Foveal input is not required for perception of crowd facial expression

Benjamin A. Wolfe University of California, Berkeley, Department of Psychology, Berkeley, CA, USA

Anna A. Kosovicheva University of California, Berkeley, Department of Psychology, Berkeley, CA, USA

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

Katherine Wood University of California, Berkeley, Department of Psychology, Berkeley, CA, USA

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

The visual system extracts average features from groups of objects (Ariely, 2001; Dakin & Watt, 1997; Watamaniuk & Sekuler, 1992), including high-level stimuli such as faces (Haberman & Whitney, 2007, 2009). This phenomenon, known as ensemble perception, implies a covert process, which would not require fixation of individual stimulus elements. However, some evidence suggests that ensemble perception may instead be a process of averaging foveal input across sequential fixations (Ji, Chen, & Fu, 2013; Jung, Bulthoff, Thornton, Lee, & Armann, 2013). To test directly whether foveating objects is necessary, we measured observers' sensitivity to average facial emotion in the absence of foveal input. Subjects viewed arrays of 24 faces, either in the presence or absence of a gaze-contingent foveal occluder, and adjusted a test face to match the average expression of the array. We found no difference in accuracy between the occluded and non-occluded conditions, demonstrating that foveal input is not required for ensemble perception. Unsurprisingly, without foveal input, subjects spent significantly less time directly fixating faces, but this did not translate into any difference in sensitivity to ensemble expression. Next, we varied the number of faces visible from the set to test whether subjects average multiple faces from the crowd. In both conditions, subjects' performance improved as more faces were presented, indicating that subjects integrated information from multiple faces in the display regardless of whether they had access to foveal information. Our results demonstrate that ensemble perception can be a covert process, not requiring access to direct foveal information.

Introduction

Our visual world is composed of complex information that is continually changing from moment to moment. Any given scene contains a wealth of visual information—pebbles on a beach, leaves on a tree, faces in a crowded room—yet limitations on our attention and short-term memory prevent us from processing every detail (Duncan, Ward, & Shapiro, 1994; Luck & Vogel, 1997; Myczek & Simons, 2008). One way in which the visual system is able to efficiently process this information is by extracting summary statistics (e.g., the average) of a given stimulus feature across an array of objects through a process known as ensemble perception (for reviews, see Alvarez, 2011; Fischer & Whitney, 2011; Haberman, Harp, & Whitney, 2009; Haberman & Whitney, 2011). A large body of evidence has shown that the visual system can rapidly extract the mean of stimulus features such as orientation (Ariely, 2001; Dakin & Watt, 1997; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), size (Ariely, 2001; Carpenter, 1988; Chong & Treisman, 2003), and motion direction (Watamaniuk & Sekuler, 1992). In recent years, further research on the topic has

Citation: Wolfe, B. A., Kosovicheva, A. A., Yamanashi Leib, A., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. *Journal of Vision, 15*(4):11, 1–13. doi:10.1167/15.4.11.

doi: 10.1167/15.4.11 Received January 7, 2015; published September 11, 2015 ISSN 1534-7362

demonstrated that observers can perceive the mean features from complex objects, such as crowd heading from point-light walkers (Sweeny, Haroz, & Whitney, 2013), emotions from sets of faces (Haberman et al., 2009; Haberman & Whitney, 2007; Ji et al., 2013; Ji, Chen, & Fu, 2014; Jung et al., 2013; Yang, Yoon, Chong, & Oh, 2013), facial identity (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007; Yamanashi Leib et al., 2014; Yamanashi Leib et al., 2012), crowd gaze direction (Cornelissen, Peters, & Palmer, 2002; Sweeny & Whitney, 2014), and auditory tone (Piazza, Sweeny, Wessel, Silver, & Whitney, 2013). However, it remains a debated question whether ensemble perception of high-level visual stimuli, such as faces, can be accomplished covertly or if it requires overt, sequential foveation of objects before an ensemble representation can be extracted.

Ensemble perception could result from a covert process in which coarse but sufficient information is gathered from the periphery to generate an ensemble percept (Fischer & Whitney, 2011; Haberman et al., 2009; Sherman, Evans, & Wolfe, 2012). Consistent with this, ensemble perception of simple features (e.g., size, orientation) has been shown with a range of brief stimulus durations (from 50 to 500 ms; cf. Ariely, 2001; Dakin & Watt, 1997; Parkes et al., 2001), providing some support for the covert account, as the stimulus durations are often shorter than the time required to plan a saccade (approximately 200 ms; Carpenter, 1988). In addition, more recent evidence demonstrates ensemble perception of more complex features with brief stimulus durations. For instance, Sweeny and colleagues (2013) demonstrated that observers can extract the average heading from groups of point-light walkers with durations as short as 200 ms, and Yang and colleagues (2013) showed ensemble processing of emotional faces with a 100 ms stimulus duration. However, in the absence of sufficient time to make an eye movement, it is impossible to determine the contribution or potential necessity of foveal input to ensemble perception.

Conversely, ensemble perception might rely on an overt process, where sequentially fixated stimuli are averaged. Indeed, recent experiments suggest a dominant, if not necessary, role for foveally presented faces when extracting ensemble expression or identity (Ji et al., 2013, 2014; Jung et al., 2013). Simply put, it has been suggested that ensemble perception of expression or identity requires sequential foveation of each face in the set, and that peripheral or global information is neither required nor used. The drawback of all of the studies above—whether they support overt or covert ensemble representations—is that they are indirect tests. The most direct test for the necessity of foveal input when extracting ensemble crowd expression is to simply block the fovea in a gaze contingent manner. If, in fact, ensemble perception of expression is dependent on individual foveation of faces within the group, subjects' ensemble percept should be less accurate without foveal information.

To test this, we performed such a series of experiments in which subjects were asked to report the average emotion of a group of faces without foveal input. Using high-speed eye tracking and gaze-contingent stimulus control, we occluded the central 2.6° of the visual field, entirely blocking foveal input. Subjects performed an ensemble perception task in which they were asked to report the mean emotion of a group of faces by matching a test face to the previously seen group. We compared subjects' performance when the foveal occluder was present to a control condition in which the occluder was absent; if foveal information is not necessary, we would expect identical performance in the two conditions. In a second experiment, we utilized a subset design to measure how much face information observers are able to integrate from the display with and without foveal input.

Experiment 1. Ensemble perception of facial emotion with and without gaze-contingent foveal occlusion

To test the role of foveal input in creating an ensemble percept, we performed an experiment in which subjects were asked to determine the mean emotion of an array of 24 emotional faces (Figure 1). In one condition, subjects were able to freely view the stimulus array without interference, during which their eye movements were recorded. In a second condition, using online gaze position data from the eye tracker, we occluded the foveal region of the visual field. The occluder (Figure 1c) consisted of a white patch with a flattened Gaussian luminance profile (to blend seamlessly into the background), resulting in a circular area (2.6° in diameter) of full occlusion. Subjects matched the mean emotion of the presented faces in both conditions by scrolling through the entire face pool (Figure 1b; 147 total faces) using the method of adjustment, and clicking on the matching face.

