model which uses the joint statistics of cells which code for position, phase, orientation and scale [73]. Any pattern, such as sets of letters, can be passed through this model, resulting in a synthetic image that is somewhat distorted, yet is statistically similar to the original. When directly viewed, the original and the synthetic image look very different. However, identification performance with these synthetic images correlates with identification performance for crowded letters in the periphery, suggesting that perception in the periphery could consist of a similar statistical representation. The relationship between ensemble representation and crowding raises important questions regarding whether ensemble coding occurs automatically and whether it is perceptual in nature (Box 2).

Other studies suggest that there could be important differences between ensemble representation and crowd-

Box 2. Automaticity and directly perceived ensemble representations

A central question is whether the visual system automatically computes ensemble representations without conscious intention or effort, or whether they are computed voluntarily based on task demands. If ensemble representations were automatically computed, then we would conclude that there are dedicated mechanisms for computing and representing them. We might then focus on identifying the core ensemble feature dimensions and assessing their tuning properties. To understand such mechanisms, we can bring to bear methods that have been employed to understand perception, such as single-cell physiology, and perceptual adaptation. If ensemble representations are not computed automatically, but instead reflect a voluntary high-level judgment, then the methods we would use, and questions we would ask, might be somewhat different. For instance, physiology and adaptation are unlikely to reveal much about these mechanisms and ensemble representations would probably depend on task incentives and observers' goals. To understand such representations, we might explore regularities in how observers make ensemble judgments and turn our attention towards identifying consistent heuristics and biases in ensemble judgments.

In addition to the distinction between automatic and voluntary, there is an important distinction between 'directly perceived' and 'read-out' ensemble representations. In some cases the observer directly perceives the ensemble representation. For example, when a collection of items is presented in the periphery, their orientations appear to be automatically averaged [28]. With such crowded items, the perceptual experience is of 'directly seeing' the average orientation (all items appear to have an orientation equal to the mean of the group), with an accompanying loss of perceptual access to the individual orientation signals. By contrast, when the same display appears at the fovea, the oriented items are not crowded and the orientation signals do not appear to be obligatorily averaged: it is clear that the items have different orientations and none of them appears to have an orientation that matches the average. However, even for uncrowded displays, it is possible that ensemble representations are automatically computed. For example, ensemble representations appear to be automatically computed when the primary task does not require it [77] and even when they impair task performance [94].

ing. For instance, crowding is greater in the upper visual field than the lower visual field, whereas under the same conditions the accuracy of ensemble judgments was the same in the upper and lower visual field [74]. Thus, although ensemble coding and crowding are closely related, there could be important dissociations between them.

Neural correlates of ensemble representation

Relatively little research has explored the neural mechanisms of ensemble representation. Perhaps the most basic question we can ask is whether there are brain regions with neurons dedicated to computing ensemble representations (above and beyond the computation of individual object representations). Extensive research suggests that the parietal cortex plays an important role in the representation of number [75]. However, much less research has been done to explore the representation of perceptual averages, such as mean size, mean facial expression or mean orientation. Future research in this area would provide important insight into the nature of ensemble coding, as well as the functional organization of the visual cortex.

Additional benefits of computing ensemble representations

The present article has focused on one primary benefit of ensemble representation: the ability to combine imprecise individual measurements to construct an accurate representation of the group, or ensemble. However, computing ensemble representations could yield many related benefits [18,76], which are discussed here.

