Correcting “confusability regions” in face morphs

Emma ZeeAbrahamsen¹ · Jason Haberman¹

Published online: 13 April 2018
© Psychonomic Society, Inc. 2018

Abstract

The visual system represents summary statistical information from a set of similar items, a phenomenon known as *ensemble perception*. In exploring various ensemble domains (e.g., orientation, color, facial expression), researchers have often employed the method of continuous report, in which observers select their responses from a gradually changing morph sequence. However, given their current implementation, some face morphs unintentionally introduce noise into the ensemble measurement. Specifically, some facial expressions on the morph wheel appear perceptually similar even though they are far apart in stimulus space. For instance, in a morph wheel of happy–sad–angry–happy expressions, an expression between happy and sad may not be discriminable from an expression between sad and angry. Without accounting for this confusability, observer ability will be underestimated. In the present experiments we accounted for this by delineating the perceptual confusability of morphs of multiple expressions. In a two-alternative forced choice task, eight observers were asked to discriminate between anchor images (36 in total) and all 360 facial expressions on the morph wheel. The results were visualized on a “confusability matrix,” depicting the morphs most likely to be confused for one another. The matrix revealed multiple confusable images between distant expressions on the morph wheel. By accounting for these “confusability regions,” we demonstrated a significant improvement in performance estimation on a set of independent ensemble data, suggesting that high-level ensemble abilities may be better than has been previously thought. We also provide an alternative computational approach that may be used to determine potentially confusable stimuli in a given morph space.

Keywords Ensemble perception · Faces · Morphs · Discriminability

The tendency to consolidate crowds of similar objects into summary representations, a phenomenon known as *ensemble perception*, is an area of active research. Work in this area has broad intuitive appeal, since it may be the means by which the visual system overcomes traditional limits of visual consciousness (Alvarez & Oliva, 2008; Demeyere, Rzeskiewicz, Humphreys, & Humphreys, 2008; Fischer & Whitney, 2014; Haberman & Whitney, 2011), such as inattention blindness (Simons & Levin, 1998) and crowding (Whitney & Levi, 2011). Recent work has even suggested that ensembles may serve to bind information across visual scenes (Fischer & Whitney, 2014; Manassi, Liberman, Chaney, & Whitney, 2017), providing a sense of visual stability in an inherently dynamic environment (Whitney, Haberman, & Sweeny, 2014).

The methods employed to explore the mechanisms of ensemble perception have varied from psychophysical (e.g., Ariely, 2001; Haberman & Whitney, 2007) to neuropsychological (e.g., Leib et al., 2012) to neuroimaging (e.g., Cant & Xu, 2012). One approach of particular relevance here is the method of continuous report, a psychophysical technique in which an observer adjusts a test stimulus to match the perceived average of the preceding set. This approach is useful because it can characterize the full distribution of ensemble representation abilities (Haberman, Lee, & Whitney, 2015b; Haberman & Whitney, 2010) and is supported by an array of robust analytical procedures (e.g., circular statistics, mixture modeling; Berens, 2009; Suchow, Brady, Fougner, & Alvarez, 2013). In continuous report, observers select a response from a continuous distribution on each trial; the difference between what the observer selects and the correct response is used as an index of precision. Continuous report in ensemble perception has effectively been used to address a number of theoretical questions, including how the visual system integrates deviant items across a scene (Haberman & Whitney, 2010) and

✉ Jason Haberman
habermanj@rhodes.edu

¹ Department of Psychology, Rhodes College, 2000 N. Parkway, Memphis, TN 38112, USA

how the cognitive architecture of ensemble perception is organized (Brady & Alvarez, 2011; Haberman, Brady, & Alvarez, 2015a).

