Journal of Experimental Psychology:
Human Perception and Performance
2009, Vol. 35, No. 3, 718–734

0096-1523/09/$12.00 DOI: 10.1037/a0013899

Seeing the Mean: Ensemble Coding for Sets of Faces

Jason Haberman and David Whitney
University of California, Davis

We frequently encounter groups of similar objects in our visual environment: a bed of flowers, a basket of oranges, a crowd of people. How does the visual system process such redundancy? Research shows that rather than code every element in a texture, the visual system favors a summary statistical representation of all the elements. The authors demonstrate that although it may facilitate texture perception, *ensemble coding* also occurs for faces—a level of processing well beyond that of textures. Observers viewed sets of faces varying in emotionality (e.g., happy to sad) and assessed the mean emotion of each set. Although observers retained little information about the individual set members, they had a remarkably precise representation of the mean emotion. Observers continued to discriminate the mean emotion accurately even when they viewed sets of 16 faces for 500 ms or less. Modeling revealed that perceiving the average facial expression in groups of faces was not due to noisy representation or noisy discrimination. These findings support the hypothesis that ensemble coding occurs extremely fast at multiple levels of visual analysis.

Keywords: set perception, visual search, face recognition, summary statistics, perception

Our seamless interaction with our surroundings gives us the impression that we have a complete and accurate representation of the visual world. Well-controlled laboratory experiments, however, have revealed that the visual system samples only sparsely, and it has limited attentional and short-term memory capacity (Luck & Vogel, 1997; Potter, 1976; Rensink, O'Regan, & Clark, 1997; Scholl & Pylyshyn, 1999; Simons & Levin, 1998). What gives us the impression that we have such a complete representation of the visual world? One possible contribution may lie in the natural design of the environment—it is highly redundant. A field of grass, for example, contains repeating and overlapping features. Although we may be able to distinguish one blade of grass from another, coding every blade of grass would be computationally overwhelming and may serve little utility. Rather, what we tend to perceive by default is the whole field, a single texture comprising many blades of grass. This kind of *ensemble coding* reflects an adaptive mechanism that allows for the efficient representation of a large amount of information—so efficient that it has been suggested that this process may be responsible for the "illusion of completeness," filling in gaps of a visual scene where detailed representations are lacking (Chong & Treisman, 2003).

Ensemble coding, whereby summary statistics are derived from a set of similar items, has been examined for low-level features such as size (Ariely, 2001; Chong & Treisman, 2003, 2005) and orientation (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). Ariely (2001), for example, demonstrated that observers precisely extract the mean size from a set of dots varying in size while losing the representation of the individual set constituents. The precision of mean extraction is not significantly compromised by changing the distribution of dot sizes within the set (Chong & Treisman, 2003), suggesting a robust and flexible averaging mechanism. In orientation perception, Parkes, Lund, Angelucci, Solomon, and Morgan (2001) showed that individuals perceive a mean orientation in a set of crowded Gabor patches presented in the periphery, despite being unable to individuate the central target. Observers' inability to correctly identify the orientation of the target is not due to interference from the crowding flankers. Rather, observers' responses reflect an implicit pooling of all the elements in the set.

Given the overwhelming influx of information, it is not entirely surprising that the visual system employs an ensemble-coding heuristic. Uniform patterns such as dots or lines possess minimal amounts of variance, making it both easy and reasonable to use a single statistic to represent the whole set. By favoring a single summary statistic over a discrete representation for each set constituent, the system dramatically reduces computational load. Such ease of coding may explain why the dominant (and more relevant) percept when viewing a surface is that of a single texture and not a jumble of low-level features. In fact, ensemble coding may actually drive texture perception (Cavanagh, 2001). This does not mean, however, that ensemble coding operates only at midlevel vision (beyond V1 but before higher level object representation; Marr, 1982; Nakayama, He, & Shimojo, 1995). In a previous study, we showed that observers precisely represented the mean emotion of a set of emotionally varying faces—a level of processing well beyond that of surface perception (Haberman & Whitney, 2007).

Our initial findings revealed that ensemble coding is precise, is flexible, and occurs for high-level objects like faces. We now further characterize the mechanisms driving ensemble coding. The first two experiments demonstrate that this process occurs implicitly. Using a paradigm similar to Ariely's (2001), we show that observers unknowingly represent a set of faces using the mean emotion despite unrelated task instructions and do so at short

Jason Haberman and David Whitney, Center for Mind and Brain and Department of Psychology, University of California, Davis.

Correspondence concerning this article should be addressed to Jason Haberman, Department of Psychology, University of California, Davis, 1 Shields Avenue, Davis, CA 95616. E-mail: jmhaberman@ucdavis.edu

stimulus durations. Modeling confirms that performance cannot be explained by observer discrimination ability and thus points to an explicit averaging process. The third experiment replicates and extends previous work, demonstrating the precision with which the mean emotion of a set of faces may be represented. We show that observers' can discriminate a mean from an array of heterogeneous faces as well as they can discriminate any two individual faces—a surprising level of precision. Experiment 4 shows that despite a precise mean representation of a set of faces, observers have almost no persistent representation of the individual faces composing that set. The final two experiments are control experiments showing that this emotion averaging is indeed the result of the high-level properties of face stimuli and does not simply operate on low-level features. Observers are unable to extract a mean from a set of inverted or scrambled faces as well as they can from upright faces. These experiments converge to suggest that ensemble coding is implicit and fast and occurs across multiple levels of object complexity.

Experiment 1A

The first experiment tested observers' knowledge of the individual set members. Despite the instruction to attend to the individual members of the set, we hypothesized that performance on this task would reflect a bias to represent sets of faces with the mean emotion.

Method

Participants. Four individuals (1 woman, 3 men; mean age = 25 years) affiliated with the University of California, Davis, participated. Informed consent was obtained for all volunteers, who were compensated for their time and had normal or corrected-to-normal vision.

Stimuli. We generated a set of 50 faces by morphing (Morph 2.5; Gryphon Software, San Diego, CA) between two emotionally extreme faces of the same person, taken from the Ekman gallery (Ekman & Friesen, 1976). The emotional expression among the faces ranged from happy to sad (or neutral to disgusted), with Face 1 being the happiest. Morphed faces were nominally separated from one another by emotional units (e.g., Face 1 was one emotional unit happier than Face 2). The larger the separation between any two faces, the easier it should be to discriminate them (we tested this in Experiment 1B).

To create the range of morphs, the starting points of several features (e.g., the corner of the mouth, the bridge of the nose, the center of the eye) on one face are matched to their corresponding end points on the other face. For happy–sad stimuli, 75 points of interest were specified. The program then linearly interpolated the two original Ekman faces, creating 50 separate morphed images (see Figure 1). All face images were gray-scaled (the average face had a 98% maximum Michelson contrast) and occupied 3.04° × 4.34° of visual angle. The background relative to the average face had a 29% maximum Michelson contrast.

We varied set size from among 4, 8, 12, and 16 items, determined randomly on each trial. The faces were presented on the screen in a grid pattern in the following format: 2 row × 2 column matrix for the set of 4 items, 2 × 4 matrix for the set of 8 items (14.68° × 9.53°), 3 × 4 matrix for the set of 12 items (14.68° × 14.68°), and 4 × 4 matrix for the set of 16 items (14.68° × 19.77°). Each face was assigned to a random position in the matrix at the start of every trial.

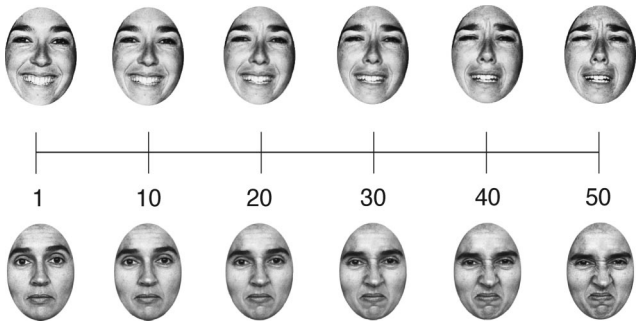

Figure 1. The spectrum of face morphs for both happy–sad and neutral–disgusted emotions. There were 50 faces for each emotional range. Observers saw only happy–sad or neutral–disgusted stimuli during a given experiment.

Procedure. On every trial there were four unique emotions displayed in the set, each of which was separated by at least six emotional units, a distance well above observers' discrimination thresholds (results discussed in Experiment 1B). In a set size of 8 there were two instances of each emotion, in a set size of 12 there were three instances of each emotion, and in a set size of 16 there were four instances of each emotion. The mean emotion of each set was randomly selected at the start of every trial. Once the mean was selected, the four unique emotions composing the set were selected surrounding the mean: two happier and two sadder. The two happier faces were three and nine units away from the mean, as were the two sadder faces (see Figure 2). The mean changed on every trial but was never a constituent of the set.

The set was displayed for 2,000 ms and was immediately followed by a single test face (0 interstimulus interval), which could be either a member or a nonmember of the preceding set. Each nonmember test face was at least 3 units away from a member face (see Figure 2). The full range of potential test faces was from 15 units below the mean to 15 units above the mean. Observers were instructed to indicate with a key press whether the test face was a member of the preceding set (a yes–no task; see Figure 2). The test face remained on the screen until a response was received.

For each of the four possible set sizes, there were 11 possible test faces (4 of which were members of the preceding set) and 5 trials for each of these test faces, for a total of 220 trials per run. Observers performed four runs for 880 trials total.