Information compression

Compression is the process of recoding data so that it takes fewer bits of information to represent that data. To the extent that the encoding scheme distorts or loses information, the compression is said to be lossy. For instance, TIFF image encoding uses a form of lossless compression, whereas JPEG image encoding is a lossy form of image compression – although the information lost occurs at such a high spatial frequency that human observers typically cannot detect this loss. Ariely [18] proposed that reducing the representation of a set to the mean, and discarding individual representations, would be a sensible form of lossy compression for the human visual system: it leaves available an informative global percept which could potentially be used to navigate and choose regions of interest for further analysis. However, this form of compression would only be economical if ensemble representations and individual representations were 'competing' in some sense. Otherwise, in terms of compression, there is no advantage to discarding the individual representations, and one might as well extract the ensemble and retain the individual representations. There is some evidence that ensemble representations take the same memory space as individual representations [39], although other studies suggest that ensemble representations and noisy individual representations are maintained concurrently and that these levels are mutually informative [77,78]. These findings suggest that ensemble representations and individual representations probably do not compete for storage, at least not in a mutually exclusive manner. However, none of these previous studies directly pitted ensemble memory versus individual memory and assessed possible trade-offs between them. Future research will be necessary to explore the extent to which ensemble representations and individual representations compete in memory. In terms of perceptual representations, it seems clear that individual and ensemble representations can be maintained simultaneously [23].

Whether ensemble coding is lossy or lossless depends on the fate of lower-level, individual representations. However, at the level of the ensemble representation, it is clear the data have been transformed into a more compressed form. It is possible that this format is more conducive to memory storage and learning. Ensemble representations are more precise than the lower-level representations composing them. Thus, there can be higher specificity of response at the ensemble level than at lower levels of representation. Such sparse coding has several advantages [79,80], including minimizing overlap between representations stored in memory [81] and learning associations in neural networks [82]. The extent to which observers can learn over ensemble representations of the type described in the present article is an important topic for future research, because it could bridge the gap between research on ensemble coding in visual cognition with the vast field of research on sparse coding and memory.

Ensemble representations as a basis for statistical inference and outlier detection

Another potential benefit to building an ensemble representation is to enable statistical inferences [83], including estimating the parameters of the distribution (mean, variance, range, shape), setting confidence intervals on those parameter estimates and classifying items into groups. A special case of classification is outlier detection, and an ensemble representation is ideal for this purpose [18,76]. For instance, if a set is well described by a distribution along an arbitrary dimension, say with a mean of 20 and standard deviation of 3, then an item with a value of 30 along this dimension is unlikely to be a member of the set. The ensemble representation would enable labeling this item as an outlier or even as a member of a different group. Outlier detection has been extensively studied using the visual search paradigm, in which the question has been whether an oddball item will instantly 'pop out' from a larger set of homogeneous items [84]. Items that are very different from the set, say a red item among green items, are said to be salient, and are easy to find in a visual search task [85,86]. Interestingly, computational models of saliency focus on 'local differences' between each item and its neighbors [87]. However, one could imagine displays in which the local context of a search target remained unchanged, but more distant items varied to either increase or decrease the degree to which the target appeared to be a member of the overall set. Finding that outlier status guides visual search above and beyond its effects on local saliency would provide strong support for the idea that ensemble representations play an important role in outlier detection.

Although it would be interesting if ensemble representations could enable rapid outlier detection, this finding is not necessary to support the idea that ensemble representations play an important role in classifying and grouping items. For instance, a face with a unique facial expression does not pop-out in a visual search task [88]. However, recent research shows that an outlier face is given reduced weight in the ensemble representation of a group of faces [58], even though observers often fail to perceive the outlier. This finding is consistent with the possibility that the ensemble representation enables labeling of items, but could also indicate that the ensemble computation gives outliers lower weight without attaching a classification label. The role of ensemble representations in determining set membership has not yet been extensively studied, and research in this area can potentially bridge the gap between study on ensemble representation, statistical inference and perceptual grouping.

Building a 'gist' representation that can guide the focus of attention

As detailed in previous sections, the power of averaging makes it possible to combine imprecise local measurements to yield a relatively precise representation of the ensemble (Figure 1). Moreover, it is possible to combine individual measurements to describe spatial patterns of information (Figure 5). A primary benefit of computing either type of ensemble representation is to provide a precise and accurate representation of the 'gist' of information outside the focus of attention. Without focused attention, our representations of visual information are highly imprecise [23]. If we were to simply discard or ignore these noisy representations, our conscious visual experience would be limited to only those items currently within the focus of attention. Indeed, some have argued that this is the nature of conscious visual experience [89,90]. In such a system, attention would be 'flying blind', without access to any information about what location or region to focus on next.