Methods

Subjects

Six subjects (including two authors, four female; mean age 26.7) participated in this experiment. All subjects reported normal or corrected-to-normal vision. Subjects provided written informed consent as required by the Institutional Review Board at the University of California, Berkeley in accordance with the Declaration of Helsinki. Aside from the two authors who

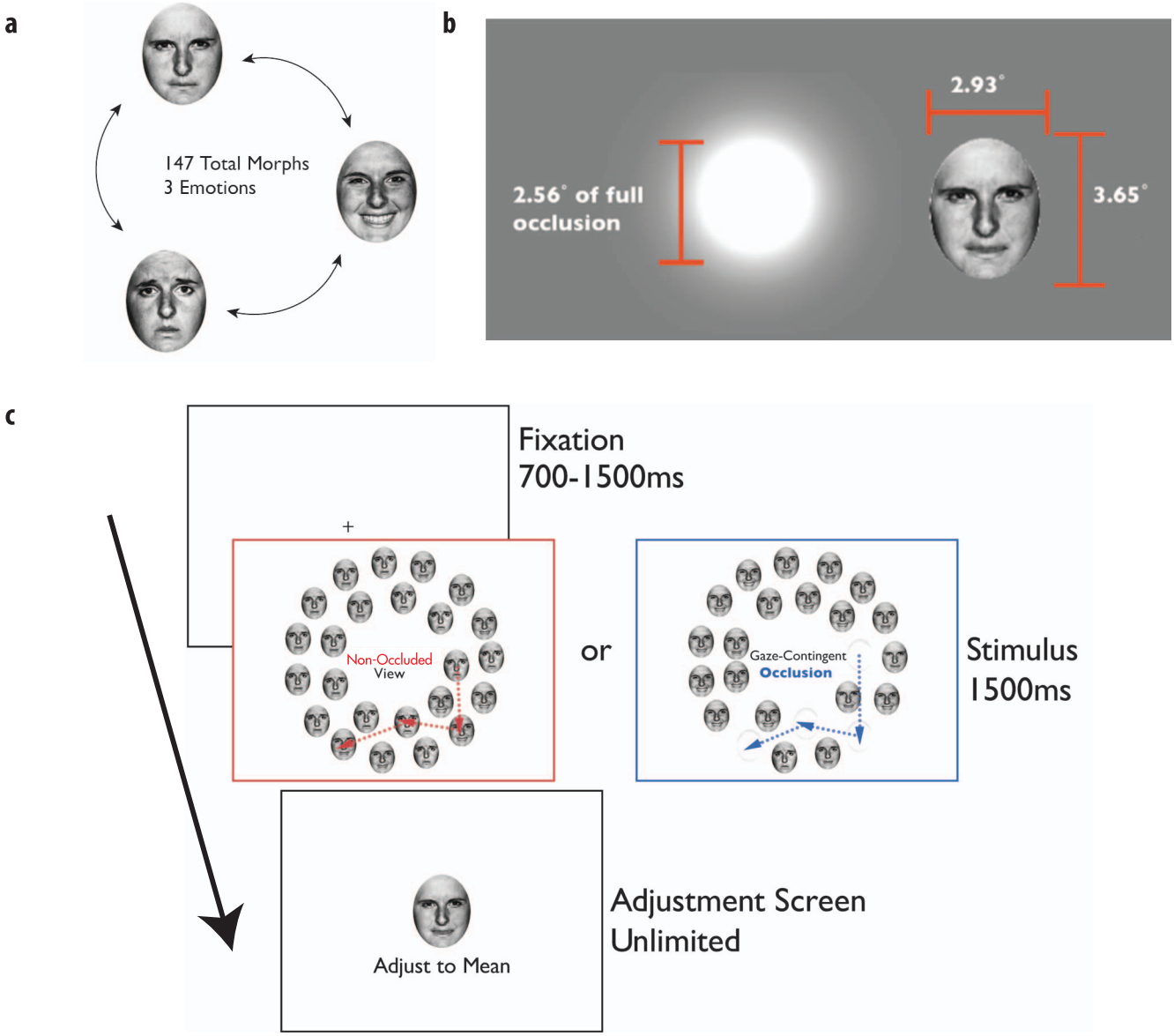

Figure 1. (a) Face pool used in all experiments (three emotions, happy, sad, and angry; 47 morphs between each emotion). (b) Illustration of foveal occluder (Gaussian blob; 1.125° *SD*) relative to size of stimulus face. (c) Stimulus sequence for Experiment 1. Subjects began each trial by fixating a cross in the center of the screen for 700–1500 ms, after which time the stimulus array (24 emotional faces, see Methods) was presented for 1500 ms of free viewing. The fovea was either occluded (blue) or non-occluded (red). After 1500 ms, the stimulus was removed and subjects were able to adjust a face presented onscreen by moving the mouse through the entire 147-face pool.

participated (AK and KW), subjects were naïve to the purpose of the experiment.

Display setup

Stimuli were presented on a 43 cm Samsung SyncMaster 997DF cathode ray tube with a monitor refresh rate of 75 Hz and a resolution of 1024 × 768. Subjects were seated in a dark booth at a viewing distance of 57 cm from the monitor and head movement was limited with a chinrest. At this distance, 30 pixels subtended approximately 1° of visual angle. The experiment was run on a Mac Mini (Apple, Cupertino, CA) and written using Matlab 2010a (MathWorks, Natick, MA) used in conjunction with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and the Eyelink Toolbox (Cornelissen et al., 2002).

Stimuli

Stimuli were morphed faces between the emotional states of happy, sad, and angry, as used by Yamanashi

Leib and colleagues (2012). The morphs were generated by starting with three images of the same individual expressing happy, sad, or angry emotional expressions, selected from the Ekman gallery (Ekman & Friesen, 1976). We then morphed the faces linearly to produce 48 morphs between each pair of basic emotions (i.e., 48 morphs between happy and sad, 48 morphs between sad and angry, and 48 morphs between angry and happy), for a total of 147 faces (Figure 1a; 144 morphs plus three original images). Note that the set forms an approximately triangular arrangement of morphs, with the maximally happy, sad, and angry faces at the "vertices" of the stimulus set. Morphs were created using Morph 2.5 (Gryphon Software, San Diego, CA). The grayscale images of the faces were ovals 2.93° wide by 3.65° high, and cropped so that hair and other background features were not visible. The mean luminance of the faces was 57.7 cd/m^2 and their mean contrast was 94%. Faces were presented on a white (141.5 cd/m^2) background.

On each trial, the set of 24 presented faces was arranged in two concentric rings (Figure 1c), consisting of an inner ring of nine faces and an outer ring of 15 faces. The inner and outer rings had radii of 6.75° and 10.5°, respectively, from the center of the display. To add random variation to the positions of the faces on each trial, the center of each face was randomly jittered around a set of evenly spaced angular locations within each ring. On each trial, faces in the inner ring were jittered by up to ±4.74° of rotation angle; faces in the outer ring were randomly jittered by up to ±2.4° of rotation angle, maintaining their distances from the center of the display. This amount of position jitter prevented any overlap or occlusion of the faces.