Results and Discussion

Figure 3A depicts the percentage of "yes" responses for each observer, collapsed across set size. A "yes" response indicates that the observer thought that the test face was a member of the previously presented set. The *x*-axis depicts separation of the test face from the mean of the set in emotional units (i.e., the mean changed from trial to trial, but this graph represents performance collapsed across all means). For all 4 observers, the probability of a "yes" response was substantially lower when the test face fell near the edge or outside of the set range, demonstrating sensitivity

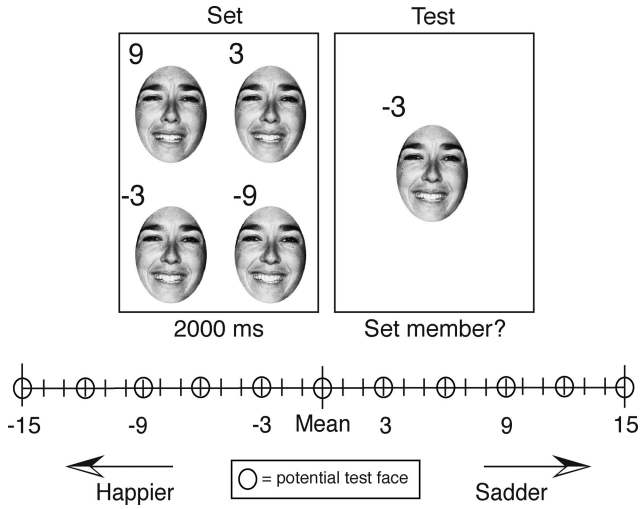

Figure 2. Task design for Experiment 1A. Observers saw four unique faces, selected on the basis of their emotional distance from the mean emotion, for 2,000 ms. Set size varied among 4, 8, 12, and 16 items. Observers had to indicate whether the test face was a member of the previously displayed set. The test face could be any of the distances indicated by the circles (numbers were not seen by participants).

to the emotional range of the set. More importantly, the probability of responding "yes" increased as the test face approached the emotional mean of the set even though the emotional mean was never actually presented in the original set. Despite the instruction to attend to the individual set members, their responses revealed a bias to indicate the mean emotion of each set. This is consistent with the findings of Ariely (2001), which demonstrated that observers unintentionally represented the average size of a set of dots. Our results suggest that observers implicitly extracted the mean of a set of faces on a trial-by-trial basis.

Observers did not act as ideal observers. An ideal observer's probability of "yes" responses would have produced a saw-toothed function (perfect accuracy). Figure 3A clearly demonstrates that this was not the case.

Experiment 1B

Is it possible that participants were simply noisy observers? Specifically, could noise at the perceptual, decision, or response stage produce something that looks like the data in Figure 3? To test this, we ran multiple simulations in which we convolved the expected performance of an ideal observer (i.e., a saw-toothed function) with observer discrimination ability. If discrimination ability determined performance on the yes–no set membership task, the resulting convolution should resemble observer performance. To create this convolution, however, we first had to determine each observer's discrimination performance, nominally referred to as homogeneous discrimination.

Method

Procedure. Each trial consisted of two intervals: a set of four identical faces simultaneously displayed for 2,000 ms in a grid pattern (6.94° × 9.53°) immediately followed by a single test face displayed in the center of the screen (see Figure 4A). The test face remained on the screen until a response was received. The emotionality of the set was randomly selected from the gallery of morphed faces. The subsequent test face was happier or sadder than the set by ±1–6 emotional units. In a method of constant stimuli two-alternative-forced-choice task (2AFC), observers were asked to indicate with a key press whether the test face was happier or sadder than the set of identical faces. Each run consisted of 20 trials at each of the six levels of separation for a total of 120 trials. Observers performed eight runs over two testing sessions for a total of 960 trials. Thus, 160 judgments were made at each of the six levels of separation. A logistic psychometric function was fit to the data with Psignifit toolbox version 2.5.6 from MATLAB (see http://bootstrap-software.org/psignifit/). Confidence intervals were derived through the bias-corrected accelerated bootstrap method based on 5,000 simulations, also implemented by Psignifit (Wichmann & Hill, 2001a, 2001b).

We also derived threshold estimates for neutral–disgusted morphs. Three observers (2 women, 1 man; mean age = 20.33 years) performed four runs of the same task described above using the alternate emotional stimuli (see Figure 4B). The separation between set emotion and test emotion was increased to ±2–12 units, with increments of 2 units between test conditions.

Results and Discussion

For each observer, we identified 75% correct discrimination in terms of units of emotional separation between set and test. Figure 5 shows the psychometric functions for all observers. Seventy-five percent correct thresholds were comparable for all observers in the happy–sad morph condition (KS: 5.3; FF: 3.8; JH: 2.6; DH: 4.4) as well as in the neutral–disgusted morph condition (KS: 4.4; JSH: 7.4; AC: 5.8). The results here reveal the precision with which observers could discriminate any two of the morphed faces (homogeneous discrimination ability). This information is critical for our modeling procedures, described below.

Modeling procedure and results. To test whether discrimination performance could predict yes–no set membership performance (i.e., simply because of observer noise), we convolved performance of an ideal observer with both the poorest discriminator of happy–sad morphs (Observer KS; see Figure 5A) and the most sensitive discriminator of happy–sad morphs (Observer JH; see Figure 5A). If the convolutions for these 2 observers mimic their yes–no set membership data, then performance on the task may be attributed to observer noise. If, however, this convolution does not replicate their yes–no set membership results, then performance may be attributed to an implicit representation of the mean emotion of each set.

As observer noise increases (i.e., discrimination ability is worse), the expected performance (i.e., the convolution) on the yes–no membership task should begin to look more like what we observed in Figure 3A—a Gaussian-shaped response probability distribution, albeit wider than observed data. As noise decreases, the expected performance should begin to resemble a saw-toothed function. Figure 6 shows KS's and JH's actual yes–no membership data and the modeled data (the ideal observer convolved with each observer's psychometric homogeneous discrimination function from Figure 5A).

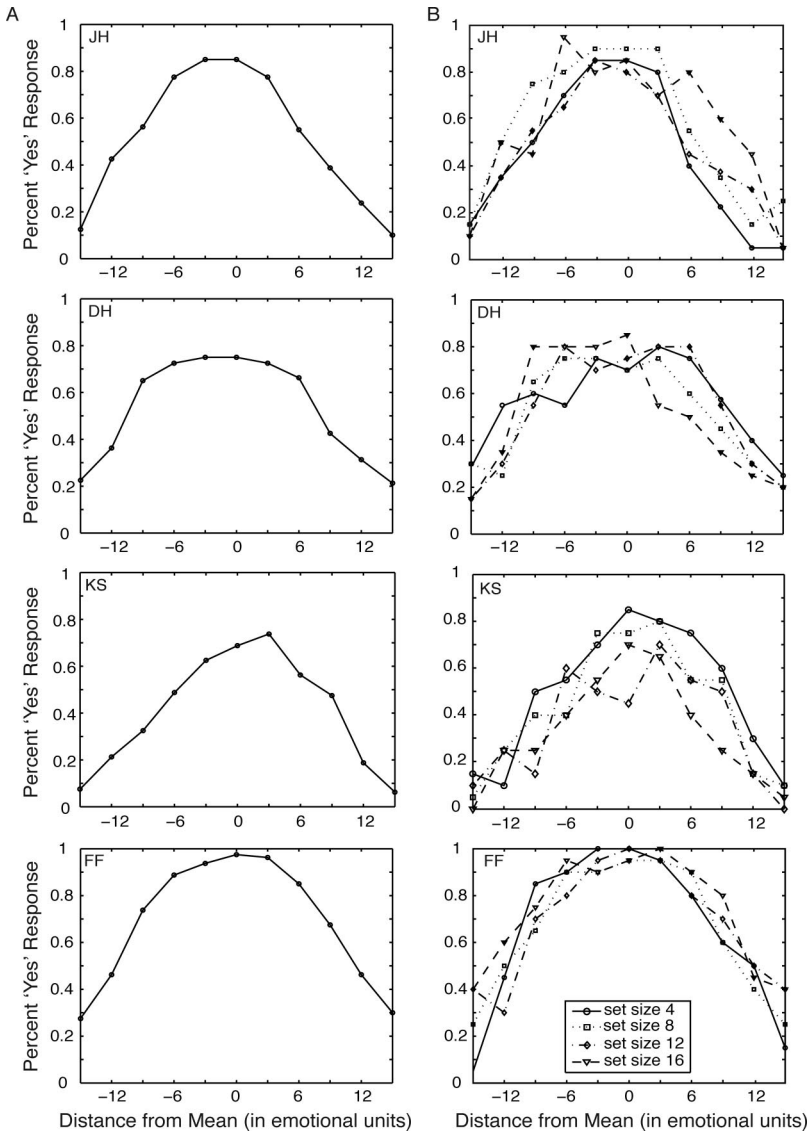

Figure 3. Probability of making a "yes" response for each subject (A) collapsed across set size and (B) broken down by set size. A "yes" response indicates that the observer believed the test face was a member of the preceding set. Probability of making a "yes" response peaked when the test face corresponded to the mean emotion of the set. (B) There were no systematic differences in probability of making a "yes" response as a function of set size.