Although locally imprecise, ensemble representations provide an accurate representation of higher-level patterns and regularities outside the focus of attention [23,31]. These patterns and regularities are highly diagnostic of the type of scene one is viewing [14], and therefore they are useful for determining which environment one is currently located within. Over experience, observers appear to learn associations between these ensemble representations and the location of objects in the visual field. For instance, observers appear to use global contextual information to guide the deployment of attention to locations likely to contain the target of a visual search task [33,91–93].Thus, rather than flying blind, the visual system can compute ensemble representations, providing a sense of the gist of information outside the focus of attention, and guiding the deployment of attention to important regions of a scene.

In terms of forming a complete representation of a scene, gist representation and outlier detection probably work in tandem. For instance, when holding a scene in working memory, observers appear to encode the gist of the scene plus individual items that cannot be incorporated into the summary for the rest of the scene (i.e. outliers) [78].

Benefits of building a hierarchical representation of a scene

There are distinct computational advantages to building a hierarchical representation of a scene. In particular, by integrating information across levels of representation, it is possible to increase the accuracy of lower-level representations. It appears that observers automatically construct this type of representation when asked to hold a scene in working memory [77,78]. For instance, when recalling the size of an individual item from a display, the remembered size was biased towards the mean size of the set of items in the same color, and towards the overall mean size of all items in the display [77]. These results were well captured by a Bayesian model in which observers integrate information at multiple levels of abstraction to inform their judgment about the size of the tested item.

Concluding remarks

Traditional research on visual cognition has typically assessed the limits of visual perception and memory for individual objects, often using random and unstructured displays. However, there is a great deal of structure and redundancy in real-world images, presenting an opportunity to represent groups of objects as an ensemble. Because ensemble representations summarize the properties of a group, they are necessarily spatially and temporally imprecise. Nevertheless, such ensemble representations confer several important benefits. Much of the previous research on ensemble representation has focused on the fact that the human visual system is capable of computing accurate ensemble representations. However, the field is moving towards a focus on investigating the mechanisms that enable ensemble coding, the nature of the ensemble representation, the utility of ensemble representations and the neural mechanisms underlying ensemble coding. This future research promises to uncover important new properties of the representations underlying visual cognition and to further demonstrate how representing ensembles enhances visual cognition.

Acknowledgments

For helpful conversation and/or comments on earlier drafts, I thank Talia Konkle, Jason Haberman and Jordan Suchow. G.A.A. was supported by the National Science Foundation (Career Award BCS-0953730).

References

1 Kersten, D. (1987) Predictability and redundancy of natural images. *J. Opt. Soc. Am. A* 4, 2395–2400
2 Field, D.J. (1987) Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* 4, 2379–2394
3 Brady, N. and Field, D.J. (2000) Local contrast in natural images: normalisation and coding efficiency. *Perception* 29, 1041–1055
4 Frazor, R.A. and Geisler, W.S. (2006) Local luminance and contrast in natural images. *Vis. Res.* 46, 1585–1598
5 Webster, M.A. and Mollon, J.D. (1997) Adaptation and the color statistics of natural images. *Vis. Res.* 37, 3283–3298
6 Hyvärinen, A. and Hoyer, P.O. (2000) Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.* 12, 1705–1720
7 Judd, D.B. *et al.* (1964) Spectral distribution of typical daylight as a function of correlated color temperature. *J. Opt. Soc. Am. A* 54, 1031–1040