On each trial, 24 face morphs were selected from a Gaussian probability distribution. The center of the distribution (i.e., the average face) was selected with uniform probability from the full set of 147 morphs, and the standard deviation of the distribution was always 18 morphs. In addition, to assess any potential effects of display arrangement on subjects' performance, we performed an additional stimulus manipulation by organizing the faces by their emotional state, similar to the work of Sherman and colleagues (2012), who found organization-based facilitation of ensemble perception of orientation. On the organized trials, the face closest to the mean was assigned a random location out of 24 possible positions with equal probability. The remaining faces were sorted by absolute morph distance from the average face and assigned to the remaining 23 slots based on angular distance from the slot containing the face closest to the mean, with morph separation increasing with angular distance. On the random trials, the selected morphs were selected randomly assigned to each slot. As we found no difference in mean error between the organized (13.28 morph units) and random (13.48 morph units) conditions at the group level ($p = 0.89$) in Experiment 1, all data were averaged across the organized and random conditions in all subsequent analyses.

Trial sequence

On each trial, subjects fixated a 0.23° black cross (1.9 cd/m^2) at the center of the screen for a random period, between 700 and 1500 ms, and were subsequently shown the array of 24 emotional faces for 1500 ms, similar to previous studies as discussed in our Introduction. Subjects were allowed to freely move their eyes around the screen (Figure 1c). After the stimulus was removed from the screen, a 200-ms interstimulus interval (ISI) elapsed before subjects were shown a single face that they were instructed to adjust to match the mean emotion of the previously presented faces. Using a mouse, subjects were able to adjust the face on the response screen to any one of the 147 morphs. Once subjects had entered their response by adjusting the face to the perceived mean, and clicking the mouse to confirm their response, an 800-ms intertrial interval (ITI) elapsed before the next trial. Subjects were given feedback on their performance. Responses within 20 morphs from the mean of the set resulted in a high-pitched (652.9 Hz) tone, indicating an accurate response, and responses more than 20 morphs from the mean resulted in a low-pitched (157.1 Hz) tone, indicating an inaccurate response. Feedback was introduced to minimize lapsing; all responses were analyzed regardless of the tone subjects heard on any given trial. With the exception of the presence of the gaze-contingent occluder at the fovea (Figure 1b), the procedure was identical across the occluded and non-occluded conditions. Subjects performed the task in six blocks of 80 trials each—three blocks in the non-occluded condition and three blocks in the occluded condition, for a total of 240 trials per condition. To avoid training effects, the sequence of conditions was randomized across subjects, with half the subjects running in the occluded condition first and the other half running in the non-occluded condition first.

Eye tracking

Subjects' eye movements were recorded throughout each run using an Eyelink 1000 (SR Research, Mississauga, ON, Canada) with a level desktop camera, recording the right eye at 1000 Hz. Subjects were calibrated using a standard nine-point grid (mean error <0.5°). For the fixation analysis (see Results), time points from the recording were parsed into fixations and saccades offline using the Eyelink parser. The beginning of a fixation interval was defined as the first

time point at which the velocity fell below 30°/s and the acceleration fell below 8000°/s^2, and saccades were defined as time points in which velocity and acceleration exceeded their respective thresholds.

In the foveal occlusion condition, we used the raw gaze position data from the eye tracker to present a white occluder with a flattened Gaussian luminance profile according to the equation:

$$f(x,y) = \begin{cases} A, & \text{if } x_0 + z > x > x_0 - z \\ 2A\exp\left(\dfrac{-(x-x_0)^2 + (y-y_0)^2}{2\sigma^2}\right), & \text{otherwise} \end{cases} \quad (1)$$

where x and y represent horizontal and vertical position, respectively, x_0 and y_0 represent gaze position, σ represents the standard deviation, A represents the amplitude (corresponding to the maximum luminance of the patch), and z represents the location of the full-width at half-maximum (FWHM) or:

$$z = \sqrt{(2\ln 2)\sigma^2 - (y-y_0)^2}. \quad (2)$$

The minimum luminance of the occluder was identical to the background (141.5 cd/m^2), allowing it to blend seamlessly in with the background, and the standard deviation was set to 1.125°, resulting in a fully occluded region approximately 2.6° in diameter. The dimensions of the occluder were determined based on both the dimensions of the stimuli and retinal anatomy. The central 2.6° of the visual field has an area approximately twice as large as the entire rod-free portion of the fovea (1.8° diameter; Polyak, 1941). In addition, at the edge of the fully occluded-region (i.e., 1.3° eccentricity), human cone density drops to approximately 22.8% of its maximum (Curcio, Sloan, Kalina, & Hendrickson, 1990). More importantly, an occluder of this size fully covers the features (eyes, nose, and mouth) of each face when fixated centrally (see Figure 1b for a scale comparison of the occluder and an example face). This way, subjects were unable to extract detailed features of the face images when fixating them directly.

Analysis

All data (behavioral and eye tracking) were analyzed offline using custom Matlab scripts and S-R Research's "edfmex" file import tool. For the behavioral responses, we calculated the absolute difference between the mean emotion of the 24 randomly selected faces for a given trial and the subject's chosen match face on each trial and then calculated the mean across trials to get a measure of subjects' errors across the different conditions. We performed nonparametric bootstrap tests in order to compare subjects' performance between the occluded and non-occluded conditions, using the method of Efron and Tibshirani (1993). Bootstrapped estimates of mean response error were calculated by resampling each subject's data 1,000 times with replacement separately for the occluded and non-occluded conditions. We separately calculated the difference in errors between the occluded and non-occluded conditions within each subject, and then averaged the bootstrapped estimates across subjects. To compare observers' performance to chance (i.e., floor) performance, we calculated a null distribution of the expected errors generated by random guessing. For each of the 1,000 permutations, we shuffled the mapping between the mean of the presented group and subjects' responses and recalculated the error. In other words, the error on each trial was calculated by comparing the mean of the presented group on one trial to the response on a different trial.

Results

Response errors

To determine the effects of display configuration on subjects' performance, we compared the mean of the absolute errors between the random and organized display conditions. There was no difference ($p = 0.22$). In addition, we find no difference between the foveal occlusion (absolute mean error, 13.69 morph units) and the non-occluded (absolute mean error, 13.17 morph units) conditions ($p = 0.874$; Figure 2a; Figure 2b illustrates an individual subject's responses in both conditions). In addition, there was no effect of block order; the difference in absolute mean errors between the foveal occlusion and non-occluded conditions was similar for subjects that performed the occluded condition first versus those that performed it second ($p = 0.246$). Importantly, the lack of difference between these two conditions is not because of chance or floor performance. Subjects were very sensitive to average expression of the crowd, and performance was significantly above the expected chance performance level of 36.75 morph units (mean of permuted distribution; permutation test, $p < 0.001$), replicating several previous studies (Haberman et al., 2009; Haberman & Whitney, 2007, 2009; Sweeny et al., 2013).