We fit Gaussian curves to both the modeled (convolved) data and KS's and JH's observed yes–no membership data (see Figure 6). We selected a Gaussian curve because the data looked roughly normally distributed, but this particular function is not critical for our conclusions. Any symmetrical function with a central peak would have adequately fit these data. The Gaussian equation was formalized as

$$a \times \exp\left(-\left[\frac{(x-b)}{c}\right]^2\right),$$

where a is the amplitude, b is the phase, and c is the full width at 75% maximum (nominally referred to as width). The parameter of most interest was the width of the Gaussian fit, as this parameter reveals the precision of mean representation—the narrower the width, the more precise the representation. We used a conservative approach and fixed the amplitude of each Gaussian curve to the average amplitude of KS's observed yes–no membership data (see Figure 3A) and KS's convolved data. With amplitude fixed, only two parameters were free to vary: curve width and phase. Figure 6 clearly shows that the Gaussian curve fit to KS's probability of "yes" responses (Experiment 1A data, solid line) was substantially narrower than the noisy observer model (dashed curve), suggesting KS's representation of the set mean was more precise than would be predicted by KS's discrimination ability. To statistically test the

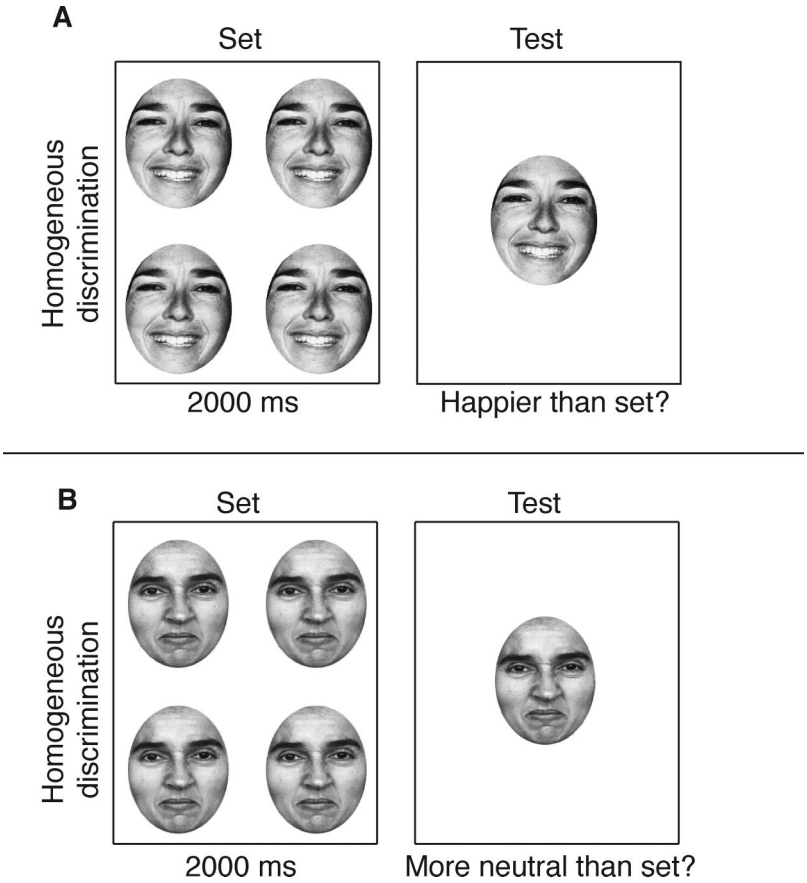

Figure 4. Task design for Experiment 1B. (A) Observers saw four instances of a randomly selected face displayed on the screen for 2,000 ms. This was immediately followed by a single test face. Observers had to determine whether the set or the test face was happier. (B) Same task as in Figure 4A, except the stimuli were neutral–disgusted morphs. Separations between set and test were doubled relative to the happy–sad condition.

difference between the width parameter estimates of the two curves, we fixed the width parameter of KS's convolved data to the width of KS's observed data and refit the Gaussian curve (now with two freely varying parameters, amplitude and phase, and fixed width). If there were a substantial decline in the goodness of fit between the original convolved Gaussian curve and the two-parameter Gaussian curve, then the two width parameters from the observed and convolved curves would be considered significantly different from one each other (i.e., one would be significantly wider than the other). There was a statistically significant difference in the quality-of-fit between these two models, $F(1, 8) = 31.01$, $p < .005$, suggesting that KS's face discrimination ability (Experiment 1B) could not account for the observed pattern of KS's responses in the set membership experiment (Experiment 1A).

Observer JH was the most precise discriminator of happy–sad face expression (see Figure 5A). Convolving JH's discrimination function with an ideal observer reveals a pattern resembling a saw-toothed function (triangles in Figure 6B). As opposed to KS's data, a Gaussian curve may not even be the appropriate function to use, making parameter comparisons between convolved and observed data moot. We therefore compared the quality-of-fit of a Gaussian distribution (a three-parameter model) with the quality-of-fit of a boxcar (a two-parameter model) distribution. A boxcar distribution would suggest insensitivity to the mean, because probability of responding that a test face is a set member would not vary as a function of distance from the mean. In this case, the less complicated model (i.e., the boxcar model with two parameters) fit the data better (sum of squares = 0.08 for boxcar vs. sum of squares = 0.13 for Gaussian). Because the same is not true of JH's observed yes–no data (first panel in Figure 3A)—that is, a Gaussian distribution describes the data better than a boxcar distribution, $F(1, 8) = 76.61$, $p < .0001$—we can conclude that JH's discrimination ability (Experiment 1B) cannot account for JH's pattern of yes–no membership data (Experiment 1A).

By systematically varying the slope and threshold of simulated psychometric discrimination functions (such as from Figure 5), a family of hypothetical noisy observers was created. The convolution technique described above was iteratively applied to each of the hypothetical noisy discrimination functions (see Figure 7). It is interesting that no realistic level of modeled noise was able to match the performance curve seen in the actual task. As the noise increased, the curves tended to become flatter and wider. Although the modeled noise still resembled Gaussian distributions, the pre-

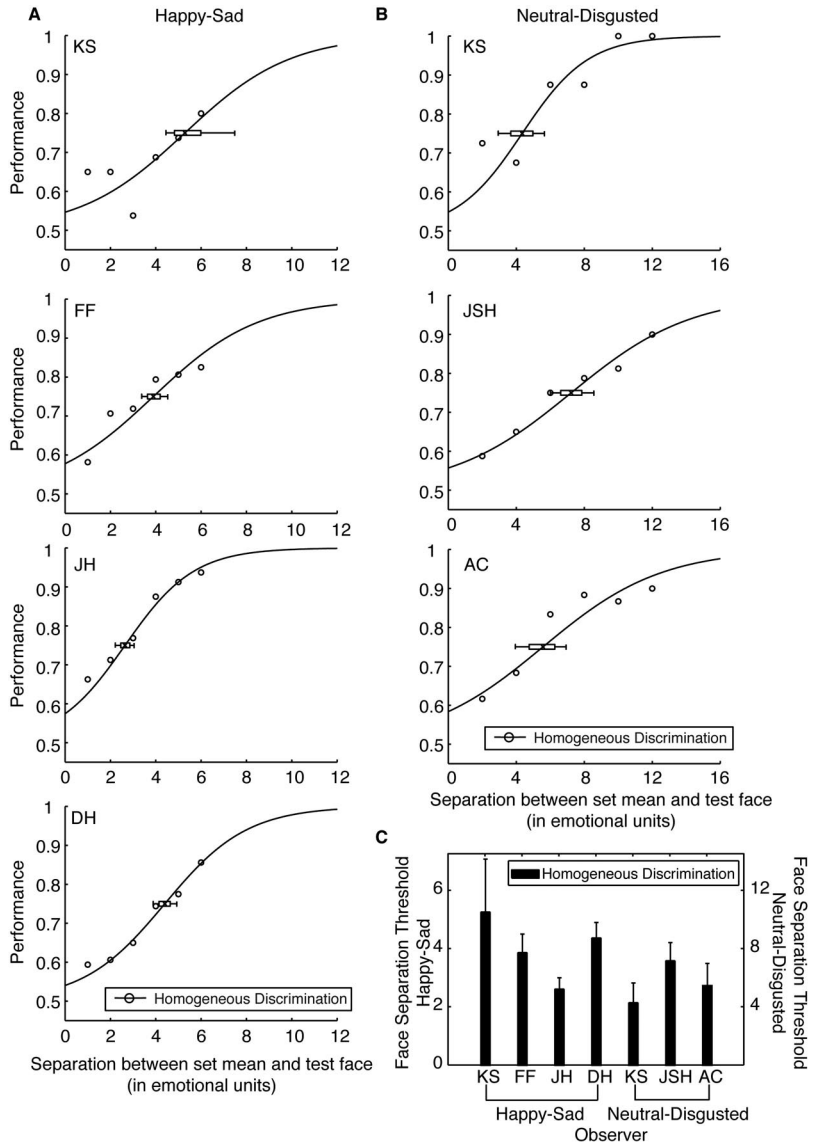

Figure 5. Results of Experiment 1B. (A–B) Psychometric functions for all observers. (C) Seventy-five percent thresholds for each observer in the happy–sad condition and the neutral–disgusted condition. (A–C) Error bars are 95% confidence intervals derived from 5,000 bootstrap simulations (Wichmann & Hill, 2001a, 2001b). For fitting purposes, we included a point at 0 separation between set and test (chance performance), which does not appear in the graph.

cision of predicted mean representation (as reflected by width) did not approach the level of precision in the actual data. As the noise decreased (i.e., became more like an ideal observer), the distribution became less like a Gaussian curve and more like a saw-toothed function. However, the saw-toothed shape does not resemble the membership data. As was the case for JH's convolved data (see Figure 6B), the saw-toothed function (see Figure 7, inverted triangles) is not well modeled by a Gaussian curve—a boxcar function is a better fit. Expressed in another way, the yes–no membership data reveal a very narrow distribution of responses centered at the average facial expression; to produce such a narrow distribution of responses based solely on noisy discrimination would have required a saw-toothed shaped yes–no distribution. Clearly, this did not happen. Therefore, no level of noise in discrimination ability, whether it was substantial (very poor discriminator, or worse than Observer KS) or minimal (close to an ideal observer, or better than Observer JH), could explain the precision of the mean representation found in observers' data. These simulations suggest that the noisy observer hypothesis cannot account for the yes–no membership data.