8 Long, F. *et al.* (2006) Spectral statistics in natural scenes predict hue, saturation, and brightness. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6013–6018
9 Maloney, L.T. (1986) Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *J. Opt. Soc. Am. A* 3, 1673–1683
10 Maloney, L.T. and Wandell, B.A. (1986) Color constancy: a method for recovering surface spectral reflectance. *J. Opt. Soc. Am. A* 3, 29–33
11 Field, D.J. (1989) What the statistics of natural images tell us about visual coding. *SPIE: Hum. Vis. Vis. Process. Digit. Display* 1077, 269–276
12 Burton, G.J. and Moorehead, I.R. (1987) Color and spatial structure in natural scenes. *Appl. Opt.* 26, 157–170
13 Geisler, W.S. (2008) Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* 59, 167–192
14 Torralba, A. and Oliva, A. (2003) Statistics of natural image categories. *Network* 14, 391–412
15 Huffman, D.A. (1952) A method for construction of minimum redundancy codes. *Proc. IRE* 40, 1098–1101
16 Shannon, C.E. and Weaver, W. (1949) *The Mathematical Theory of Communication*, The University of Illinois Press
17 Atick, J.J. (1992) Could information theory provide an ecological theory of sensory processing? *Network: Comput. Neural Syst.* 3, 213–251
18 Ariely, D. (2001) Seeing sets: representation by statistical properties. *Psychol. Sci.* 12, 157–162
19 Chong, S.C. and Treisman, A. (2003) Representation of statistical properties. *Vis. Res.* 43, 393–404
20 Bauer, B. (2009) Does Steven's power law for brightness extend to perceptual brightness averaging? *Psychol. Rec.* 59, 171–186
21 Parkes, L. *et al.* (2001) Compulsory averaging of crowded orientation signals in human vision. *Nat. Neurosci.* 4, 739–744
22 Dakin, S.C. and Watt, R.J. (1997) The computation of orientation statistics from visual texture. *Vis. Res.* 37, 3181–3192
23 Alvarez, G.A. and Oliva, A. (2008) The representation of simple ensemble visual features outside the focus of attention. *Psychol. Sci.* 19, 392–398
24 Haberman, J. and Whitney, D. (2007) Rapid extraction of mean emotion and gender from sets of faces. *Curr. Biol.* 17, R751–R753
25 de Fockert, J. and Wolfenstein, C. (2009) Rapid extraction of mean identity from sets of faces. *Q. J. Exp. Psychol. (Colchester)* 62, 1716–1722
26 Spencer, J. (1961) Estimating averages. *Ergonomics* 4, 317–328
27 Smith, A.R. and Price, P.C. (2010) Sample size bias in the estimation of means. *Psychon. Bull. Rev.* 17, 499–503
28 Morgan, M. *et al.* (2008) A 'dipper' function for texture discrimination based on orientation variance. *J. Vis.* 8, 1–8
29 Peterson, C.R. and Beach, L.R. (1967) Man as an intuitive statistician. *Psychol. Bull.* 68, 29–46
30 Pollard, P. (1984) Intuitive judgments of proportions, means, and variances: a review. *Curr. Psychol.* 3, 5–18
31 Alvarez, G.A. and Oliva, A. (2009) Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7345–7350
32 Oliva, A. and Torralba, A. (2006) Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.* 155, 23–36
33 Oliva, A. and Torralba, A. (2007) The role of context in object recognition. *Trends Cogn. Sci.* 11, 520–527
34 Oliva, A. and Torralba, A. (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175
35 Navon, D. (1977) Forest before trees: the precedence of global features in visual perception. *Cognit. Psychol.* 9, 353–383
36 Kimchi, R. (1992) Primacy of wholistic processing and global/local paradigm: a critical review. *Psychol. Bull.* 112, 24–38
37 Thompson, P. (1980) Margaret Thatcher: a new illusion. *Perception* 9, 483–484
38 Young, A.W. *et al.* (1987) Configurational information in face perception. *Perception* 16, 747–759
39 Halberda, J. *et al.* (2006) Multiple spatially overlapping sets can be enumerated in parallel. *Psychol. Sci.* 17, 572–576
40 Albrecht, A.R. and Scholl, B.J. (2010) Perceptually averaging in a continuous visual world: extracting statistical summary representations over time. *Psychol. Sci.* 21, 560–567
41 Galton, F. (1907) Vox populi. *Nature* 75, 450–451