Eye tracking

In addition to examining whether the presence of a foveal occluder affected performance, we compared fixation behavior between the occluded and non-occluded conditions. For the eye tracking analysis, we analyzed the fixation locations for each trial and determined whether they corresponded to a fixation on or off of one of the faces onscreen during that time, and then calculated the proportion of trials where subjects

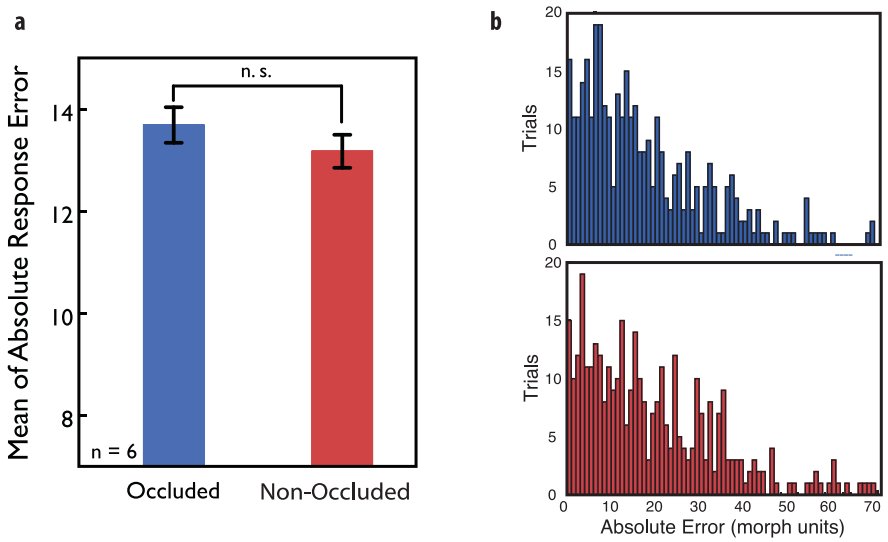

Figure 2. (a) Mean absolute response error result for Experiment 1. In Experiment 1, we found no significant difference with six subjects ($p = 0.874$; bootstrapped, two-tailed) between the mean of absolute response error in the foveal occlusion case and the non-occluded case. (b) The distribution of absolute response error in the occluded (upper) and non-occluded (lower) conditions for a single exemplar subject. Error bars represent ± 1 bootstrapped *SD*.

did not directly foveate a face. There was a slight difference (at a Bonferroni-corrected $\alpha = 0.0125$) between the occluded and non-occluded conditions in the mean number of total fixations per trial (occluded = 5.31, non-occluded = 5.39; $p = 0.006$, non-occluded > occluded). There was also a slight difference in the mean fixation duration, calculated by taking the average duration of each parsed fixation event (regardless of fixation location) within a trial, and then averaging across trials (occluded = 186.79 ms, non-occluded = 189.82 ms; $p = 0.01$, non-occluded > occluded). In addition to these common saccade and fixation metrics, we also examined whether there were any differences in gaze position relative to the faces. In particular, we classified each time point recorded during the 1500 ms stimulus presentation by whether or not the subject's point of gaze overlapped with any of the 24 faces. We then calculated the proportion of the stimulus duration (out of the 1500 ms of the trial) during which subjects fixated the space *in between* the faces in the display. This duration was longer in the foveal occlusion condition versus the non-occluded condition (occluded = 513.74 ms, non-occluded = 385.65 ms; $p < 0.001$), indicating differences in saccade behavior across the two conditions. Similarly, we classified each parsed fixation event based on whether it overlapped with one of the 24 faces, and for each trial calculated the mean proportion of fixations that were not directly on a face. Consistent with the mean duration result, there was a greater proportion of fixations in the gaps between the presented faces in the occluded condition compared to the non-occluded condition, (occluded = 0.3361, non-occluded = 0.2392; $p < 0.001$).

Discussion

Experiment 1 compared observers' accuracy in reporting the mean emotion of a set of faces, with and without a foveal occluder, to test whether foveal information is necessary for ensemble perception. In the occlusion condition, subjects were prevented from extracting detailed foveal information from any face, yet they were able to perform our ensemble perception task just as accurately as when no aspect of the stimulus was occluded. If ensemble perception relied on averaging foveal information from sequential fixations, we would have expected the occluder to have significantly impaired subjects' performance. Therefore, our results suggest that ensemble perception of facial emotion does not require foveal input, in contrast to previous reports (Ji et al., 2013, 2014; Jung et al., 2013).

However, we do find a significant difference in the amount of time subjects spend fixating faces directly between our two conditions. Given that subjects are not able to acquire detailed information about facial expression when fixating the faces in the occlusion condition, it is unsurprising that subjects opt to maximize the available information by fixating in the interstitial space between the faces. This is not to say that subjects exclusively fixated between faces in the occlusion condition; the majority of their fixations (66.39%) remained targeted at faces, rather than interstitial space. Despite this change in behavior or strategy, the results suggest that foveal detail is simply not required to process ensemble information.

Experiment 2: Do observers integrate information from multiple faces?

While our results in Experiment 1 suggest that a lack of foveal input does not necessarily impair observers' ability to perceive the mean emotion of an array of faces, the amount of information that observers use to compute this average in the occluded relative to the non-occluded condition remains an open question. One possibility is that observers extract a representation of the average emotion of the crowd by integrating information across the group of faces or a subset of that group, as suggested by the covert account of ensemble perception. In order to test whether subjects integrate multiple faces into their ensemble judgments, we modified Experiment 1 to present a subset of the total faces and calculated subjects' errors relative to the entire set of 24 faces. If subjects only use one face from the set of 24 to make their judgment, performance (when errors are calculated relative to the full set) should be the same when only one random face is visible compared to when all 24 faces are visible. If they integrate a larger number (e.g., eight faces) to make their judgment, performance with a random subset of eight should be better than when only one face is visible. In other words, we expect that if subjects integrate information from multiple faces, performance should improve with an increasing number of faces presented.

Methods

Subjects

Four subjects (two authors, AK and KW; all female; mean age 25 years) who participated in Experiment 1 also participated in this experiment. Subjects provided written informed consent as required by the Institutional Review Board at the University of California, Berkeley in accordance with the Declaration of Helsinki. Aside from the two authors who participated (AK and KW), subjects were naïve to the purpose of the experiment.

Stimuli and procedure

The stimuli (Figure 3a) and procedure in Experiment 2 were identical to those of Experiment 1 aside from the elimination of the organized configuration and the addition of a subset design. First, given the lack of a display configuration (random vs. organized) effect in Experiment 1, only the random condition was used in Experiment 2. In other words, the faces were randomly assigned to the 24 "slots" on each trial, and not arranged by proximity to the mean face. As in Experiment 1, 24 faces were selected from a Gaussian distribution (standard deviation of 18 morphs). However, on each trial, subjects always viewed only a subset of the 24 faces, drawn by randomly subsampling 1, 2, 4, 8, or 12 faces from the set of 24. For each occluder condition (occluded fovea vs. non-occluded fovea), there were 72 trials for each of five subset conditions: 1, 2, 4, 8, or 12 faces visible, which were randomly interleaved for subsets, and blocked by occlusion condition (Figure 3b). The faces presented on each trial were randomly drawn from the full set of 24, and as before, subjects were instructed to judge the mean emotion of the presented faces. The stimulus timing and response procedure remained identical to Experiment 1. Subjects performed three runs each in the non-occluded and occluded conditions for a total of 360 trials per condition. The stimuli are illustrated in Figure 3.

