Ensemble Perception of Temporal Crowds: Detection of Facial Outliers

Johnson C., Nakhkoob S., Pu K., and Zhang J.
Whitney Lab, Department of Psychology: University of California, Berkeley.

1. ABSTRACT

In his or her lifetime, an adult perceives and is in contact with an average of approximately 9,000 faces, often viewing them in multiple crowds simultaneously. Despite this complexity, however, the human visual system is remarkably sensitive, such that it is able to identify specific faces in a crowd within only a matter of milliseconds.[1] While single face perception has been widely studied, ensemble perception, or the perception of crowds as a whole, is a newly emerging field. To date, ensemble perception has been studied using a limited display of single crowds, hence failing to simulate the visual intricacy of real-life interactions. This experiment increases ecological validity by examining whether participants perceive statistical differences when multiple crowds are presented simultaneously in the visual field. Our initial findings suggest that participants are indeed sensitive to statistical variances in multiple crowds. Overall, this investigation provides preliminary evidence that ensemble perception operates efficiently across multiple groups in a scene, potentially providing a plethora of practical applications, ranging from illness detection in the healthcare industry to threat detection in public areas.

2. KEYWORDS

Ensemble perception, temporal sequence, crowds, facial outliers, emotional percepts, statistical summary, visual system, ensemble coding, facial stimuli

3. INTRODUCTION

Ensemble perception, or the ability to obtain summary information about groups of objects within brief amounts of time, is capable of detecting a group average and other statistical features such as the direction of variance which was investigated in this study. With this taken into account, there are two types of ensemble perception: low- and high-level.

[1] Hampton, C., Purcell, D. G., Bersine, L., Hansen, C. H., & Hansen, R. D. (1989). Probing "pop-out": Another look at the face-in-the-crowd effect. *Bulletin of the Psychonomic Society*, *27*(6), 563-566.

Low-level ensemble perception examines the processing of image features from the retina. It has been studied for the trivial motion of dots,[2,3] as well as static features such as average orientation[4,5,6] and average brightness.[7]

High-level ensemble perception, however, examines how perceptual organizations in the visual system function. The study of high-level ensemble perception originated with Whitney and Haberman's 2007 research regarding how well people can extract emotional and gender means.[8] This spurred a new field of research that explored how psychological phenomena act when applied to groups of stimuli. Whitney and Sweeny then studied how well individuals can use ensemble perception to measure the average direction of a crowd's gaze.[9] Expanding on our ability to predict other humans' impending actions, Leib et al. studied how well people can perceive a mean crowd identity.[10] Moving away from ensemble perception in a spatial context, Haberman et al. tested how well individuals can measure the average facial expression of a 56single person separated temporally.[11] Up until now, the percept of facial expressions has not been investigated in a multiple-crowd context; rather, past research has focused almost entirely on ensemble perception in a single-crowd context. This experiment was aimed at proving whether or not ensemble perception operates across complex natural scenes. In order to investigate

[2] Watamaniuk SN, McKee SP. 1998. Simultaneous encoding of direction at a local and global scale. Percept. Psychophys. 60(2):191–200

[3] Watamaniuk SN, Sekuler R, Williams DW. 1989. Direction perception in complex dynamic displays: the integration of direction information. Vis. Res. 29(1):47–59

[4] Dakin SC, Watt RJ. 1997. The computation of orientation statistics from visual texture. Vis. Res. 37(22):3181– 92

[5] Miller AL, Sheldon R. 1969. Magnitude estimation of average length and average inclination. J. Exp. Psychol. 81(1):16–21

[6] Parkes L, Lund J, Angelucci A. 2001. Compulsory averaging of crowded orientation signals in human vision. Nat. Neurosci. 4(7):739–44

[7] Bauer B. 2009. Does Stevens's power law for brightness extend to perceptual brightness averaging? Psychol. Res. 59:171–86

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

[9] Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological science*, *25*(10), 1903-1913.

[10] Leib, A. Y., 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), 26-26.

[11] Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of vision*, *9*(11), 1-1.

this question, the accuracy of humans' perception of complex displays, in this case, multiple crowds were measured, ultimately deriving a connection between the two elements of high-level visual processing.

Thus, the novel aspect of this study is that we presented the temporal ensemble display in four locations across the computer monitor, rather than a single crowd. The crowds differed in their statistical properties; specifically, three crowds were the same , while one crowd was an outlier. If ensemble perception successfully operates across more complex displays, we predicted that the participants will be able to detect an outlier temporal crowd in a display of multiple crowds with a high degree of accuracy.

