Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity

Allison Yamanashi Leib University of California, Berkeley, Berkeley, CA, USA

Jason Fischer University of California, Berkeley, Berkeley, CA, USA
Massachusetts Institute of Technology, Cambridge, MA, USA

Yang Liu University of California, Berkeley, Berkeley, CA, USA

Sang Qiu University of California, Berkeley, Berkeley, CA, USA

Lynn Robertson University of California, Berkeley, Berkeley, CA, USA

David Whitney University of California, Berkeley, Berkeley, CA, USA

Individuals can rapidly and precisely judge the average of a set of similar items, including both low-level (Ariely, 2001) and high-level objects (Haberman & Whitney, 2007). However, to date, it is unclear whether ensemble perception is based on viewpoint-invariant object representations. Here, we tested this question by presenting participants with crowds of sequentially presented faces. The number of faces in each crowd and the viewpoint of each face varied from trial to trial. This design required participants to integrate information from multiple viewpoints into one ensemble percept. Participants reported the mean identity of crowds (e.g., family resemblance) using an adjustable, forward-oriented test face. Our results showed that participants accurately perceived the mean crowd identity even when required to incorporate information across multiple face orientations. Control experiments showed that the precision of ensemble coding was not solely dependent on the length of time participants viewed the crowd. Moreover, control analyses demonstrated that observers did not simply sample a subset of faces in the crowd but rather integrated many faces into their estimates of average crowd identity. These results demonstrate that ensemble perception can operate at the highest levels of object recognition after 3-D viewpoint-invariant faces are represented.

Introduction

There is a duality to perceptual processing. Our visual system is severely limited, and yet we have a rich phenomenological impression of the world. The limitations we face include attentional capacity, speed of neural processing, short-term memory, visual crowding, temporal crowding, and change blindness (Bentin, Allison, Puce, Perez, & McCarthy, 1996; Bonneh, Sagi, & Polat, 2007; Duncan, Ward, & Shapiro, 1994; Luck & Vogel, 1997; Simons & Levin, 1997; Whitney & Levi, 2011). Despite the striking limitations of vision (and perception in general) that have been uncovered experimentally, our subjective visual experience seems rich with detail. What is the content of this rich perception? Gist information, which is readily and quickly perceived in scenes (Oliva & Torralba, 2006; Potter, 1975; Rousselet, Joubert, & Fabre-Thorpe, 2005), may underlie the subjective richness of perception, and ensemble or summary statistical information may be the basic unit of gist perception (Alvarez, 2011; Haberman & Whitney, 2012).

The visual system takes advantage of redundancies in the scene by extracting summary statistics from groups of similar items. For example, a person viewing a complex outdoor scene will probably not examine every leaf on every tree. Instead, his or her visual

Citation: Yamanashi Leib, A., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, *14*(8):26, 1–13, http://www.journalofvision.org/content/14/8/26, doi:10.1167/14.8.26.

doi: 10.1167/14.8.26 Received September 3, 2013; published July 29, 2014 ISSN 1534-7362

system will take advantage of redundant leaves and efficiently compute average statistics, such as mean leaf color, shape, or size. This type of group statistical analysis is referred to as ensemble coding. Such ensemble information may allow the observer to recognize that he or she is viewing a forest or even categorize a tree (e.g., conifer or deciduous).

Importantly, statistical summaries can be generated very rapidly before the visual system has time to localize or discriminate any particular individual item in the scene (Ariely, 2001; Haberman & Whitney, 2007, 2011). As such, ensemble codes are functionally very useful. Observers may achieve an accurate ensemble percept while distracted by another task (Alvarez & Oliva, 2008). Similarly, observers can effectively ensemble code while experiencing change blindness (Haberman & Whitney, 2011) or while experiencing visual crowding (Fischer & Whitney, 2011). Even individuals with neurological impairments, such as prosopagnosia or unilateral neglect, may gain access to useful ensemble information although discrimination of individual faces/objects is impaired (Demeyere, Rzeskiewicz, Humphreys, & Humphreys, 2008; Pavlovskaya, Bonneh, Soroker, & Hochstein, 2010; Yamanashi Leib, Landau, Baek, & Chong, 2012; Yamanashi Leib, Puri, et al., 2012). Because ensemble information is achieved so rapidly and is unhindered by many perceptual limitations, it is theorized that ensemble percepts contribute significantly to our perceptual awareness of the world, including the updating of visual working memory (Brady & Alvarez, 2011), guiding attention (Alvarez, 2011), outlier detection (Haberman & Whitney, 2010, 2012), and hierarchical organization in scene perception (Alvarez, 2011).

Importantly, ensemble coding can be successfully accomplished across numerous perceptual domains. For instance, observers can accurately estimate the average speed of moving objects, the average orientation and position of targets, and the average size of items in a set (Ariely, 2001; Chong & Treisman, 2003; Dakin & Watt, 1997; M. Morgan, Chubb, & Solomon, 2008; M. J. Morgan & Glennerster, 1991; M. J. Morgan, Watamaniuk, & McKee, 2000; Motoyoshi & Nishida, 2001; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). Statistical summary also occurs for faces (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007), suggesting that summary statistical information may be calculated even at the highest level of individual object recognition. However, it remains unclear from previous research whether statistical summaries are computed on 2-D image information or on 3-D viewpoint-invariant representations of objects.

Some previous work began to approach this question. For example, Chong and Treisman (2003) first addressed the question of whether statistical summary is based on merely the physical attributes of an object or whether it is based on the perception of an object. They asked participants to extract the mean size from a group of circles and found that participants' estimates were based on a psychological scale (Teghtsoonian, 1965) rather than on the geometric area of the circle. Additionally, Im and Chong (2009) required participants to make mean judgments of size using the Ebbinghaus illusion and found that extraction of the mean was based on the perceived size, not the physical size, of the objects.

These results are consistent with the idea that ensembles are formed on object-centered viewpoint-invariant representations. However, size illusions (including the Ebbinghaus illusion) may occur early in visual processing (Murray, Boyaci, & Kersten, 2006; Schwarzkopf, Song, & Rees, 2011), operating on 2-D image properties. Im and Chong's (2009) results leave open the possibility, then, that summary statistical processes are restricted to 2-D representations (and their 2-D context).

To address the question of whether statistical summary operates on viewpoint-invariant representations, one approach is to use real objects or faces. We and others have explored statistical summary in faces and demonstrated that participants are able to precisely and efficiently estimate the average expression and identity of a crowd (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007, 2009; Neumann, Schweinberger, & Burton, 2013). Haberman and Whitney (2007, 2009) found that participants' performance degraded when faces in the crowd were inverted, scrambled, or contained added noise. This result may suggest ensemble coding of high-level face information. However, in all of these experiments, the face images were 2-D, and a summary statistical process that operates over 2-D holistic descriptions of the faces could account for these results. Similarly, a recent study by Neumann and colleagues (2013) also demonstrated ensemble identity perception across different photographs of celebrities, suggesting that ensemble coding is based on the identity of the members in the crowd and not the photos themselves. However, all of the stimuli had similar "head angle and gaze direction" (Neumann et al., 2013). Thus, the question of whether ensemble coding can operate on viewpoint-invariant representations remains unanswered.