42 Palmer, J. (1990) Attentional limits on the perception and memory of visual information. *J. Exp. Psychol. Hum. Percept. Perform.* 16, 332–350
43 Alvarez, G.A. and Franconeri, S.L. (2007) How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *J. Vis.* 7, 1–10
44 Franconeri, S.L. *et al.* (2007) How many locations can be selected at once? *J. Exp. Psychol. Hum. Percept. Perform.* 33, 1003–1012
45 Titchener, E.B. (1908) *Lectures on the Elementary Psychology of Feeling and Attention*, Macmillan
46 Carrasco, M. *et al.* (2004) Attention alters appearance. *Nat. Neurosci.* 7, 308–313
47 Carrasco, M. *et al.* (2002) Covert attention increases spatial resolution with or without masks: support for signal enhancement. *J. Vis.* 2, 467–479
48 Yeshurun, Y. and Carrasco, M. (1998) Attention improves or impairs visual performance by enhancing spatial resolution. *Nature* 396, 72–75
49 Mack, A. and Rock, I. (1998) *Inattentional Blindness*, The MIT Press
50 Neisser, U. and Becklen, R. (1975) Selective looking: attending to visually specified events. *Cognit. Psychol.* 7, 480–494
51 Most, S.B. *et al.* (2005) What you see is what you set: sustained inattentional blindness and the capture of awareness. *Psychol. Rev.* 112, 217–242
52 Setic, M. *et al.* (2007) Modelling the statistical processing of visual information. *Neurocomputing* 70, 1808–1812
53 Kinchla, R.A. and Wolfe, J.M. (1979) The order of visual processing: "Top-down", "bottom-up", or "middle-out". *Percept. Psychophys.* 25, 225–231
54 Myczek, K. and Simons, D.J. (2008) Better than average: alternatives to statistical summary representations for rapid judgments of average size. *Percept. Psychophys.* 70, 772–788
55 Chong, S.C. and Treisman, A. (2005) Attentional spread in the statistical processing of visual displays. *Percept. Psychophys.* 67, 1–13
56 Haberman, J. *et al.* (2009) Averaging facial expression over time. *J. Vis.* 9, 1–13
57 Chong, S.C. *et al.* (2008) Statistical processing: not so implausible after all. *Percept. Psychophys.* 70, 1327–1334
58 Haberman, J. and Whitney, D. (2010) The visual system discounts emotional deviants when extracting average expression. *Atten. Percept. Psychophys.* 72, 1825–1838
59 Kersten, D. and Yuille, A. (2003) Bayesian models of object perception. *Curr. Opin. Neurobiol.* 13, 150–158
60 Vul, E. and Pashler, H. (2008) Measuring the crowd within: probabilistic representations within individuals. *Psychol. Sci.* 19, 645–647
61 Vul, E. and Rich, A.N. (2010) Independent sampling of features enables conscious perception of bound objects. *Psychol. Sci.* 21, 1168–1175
62 Mareschal, I. *et al.* (2010) Attentional modulation of crowding. *Vis. Res.* 50, 805–809
63 de Fockert, J.W. and Marchant, A.P. (2008) Attention modulates set representation by statistical properties. *Percept. Psychophys.* 70, 789–794
64 Dehaene, S. *et al.* (1998) Abstract representations of numbers in the animal and human brain. *Trends Neurosci.* 21, 355–361
65 Feigenson, L. *et al.* (2004) Core systems of number. *Trends Cogn. Sci.* 8, 307–314
66 Whalen, J. *et al.* (1999) Nonverbal counting in humans: the psychophysics of number representation. *Psychol. Sci.* 10, 130–137
67 Burr, D. and Ross, J. (2008) A visual sense of number. *Curr. Biol.* 18, 425–428
68 Geisler, W.S. *et al.* (2001) Edge co-occurrence in natural images predicts contour grouping performance. *Vis. Res.* 41, 711–724
69 Chandler, D.M. and Field, D.J. (2007) Estimates of the information content and dimensionality of natural scenes from proximity distributions. *J. Opt. Soc. Am. A* 24, 922–941
70 Barlow, H.B. and Foldiak, P. (1989) Adaptation and decorrelation in the cortex. In *The Computing Neuron* (Durbin, R. *et al.*, eds), pp. 54–72, Addison-Wesley
71 Lewicki, M.S. (2002) Efficient coding of natural sounds. *Nat. Neurosci.* 5, 356–363
72 Olshausen, B.A. and Field, D.J. (1996) Natural image statistics and efficient coding. *Network* 7, 333–339