We also examined performance as a function of set size, depicted in Figure 3B. We should note that viewing four iterations of 4 emotionally varying faces (set size 16) is not the same as viewing only 4 faces. Increasing the number of items effectively increases

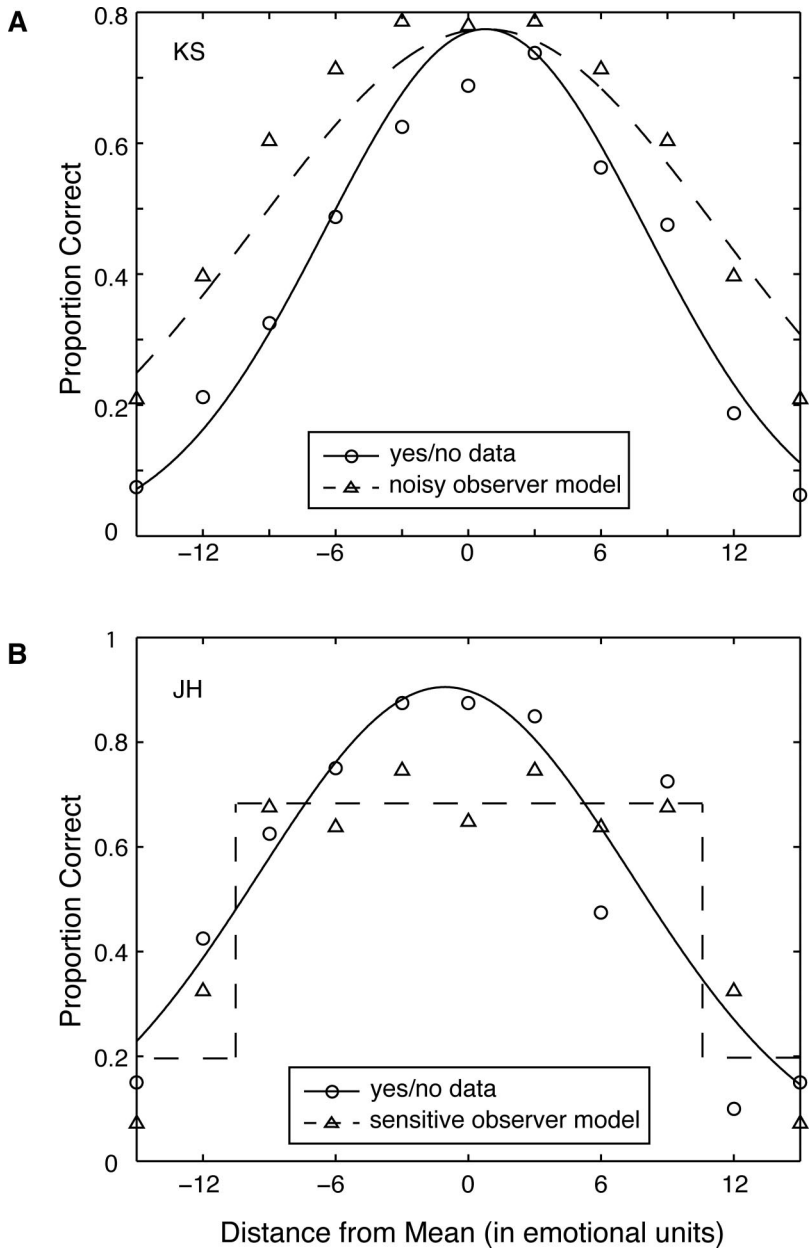

Figure 6. (A) Comparison of expected performance for KS on the yes–no membership task (happy–sad stimuli) to KS's actual performance. The triangles are KS's discrimination performance (Experiment 1) convolved with performance of an ideal observer in the yes–no membership task (noisy observer model); this reflects expected performance on the yes–no membership task. The circles indicate KS's actual performance. The width of the Gaussian curve fit to the actual yes–no membership data reveals a greater level of mean precision than expected on the basis of KS's discrimination performance alone (narrower fitted Gaussian curve). This suggests that KS's relatively poor face discrimination ability (from Figure 5A) cannot account for KS's actual sensitivity to the mean emotion of a set of faces. (B) Comparison of expected performance for JH on the yes–no membership task (happy–sad stimuli) to JH's actual performance. The triangles represent JH's discrimination performance (see Figure 3A) convolved with performance of an ideal observer in the yes–no membership task. A boxcar function approximates the simulated data better than a Gaussian curve. However, JH's actual yes–no membership data are better captured by a Gaussian curve. Therefore, JH's relatively precise face discrimination (see Figure 5A) cannot account for JH's actual sensitivity to the mean emotion of a set of faces.

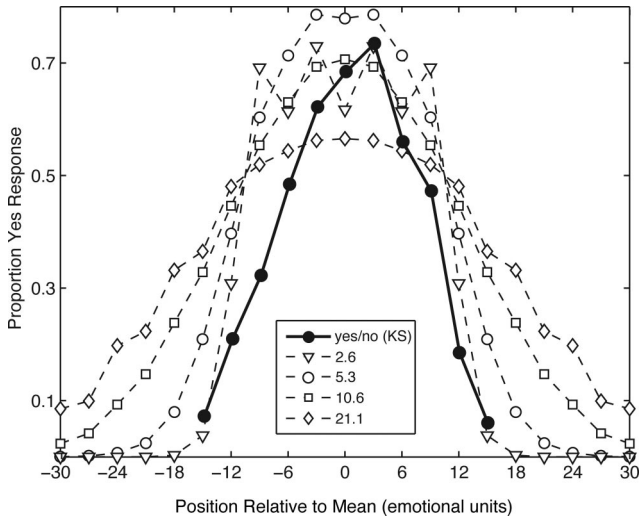

Figure 7. A family of simulated data sets based on various levels of face discrimination ability. Each curve was generated by convolving ideal observer performance with some degree of discrimination performance (starting with KS's discrimination data in Figure 5A). The solid circles represent KS's actual yes–no membership performance. To generate the family of models, noise was increased or decreased by parametrically multiplying the *x*-axis of KS's discrimination data (see Figure 3A) by one of several different gain values. Increasing noise (making the simulated observer a less precise face discriminator) increased curve width and flattened its overall appearance (diamonds), whereas decreasing noise (making the simulated observer a more precise face discriminator) created a curve that looks more like that of an ideal observer (a saw-toothed function; triangles). This simulation demonstrates that KS's actual yes–no membership performance (solid circles) cannot be generated through the direct manipulation of discrimination noise. The legend shows 75% thresholds (cf. Figure 5A) in order of increasing noise; the threshold of 5.3 (open circles) was KS's threshold discrimination from Figure 5A.

processing load. The greater the number of faces in a set (even if there are duplicates), the greater the number of faces one must track to obtain the same level of performance seen for set size of 4. That said, it is clear that there are no systematic differences in the yes–no membership data as a function of set size, consistent with the findings of Ariely (2001). Therefore, within at least the range of 4–16 faces, set size does not appear to influence the ability of observers to extract the mean of the set.

The fact that the membership data (e.g., Figures 3, 6, and 7) followed a Gaussian rather than a saw-toothed distribution suggests that observers had a poor representation of the individual emotions present in the set of faces, and hints at the possibility that individuals are implicitly extracting a mean representation of emotion. This occurs rapidly and flexibly, as observers perceived a different set mean on every trial and could do so even with 16 faces on the screen.

Experiment 2

The previous experiment demonstrated that individuals implicitly perceived a mean emotion in a set of heterogeneous faces. Observers were exposed to the sets for 2,000 ms. To examine the speed and time course of mean emotion perception, we repeated the yes–no membership experiment and manipulated stimulus duration.

Method

Participants. Five individuals (3 women, 2 men; mean age = 20.67 years) affiliated with the University of California, Davis, participated in this experiment. Two of these observers did not participate in the prior experiment. Observer KS viewed both happy–sad morphs and neutral–disgusted morphs, Observer FF viewed only happy–sad morphs, and Observers AC, TH, and JSH viewed only neutral–disgusted morphs. Informed consent was obtained for all volunteers, who were compensated for their time and had normal or corrected-to-normal vision.

Procedure. Observers performed the same task as described in Experiment 1 but were exposed to the sets for 2,000 ms (as before), 500 ms, or 50 ms. We used a block design for stimulus presentation, such that observers were tested at a single duration for the entirety of a run. For this follow-up study, the 5 observers ran three runs (660 trials) at each duration. Set size was evenly divided between 4 and 16 faces, presented in random order.

Results and Discussion

As in Experiment 1A, we examined the proportion of "yes" responses—an indication that the observer thought the test face was a member of the preceding set. To quantify the effect of set duration on mean representation, we fit a Gaussian curve to the probability of "yes" responses for each condition, independently for each observer (Figure 8A shows one representative observer). As described above, curve fitting provides information regarding the precision of observers' mean extraction ability. For example, a narrower curve and higher amplitude reflect greater precision. Thus, we report full width at 75% maximum (as before), as well as curve amplitude, as a function of stimulus duration (see Table 1). Nearly all Gaussian fits were significant (as indicated by the goodness-of-fit statistic, R^2; see Table 1).

The width of the Gaussian curves should approach infinity as set duration approaches zero. Similarly, the amplitude of the curves must approach zero as set duration approaches zero (i.e., response trends become a flat line). On the basis of these limits, we used a power function, $f(x) = (ax^b + c)$, to examine the trends in the parameter estimates as a function of set duration. Figure 8B shows the width and amplitude parameters of the Gaussian curve fits for each observer as a function of set duration. As expected, with decreasing presentation time, there was a significant increase in Gaussian curve width, $F(2, 10) = 4.47$, $p = .04$, and a significant decrease in curve amplitude, $F(2, 10) = 5.33$, $p = .03$. This suggests that the precision of the mean emotion representation depends upon set exposure time.

The data in Figure 8A demonstrate that curve width increases and curve amplitude decreases with decreasing set duration. However, the quality of each Gaussian fit (see Table 1), reflected by the R^2 for each curve, does not substantially decline. The fact that a Gaussian distribution represents these data well at all levels suggests that observers still represented mean facial expression even at 50 ms, albeit more coarsely. A trend of increasing width or decreasing amplitude across set durations simply implies a reduction in the precision of the mean representation with decreasing set duration, not a complete lack of mean representation.