The goal of our study was to test whether ensemble percepts of crowds are based on viewpoint-invariant representations. We tested this by presenting rotated faces one at a time during a 900-ms window. This serial presentation served to approximate the natural scanning humans engage in when evaluating crowds. This type of presentation may also serve to simulate a crowd streaming past the observer (e.g., students coming out of a classroom, passengers disembarking an airplane, etc.). Our paradigm required participants to incorpo-

rate faces from multiple viewpoints into the ensemble percept. We found that observers were able to quickly and efficiently perceive the average facial identity in a crowd even when the faces were displayed in different orientations. The results demonstrate that summary statistical perception operates on viewpoint-invariant representations of faces. This is the strongest evidence to date that ensemble perception can occur at the highest levels of visual object processing.

Methods

Participants

In Experiment 1, we tested four participants. Participants' ages ranged between 24 and 34 ($M = 31$, $SD = 4.96$). In Experiment 2, we tested four participants as well (two participants who were also in Experiment 1). Participants' ages ranged between 21 and 35 ($M = 28$, $SD = 6.58$). In Experiment 3, we tested three participants (three participated in one or more of the previous experiments). Participants' ages ranged between 24 and 35 ($M = 27.33$, $SD = 4.93$). Each participant provided informed consent in accordance with the institutional review board guidelines of the University of California, Berkeley. All participants were familiar with the three identities of the photographed individuals.

Stimuli

To create our stimuli, we began with three distinct identities (Identity #1, Identity #2, Identity #3). We linearly morphed these identities using Fantamorph Deluxe, creating 47 morphs between each identity. There were 47 morphs between Identity #1 and Identity #2, 47 morphs between Identity #2 and Identity #3, and 47 morphs between Identity #3 and Identity #1 (see Figure 1). The original photos were created by photographing the individuals rotated at different orientations under uniform lighting conditions. This yielded a stimulus array with 144 pictures in total, including the original photos. We created four different arrays of stimuli. In one array, the faces were forward oriented (0°); in a second array, the faces were oriented at 22.5° rightward; in a third array, the faces were oriented at 22.5° leftward; and in the fourth array, the faces were oriented at 90° leftward (see Figure 1). (Different models were used in the actual experiment, but these models preferred not to have their photos published.) The maximum and minimum luminance in the pictures was 44.65 and 213.70 cd/m^2, respectively. The average maximum Michelson contrast was 0.60. Each face subtended 5.06° × 3.53° of the visual angle. All stimuli were viewed on a Macbook Pro laptop monitor with a resolution of 1152 × 720 pixels and a 60-Hz refresh rate. We used Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) in Matlab to present the stimuli.

Experiment 1

Experiment 1 task

During each trial, the computer program selected 18 faces surrounding a randomly chosen mean value. Importantly, the mean face was never displayed; rather, the faces surrounding the mean value were displayed to participants. Displayed faces ranged from −25 to +25 steps away from the mean in increments of 10 units (−25, −15, −5, 5, 15, 25 units around the mean). The temporal order of the displayed faces was randomized. Faces of a given value were repeated three times in an 18-face display whereas faces of a given value were not repeated when the set size was below 18 (see further description of varying set sizes below). In Experiment 1, participants viewed sequentially presented faces oriented at 22.5° leftward. The faces were presented on a white background in the center of the screen with a maximum spatial jitter of 2.63° on the x-axis and 1.85° on the y-axis. The faces were drawn from the stimulus array in Figure 1b, and the participants were asked to judge the average identity of the sequentially presented faces. Each face was presented for 50 ms with a 50-ms interstimulus interval (ISI). Three hundred thirty-four milliseconds after the display disappeared, a single random test face was presented centrally, and participants adjusted the test face to match the mean identity of the crowd by using the computer mouse to scroll through the array of stimuli (144 choices in all). Importantly, the array of possible test faces were forward oriented. Although there were 18 faces in each set, in each trial, we varied the proportion of the faces that were visible such that either one, two, four, or 18 faces were visible (6%, 11%, 22%, or 100% of the set). There were 100 trials of each subset condition in Experiment 1, 200 trials of each subset condition in Experiment 2, and 100 trials of each subset condition in Experiment 3. Our experimental design is similar to a paradigm employed by Haberman, Harp, & Whitney (2009) used to explore temporal ensemble coding. The notable exception is that the orientation of the display and test faces was altered in our design. By manipulating the proportion of faces presented, we were able to evaluate whether participants integrated more and more faces as they became available. This has the power to rule out random guessing or judging the set of