73 Balas, B. *et al.* (2009) A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* 9, 13–18
74 Bulakowski, P.F. *et al.* Reexamining the possible benefits of visual crowding: dissociating crowding from ensemble percepts. *Atten. Percept. Psychophys.* (in press)
75 Piazza, M. and Izard, V. (2009) How humans count: numerosity and the parietal cortex. *Neuroscientist* 15, 261–273
76 Cavanagh, P. (2001) Seeing the forest but not the trees. *Nat. Neurosci.* 4, 673–674
77 Brady, T.F. and Alvarez, G.A. Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychol. Sci.* (in press)
78 Brady, T.F. and Tenenbaum, J.B. (2010) Encoding higher-order structure in visual working-memory: a probabilistic model. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (Ohlsson, S. and Catrambone, R., eds), pp. 411–416, Cognitive Science
79 Olshausen, B.A. and Field, D.J. (2004) Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487
80 Olshausen, B.A. and Field, D.J. (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* 37, 3311–3325
81 Willshaw, D.J. *et al.* (1969) Non-holographic associative memory. *Nature (Lond.)* 222, 960–962
82 Zetzsche, C. (1990) Sparse coding: the link between low level vision and associative memory. In *Parallel Processing in Neural Systems and Computers* (Eckmiller, R. *et al.*, eds), pp. 273–276, Elsevier Science
83 Rosenholtz, R. (2000) Significantly different textures: a computational model of pre-attentive texture segmentation. In *Proceedings of the 6th European Conference on Computer Vision* (Vernon, D., ed.), pp. 197–211, Springer-Verlag
84 Rosenholtz, R. (1999) A simple saliency model predicts a number of motion popout phenomena. *Vis. Res.* 39, 3157–3163
85 Itti, L. and Koch, C. (2001) Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203
86 Wolfe, J.M. (1994) Guided search 2.0: a revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238
87 Itti, L. and Koch, C. (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. *Vis. Res.* 40, 1489–1506
88 Nothdurft, H.C. (1993) Faces and facial expressions do not pop out. *Perception* 22, 1287–1298
89 Noë, A. and O'Regan, J.K. (2000) Perception, attention and the grand illusion. *Psyche* 6 (http://psyche.cs.monash.edu.au/v6/psche-6-15-noe.html)
90 O'Regan, J.K. (1992) Solving the "real" mysteries of visual perception: the world as an outside memory. *Can. J. Psychol.* 46, 461–488
91 Torralba, A. *et al.* (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol. Rev.* 113, 766–786
92 Ehinger, K.A. *et al.* (2009) Modeling search for people in 900 scenes: a combined source model of eye guidance. *Vis. Cogn.* 17, 945–978
93 Chun, M.M. (2000) Contextual cueing of visual attention. *Trends Cogn. Sci.* 4, 170–178
94 Haberman, J. and Whitney, D. (2009) Seeing the mean: ensemble coding for sets of faces. *Hum. Percept. Perform.* 35, 718–734