Although continuous report has yielded fruitful results in understanding ensemble representations, an inherent concern exists with its implementation in perceiving average faces. Although this concern does not undermine the conclusions to date regarding how humans perceive crowds of faces, it may ultimately lead to an underestimation of face ensemble abilities, therefore making it difficult to detect subtle differences between conditions. Often in continuous report, the stimuli span a circular distribution (this is not a requirement, it is just often the case; Haberman & Whitney, 2010). This stimulus design works exceptionally well for domains such as orientation, in which the stimulus space naturally lies along a circular continuum. In face space, however, the circular distribution must be artificially constructed. Typically, this entails morphing between multiple exemplars from a single person (e.g., a single individual displaying a happy, sad, and angry expression, if the domain of interest is facial expression). Mathematically, the relationship between any two morphs on the continuum is well characterized, since the morphs are simple linear interpolations. Perceptually, however, the morph space may be heterogeneous, whereby some elements may be more difficult to discriminate than others (note that this concern is true even within orientation space, in which vertical and horizontal orientations are more easily discriminable than orientations around oblique meridians; Andrews, 1967). The more critical concern, which the present article seeks to mitigate, is that some faces along one section of the morph wheel may be perceptually confusable with faces from an entirely different section of the wheel.

An example of this is displayed in Fig. 1. The two faces were selected from a morph continuum created from the Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998; Fig. 2). In stimulus space, the two images would be considered rather distant from one another. In perceptual space, however, it is evident that the images are quite similar. This discrepancy is potentially problematic when attempting to estimate performance on an ensemble task, in that an observer selecting a perceptually similar morph that resides a great distance from the “correct” response would have a large error. Thus, the purpose of the present experiments is to delineate the nature of this perceptual space for one particular morph sequence and to correct for perceptual inhomogeneity when estimating the ensemble representation of expression. The implementation of the prescribed corrective procedures is flexible, but the critical point is that some form of corrective measure be taken when considering a circular morph of this nature, since not doing so will result in an underestimation of ensemble abilities (i.e., by adding noise to the measure).

Fig. 1 Images 80 and 155 provide an example of a perceptually confusable face pair on opposite sides of the morph wheel. Though the two images are 75 faces away from one another on the morph continuum, our results suggest they are much closer in perceptual space

Experiment 1

Method

The purpose of this experiment was to better estimate ensemble expression ability by accounting for faces that might be confused with one another within a commonly used stimulus set. The first step was to identify any and all such face regions along the morph continuum by having observers evaluate whether any two images displayed were the same or different.

Participants Eight observers participated in this experiment (average age = 21.3 years), seven of whom were naïve to its purposes. All participants gave informed consent and had normal or corrected-to-normal vision. This research and all research described herein was approved by and conducted in

Fig. 2 Face morph wheel consisting of 360 faces, from angry to sad to happy and back to angry again. The original angry, sad, and happy faces were taken from the KDEF database.

accordance with the Institutional Review Board at Rhodes College.

Stimuli and design The stimulus set originated from a single individual taken from the Karolinska Directed Emotional Faces database (KDEF; Lundqvist et al., 1998) displaying three emotional expressions: angry, happy, and sad. The images were first gray-scaled and then morphed from one expression to the next using linear interpolation (MorphAge, version 4.1.3, Creaceed). This morphing procedure generated a circular distribution of 360 images going from angry to happy to sad and back to angry again (Fig. 2).

To test for confusability, 36 “anchor images” (every 10th face on the wheel starting from “Image 1,” including all three pure expressions) were compared to every other face on the continuum. All observers judged the same set of anchor images in blocks (i.e., one anchor image per run). The order of the blocks was randomized for each observer.

On each trial, the anchor image was presented adjacent to an identical face or a different face. The order of each trial type was randomized for each participant, without trial order optimization (i.e., the same trial type could occur multiple times in succession). In the “different” condition, the anchor image was presented with a randomly selected face (without replacement) from the morph wheel, and in the “same” condition, the anchor image was compared to itself. Each image subtended $6.5^\circ \times 8.2^\circ$ of visual angle, and was displayed 4.2° on either side of the screen along the horizontal meridian. Participants judged whether the images in the presented pair were the “same” or “different.”

Procedure For each trial, observers viewed a pair of faces from the morph wheel and had to judge whether the faces were the same or different. One face served as the anchor image for the entirety of a given block (i.e., the standard by which all other faces would be compared). The other face was either identical to the anchor image or one of the other 359 faces from the continuum.