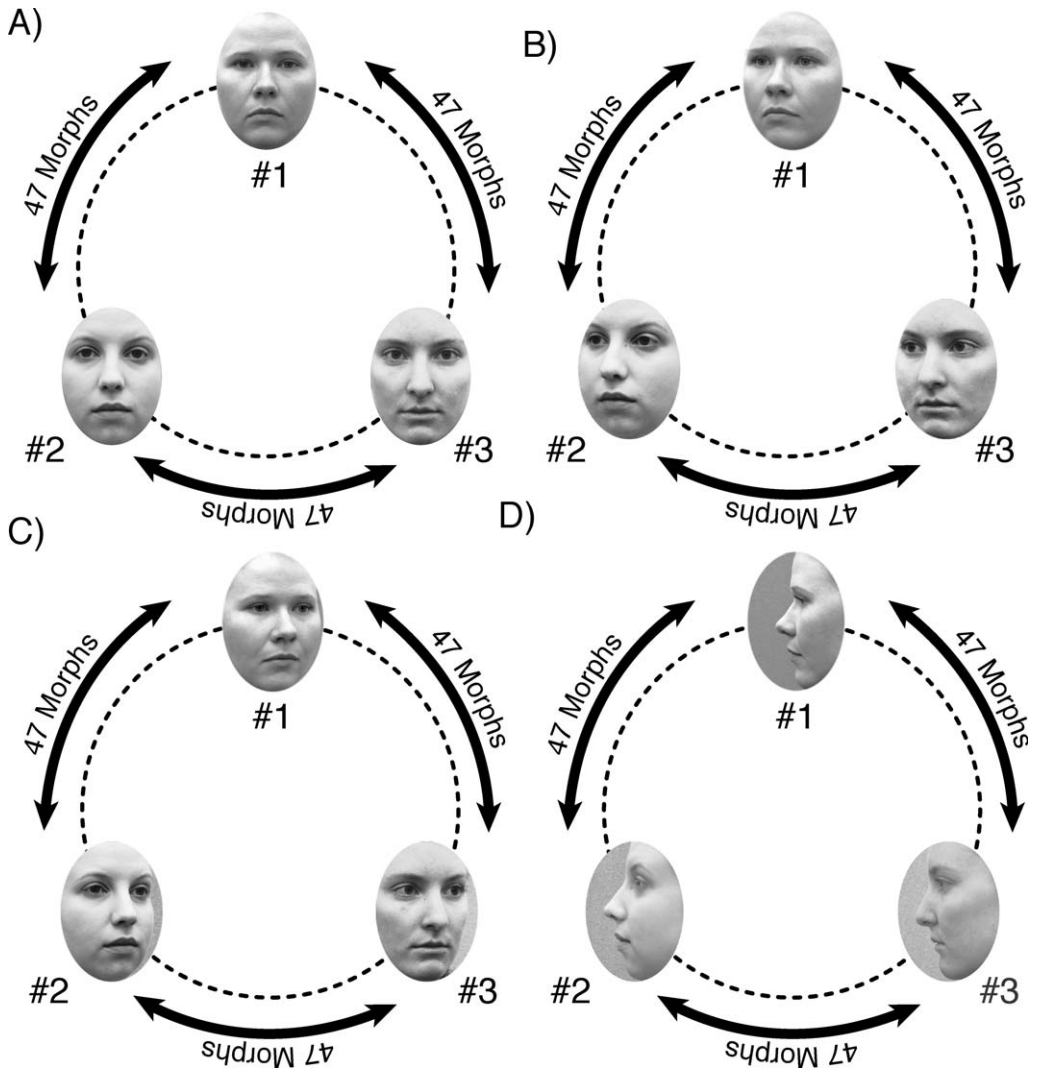

Figure 1. Illustration of the four different stimulus arrays used in the experiments. Each stimulus array began with three original pictures. We created 47 morphs in between each picture for a total of 144 pictures in each array. (A) The forward-facing stimulus array. (B) The leftward-facing 22.5° stimulus array. (C) The rightward-facing 22.5° stimulus array. (D) The leftward-facing 90° stimulus array.

18 faces based on just a small number of displayed faces. To the extent that observers integrated multiple rotated faces into a summary statistical percept, their sensitivity to the average of a set of 18 faces would have improved with more face samples (i.e., sensitivity to the mean of 18 should improve with increasing proportion of the set available).

Experiment 1 analysis

In order to analyze participants' accuracy for each trial, we used the following equation: Error = Mean of the Whole Display (in morph units) – Participants' Response (in morph units). By calculating the error for each trial in this manner, we were able to obtain an error distribution for the each condition. Next, we computed the mean of the error distribution using the following equation: Average Error (AE) = $\bar{x}$ (Absolute Value Error Distribution) and the standard deviation of the distribution of error using the following equation: Standard Deviation of Error (SDE) = σ (Error Distribution). This allowed us to assess the accuracy and precision of participants' responses respectively.

In each trial, the computer program calculated a mean for 18 faces. If the participant based his or her estimate of the mean on all of the available information, the error distribution should systematically decrease as more information became available.

Whereas, if the participant used only a small subset of faces to determine their estimate of the mean, their error distribution would remain relatively constant even when more information (i.e., larger number of

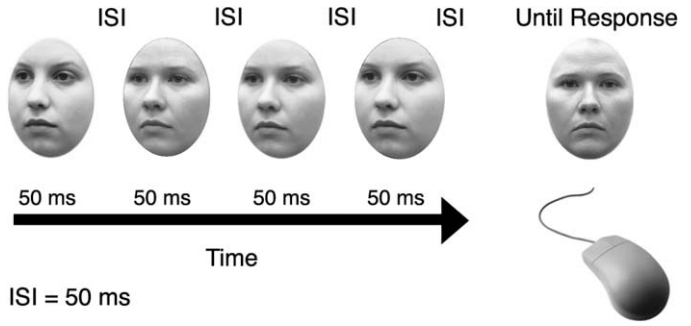

Figure 2. The sequence of trial events in Experiment 1. Subjects viewed sequentially presented faces oriented 22.5° leftward (this example shows a set size containing four faces). Each face was presented for 50 ms with a 50-ms ISI. After the stimuli disappeared, the participant could access all 144 faces in the forward-facing stimulus array by scrolling the mouse left or right. The participant had unlimited time to choose the mean identity of the crowd via mouse scroll.

faces) was revealed. The predicted pattern of error is shown in Figure 3.

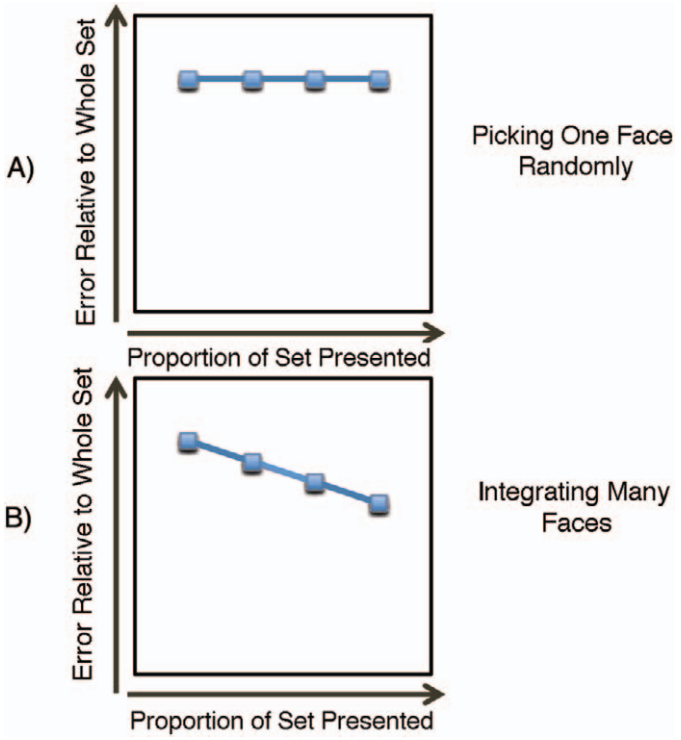

Figure 3. Participants' predicted error relative to the average of the 18 faces in the set as a function of the proportion of faces that are presented. The x-axis in each graph shows the proportion of faces from the set of 18 that were visible to observers. The y-axis shows the error (inversely proportional to sensitivity). (A) The predicted pattern of errors if the participant bases his or her estimate of the set mean on a single randomly viewed face. In this scenario, if the observer uses one face on which to base a judgment, he/she will not take new information into account. As a result, the participants' error remains constant even when more faces are revealed in the set. (B) The predicted pattern of errors if the participant integrates several faces into his or her estimate of the mean. In this scenario, the participant is incorporating much of the available information into his or her estimate of the mean. As a result, the error systematically decreases as a larger proportion of the set of faces is presented.

Experiment 1 results

We analyzed participants' performance relative to the mean of the whole set and found that the averaged group data matched the predicted pattern for ensemble coding.