4. METHODS

Participants

All procedures and experiments were conducted in accordance with the Institutional Review Board (IRB) at the University of California, Berkeley. Four subjects were tested for accuracy, all of whom were right-handed female adults. Subject A was 26 years old and of Caucaision race, B was 19 years old and of Chinese ethnicity, C was 23, also of Chinese ethnicity, and D was 25 and of Asian race. The average age of the four participants was 23.25, and the standard deviation was 2.68.

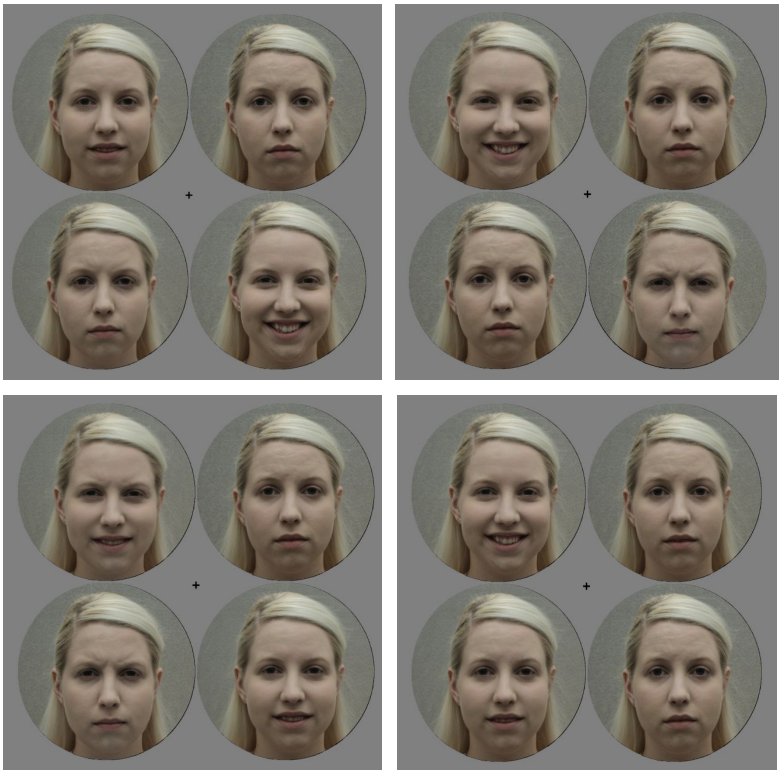

Figure 1: Experimental Design - Display Window of Four Flashing Morphed Faces. Sample screen of what the test subjects view during each trial. Every two milliseconds, a 2x2 grid of four morphed faces

flashed with one outlier. In this specific example, the bottom right image in all four displays is the outlier due to the fact that it morphs the most throughout the flashes.

5. MATERIALS

In order to test the subjects, a Macintosh desktop with monitor size/resolution 1920x1080 and 60 Hz refresh rate was used. The four subjects all sat 57 cm from the screen in a controlled head-chin rest. The experiments were performed using Matlab 2019a[12] the Psychtoolbox (Brainard, 1997; Pelli, 1997).