Each trial was preceded by a 250-ms preparatory fixation cross, followed by the two faces displayed simultaneously for 500 ms. A response screen then appeared with instructions to participants: “F for Same J for Different.” Observers had unlimited time to make their responses (see Fig. 3).

Observers each participated in 36 blocks over the course of several months. Each block consisted of 720 trials (360 “same” and 360 “different”). With 36 blocks (i.e., 36 anchor images), this amounted to 25,920 trials per observer.

Results

Observer performance is displayed in the “confusability matrix” in Fig. 4. Each data point indicates the probability of reporting the images as “different,” averaged across observers

Fig. 3 Example trial from the experiment. Observers viewed randomly interleaved conditions and responded whether the faces were the “same” or “different”

(the x -axis is the anchor image, and the y -axis is every other image in the morph continuum). In this visualization, only trials in which the images were actually different are shown, except for the diagonal (where the anchor image was compared to itself). This visualization may be used to identify “confusability regions,” areas of the morph continuum where observers were more or less likely to report that two different images were the same. Thus, the “confusability regions” represent face pairs that were perceptually confusable. The matrix indicates that some face morphs are quite likely to be confused for other faces far away on the wheel (e.g., Face 80 seems to be confusable with a lot of other distant morphs).

An example of a confusable face pair may be seen in Fig. 1. In stimulus space, these images are far from one another, well beyond what should be the just noticeable difference (JND)¹ between any two images. In an ensemble task in which observers are asked to select the average expression of a set, the selection of image 155 when the correct response was Image 80 would grossly overestimate the perceptual error. That is, even though in stimulus space these two images (the “correct” image and the observer’s selection) are separated by 75 units, their *perceptual* similarity is much closer. The confusability matrix allows for the correction of items that are distant in stimulus space but close in perceptual space, which can provide a more accurate assessment of ensemble abilities.

¹ JNDs were determined in a separate pilot experiment. Six observers were asked to determine which of two images differed from a template image. The images were simultaneously displayed in a triangle formation for 2 s. The template was randomly selected on every trial, and the comparison images varied from among the faces 10–60 emotional units from the template. Observers performed 240 trials. The average 75% JND across observers was 27° .

Fig. 4 The “confusability matrix,” depicting the proportions of the times that observers reported two morphs as being “different”

Experiment 2

To test whether accounting for confusability regions significantly improves ensemble performance, we compared accuracy before and after implementing a correction to an unpublished ensemble dataset that utilized the same morph continuum. Variants of this task have been published and extensively described elsewhere (e.g., Haberman & Whitney, 2010)

Method

Participants Nine naïve participants from the Harvard University community participated for course credit or cash compensation. This research was approved by and conducted in accordance with the Institutional Review Board at Harvard University.

Stimuli and design The same face morphs described in Experiment 1 were used to generate ensembles in this experiment. The mean was randomly selected on every trial. Sets were composed of four faces ± 10 and ± 30 emotional units from the mean. Each face within a set subtended $1.9^\circ \times 2.5^\circ$ of visual angle and was presented radially 4.7° from fixation in a square formation. Following each set, observers adjusted a single test face, randomly selected from the morph wheel, presented in the center of the screen at the same size as the faces in the set.

Procedure On each trial, observers viewed a set of four faces for 750 ms and then adjusted a test face to match the average of the preceding set using continuous report. Observers altered the appearance of the test face by moving the mouse along the

x -axis. This movement was yoked to the morph wheel. Observers scrolled through the morph wheel until they found what they perceived to be the average expression and locked in their selection by pushing the space bar.