When averaging across subjects, we used the formula for pooled standard deviation. The formula for pooled standard deviation is as follows:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \ldots + (n_k - 1)S_k^2}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1)}$$

Participants' accuracy and sensitivity increased as more information (i.e., more faces) became available. A one-way ANOVA revealed a significant main effect of set size: Accuracy = $F(3, 9) = 19.224$, $p < 0.001$, $\eta p^2 = 0.865$; sensitivity = $F(3, 9) = 20.623$, $p < 0.001$, $\eta p^2 = 0.873$. Participants performed better as set sizes increased (AE Set Size 1 = 25.97, AE Set Size 2 = 25.37, AE Set Size 4 = 21.72, AE Set Size 18 = 16.95; SDE Set Size 1 = 32.19, SDE Set Size 2 = 31.93, SDE Set Size 4 = 27.74, SDE Set Size 18 = 21.77). To determine whether participants were gaining information past four faces, we compared bootstrapped samples (Efron, 1986). A comparison of four and 18 set size bootstrapped samples revealed that participants were performing significantly better in the 18-face condition (six identities repeated three times) compared to the four-face condition for both accuracy ($p < 0.001$) and precision ($p < 0.001$). This indicated that participants increasingly integrated the available information into the ensemble code beyond four stimuli, suggesting that much of the multidirectional information was assimilated into the ensemble percept.

We were primarily interested in exploring whether participants can ensemble code a large crowd (up to 18 faces), and the subset design allowed us to confirm that participants were integrating the available information into their large crowd judgment.

However, a beneficial property of the subset design is that it also allows us to measure the participants' accuracy and sensitivity when discriminating a single face. The equation used for this analysis is Display Error = Display Face Value – Participant's Response. We compared participants' performance when they were engaged in an ensemble coding task compared to

Comparison of bootstrapped samples			
	One versus two faces Displayed	Two versus four faces Displayed	Four versus 18 faces Displayed
Accuracy	$p < 0.001$	$p < 0.001$	$p < 0.001$
Precision	$p = 0.008$	$p = 0.026$	$p = 0.006$

Table 1. Performance between set size conditions in Experiment 2. *Notes*: The table shows the p values for each comparison. Error drops significantly as more faces are displayed to the participants. This result suggests that participants are integrating new information as it becomes available and not basing their response on one or two randomly selected faces.

a single face discrimination task. In this comparison, the participant's response was not compared to the mean of the 18 faces, but rather, the participant's response was directly compared to the individual face presented. We compared the bootstrapped distributions of single face discrimination versus ensemble coding. These revealed that participants were more accurate ($p < 0.001$) and precise ($p < 0.001$) when judging crowd characteristics compared to single face discrimination. These results highlight the benefits of ensemble coding. When participants were shown a single face for a limited period of time, they only achieved a noisy representation of the face. However, as sample size increased, presumably, noise was cancelled out and greater precision was achieved.

An alternative explanation is that increased exposure duration affected participants' performance in both conditions. Although each face was shown for the same exposure duration (50 ms), as set size increased, the set of faces was shown for longer and longer durations (set size $\times$ 50 ms). Thus, the set of 18 faces was shown for the longest total duration. Therefore, we ran a second experiment and equalized exposure duration.

Experiment 2

Experiment 2 task

This experiment was similar to Experiment 1 except that we varied the amount of exposure time to the individual faces. One face was shown for 850 ms, two faces were shown for 434 ms each, four faces were shown for 217 ms each, and 18 faces (six identities repeated three times) were shown for 50 ms each. The ISI was 50 ms across all conditions. The design minimized the differences in total exposure time to faces for all conditions. If subjects incorporate multiple faces into their judgment of the set mean, we would still expect to find a downward slope in their response error, similar to the slope observed in Experiment 1. This would indicate that participants integrated additional information into their estimate of the mean as more faces became available, regardless of exposure duration.

Experiment 2 results

Again, we analyzed participants' performance in relation to the mean of the 18-face set. We used a one-way ANOVA with set size as the main factor and again found a significant main effect of set size. Participants' accuracy and precision increased as more information (i.e., more faces) became available: Accuracy, $F(3, 9) = 75.157$, $p < 0.001$, $\eta p^2 = 0.962$ (AE Set Size 1 = 18.90, AE Set Size 2 = 15.83, AE Set Size 4 = 13.83, AE Set Size 18 = 11.50); precision, $F(3, 9) = 53.132$, $p < 0.001$, $\eta p^2 = 0.947$ (SDE Set Size 1 = 23.28, SDE Set Size 2 = 20.36, SDE Set Size 4 = 17.76, SDE Set Size 18 = 15.10). We explored this main effect by comparing bootstrapped samples and found that there was a significant difference between each set size with performance systematically increasing beyond four faces. See Table 1.

Once again, although our primary question was to determine if participants exhibit ensemble coding behavior when viewing faces displayed in multiple orientations, we also compared participants' performance when they were engaged in an ensemble coding task versus a single face discrimination task using the separate display error analysis. Although the group average for single face discrimination was higher compared to crowd discrimination, this difference was only trending toward significance for both accuracy and precision when we compared the bootstrapped samples ($p = 0.064$, $p = 0.058$).

Experiment 3

In the first two experiments, we tested our participants' abilities to integrate information from one orientation (22.5°) and choose the mean identity from a different orientation (forward facing). Participants successfully chose the mean identities of the sets even though the test faces were presented in a different viewpoint. However, it is possible that participants encoded the individual faces as 2-D images, averaged those images, and then mentally rotated the ensemble. In this case, the ensemble is still calculated on the basis of the retinal image, and only the ensemble representation itself would be transformed into a different viewpoint. Alternatively, participants may have encoded the individual faces as 3-D representations and then integrated these representations into one ensemble percept. In Experiment 3, we sought to determine if