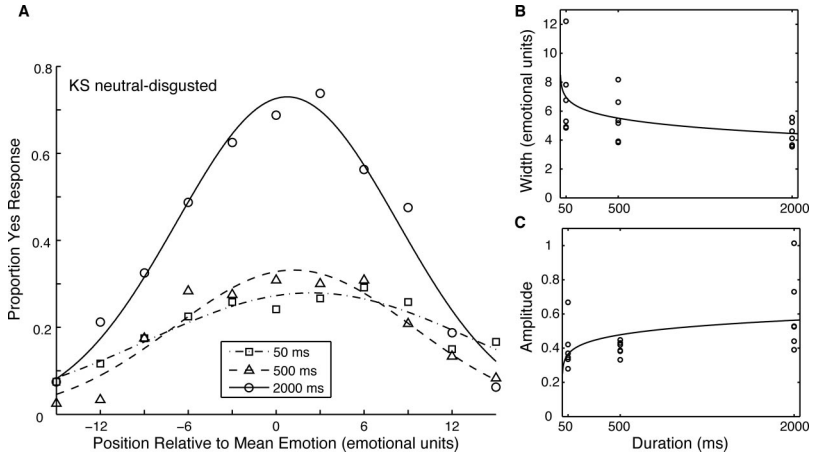

Figure 8. Results from Experiment 2 (duration manipulation). (A) Gaussian fits for one representative observer at three durations. The quality of the fit was comparable for each set duration (see Table 1), although there is an effect of set duration (lower amplitude and greater curve width at shorter durations). (B–C) The general trend of curve parameters width and amplitude as a function of duration. Each circle represents one observer's parameter estimates. A power function best captures these data. (B) As set duration decreased, curve width increased. For set durations of 50 ms, observers had a coarser representation of the mean than at 2,000 ms. (C) Similar results were seen for amplitude. As set duration decreased, curve amplitude decreased.

To illustrate this principle more concretely, we examined the one case in which an observer was unable to represent the mean emotion at 50 ms set exposure. Unlike that of other observers, a Gaussian curve does not adequately capture Observer JSH's performance at 50 ms (see Table 1). There may be an alternative function that better represents his pattern of performance. If JSH was simply unable to extract anything meaningful from the set in such a short amount of time, a linear function (i.e., a flat line) would fit JSH's data better than a Gaussian curve. This would suggest that JSH's responses were essentially random, no longer dependent upon the distance of the test face from the set mean. We used Akaike's information criterion (AIC) method for comparing the likelihood of two models: Gaussian versus linear. In this method of model comparison, a lower AIC indicates a better model fit (Motulsky & Christopoulos, 2003). The AIC value for one model by itself, however, is meaningless without the AIC value for another comparison model. The difference in AIC between two models is computed, and from this an information ratio (IR) is derived, which reflects the likelihood that one of the two models is correct. For Observer JSH at 50 ms, a flat line was more likely the correct model (Difference in AIC = 14.37, IR = 26.23; the linear fit was 26.23 times more likely to be correct). This suggests that at 50 ms, JSH was unable to derive a mean from the set of faces. However, this was not the case for the other 5 observers at the same short duration, whose AICs indicated that a Gaussian curve was likely the better fitting model (KS, happy–sad: Difference in AIC = 15.68, IR = 2536.04; FF, happy–sad: Difference in AIC = 17.38, IR = 5952.20; KS, neutral–disgusted: Difference in AIC = 14.37, IR = 1318.46; AC, neutral–disgusted: Difference in AIC = 5.46, IR = 15.33; TH, neutral–disgusted: Difference in AIC = 14.30, IR = 1273.19). Therefore, all subjects but JSH were able to extract a mean representation even at 50 ms, albeit a coarser representation than in the 2,000-ms condition.

Although a Gaussian distribution fits the data better than a flat line for 5 out of 6 observers, this does not rule out the possibility there might be another function better fitting than a Gaussian curve—perhaps one that does not exhibit a peak at the center of the distribution. Examination of the curves from this experiment, as well as those from Experiment 1A, reveals a precipitous drop in the probability of responding "yes" when the test face falls beyond the emotional range of the set. Is it possible that observers possessed an implicit knowledge of the *range* of the set rather than the mean? If this were the case, a boxcar function should fit these data better than a Gaussian distribution; the probability of "yes" responses would be minimized and equal for test faces beyond the emotional range of the set, and the probability of "yes" responses would be maximized and equal for those test faces within the emotional range of the set (see Figure 9). To test this alternative, we computed the IR by comparing the fit of a boxcar function to the fit of a Gaussian distribution for each observer at the 50-ms set duration. For 5 out of 6 observers, the Gaussian distribution was more likely the better fit than the boxcar function (KS, happy–sad: Difference in AIC = 5.22, IR = 13.62; FF, happy–sad: Difference in AIC = 10.17, IR = 161.81; KS, neutral–disgusted: Difference in AIC = 4.82, IR = 11.14; AC, neutral–disgusted: Difference in AIC = 1.49, IR = 2.11; TH, neutral–disgusted: Difference in AIC = 13.06, IR = 686.48). These results indicate that observers were extracting more than just the emotional range of the set of faces.

We may conclude from these results that individuals perceive a mean emotion (implicitly) across multiple set durations, although this representation becomes noisier as set duration decreases. The fact that the width of the fitted Gaussian curves increased (and amplitude decreased) as set duration decreased implies only a reduction in the overall precision of mean emotion representation, and this is expected (i.e., one cannot represent anything when the set duration is 0). Therefore, observers still extract a coarse representation of the mean in a short amount of time.

Table 1
Curve Parameters for Observers at Three Set Durations

Observer	R^2	Width	Amplitude
	50 ms		
Happy–sad			
KS	.90**	4.90	0.37
FF	.91**	4.85	0.67
Neutral–disgusted			
KS	.88**	5.29	0.28
AC	.74**	6.76	0.42
JSH	.22	12.21	0.35
TH	.88**	7.82	0.34
	500 ms		
Happy–sad			
KS	.92**	3.91	0.38
FF	.96**	5.17	0.45
Neutral–disgusted			
KS	.90**	3.84	0.33
AC	.54*	8.18	0.43
JSH	.82**	6.62	0.42
TH	.87**	5.35	0.39
	2,000 ms		
Happy–sad			
KS	.97**	3.66	0.39
FF	.98**	4.62	1.01
Neutral–disgusted			
KS	.90**	3.55	0.73
AC	.70**	5.56	0.44
JSH	.87**	5.23	0.53
TH	.92**	4.12	0.53

* $p < .05$. ** $p < .005$.

Experiment 3

In the previous experiments, observers implicitly perceived the mean emotion of a set of heterogeneous faces even though they were instructed to attend to the individual constituents. We were able to measure the precision of mean face representation by fitting a Gaussian curve to the data at multiple stimulus durations. However, that method does not provide specific thresholds for mean extraction ability. In the following experiment, we explicitly asked observers whether a test face was happier or sadder (or more neutral or more disgusted, depending on task condition) than the mean emotion of the preceding set, thus deriving a concrete assessment of observer precision. This replicates and extends the work of Haberman and Whitney (2007).

Method

Participants. All observers had participated in some or all of the previously described experiments. Four observers viewed happy–sad morphs, 3 observers viewed neutral–disgusted morphs, and 1 observer viewed both (in separate runs).

Procedure. We presented sets of 4 and 16 faces. The sequence of events was nearly identical to that of Experiment 2, only the task instructions changed. The range of potential test faces was ±5, ±4, ±2, or ±1 unit happier or sadder (or ±10, ±8, ±4, or ±2 units for neutral–disgusted morphs) than the mean. There were 208 trials per run, with equal number of trials at each of the two set sizes. Each observer performed at least three runs for a minimum total of 624 trials (78 presentations of each of the eight possible test face separations). Some observers performed four runs. The mean emotion of the set of faces was randomly selected on every trial.

For each observer, separate logistic functions were fit for observers' homogeneous discrimination data (Experiment 1B) and their mean discrimination data. These two curves were then subjected to Monte Carlo simulations with the Psignifit toolbox (Wichmann & Hill, 2001b), associated with MATLAB, to test the null hypothesis that both curves came from the same underlying distribution. Psignifit created a density plot containing 5,000 simulated values corresponding to slope and threshold differences between our curves of interest. Significance regions were derived based on this plot, and a p value calculated reflecting how aberrant the actual observed difference was relative to the simulated differences.

In a method of constant stimuli 2AFC task, observers indicated with a button press whether the test face was happier or sadder (or more neutral or more disgusted) than the mean emotion of the previously presented set (see Figure 10).

Results and Discussion

Observers were explicitly asked to indicate whether a test face was happier or sadder (or more neutral or more disgusted) than the mean emotion of the previously displayed set of faces. Judgments were made for set sizes of 4 and 16. As Figure 11 indicates,

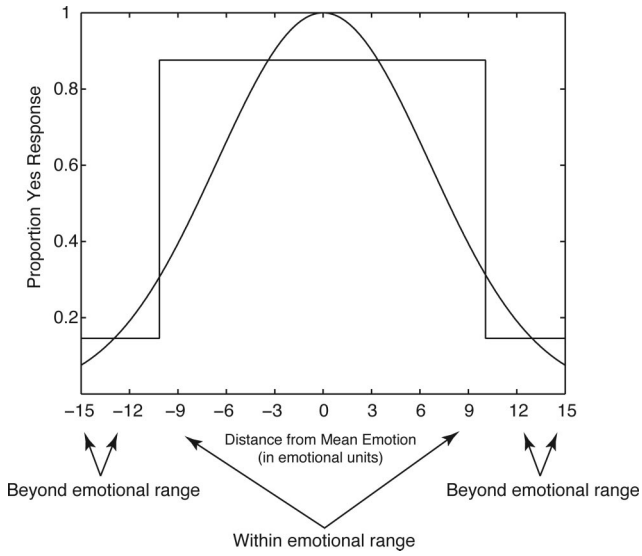

Figure 9. Depiction of two possible models describing the membership data for 50-ms duration (from Figure 8). The Gaussian distribution depicts the hypothesis that observers' "yes" and "no" responses are dependent upon the proximity of the test face to the mean emotion of the set. The boxcar alternative suggests that the range of the set influences "yes" and "no" responses, where observers are most likely to reject a test face as a set member when it falls beyond the emotional range of the set and most likely to accept a test face as a set member when it falls within the emotional range of the set. The boxcar was the better fitting model for only 1 observer (JSH).