Analysis

Using a subset method, subjects in Experiment 2 were asked to report the perceived mean emotion of a subset of the total set of faces. Subjects viewed 1, 2, 4, 8, or 12 faces and reported the average emotion. On a trial-by-trial basis, their responses were compared to the true emotional mean of a set of 24 faces, although they were only shown a subset of faces on any given trial. For each subject, we calculated the mean of the absolute response errors, using the same method as in Experiment 1, within each of the five subset conditions (1, 2, 4, 8, and 12) and fit a line to these data using a least-squares procedure. A negative slope would indicate that subjects' performance improves as more faces are visible, and that subjects integrate multiple faces from the display. We followed a similar bootstrapping procedure to that used in Experiment 1 to determine whether the slope of the linear fit was significantly below zero. For each subject, the mean of the absolute errors was bootstrapped by resampling the single-trial data from each subset condition 1,000 times with replacement. We estimated the linear fit for each of the 1,000 bootstrapped iterations for each subject, and the bootstrapped slope estimates were averaged across subjects.

Results

In Experiment 2, we measured subjects' errors in estimating the mean of the array of faces when they were only able to use a subset (1, 2, 4, 8, and 12) of the entire group to make their judgment. We found an overall decrease in mean response error with increasing set size in both the foveal occlusion and the non-occlusion conditions in this experiment (Figure 4). The linear fits of mean response errors as a function of subset number had significant negative slopes in both the foveal occlusion condition (slope $= -0.22$, $p = 0.004$) and in the non-occluded condition (slope $= -0.24$, $p = 0.002$; Figure 4). We find no significant difference

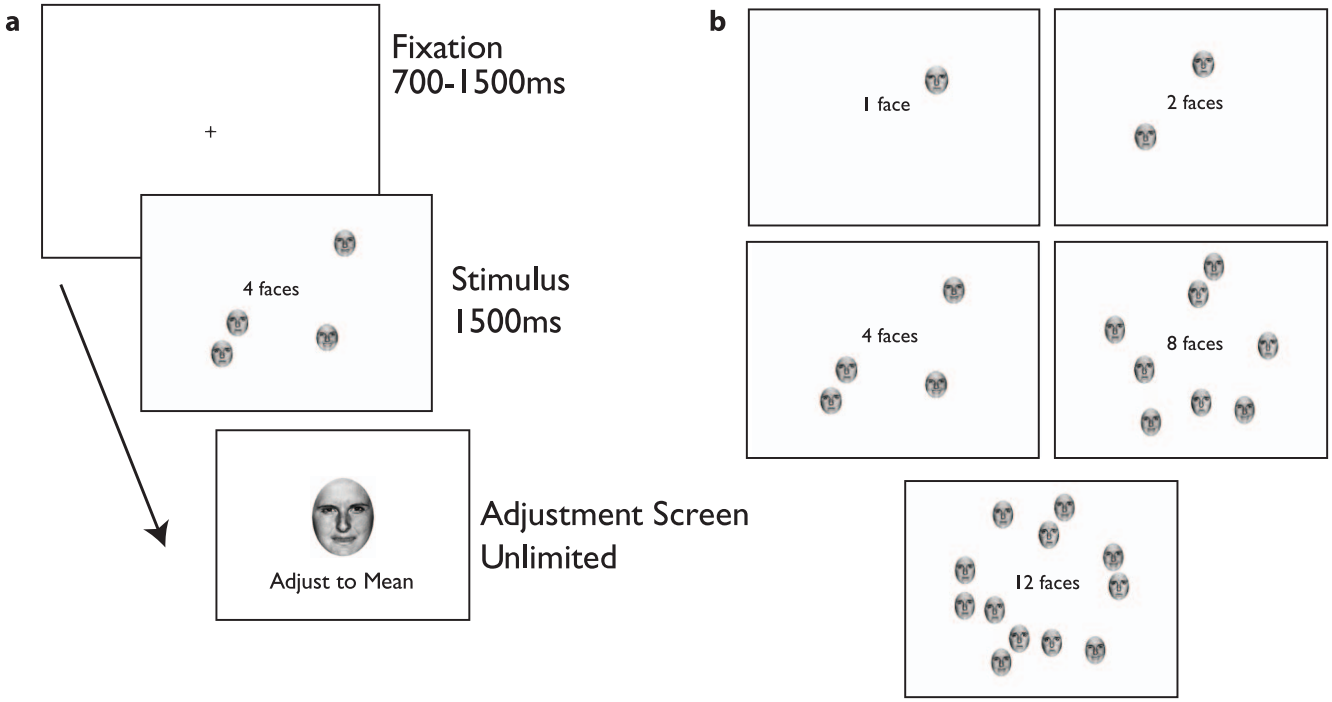

Figure 3. Experiment 2 stimuli. On any given trial (a), subjects could be presented with 1, 2, 4, 8, or 12 faces (b) from a total set of 24 and asked to report the mean emotion of the set. The methods were otherwise similar to Experiment 1.

between the two slopes ($p = 0.91$). Collapsed across both occluder conditions, compared to a subset of four faces, response error was significantly smaller with subsets of eight ($p = 0.03$) and 12 ($p = 0.016$) faces. This indicates that subjects integrated information from multiple faces when available.

We also confirmed that in the one-face condition, subjects were more accurate in judging the emotion of the single face itself when they were able to foveate it in the non-occluded condition than when they could only view it in the periphery in the occluded condition (absolute mean error, 8.52 morph units in the non-occluded condition, 9.89 morph units in the occluded condition; $p = 0.09$). Note that this value does not reflect subjects' performance relative to the mean of the full, unseen, set of 24, but rather it shows subjects' performance on assessing the emotion of a single face with and without foveal input. This is not surprising, but serves as a control, as it simply confirms that, when available, detailed foveal information about a single face is useful.

Discussion

The purpose of Experiment 2 was to determine whether subjects are able to extract ensemble information from multiple faces in the presence and absence of a foveal occluder. In particular, the subset manipulation allows us to test whether subjects' performance improves as more information is available. If subjects use multiple faces in their ensemble estimates, accuracy should improve as more faces are added. In Experiment 2, the progressive decrease in mean error, as shown by the negative slopes in each condition, indicates that subjects were able to use more faces when those faces were presented, and were able to achieve a more accurate representation of the ensemble emotion of the set, confirming previous studies (Sweeny et al., 2013; Yamanashi Leib et al., 2014). It is possible that subjects' strategies may vary as a function of the number of faces present in a given trial, but the overall effect of the presence of more faces is an improvement in task performance. The addition of our foveal occlusion manipulation allows us to extend our demonstration of ensemble perception with emotional faces; we found no differences between the foveal occlusion and nonocclusion conditions in Experiment 1 or Experiment 2. Therefore, the efficiency of ensemble face perception does not hinge on foveal input. Different strategies may be at play with and without foveal information, and different weights might be assigned to foveally viewed faces, but the effective amount of information that subjects integrate into their ensemble percept (number of integrated faces) is consistent with and without foveal input. Our results are consistent with prior demonstrations that peripheral input is sufficient to identify single faces (Mäkelä, Näsänen, Rovamo, & Melmoth, 2001; McKone, Kanwisher, & Duchaine, 2007; Melmoth, Kukkonen, Mäkelä, & Rovamo, 2000), and that this input is also useful for ensemble processing of groups of faces (Farzin, Rivera, &