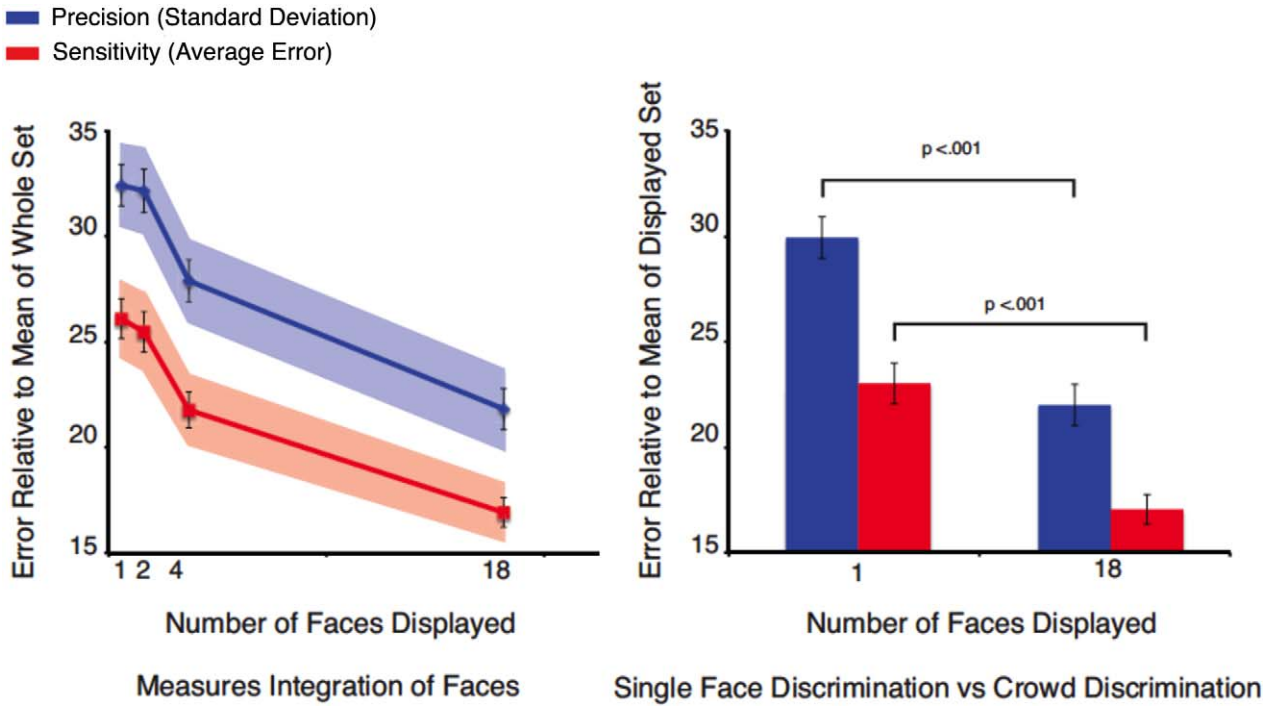

Figure 4. Group results for Experiment 1. (A) Subjects' accuracy and precision relative to the entire set of 18 faces. The negative slope shows the integration of information into the ensemble percept. As more information (i.e., faces) became available (x-axis), the reported ensemble face approached the mean of the 18 faces. Improvement in performance continued beyond four faces, indicating that at least four faces were integrated into the ensemble percept. (B) Subjects are significantly more precise and accurate when judging the average identity of a group of faces compared to judging the identity of a single face. Participants benefit from redundancy or noise reduction in ensemble coding by averaging out the error that might be present in single face discrimination. Error bars represent the standard deviation of 1,000 bootstrapped samples. The shaded regions represent the 95% confidence intervals of the bootstrapped distributions. The formula and description of bootstrapping are included in the supplementary material.

participants could integrate faces from multiple viewpoints into one ensemble code. We minimized the possibility that participants would use purely retinal images by presenting faces of different orientations in rapid succession. Because the displayed faces were presented leftward facing, rightward facing, and full profile, it was not advantageous for participants to average the retinal images. Averaging of 2-D images in multiple orientations would yield a distorted image that provides minimal information. In order to achieve successful performance in the task, participants would need to encode the individual faces as view-invariant representations.

Experiment 3 task

The third experiment was identical in design to Experiment 2 except that the individual faces in the display appeared in multiple orientations: leftward oriented at 22.5°, rightward oriented at 22.5°, and leftward oriented at 90° (Figure 6). As in both of the previous experiments, face values (distance from the mean) could be repeated up to three times in the 18-face set. However, the repeated face values were drawn randomly from three orientations; therefore, viewpoint homogeneity was minimized. Although viewpoint was chosen randomly, there was one constraint: Identical viewpoints could never be repeated sequentially. Thus, in the smaller subsets, no condition contained homogenous orientations. The exposure time also was identical to Experiment 2. This design ensured that improvements in performance observed in large set sizes were not merely due to exposure duration effects. In Experiments 1 and 2, a 2-D image-based averaging of the face images could result in a potentially meaningful image. However, in Experiment 3, averaging the 2-D face images would yield an identity that was not meaningful.

Experiment 3 results

We conducted a one-way ANOVA identical to the analysis in the previous experiments. Once again, we explored participants' performance in relation to the mean of the set of 18 faces, using set size as the main factor. As before, we found a significant main effect of

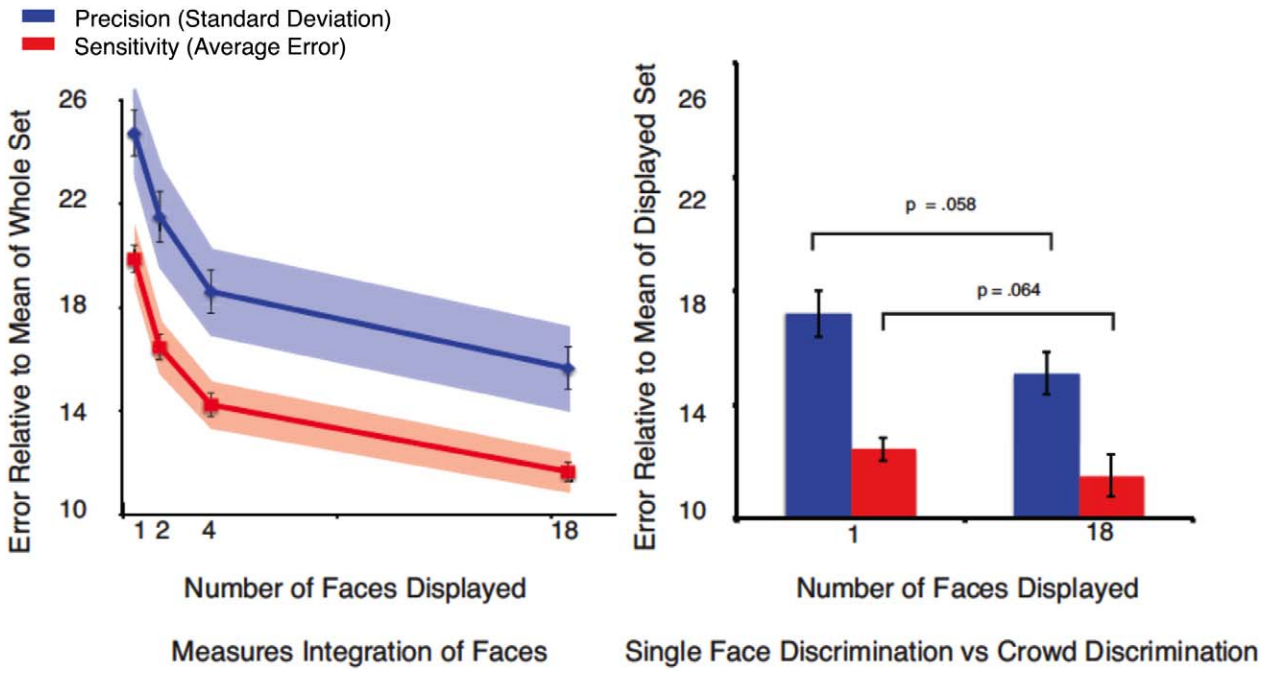

Figure 5. Group performance for Experiment 2. (A) Sensitivity and accuracy calculated relative to the entire set of 18 faces. The negative slope shows the integration of information into the ensemble percept. As more information (i.e., faces) became available (x-axis), perceived ensemble identity approached the mean of the 18 faces. Improvement in performance continued beyond four faces, indicating that at least four faces were integrated into the ensemble percept. Error bars represent the standard deviation of 1,000 bootstrapped samples. The shaded regions represent the 95% confidence intervals of the bootstrapped distributions.