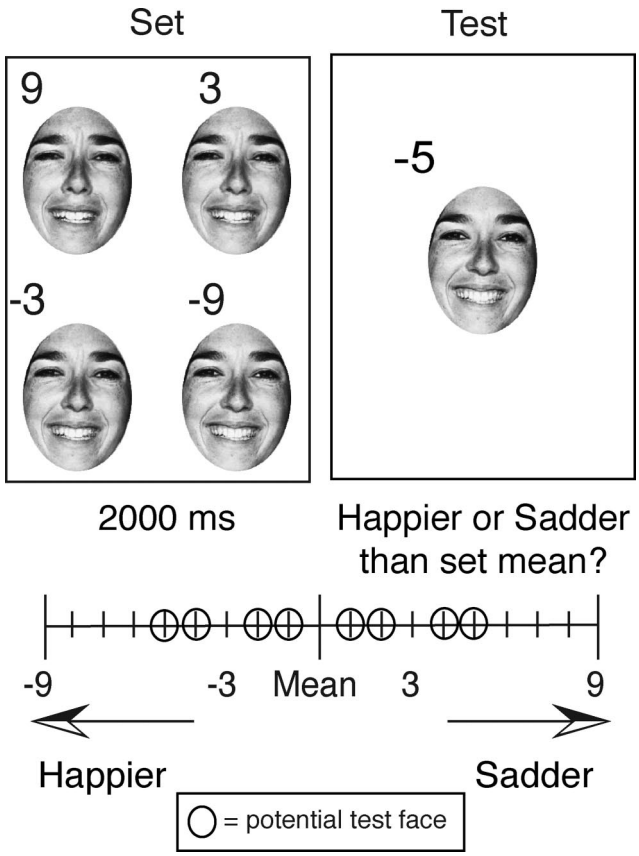

Figure 10. Task design for Experiment 3. Observers had to indicate whether the test face was happier or sadder (or more neutral or more disgusted) than the mean emotion of set. The test face could be any of the distances indicated by the circles (numbers were not seen by participants).

observers were remarkably precise in their assessment of the mean emotion of a set (i.e., mean discrimination), performing just as well as when they were asked to discriminate between two faces (homogeneous discrimination, Experiment 1B). Just how good are individuals at representing the mean emotion of a set? The Monte Carlo simulations revealed that for 6 out of 7 observers, thresholds between mean discrimination and regular discrimination did not differ, suggesting that the two psychometric functions conceivably came from the same underlying distribution. Observers were equally good at representing a mean emotion from a set of heterogeneous faces as they were at indicating which of two faces was happier. This is particularly striking, especially when one considers that Experiment 1A indicated that observers were unable to explicitly represent the individual set members. It appears that the visual system favors a summary representation of a set of faces over a representation of each set constituent.

Experiment 4A

Experiment 3 demonstrated that observers were able to extract a precise representation of the mean emotion of a set of faces. Further, Experiment 1A suggested that observers disregarded the individual set members in favor of this mean representation. In the following experiments we tested to what extent observers represented the individual members of the set.

Method

Participants. Three observers (2 women, 1 man; mean age = 23.33 years) affiliated with the University of California, Davis, participated in this experiment.

Procedure. Observers were instructed to identify the location in which a test face had appeared within the previously displayed set. Set size varied from one to four items. The sets of faces were created in exactly the same way as in Experiment 1A, which ensured a minimum separation of six emotional units among set members (a separation above discrimination threshold). Each face was randomly assigned to one of four locations on the screen. Set size of four looked as it did in Figure 10, whereas smaller set sizes had one or more gaps. Here we used neutral–disgusted morphs, although we have reported a different version of this task using happy–sad morphs (Haberman & Whitney, 2007). Sets were presented for 2 s, followed by a single test face in the center of the screen that remained until a response was received. The test face was surrounded by four letters (*A, B, C*, and *D*) that corresponded to the possible locations of the faces in the previous set. Observers had to indicate where in the set the test face had appeared. Within a given run there were 160 trials, and observers performed three runs for a total of 480 trials.

Results and Discussion

As expected, location identification declined as a function of set size (see Figure 12A). For set size of 4 items, the group average was only 50% correct. Observers derived some information from the set, but how much? For the purpose of comparison, we estimated expected performance when a hypothetical observer explicitly remembered only one face from the set (solid line in Figure 12). Two of the observers (JH and PL) were at or below this level of performance, suggesting that they could remember only one face (or less) from the sets, and this is consistent across all set sizes (see Figure 12A). Observer AD performed at a level of accuracy that would be expected if AD were able to remember between one and two faces. This amount of information cannot explain the level of precision on mean discrimination (see Figure 11). Despite explicitly remembering only one of the faces in the set, observers were still able to precisely represent the mean emotion of an array of faces containing up to 16 items.

Experiment 4B

In the previous task, observers were asked to identify the location of a face within the set. It is possible that observers had a high-fidelity representation of each set member but simply lost its corresponding location information and therefore performed poorly on the task. To test this, observers performed a 2AFC judgment to identify which of two test faces was a member of the previously viewed set.

Method

Participants. The same 3 observers from the previous experiment participated.

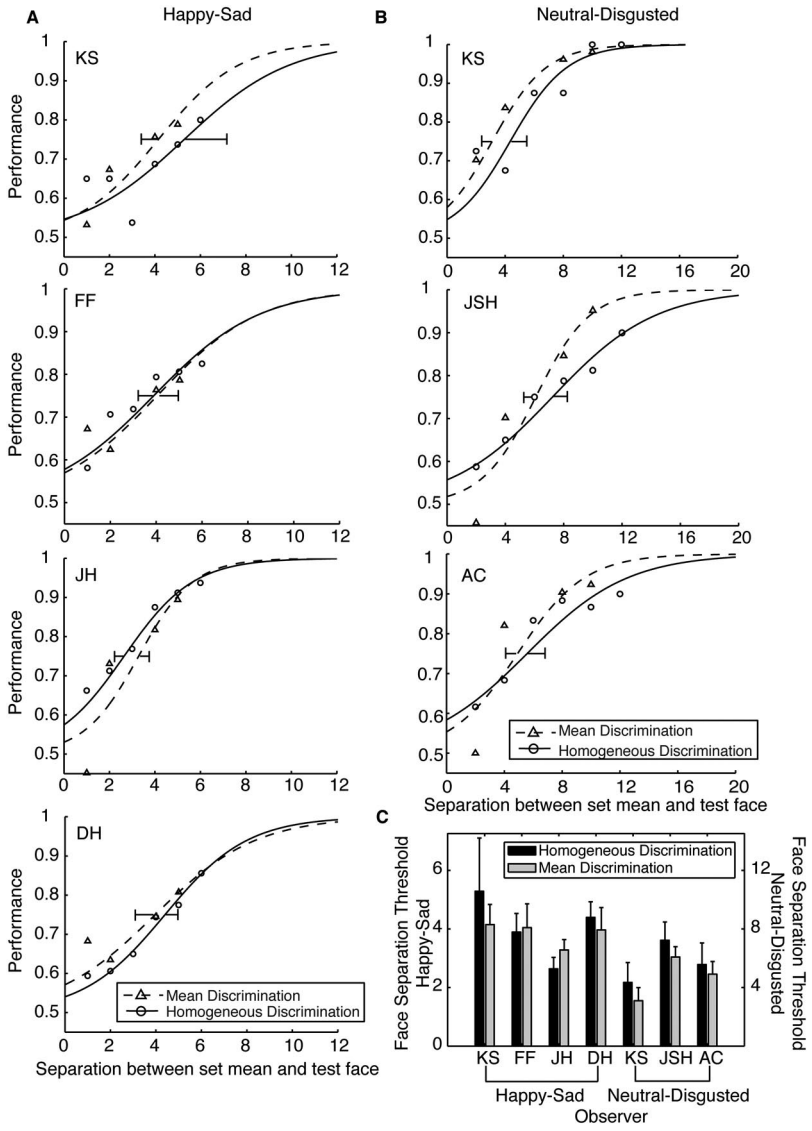

Figure 11. Results for Experiment 3. Figures 11A and 11B show individual psychometric functions for mean discrimination for happy–sad morphs and neutral–disgusted morphs, respectively. Superimposed on each graph is that observer's discrimination performance (from Figure 5). Mean discrimination performance was as precise as regular discrimination performance for nearly all observers. Figure 11C depicts the 75% thresholds for each observer. Error bars in Figures 11A–11C are 95% confidence intervals derived from 5,000 bootstrap simulations (Wichmann & Hill, 2001a, 2001b). For fitting purposes, we included a point at 0 separation between set and test (chance performance), which does not appear in the graph.

Procedure. We used an unbiased, 2AFC paradigm to examine how well observers represented the set members. We varied set size from one to three faces and asked participants to identify which of two subsequently presented test faces had appeared in the set. The sets of faces were the same as in Experiment 4A (with the exception that the largest set size was three faces, which kept the minimum separation between any test and/or set faces at six or more units). Sets were presented for 2 s (as before), followed by two simultaneously presented test faces, one of which was a target. The target was randomly selected from among the set items. The lure was at least six units away from the test face and any other member of the set (often this separation was larger). The position (top or bottom) of the target face was randomized on every trial. Observers were instructed to indicate with a key press which of the two test faces was a member of the preceding set. Within a given run there were 160 trials, and observers performed three runs for a total of 480 trials.