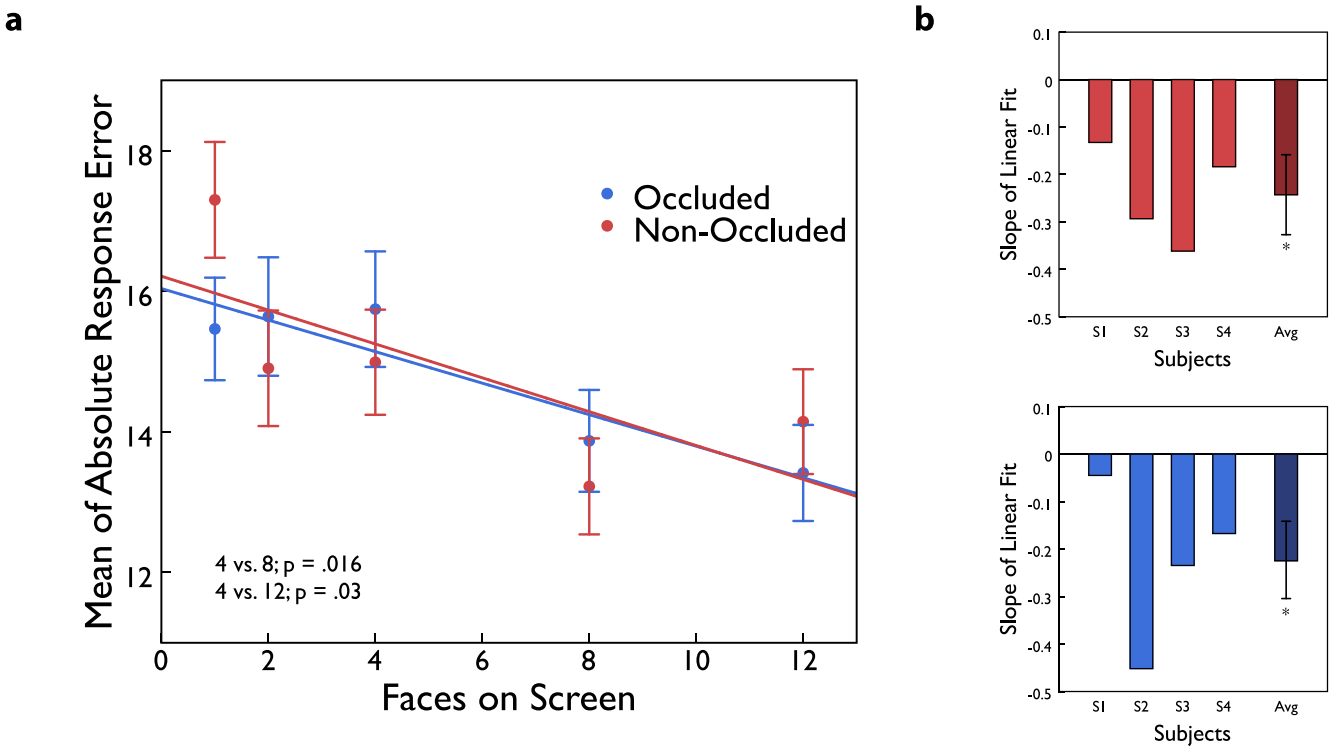

Figure 4. Experiment 2 results. (a) Mean of absolute response error for five subsets with foveal occlusion (blue) and without occlusion (red). We find no difference between the slopes for the two conditions; $p = 0.914$. There was a significant negative slope for both the non-occluded and occluded conditions ($p = 0.004$ and $p = 0.002$, respectively, bootstrapped, two-tailed) across the five subsets tested, indicating that subjects integrated more faces into their ensemble percept when available. (b) Slopes of linear fit for individual subjects. Average performance is identical to results shown in (a). Error bars represent ±1 bootstrapped *SD*.

Whitney, 2009; Fischer & Whitney, 2011; Haberman et al., 2009; Louie, Bressler, & Whitney, 2007).

Experiment 3: Determining the impact of response method

While our results in Experiments 1 and 2 suggest that ensemble perception of emotion does not require foveal input, subjects in those experiments were only able to select matching faces from the pool of morphed faces. That is, the response method allowed subjects to choose a morph between any two emotions as their response for a given trial, but it excluded the possibility that all three emotions may have been represented in a given trial, which could have resulted in less accurate responses. To address this concern, we performed Experiment 3, which reproduced Experiment 2 with a response method that averaged from all three canonical emotions in the pool. To put it another way, in Experiments 1 and 2, subjects responded with a morph between two of the three possible emotions. In Experiment 3, subjects were able to choose a response face that was generated from a weighted average from all three emotions (Figure 5). In all other respects, this experiment was identical to Experiments 1 and 2.

Methods

Subjects

Four observers—one author (AK) and three naïve observers—participated in Experiment 3 (three female; mean age 26.25 years).

Stimuli and procedure

The stimuli and procedures in Experiment 3 were identical to those in Experiment 2, with the addition of trials containing the full set of 24 faces, and a modification to the response procedure described below.

As shown in Figure 5a, on the adjustment screen, subjects were presented with a 3.65° × 2.93° adjustment face (similar to Experiments 1 and 2) positioned 4° above screen center, with a downward-pointing equilateral triangle (9° on each side) positioned 4° below screen center. The color of each pixel inside the triangle was determined by its distance from each of the vertices. The upper-left, upper-right, and bottom vertices corresponded to maximally saturated red,