set size with participants performing more accurately and precisely as more information became available: Accuracy, $F(3, 9) = 40.584$, $p < 0.001$, $\eta p^2 = 0.953$ (AE Set Size 1 = 19.231, AE Set Size 2 = 13.83, AE Set Size 4 = 12.70, AE Set Size 18 = 10.64); precision, $F(3, 9) = 23.140$, $p < 0.001$, $\eta p^2 = 0.920$ (SDE Set Size 1 = 24.01, SDE Set Size 2 = 17.96, SDE Set Size 4 = 16.91, SDE Set Size 18 = 14.06). One again, participants' performance continued to improve between four- and 18-face set size conditions, suggesting that more than four faces were integrated into the ensemble code.

We compared bootstrapped samples and found that participants performed significantly more accurately ($p < 0.008$) and precisely ($p = 0.008$) in the 18 set size condition compared to the four set size condition. This analysis addressed whether participants ensemble coded faces displayed in multiple viewpoints, and we found that participants did ensemble code faces even when orientation were divergent.

We also compared performance in Experiment 2 versus performance in Experiment 3. There were no significant differences between performance in the 18 set size between the two experiments for either accuracy ($p = 0.774$) or precision ($p = 0.872$), suggesting that diverse orientations in the display set in Experiment 3 did not hinder ensemble coding performance. Although our primary interest was whether participants could ensemble code faces displayed in multiple viewpoints, we also compared participants' performance when they were engaged in an ensemble coding task compared to a single face discrimination task using the separate display error analysis. Just as in previous experiments, we compared bootstrapped samples of performance during single face discrimination versus 18-face crowd discrimination. In Experiment 3, participants were significantly more accurate ($p = 0.008$) and more precise ($p = 0.008$) in the crowd condition. The results

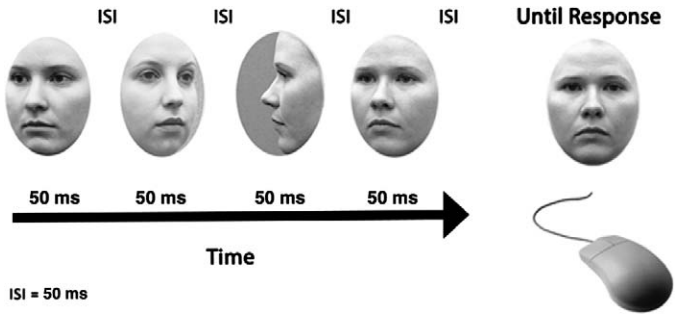

Figure 6. The sequence of trial events in Experiment 3. Participants viewed a sequence of faces oriented 22.5° leftward, 22.5° rightward, and 90° leftward (pseudorandom sequence in each trial). This particular example shows a set size of four faces although participants can view set sizes up to 18. After the stimuli disappeared, participants chose the mean identity of the crowd via mouse scroll (identical to Experiments 1 & 2).

from Experiment 3 suggest that participants can incorporate 3-D images into the ensemble percept. Participants' performance clearly increased as more information became available even when the combination of 2-D images was minimally informative.

Discussion

Previous work on ensemble or summary statistical perception has not clarified whether these percepts can be formed from viewpoint-invariant object representations. If summary statistical perception operates over the viewpoint-invariant, 3-D representations of objects, this would broaden the applicability and usefulness of ensemble coding throughout natural scenes, including faces in a crowd. Experiments 1 and 2 combined demonstrate that participants can transform either an object or an ensemble percept from one orientation into a new orientation. This potentially minimizes the need to resample ensemble codes after viewing orientation has changed. Experiment 3 demonstrates that participants can integrate multiple orientations into one ensemble percept. Experiment 3 also demonstrates that it is possible to formulate the ensemble percept based on 3-D representations of objects and not merely 2-D images. Finally, Experiment 3 demonstrates that participants can not only successfully ensemble code faces angled slightly away from the observer (22.5°), but they can also presumably integrate faces angled a full 90° from the observer. This is especially indicative of high-level processing as previous studies show that 90° profiles are not recognized by feature-based processing alone (Hill & Bruce, 1996). Taken together, these experiments suggest that the ensemble percept is not strictly image-based but can operate on viewpoint-invariant representations.

Our data highlights the precision of the ensemble-coding process. Participants were more precise at identifying the average of a group of faces compared to discriminating a single face. This result is intriguing given that faces in the group set condition were shown for a shorter duration and in multiple orientations whereas the single face was shown in one orientation and for a longer duration. Ariely (2001), using low-level objects, first hypothesized that ensemble coding could be more precise or at least equivalent to individual member identification. Our results show that ensemble coding precision trumps individual discrimination in higher-level object representations (i.e., faces), consistent with ensemble coding of crowd biological motion (T. Sweeny, Haroz, & Whitney, 2011; T. D. Sweeny, Wurnitsch, Gopnik, & Whitney, 2013). Furthermore, our results indicate that ensemble coding precision is preserved in the midst of increased processing demands, such as diverse orientation and briefer exposure times.

In any averaging process, noise is reduced with a greater number of samples. Many speculate that the process of ensemble coding similarly benefits from larger set sizes because of noise reduction, assuming noise is uncorrelated (Alvarez, 2011). Our finding that ensemble coding performance is often better than single face discrimination may be a result of noise cancellation. Robitaille and Harris (2011) offer direct evidence for this assertion by showing that reaction time and accuracy improve with larger sets using size/orientation ensemble coding tasks. Robitaille and Harris's results pertained to low-level ensemble discriminations. Our results complement and extend their observations as we also see an improvement in performance as set sizes increase but for higher-level objects.

Could the improvement in performance simply reflect redundancy contained within the displayed crowds? Experiments 1 and 2 allowed for the repetition of faces. For instance, in a set of 18 faces, three face values could be repeated. Thus, it is possible that subjects' enhanced performance in larger sets reflects the benefits of redundancy (Haberman & Whitney, 2009). In our experiment, we did not directly test the effect of redundancy; therefore, we cannot rule out that redundancy played a role. However, significant improvement was still observed between set sizes of two and four, which contained no repetition. Additionally, in Experiment 3, although the face value (distance from the mean) was repeated, the orientation often varied. Thus, it is very unlikely that the improvement observed in Experiment 3 was strictly due to repetition of the photographic images.