6. RESULTS:

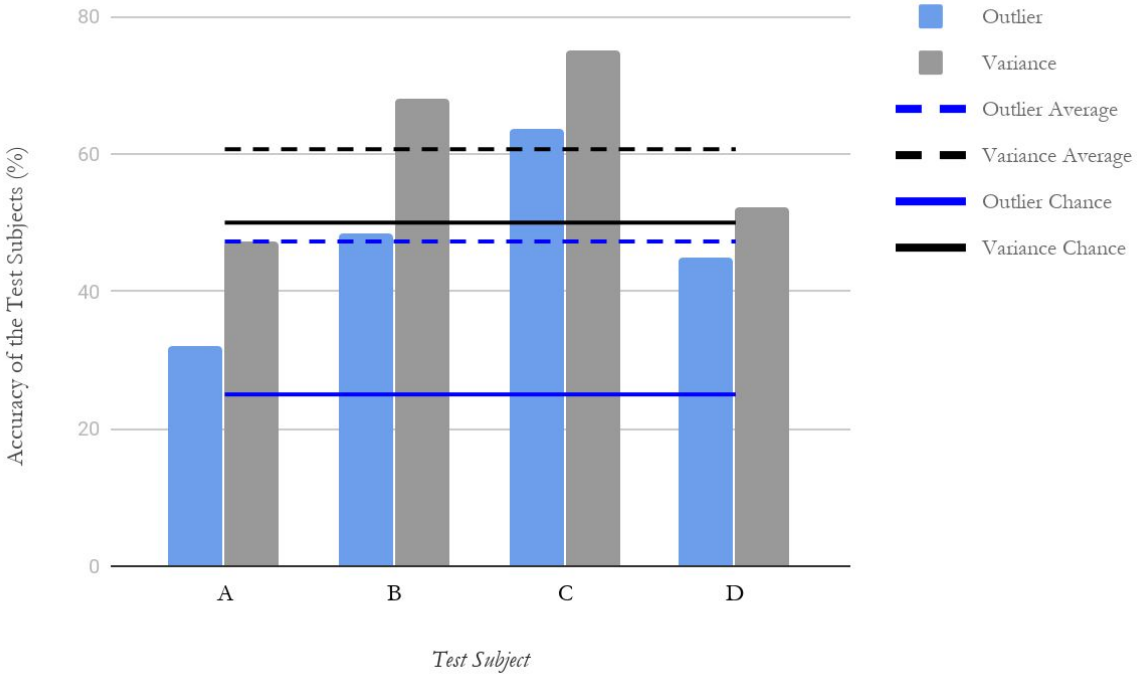

Figure 2: Results of the Four Test Subjects (Outlier Accuracy, Variance Accuracy, Average). The accuracy of response in both the locale of outlier and degree of variance were measured. As observed by the data, the outlier accuracy was above chance for all four participants, yet only the third test subject had significant results; the variance, though, was relatively more significant, as the second and third participants had accuracies of greater than 60%. Moreover, the average accuracy in both categories shows that, albeit not extremely precise, the test subjects were more accurate than random choice (25%, 50%) would suggest. However, as the standard deviation illustrates, no subjects were able to reach a level of accuracy significant enough for everyday reliance on the temporally-framed ensemble perception.

[12] MATLAB. (2019). *version 9.6.0 (R2019a)*. Natick, Massachusetts: The MathWorks Inc.

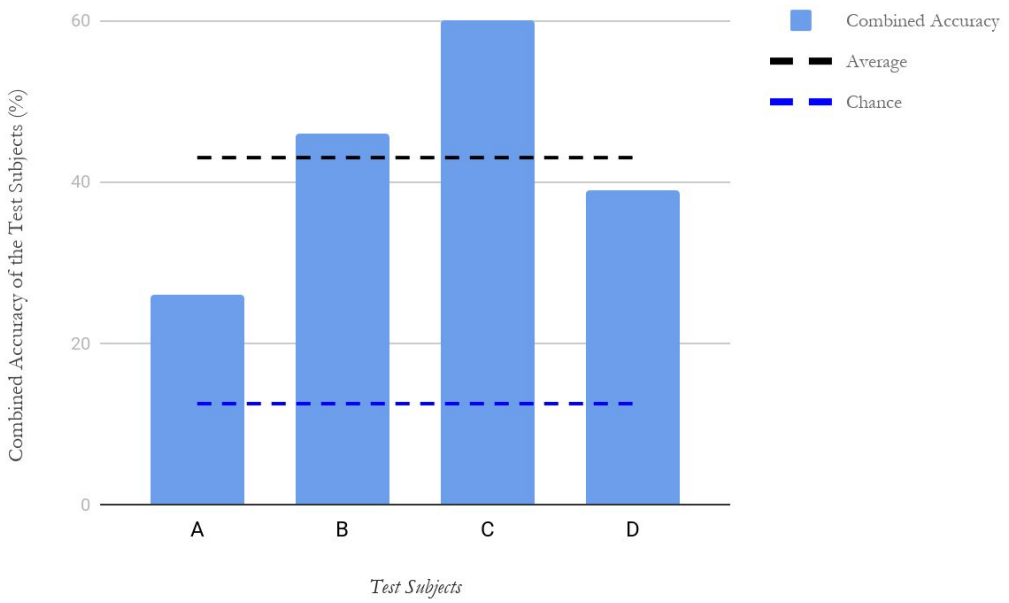

Figure 3: Combined Accuracy of the Four Test Subjects. According to the raw data, all four participants had a combined accuracy above chance, with an average accuracy of approximately 43%. Once this combined accuracy was imputed into an impaired t-test, a p-value of 0.005 was obtained, indicating that the results of the participants were statistically significant.

7. STIMULI

The stimuli was generated from a dataset of 147 images depicting a caucasian female in consecutive facial morphs.[13] Each image depicted the woman with a slightly altered emotional expression, ranging from angry to cheerful, neutral, and sad.[14,15]

The stimuli was originally displayed on a white background but later changed to a gray background because of eye strain caused by fixating on the screen for prolonged periods of time. In addition, the fixation point was altered to persist during crowd flashes instead of appearing intermittently as our stimuli was displayed. These changes were a direct result of incorporating feedback in early stages of testing.