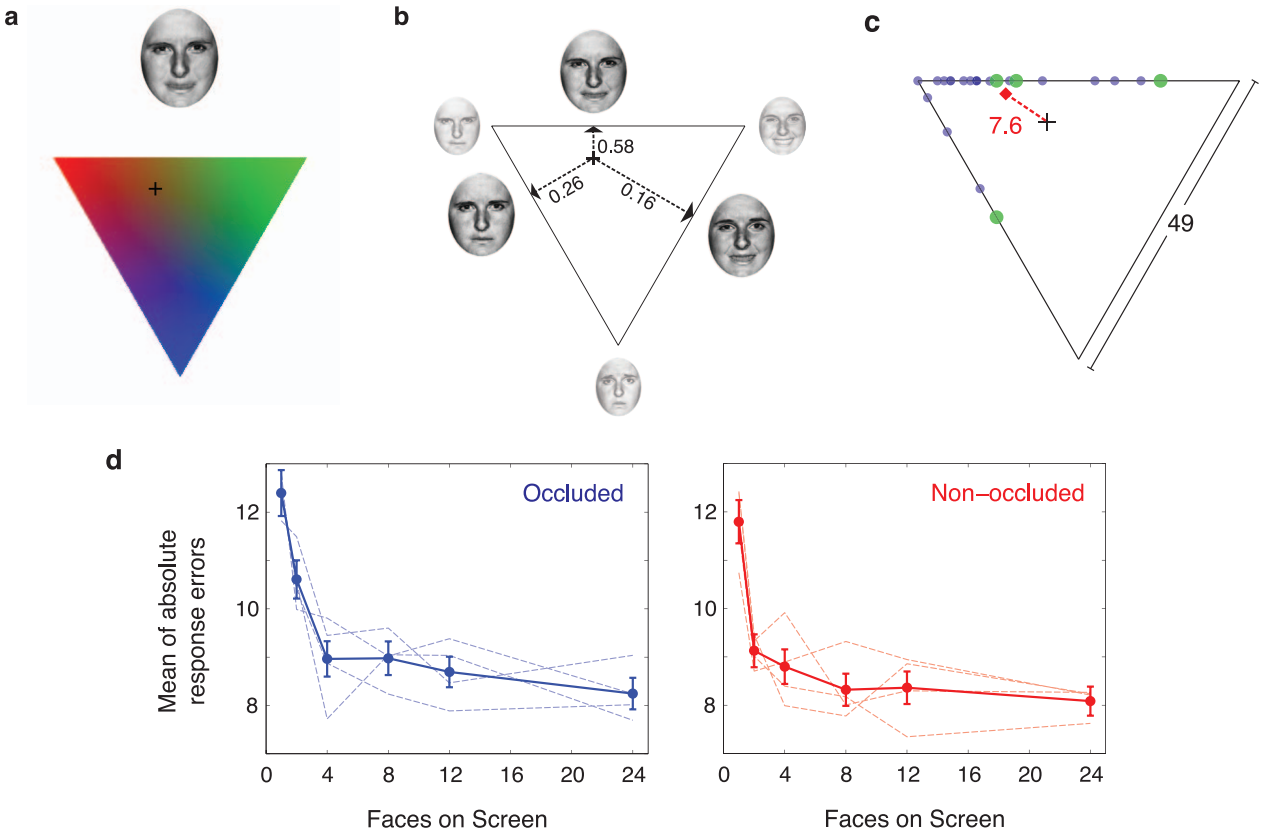

Figure 5. Experiment 3 response space (a–c) and results (d). During the response period in each trial in Experiment 3, subjects moved a crosshair inside the triangle shown in (a) to manipulate the expression of the face directly above it. The expression of each possible crosshair position was generated from a weighted average of three faces, shown in (b). The small faces in the upper-left, upper-right, and bottom vertices depict the maximally angry, happy, and sad expressions, respectively. The larger three faces correspond to morphs in between the three canonical expressions (see Experiment 1 and Figure 1a). These three faces were morphs selected from the most proximate locations on each of the three edges. The numbers represent the weights (inversely proportional to distance from the edge) used to generate pixel-wise averages the three faces. The faces and weights in (b) were only used to generate the morph in (a) and were not visible to subjects. Calculation of each set means and errors, in an example trial, shown in (c). Circles represent individual faces from the set of 24. Large green circles represent the subset visible to observers and used by the observer to make a response, shown by the black crosshair. The red diamond represents the true mean of the full set (mean of *x*- and *y*-locations). The error is calculated as the distance between the response and true mean (7.6 morph units; maximum error: 49 morph units). (d) Mean errors for each subset condition for four observers, for both the occluded (blue) and non-occluded (red) conditions. Dashed lines show mean error for each observer. There was no significant difference in mean response error between the occluded and non-occluded conditions at any subset condition, except the two-face subset condition ($p = 0.002$; all other p-values > 0.16). Error bars represent ± 1 bootstrapped *SD*.

green, and blue intensity values, and the middle of the triangle corresponded to an equal mixture of red, green, and blue (i.e., medium gray).

Subjects changed the expression of the adjustment face by using the mouse to move a 0.51° black crosshair to a position constrained inside the triangle. As shown in Figure 5b, crosshair positions at the upper-left, upper-right, and bottom vertices corresponded to the maximally angry, happy, and sad faces, respectively. Each of the three edges corresponded to the 47 face morphs (i.e., 147 morphs in total) between the three canonical faces. In other words, moving clockwise, the top, bottom-right, and bottom-left edges transitioned from angry-to-happy, happy-to-sad, and sad-to-angry, respectively. Each possible cursor position inside the triangle corresponded to a weighted mixture of three face morphs, determined by the face at the nearest location on each of the three edges (Figure 5b). The weights were calculated by taking the inverse of the three distances from the points on the nearest edges, which were normalized to sum to 1. For example, the face corresponding to the cursor position at the center of the triangle was calculated by averaging three face morphs (the three faces exactly halfway between angry–happy, happy–sad, and sad–angry), each of which was assigned a weight of 1/3. These weights were used to

generate a pixel-wise weighted average of the three morphs, which corresponded to the adjustment face shown above the triangle.

As before, subjects were instructed to match the adjustment face as closely as possible to the mean of the faces presented on the previous screen by moving the cursor. Subjects were told that the vertices corresponded to the maximally happy, angry, and sad faces, and that they could click anywhere within the bounds of the triangle (i.e., at the edges, vertices, or inside the triangle) to make their response. As in Experiments 1 and 2, subjects were given feedback on the accuracy of their response. Responses within 14 morph units of the mean resulted in a high-pitched tone, and responses more than 14 morph units from the mean resulted in a low-pitched tone. Subjects completed three blocks of 144 trials for both the occluded and non-occluded conditions, resulting in 72 trials per subset condition for each occlusion condition.

Data analysis

The same procedure described in Experiments 1 and 2 was used to select the set of 24 faces on each trial. To determine the mean of the full set of 24 faces in the triangular response configuration, each of the 24 faces was first assigned a location along the edges of the triangle (see Figure 5c; e.g., the maximally angry face corresponded to the upper-left corner, and the face halfway between happy and angry was halfway along the top edge). Then, the x- and y-locations of the entire set of 24 were averaged to determine the x- and y-coordinates of the mean face inside the triangle. Subjects' errors were calculated by taking the linear distance between this mean location and the location that the subject clicked (see Figure 5c).

The use of a triangular response configuration also resulted in a compressed range of response errors. With a response triangle, the maximum error theoretically possible was 49 morph units (i.e., the true "mean" was exactly at one of the vertices, and the subject's response was at one of the other two vertices). In contrast, the maximum possible error in Experiments 1 and 2 was 73 morph units (i.e., points on opposite ends of the set of morphed faces). As a result of this calculation procedure, subjects' errors in Experiment 3 are lower than those in Experiments 1 and 2.

Results

As shown in Figure 5c, on each trial, we calculated subjects' errors relative to the full set of 24 faces, regardless of how many were visible on a single trial (1, 2, 4, 8, 12, or 24 faces). As expected, the error relative to the full set decreases as observers are given more information about the full set of 24 faces (Figure 5d). This is consistent with the expected pattern of performance if observers were averaging multiple faces from the group. Consistent with the results from Experiments 1 and 2, the pattern of errors in the occluded and non-occluded conditions were similar. Mean response errors were not significantly different between the occluded and non-occluded conditions for each subset condition (at a Bonferroni-correct $\alpha = 0.0083$) except the two-face subset condition ($p = 0.002$; all other p-values > 0.16).