Our experiments also demonstrate the efficiency of ensemble coding. Our participants integrated objects into the ensemble code rapidly—much more rapidly than the response times reported for mental rotation in depth (Duncan et al., 1994; Marotta, McKeeff, & Behrmann, 2002; Tarr & Pinker, 1989). Mental rotation of faces commonly occurs within 1–3 s (Marotta et al., 2002), whereas the brief exposure times and ISIs in our experiments did not allow for mental rotation of individual faces before the subsequent face appeared. Our results complement previous findings that showed a dissociation between mental rotation and viewpoint invariance. For instance, Farah, Hammond, Levine, and Calvanio (1988) reported that a neurological patient accurately recognized misoriented objects, yet the patient was completely unable to perform mental rotation. Conversely, Turnball and McCarthy (1996) reported that another neurological patient was able to successfully mentally rotate objects but was unable to recognize objects that were misoriented. Thus, our findings extend this dissociation into the domain of

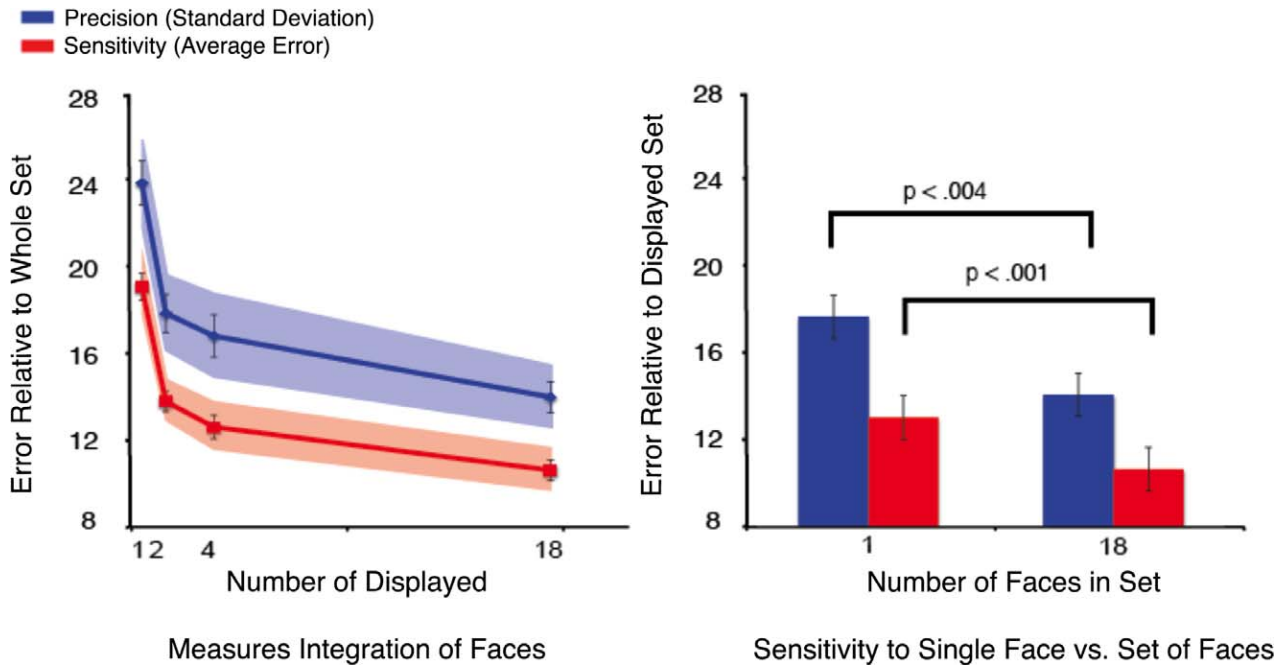

Figure 7. Group results for Experiment 3. (A) Sensitivity and accuracy calculated relative to the entire set of 18 faces. The negative slope clearly shows the integration of information into the ensemble percept. As more information (i.e., faces) became available (x-axis), subjects got closer to the mean of the 18 faces. Improvement in performance continued beyond four faces, indicating that at least four faces were integrated into the ensemble percept. (B) Subjects are significantly better at judging the average identity of a group of faces than they are at judging the identity of a single face. Notably, we replicated the result from Experiment 1 even though the group face judgment requires the integration of multiple viewpoints. Error bars represent the standard deviation of 1,000 bootstrapped samples. The shaded regions represent the 95% confidence intervals of the bootstrapped distributions.

ensemble coding and highlight the efficiency of ensemble coding even when items are diversely oriented.

The efficiency of ensemble coding also becomes apparent when our findings are compared to results in the visual search literature. For instance, it is commonly reported that individuals require 70–150 ms to find a particular face in a display (Nothdurft, 1993; Tong & Nakayama, 1999). Moreover, attentional capacities are strained when items are presented at speeds greater than 4–8 Hz (Verstraten, Cavanagh, & Labianca, 2000) with participants reporting interference when items are presented up to 300 ms apart (Duncan et al., 1994). In contrast, participants in our experiments integrated morphed faces when they were displayed for as little as 50 ms each. Although we cannot determine which face(s) were weighted more heavily during the integration process, our participants performed well when faces were displayed at a speed of 10 Hz, and participants exhibited increased accuracy at ensemble coding with larger set sizes, suggesting that interference was minimal even when stimuli were presented 50 ms apart. Thus, ensemble coding may successfully operate at the outer limits of attentional capacity.

Previous research suggests that ensemble coding can effectively operate even when perceptual and attentional processing is limited. For instance, Haberman and Whitney (2011) report that participants can accurately ensemble code faces even when experiencing change blindness. Additionally, many have reported that participants can effectively ensemble code with limited or impaired attention (Alvarez & Oliva, 2008, 2009; Yamanashi Leib, Landau, et al., 2012). Although our experiment cannot explain how ensemble coding bypasses the bottleneck of attention, our results complement these previous reports by confirming that ensemble coding of faces is not restricted by common limitations of visual processing and visual attention. Our experiments extend these finding by showing that rapid processing is feasible even when ensemble coding tasks demand the recruitment of high-level resources (i.e., viewpoint-invariant mechanisms).

Although the goal of our experiments was not to identify brain regions associated with ensemble coding, our data suggest that it is possible for ensemble coding to occur at the highest levels of visual object processing. Single unit recording studies indicate that viewpoint-invariant processing of objects occurs in extrastriate areas (Booth & Rolls, 1998). Similarly, single unit recordings of face-specific neurons suggest that viewpoint-invariant processing of faces is associated with neurons in ventral face-selective patches (Freiwald & Tsao, 2010; Perrett, Rolls, & Caan, 1982). Given this information, it is reasonable to conclude that our

participants utilized input from ventral visual cortex areas to achieve successful performance during the ensemble coding tasks. Although previous data suggest that ensemble coding likely occurs beyond primary visual areas (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007; Haberman & Whitney, 2009), our results are the first to suggest that ensemble coding utilizes input from cortical regions associated with viewpoint invariance.