8. PROCEDURE

Participants were provided instruction and familiarized with the task with 3-5 practice trials. In each trial, a grid of four different facial expressions of the same individual was

[13] Dataset provided by the Whitney Lab.
[14] Ji, L., & Pourtois, G. (2018). Capacity limitations to extract the mean emotion from multiple facial expressions depend on emotion variance. *Vision research*, *145*, 39-48.
[15] 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.

shown, one of which was an outlier. The experiment varied between low variance or high variance as the outlier: if the outlier was low variance, the other three facial expressions were, in respect to one another, high variance; if the outlier was high variance, the other We thank Dr. Allison Yamanashi Leib of the Whitney Lab for assisting us with obtaining user information through a MATLAB dialog box, debugging our code, and finding experimental subjects. Additionally, the square display of the crowd of four faces was repeated six times in rapid succession (200 ms display with 100 ms blank breaks between them), with the location of the outlier remaining constant. The user was then prompted with a 2x2 grid, in which they were instructed to click the location of the facial outlier throughout all of the six trials. After the user clicked on one of the grid's squares, he or she was asked if the variance of the set of crowds was high or low. There were a total of 300 trials, with one-minute breaks every 50 trials.

9. DISCUSSION

The study results pose promising figures in that both average accuracies (47.25% and 60.67% for outlier locating and variance detection, respectively) are above chance (25% and 50%). Moreover, an analysis using a one sample t-test proved that the combined accuracy was statistically significant, as the two-tailed p-value was equal to 0.0053. However, neither the accuracy for outlier locating nor the accuracy for variance detection alone provided statistically significant results.

As a whole, the average results obtained from the experiment indicate that humans can indeed detect ensemble outlier variances as well as the direction of variance in temporal crowd sequences, which confirmed the hypothesis of the experiment.

Examining past literature reveals that little research has been dedicated to analyzing ensemble perception in multiple temporally separated crowds. Because time was limited, we were only able to examine crowds generated with one type of variance, using one stimulus set, and one type of spatial arrangement. However, our mentors and collaborators in the Whitney lab were able to concurrently run experiments using additional methods. These additional experiments provided further validation for our preliminary findings.

Our collaborators' experiments also utilized a different variance algorithm (that was perceptually harder to detect) and they spatially jittered each stimulus (controlling for low-level flicker effects), and they incorporated a slightly different stimulus set. Even with these experimental manipulations, participants can still successfully detect an outlier temporal crowd among multiple temporal crowds.

Taken together, these findings provide a valuable foundation for new research paradigms focusing on ensemble perception in multiple crowd context. With the knowledge that ensemble perception in multiple crowds can be accurately detected by the visual system, this newly emerging field can be more rigorously studied in order to further increase ecological validity and capture the visual complexity of human interactions.

10. FURTHER DIRECTIONS

In subsequent tests, it would be beneficial to vary the ethnicity, gender, and age of the displayed individual. Additionally, constructing a multi-individual grid—rather than simply displaying morphs of a single individual—would provide insight into the ensemble processing of more realistic crowds. Since the observation and evaluation of multiple crowds is a necessity of public behavior, a myriad of practical applications exist, including but not limited to illness detection in the healthcare industry. Currently, medical professionals must meet with patients face-to-face in order to accurately detect a disease, which makes hospital overcrowding remarkably common, especially in the emergency room. In fact, 25-50% of ER visits could be safely treated in a care clinic; consequently, far too many patients who require urgent care are left untreated for a long period of time. [16] Considering that ensemble perception in multiple crowds has the potential to replicate the visual intricacy of real-life interactions, doctors can theoretically determine which patient has the most severe health condition compared to the others based on symptoms from facial features, such as eye swelling, dry skin, or facial asymmetry (early signs of a stroke). As such, healthcare professionals could accurately identify patients with poor health conditions and treat them in the ER room first, while sending those with less-severe illnesses to UCCs—which would benefit both the patients and the efficiency of the healthcare industry. Moreover, another application is threat detection in dense areas such as airports. Currently, TSA agents are trained to spot individuals who appear to be emotionally out of place (an outlier); as such, using this research, computers can be programmed to perform this same task more accurately than humans.

11. ACKNOWLEDGEMENTS

We thank Dr. Allison Yamanashi Leib of the Whitney Lab for assisting us with obtaining user information through a MATLAB dialog box, debugging our code, and finding experimental subjects. Additionally, we appreciate the assistance of the undergraduate research assistants who provided general guidance and patiently tested our experiment.

[16] Weinick, R. M., Burns, R. M., & Mehrotra, A. (2010). Many emergency department visits could be managed at urgent care centers and retail clinics. *Health affairs (Project Hope)*, *29*(9), 1630–1636. doi:10.1377/hlthaff.2009.0748