Finally, we confirmed that the presence of a foveal occluder reduced observers' performance in matching the expression of a single face; the fovea helps when perceiving a single face. To do this, we calculated observers' errors in the one-face subset condition relative to the face that was presented (rather than the full set of 24). As expected, when responding to a single face, mean errors were higher in the occluded condition than the non-occluded condition (8.40 vs. 7.00 morph units), $p < 0.001$.

Discussion

Experiment 3 introduced a novel response method, in which subjects could choose a face from three-dimensional space to match the average expression in a crowd. The results replicated the first two experiments, confirming that subjects can extract an ensemble representation of crowd expression.

General discussion

Ensemble perception—the ability to accurately extract the mean of a given stimulus feature from a set of objects—enables the visual system to summarize complex information in a scene in an efficient manner. Generating an ensemble representation could result from a covert process, not requiring fixation of individual objects, or an overt process of averaging foveal information across fixations. Our first experiment determined that foveal information was not necessary to perceive ensemble expression, the second determined that, when available, multiple faces were integrated into the ensemble representation and the third verified that our initial response method did not influence our results in the previous experiments.

These results suggest that ensemble perception is not necessarily reliant on subjects foveating individual stimuli, as the inability to do so did not adversely impact performance. In contrast, (Ji et al., 2014) demonstrate that subjects weight information near the fovea more heavily when judging the average of the group. It is possible that foveal information may be given greater weight under certain circumstances—when detailed foveal information is available, when stimuli are

presented briefly, or when the variability of the stimulus features is large. Even if foveal information is given greater weight than extrafoveal information under some conditions, our results demonstrate that ensemble processes are just as accurate, and just as efficient, in the absence of foveal information.

The ensemble face perception we find here, based on lower-resolution peripheral information, is consistent with previous results on ensemble perception of faces, including the fact that prosopagnosic subjects can still perceive average crowd identity (Yamanashi Leib et al., 2012), and the fact that peripherally crowded faces can contribute to an ensemble percept (Fischer & Whitney, 2011; Whitney & Levi, 2011). Evidently, ensemble perception allows humans to compensate for poor resolution—whether introduced because of retinally eccentric stimulation, crowding, perceptual deficits, or other sources. Our results hint that the visual perception of gist, in the form of ensemble information from the periphery, might be more precise and accurate than the coarse resolution limits set by acuity and crowding would suggest (Levi, 2008; Whitney & Levi, 2011). Given the effectiveness of peripheral information in generating ensemble percepts in this and previous studies, an intriguing possibility is that macular degeneration patients may show few, if any, deficits in ensemble perception, even with the total absence of foveal and parafoveal input. This will be an interesting avenue for future research.

Keywords: ensemble perception, face perception, foveal occlusion, statistical summary

Acknowledgments

This material is based upon work supported by NSF-GRFP awards to BW and AK, respectively, and NIH EY018216 and NSF 0748689 to DW.

Commercial relationships: none.
Corresponding author: Benjamin Arthur Wolfe.
Email: bwolfe@berkeley.edu.
Address: University of California, Berkeley, Department of Psychology, 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. doi:10.1016/j.tics.2011.01.003.

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

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Carpenter, R. H. S. (1988). *Movements of the eyes* (2nd rev. & enlarged ed.). London: Pion Limited.

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

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc, 34*(4), 613–617.

Curcio, C. A., Sloan, K. R., Kalina, R. E., & Hendrickson, A. E. (1990). Human photoreceptor topography. *The Journal of Comparative Neurology, 292*(4), 497–523. doi:10.1002/cne.902920402.

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. *Quarterly Journal of Experimental Psychology (2006), 62*(9), 1716–1722. doi:10.1080/17470210902811249.

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior, 1*(1), 56–75.

Farzin, F., Rivera, S. M., & Whitney, D. (2009). Holistic crowding of Mooney faces. *Journal of Vision, 9*(6):18, 1–15, doi:10.1167/9.6.18. [PubMed] [Article]

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology, 106*(3), 1389–1398. doi:10.1152/jn.00904.2010.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), 751–753.

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. doi:10.1037/a0013899.

Haberman, J., & Whitney, D. (2011). 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*. New York: Oxford University Press.

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

Ji, L., Chen, W., & Fu, X. (2013). Was "seeing the mean emotion" indeed a high level analysis? *Journal of Vision, 13*(9): 591, doi:10.1167/13.9.591. [Abstract]

Ji, L., Chen, W., and Fu, X. (2014). Different roles of foveal and extrafoveal vision in ensemble representation for facial expressions. *Engineering Psychology and Cognitive Ergonomics Lecture Notes in Computer Science*, *8532*, 164–173.

Jung, W. M., Bulthoff, I., Thornton, I., Lee, S. W., & Armann, R. (2013). The role of race in summary representations of faces. *Journal of Vision, 13*(9): 861, doi:10.1167/13.9.861. [Abstract]

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research, 48*(5), 635–654.

Louie, E. G., Bressler, D. W., & Whitney, D. (2007). Holistic crowding: Selective interference between configural representations of faces in crowded scenes. *Journal of Vision, 7*(2):24, 1–11, doi:10.1167/7.2.24. [PubMed] [Article]

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281. doi:10.1038/36846.

Mäkelä, P., Näsänen, R., Rovamo, J., & Melmoth, D. (2001). Identification of facial images in peripheral vision. *Vision Research, 41*(5), 599–610.

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences, 11*(1), 8–15. doi:10.1016/j.tics.2006.11.002.

Melmoth, D. R., Kukkonen, H. T., Mäkelä, P. K., & Rovamo, J. M. (2000). The effect of contrast and size scaling on face perception in foveal and extrafoveal vision. *Investigative Ophthalmology & Visual Science, 41*(9), 2811–2819. [PubMed] [Article]

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception and Psychophysics, 70*(5), 772–788. doi:10.3758/PP.70.5.772.

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. doi:10.1038/89532.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science, 24*(8), 1389–1397. doi:10.1177/0956797612473759.

Polyak, S. L. (1941). The retina. Chicago, IL: The University of Chicago Press.

Sherman, A. M., Evans, K. K., & Wolfe, J. M. (2012). The gist of the organized is more precise than the gist of the random. *Journal of Vision, 12*(9): 841, doi:10.1167/12.9.841. [Abstract]

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological Science, 25*(10), 1903–1913. doi:10.1177/0956797614544510.

Sweeny, T. D., 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. doi:10.1037/a0028712.

Watamaniuk, S. N., & Sekuler, R. (1992). Temporal and spatial integration in dynamic random-dot stimuli. *Vision Research, 32*(12), 2341–2347.

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. doi:10.1016/j.tics.2011.02.005.

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, doi:10.1167/14.8.26. [PubMed] [Article]

Yamanashi Leib, A., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia, 50*(7), 1698–1707. doi:10.1016/j.neuropsychologia.2012.03.026.

Yang, J.-W., Yoon, K. L., Chong, S. C., & Oh, K. J. (2013). Accurate but pathological: Social anxiety and ensemble coding of emotion. *Cognitive Therapy and Research, 37*(3), 572–578.