Because our experiment involved an explicit ensemble coding task, we can only conclude that participants are able to utilize information from multiple viewpoints to formulate an ensemble code when required to. It does not necessarily follow that participants will automatically utilize information from multiple viewpoints to formulate an ensemble percept. Future experiments should explore whether similar results can be achieved during implicit ensemble coding tasks. Additionally, our experiment involved temporal processing of faces in a crowd. Although temporal processing of crowds is an integral aspect of daily visual perception, spatial processing of crowds is an equally useful aspect of visual perception. Future experiments should investigate whether ensemble coding is equally precise when multioriented faces are viewed in a spatial array.

Many experiments have shown that participants can ensemble code faces in crowds in a uniform orientation (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007; Haberman & Whitney, 2009). However, in natural scenes, items are rarely arranged in a homogeneous orientation. Our results may help provide a bridge between low-level image-based ensemble coding and high-level scene gist perception by showing that viewpoint-invariant ensemble representations can be accomplished. The results show that ensemble-coding a large number of items can yield increased precision compared to discriminating a single item. Furthermore, we show that ensemble coding is achieved very efficiently, much faster than individuating, attentionally dwelling upon, or mentally rotating a face. Most importantly, our results are the first demonstration that ensemble coding operates not merely by incorporating 2-D images, but also by incorporating 3-D, viewpoint invariant representations.

Keywords: ensemble coding, face perception, statistical summary

Acknowledgments

This research was supported by funding from the NIH EY018216 and NSF 0748689 awarded to David Whitney. The authors would like to thank Michael Passaglia for stimuli creation help. We would also like to thank Kelly Chang for help with figures and editing. Finally, we would like to thank Anna Kosovicheva for helpful discussions.

Commercial relationships: none.
Corresponding author: Allison Yamanashi Leib.
Email: ayleib@gmail.com.
Address: University of California, Berkeley, Berkeley, CA, USA.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 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, USA, 106*(18), 7345–7350.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8*(6), 551–565.

Bonneh, Y. S., Sagi, D., & Polat, U. (2007). Spatial and temporal crowding in amblyopia. *Vision Research, 47*(14), 1950–1962.

Booth, M. C., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex, 8*(6), 510–523.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory ensemble statistics bias memory for individual items. *Psychological Science, 22*(3), 384–392.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404.

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research, 37*(22), 3181–3192.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The*

Quarterly Journal of Experimental Psychology, 62(9)1716–1722.

Demeyere, N., Rzeskiewicz, A., Humphreys, K. A., & Humphreys, G. W. (2008). Automatic statistical processing of visual properties in simultanagnosia. *Neuropsychologia, 46*(11), 2861–2864.

Duncan, J., Ward, R., & Shapiro, K. (1994). Direct measurement of attentional dwell time in human vision. *Nature, 369*(6478), 313–315.

Efron, B. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*(1), 54–75.

Farah, M. J., Hammond, K. M., Levine, D. N., & Calvanio, R. (1988). Visual and spatial mental imagery: Dissociable systems of representation. *Cognitive Psychology, 20*(4), 439–462.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology, 106*(3), 1389–1398.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science, 330*(6005), 845–851.

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9*(11):1, 1–13, http://www.journalofvision.org/content/9/11/1, doi:10.1167/9.11.1. [PubMed] [Article]

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(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*(7), 1825–1838.

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review, 18*(5), 855–859.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), *From perception to consciousness: Searching with Anne Treisman* (pp. 339–349). New York: Oxford University Press.

Hill, H., & Bruce, V. (1996). The effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 986–1004.

Im, H., & Chong, S. (2009). Computation of mean size is based on perceived size. *Attention, Perception, & Psychophysics, 71*(2), 375–384.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1, https://psychtoolbox.org/PsychtoolboxCredits

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

Marotta, J. J., McKeeff, T. J., & Behrmann, M. (2002). The effects of rotation and inversion on face processing in prosopagnosia. *Cognitive Neuropsychology, 19*(1), 31–47.

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A "dipper" function for texture discrimination based on orientation variance. *Journal of Vision, 8*(11):9, 1–8, http://www.journalofvision.org/content/8/11/9, doi:10.1167/8.11.9. [PubMed] [Article]

Morgan, M. J., & Glennerster, A. (1991). Efficiency of locating centers of dot clusters by human observers. *Vision Research, 31*(12), 2075–2083.

Morgan, M. J., Watamaniuk, S. N. J., & McKee, S. P. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision Research, 40*(17), 2341–2349.

Motoyoshi, I., & Nishida, S. (2001). Temporal resolution of orientation-based texture segregation. *Vision Research, 41*(16), 2089–2105.

Murray, S. O., Boyaci, H., & Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. *Nature Neuroscience, 9*(3), 429–434.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128*(1), 56–63.

Nothdurft, H. C. (1993). Faces and facial experssions do not pop out. *Perception, 22,* 1287–1298.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155,* 23–36.

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.

Pavlovskaya, M., Bonneh, Y., Soroker, N., & Hochstein, S. (2010). Processing visual scene statistical properties in patients with unilateral spatial neglect. *Journal of Vision, 10*(7):280, http://www.

journalofvision.org/content/10/7/280, doi:10.1167/10.7.280. [Abstract]

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Perrett, D. D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47*(3), 329–342.

Potter, M. C. (1975). Meaning in visual search. *Science, 187,* 965–966.

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12*(6), 852–877.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12):18, 1–8, http://www.journalofvision.org/content/11/12/18, doi:10.1167/11.12.18. [PubMed] [Article]

Schwarzkopf, D. S., Song, C., & Rees, G. (2011). The surface area of human V1 predicts the subjective experience of object size. *Nature Neuroscience, 14*(1), 28–30.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*(7), 261–267.

Sweeny, T., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 329–337.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2013). Sensitive perception of a person's direction of walking by 4-year-old children. *Developmental Psychology*, *49*(11), 2120–2124.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21*(2), 233–282.

Teghtsoonian, M. (1965). The judgment of size. *The American Journal of Psychology, 78*(3), 392–402.

Tong, F., & Nakayama, K. (1999). Robust representations for faces: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance, 25*(4), 1016–1035.

Turnball, O. H., & McCarthy, R. A. (1996). When is a view unusual? A single case study of orientation-dependent visual agnosia. *Brain Research Bulletin, 40*(5), 497–502.

Verstraten, F. A., Cavanagh, P., & Labianca, A. T. (2000). Limits of attentive tracking reveal temporal properties of attention. *Vision Research, 40*(26), 3651–3664.

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences, 15*(4), 160–168.

Yamanashi Leib, A., Landau, A. N., Baek, Y., & Chong, S. C. (2012). Extracting the mean size across the visual field in patients with mild, chronic unilateral neglect. *Frontiers in Human Neuroscience, 6*, 267.

Yamanashi Leib, A., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia, 50*(7), 1698–1707.