Results and Discussion

Observers performed at a high level for set size of one (see Figure 12B), as expected. Performance was not at 100%, however,

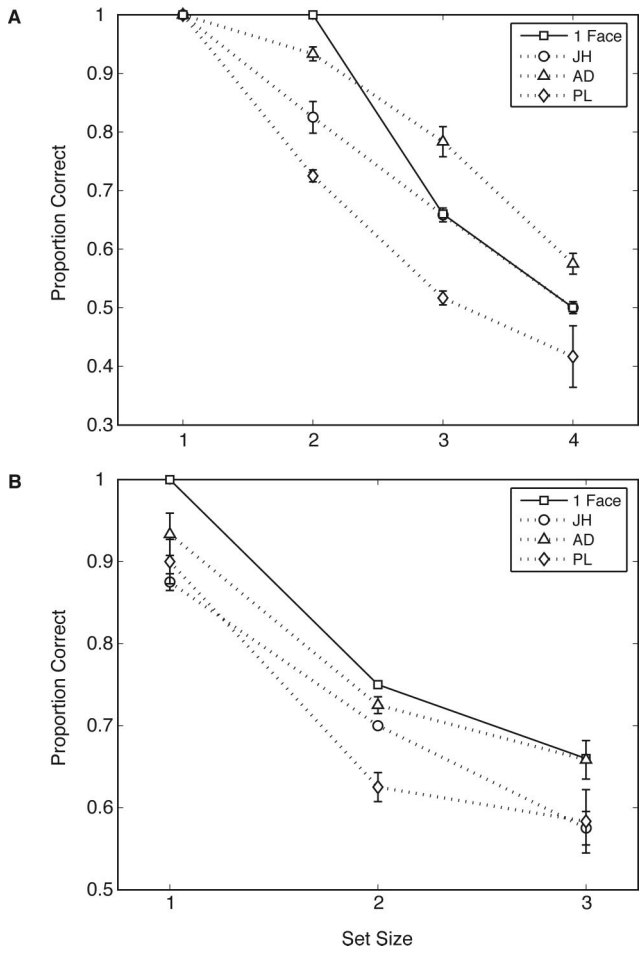

Figure 12. Results for Experiments 4A and 4B. The solid line represents expected performance if observers could remember only one face from each set. Actual performance, represented by the dotted lines, was at or below this solid line for all but 1 of the observers (AD at set size of three and four). Error bars are plus or minus one standard error of the mean.

reflecting limitations in discrimination ability. Importantly, performance declined with the introduction of just one additional face (see Figure 12B), and this trend continued through a set size of three faces. The solid line in Figure 12B indicates expected performance if subjects were only able to code one face from the set. As in Experiment 4A, observers remembered only one face (or less) from the set. Explicitly remembering one face from the set cannot explain the level of precision observed for mean discrimination (see Figure 11). This suggests that individuals lacked high-fidelity representations of the set members and not simply the locations of the members. Despite this, observers had enough coarse information about the set members to derive a precise estimate of the mean emotion (Experiment 3).

When set size was greater than just one item, participants were unable to code and retain explicit information about individual identities of the set members. Numerous studies on visual working memory have demonstrated the striking limitations in attentional capacity (Luck & Vogel, 1997; Simons & Levin, 1998). Taken in conjunction with research suggesting that searching for faces within an array is a slow and difficult process (Brown, Huey, & Findlay, 1997; Kuehn & Jolicœur, 1994; Nothdurft, 1993), the poor performance seen in this experiment may not be entirely surprising. What is surprising, however, is that despite poor performance on set membership identification, there was still precise mean discrimination of average facial expression (Experiment 3).

Experiments 5 and 6

It is widely accepted that whole upright faces, such as those used in the experiments above, are processed in a configural or holistic manner (Farah, Wilson, Drain, & Tanaka, 1998; Kanwisher, McDermott, & Chun, 1997). Inverted and scrambled faces, in contrast, are processed in a more part-based manner (Maurer, Le Grand, & Mondloch, 2002; Moscovitch, Winocur, & Behrmann, 1997; Robbins & McKone, 2003). If the summary statistical representation found above for sets of faces is specific to faces, and not the low-level features within the faces, then we should find more precise mean extraction for whole upright faces than for inverted or scrambled faces. The purpose of Experiments 5 and 6 was to test this.

Method

Participants. Three observers (KS, TH, JH) from previous experiments participated.

Stimuli. The same happy–sad and neutral–disgusted morphs from prior experiments were used here, except that they were either inverted (Experiment 5; see Figure 13A) or scrambled (Experiment 6; see Figure 13B). For scrambling, we used MATLAB to divide each of the original morphs into a 4 × 3 matrix of randomly arranged squares. The same scrambling algorithm was applied to each morph, such that the face pieces were rearranged in the same order across faces.

Procedure. With the exception of the stimuli, the procedures for Experiments 5 and 6 were identical to those used in the mean discrimination experiment (Experiment 3). Set duration was fixed at 500 ms. The recognition of inverted and scrambled faces is thought to require feature-based strategies that differ from the configural strategy associated with upright face processing (Farah et al., 1998; Kanwisher et al., 1997; Maurer et al., 2002; Mosco-

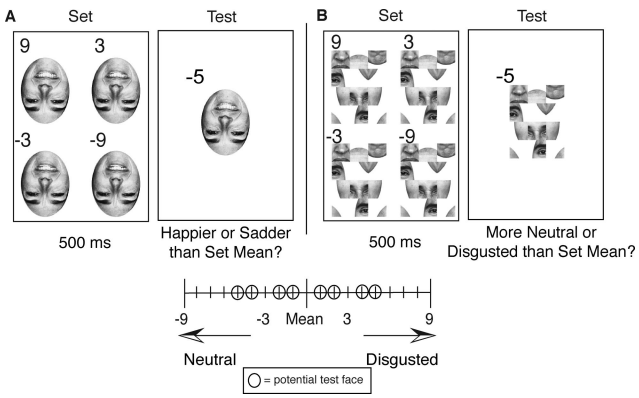

Figure 13. Task design and stimuli for Experiments 5 and 6. Procedure was identical to that of Experiment 3. Observers saw either (A) inverted or (B) scrambled faces.

vitch et al., 1997; Robbins & McKone, 2003). If the mean face extraction were based on configural or holistic facial information, then we would expect that observers' mean discrimination of inverted and scrambled faces would be poorer than for upright faces.

Results and Discussion

Figure 14 shows fitted psychometric curves for each observer on mean discrimination tasks when viewing inverted or scrambled stimuli compared with upright stimuli. Mean discrimination performance for both inverted and scrambled faces suffered relative to upright faces. The inverted and scrambled faces had the same feature information available that upright faces had, and yet 5 of the 6 observers were significantly worse at extracting the mean emotion. On the basis of 5,000 Monte Carlo simulations run with Psignifit (described in Experiment 3), the only nonsignificant result had $p = .07$ (Observer KS, upright vs. inverted; see Figure 14A). For all other observers, the difference between the curves was significant ($p < .05$). This supports the conclusion that the perception of mean facial expression in sets of faces is distinct from low-level feature averaging. The dissociation in mean extraction performance (see Figure 14) suggests that the information used to perceive mean upright facial expression was not available when the faces were inverted or scrambled. Therefore, perceiving sets of upright faces relies on a distinct and more precise facial representation. Evidently, ensemble coding occurs not only for low-level features (dots and gratings) but also for high-level, complex objects.

General Discussion

We have demonstrated that observers quickly recognize the mean expression of a set of faces with remarkable precision, despite lacking a representation of the set constituents. In fact, observers were able to discriminate the mean emotion of a set at least as well as they were able to discriminate the expression of two faces (see Figure 11). The results cannot be explained simply by observers' perceptual or decision noise. Further, the summary statistic was not simply the range of the set but approximated the set mean. Feature-based processing strategies cannot account for our findings, as observers were able to extract the mean emotion of a set of upright faces with significantly more precision than they were able to extract the mean emotion from a set of inverted or fractured faces.

Ensemble Coding and Scene Perception

Statistical representation for low-level features makes sense—an array of dots or a collection of gratings naturally combines to create a single texture. Some speculate that statistical set representation may serve, then, to promote texture perception (Cavanagh, 2001). However, we show statistical representation for face-specific processing, a level of processing well beyond that required for dots, bars, or gratings. Although it is conceivable and even probable that statistical set representation plays some role in texture perception, this cannot be its only function given that it is operating on high-level, complex objects. It is likely that set representation serves a more general role in deriving gist from a complex scene. This point is particularly compelling when one considers the speed with which individuals perceived the statistical representation—noisy mean extraction occurred in as little as 50 ms for sets of up to 16 faces. Previous studies that have reported gist perception in extremely brief displays (Biederman, Glass, & Stacy, 1973; Navon, 1977; Potter, 1976) may have tapped the set representation mechanism found here. We speculate that the impression we get of a complete and wholly accurate representation of the visual world is not actually a "grand illusion" (Noe, Pessoa, & Thompson, 2000). Rather, a great deal of condensed information arrives in the form of summary statistics. This information, though not necessarily high fidelity, is useful and may drive the impression that we "see" everything in our environment.

Groups of Faces Are Special

The fact that we perceive the mean emotion in a set of faces may not seem intuitive at first. Although deriving an average texture directly benefits surface perception, it is not clear whether similar mechanisms are at work for average emotion perception. For example, in our effort to derive a mean, do we perceive a texture of faces the same way we perceive a texture of Gabor patches? Despite this quandary, high-level ensemble coding makes sense from an evolutionary perspective. A rapid and precise assessment of a crowd of faces is useful for determining the intent of the mob. Summary statistical face processing may therefore be a unique phenomenon, at a unique level of processing. Taken with the body of work showing ensemble coding for low-level objects such as dots (Ariely, 2001; Chong & Treisman, 2003) and gratings (Parkes et al., 2001), we can conclude that some form of averaging occurs across multiple visual domains at different levels of analysis. Unlike perceiving average size, orientation, and motion, perceiving average facial expression is not mediated by low-level features, luminance cues, or other nonconfigural cues (Experiments 5 and 6). Further, the sensitivity to average facial expression is remarkably precise: Subjects were able to discriminate mean facial expression at least as well as they were able to discriminate an individual face. This degree of sensitivity is not found for average size (Ariely, 2001), orientation (Parkes et al., 2001), or other low-level features.

Parallel or Serial

Of significant interest is the exploration of whether mean extraction is a parallel or serial process. Do observers automatically extract the mean from large arrays of items, the same way they do for global motion, global orientation, and some other texture segmentation tasks (Landy & Graham, 2004; Movshon & Newsome, 1996; Newsome & Pare, 1988; Parkes et al., 2001; Regan, 2000; Watamaniuk & Duchon, 1992; Williams & Sekuler, 1984)? This remains an ongoing debate, and currently there is support for both sides. Chong and Treisman (2005) suggested that mean extraction is automatic and parallel, because neither the number of items nor cuing seemed to affect the accuracy of mean size discrimination. Alvarez and Oliva (2008) reached a similar conclusion, showing that summary statistical representation for the location of a group of objects (average or centroid position) occurred even when observers did not attend to the objects. Our

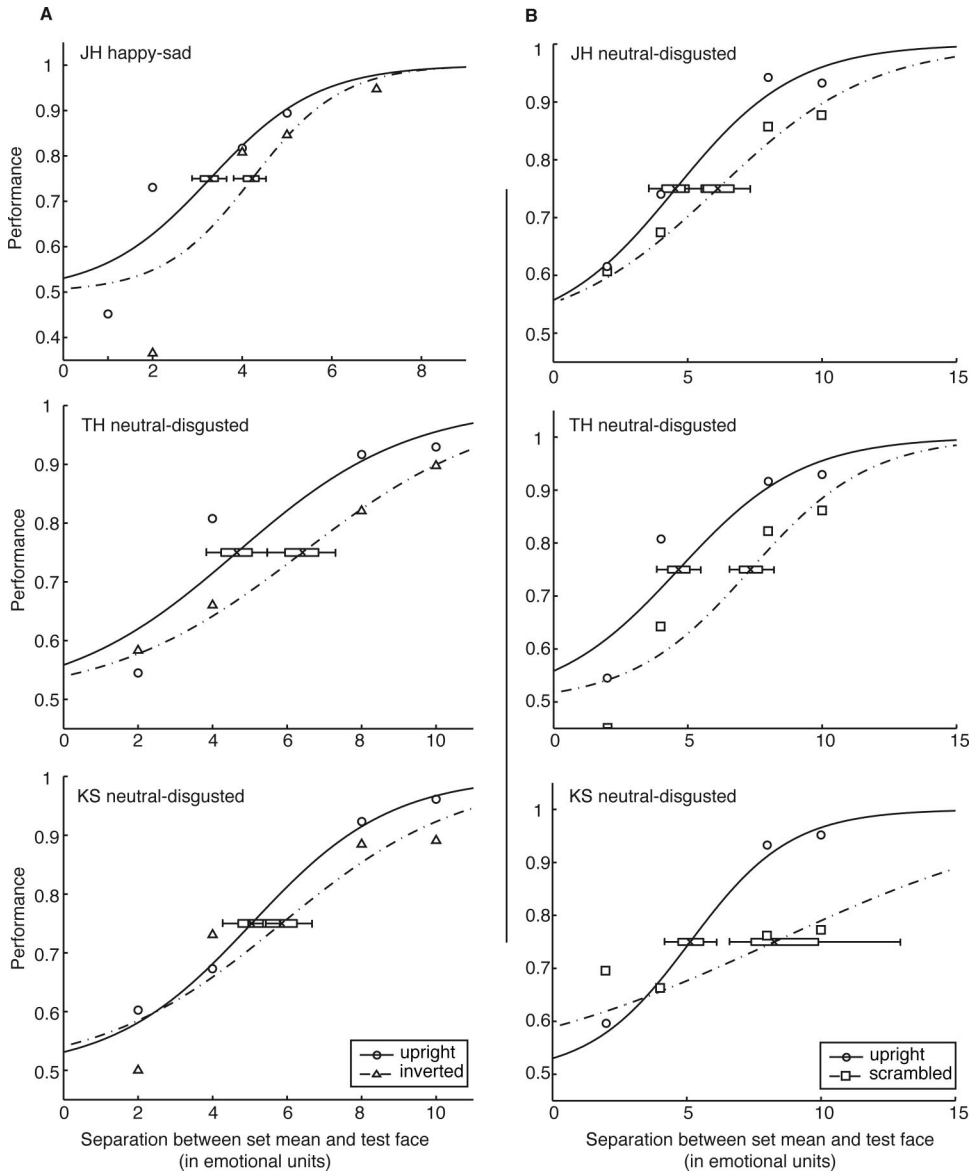

Figure 14. Psychometric curves for (A) upright mean discrimination versus inverted mean discrimination and (B) upright mean discrimination versus scrambled mean discrimination. Upright mean discrimination was significantly better ($p < .05$) than either inverted or scrambled mean discrimination for 5 out of 6 observers. Error bars are 95% confidence intervals derived from 5,000 bootstrap simulations (Wichmann & Hill, 2001a, 2001b). For fitting purposes, we included a point at 0 separation between set and test (chance performance), which does not appear in the graph.

data from Experiments 1 and 2 suggest that ensemble coding of facial expression occurs implicitly, as observers unknowingly possessed knowledge of the mean while disregarding the task instructions to attend to the set members. However, demonstrating an implicit representation is not the same as demonstrating automaticity. Even though observers derived a mean emotion in brief stimulus displays, they showed a reduction in precision as set exposure decreased. Therefore, we cannot make a strong claim regarding the automaticity of ensemble coding, only that it can occur implicitly.

In response to the suggestion that there is average size perception (Ariely, 2001; Chong & Treisman, 2003), Myczek and Simons (2008) have demonstrated that sparse sampling of set items is sufficient for accurate discrimination of average object size. Their model tested the precision with which observers could represent average size, assuming they track some number of items in the set. In other words, the authors investigated whether existing models of directed attention could explain ensemble coding performance, eliminating the need for a separate averaging module that operates in parallel. Their

model is theoretically capable of explaining much of the existing set representation data, at least when the stimuli are dots (Ariely, 2001; Chong & Treisman, 2003). However, our data depart from the average size data in one key respect: Compared with baseline discrimination performance in the two respective tasks, mean discrimination of emotion is more precise than mean discrimination of size. Whereas average size perception is worse than discrimination of two dots, average emotion perception is at least as good as discrimination of two faces (homogeneous discrimination, Experiment 1B). This distinction would boost the number of items necessary for Myczek and Simons's model to achieve behavioral levels of performance in our task.

Although Myczek and Simons's (2008) model supports a directed attention strategy for ensemble coding, it does not preclude the existence of a parallel mechanism as well. For example, a directed attentional mechanism cannot explain the results of Parkes et al. (2001), where crowded central targets (i.e., the targets were impossible to individuate because of the surrounding flankers) influenced the perceived average orientation of the entire set of Gabor patches. Because the crowded Gabor patches could not be individuated, attention could not be directed to those single items. This indicates that attention may not be necessary to derive the mean orientation (i.e., a parallel process may be at work). Thus, the jury is still out on whether ensemble coding is primarily subserved by serial or parallel processes. However, both mechanisms are capable of achieving the same end—a summary statistical representation.

Conclusions

The experiments here demonstrate the existence of summary statistical representations for groups of high-level objects: Observers perceived the mean facial expression in a group of heterogeneous faces. This reveals an efficient ensemble coding mechanism that processes and represents large crowds but is distinct from the mechanism responsible for low-level ensemble coding. The results further demonstrate that ensemble coding operates at multiple levels in the visual system.

References

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392–398.

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

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology, 97*(1), 22–27.

Brown, V., Huey, D., & Findlay, J. M. (1997). Face detection in peripheral vision: Do faces pop out? *Perception, 26*(12), 1555–1570.

Cavanagh, P. (2001). Seeing the forest but not the trees. *Nature Neuroscience, 4*(7), 673–674.

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

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect.* Palo Alto, CA: Consulting Psychologists Press.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105*(3), 482–498.

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

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311.

Kuehn, S. M., & Jolicœur, P. (1994). Impact of quality of the image, orientation, and similarity of the stimuli on visual search for faces. *Perception, 23*(1), 95–122.

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

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

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

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255–260.

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience, 9*(5), 555–604.

Motulsky, H. J., & Christopoulos, A. (2003). *Fitting models to biological data using linear and nonlinear regression: A practical guide to curve fitting.* San Diego, CA: GraphPad Software.

Movshon, J. A., & Newsome, W. T. (1996). Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *Journal of Neuroscience, 16*(23), 7733–7741.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70*(5), 772–788.

Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. In S. M. Kosslyn & D. N. Osherson (Eds.), *An invitation to cognitive science* (2nd ed., Vol. 2, pp. 1–70). Cambridge, MA: MIT Press.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353–383.

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience, 8*(6), 2201–2211.

Noe, A., Pessoa, L., & Thompson, E. (2000). Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition, 7*(1–3), 93–106.

Nothdurft, H. C. (1993). Faces and facial expressions do not pop out. *Perception, 22*(11), 1287–1298.

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.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 509–522.

Regan, D. (2000). *Human perception of objects: Early visual processing of spatial form defined by luminance, color, texture, motion, and binocular disparity.* Sunderland, MA: Sinauer.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*(5), 368–373.

Robbins, R., & McKone, E. (2003). Can holistic processing be learned for inverted faces? *Cognition, 88*(1), 79–107.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology, 38*(2), 259–290.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5*(4), 644–649.

Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research, 32*(5), 931–941.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63*(8), 1293–1313.

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63*(8), 1314–1329.

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *Vision Research, 24*(1), 55–62.

Received December 14, 2007
Revision received June 21, 2008
Accepted August 13, 2008 ■