Model implementation The confusability matrix (Fig. 4) was used as the basis for the model implementation (i.e., correcting for the confusable faces). The 36 anchor images were used to define a 10-unit range on which a particular correction was applied—if the anchor image was 10, the correction derived from the confusability matrix was applied to Faces 6 through 15. This range was selected because ± 5 faces falls well within one JND, such that if face X was confusable with face Y , it should follow that face $X - 5$ would also be confusable with face Y . Without this assumption, we would have had to collect an untenable 225,000 trials per participant (i.e., 360 anchor images rather than 36). The basic approach for implementation was to correct for all regions in the morph space that were perceptually confusable. This approach rests upon the assumption that the observer intended to select a morph that was closer to the actual mean (i.e., one that was perceptually similar), and not the one that was more distant in stimulus space. Note that there were multiple decision points in the implementation of this model—we were not committed to this particular instantiation, but rather sought to demonstrate proof of concept. In this instance, we adopted a conservative approach.

First, we identified trials with an error greater than 30° in the independent ensemble task, since these were the candidate trials likely to reveal a large performance disconnect between the stimulus and perceptual spaces (i.e., large error according to the morph wheel, and small error according to our visual system). The confusability matrix (Fig. 4) was used to identify

whether the observer response was perceptually confusable with the *correct* response (i.e., whether the morph selected by the observer had been incorrectly identified as the same as the correct response by at least 50% of the participants in Exp. 1). If so, that response was replaced with the correct answer plus Gaussian noise with a standard deviation of 20 emotional units. After all large, confusable errors had been identified and replaced, the error for the independent ensemble task was recalculated.

Results

For the ensemble task, the average absolute error was calculated for each observer (Fig. 5A), which was then averaged across all observers. This was compared to performance after implementing the model corrections, as described above (Fig. 5B). As is shown in Fig. 5B, the margin of error decreases and the distribution is tighter for this representative observer, and for all others observers as well. A within-subjects *t* test revealed that the average absolute error was significantly larger for the uncorrected than for the corrected data, $t(8) = 9.1$, $p < .001$, $\eta^2 = .91$, suggesting that the model significantly improved the estimation of face ensemble ability (mean_{uncorr} = 63.1°, SD_{uncorr} = 9.1°; mean_{corr} = 59.4°, SD_{corr} = 9.7°).

Discussion

These experiments were designed to provide a more precise estimation of high-level ensemble perception ability. Although continuous report has been effectively used to make strong conclusions about the nature of ensemble perception (e.g., Haberman et al., 2015a), the stimuli traditionally used to assess ensemble face representation introduce noise into the estimation. For example, within a particular morph

continuum, we identified multiple “confusability regions,” whereby faces far apart in morph space were indistinguishable from one another in perceptual space. We corrected for confusability regions in an independent ensemble task by replacing the observer responses with the responses they were commonly confused with (and that were closer to the correct answer). In other words, our model assumed that while observers selected one face morph, they actually meant to select a different, more accurate face morph. This is akin to distorting the morph continuum, in essence reshaping it on the basis of the perceptual relations among the faces, not the mathematical ones. After model implementation, participants showed an average improvement of 3.7° ($SD = 1.4^\circ$) in ensemble estimation, pointing to the importance of accounting for perceptual error.

Perceptual similarity versus physical similarity

It is clear that accounting for stimulus confusability has an impact on the precision of ensemble ability estimation. However, the process by which confusability regions may be identified is laborious, and perhaps even impractical. Might there be a method by which researchers can efficiently estimate potential confusability regions, without having to collect tens of thousands of trials worth of data? A substantial body of work has examined the relationship between perceptual and physical similarity (e.g., Folstein, Gauthier, & Palmeri, 2012; Yue, Biederman, Mangini, von der Malsburg, & Amir, 2012). If we could determine the physical similarity of the morphs in our continuum, and this physical similarity correlated highly with the results of our discrimination task (Exp. 1), it could provide an alternative method by which morph confusability could be measured.

Physical similarity among the morphs was determined by simulating tiled simple cell responses to each image, a procedure implemented by Yue et al. (2012). For this instantiation,

Fig. 5 (A) A representative observer’s uncorrected response errors. (B) The same observer’s corrected response errors. Note the tightened distribution and smaller average absolute error (upper left of each graph)

each image was first run through a series of Gabor filters (8 orientations $\times$ 5 spatial frequencies $\times$ 100 points). The magnitude of response of simulated simple cells was determined for each image, and the Euclidean distance between these vectors was calculated (Margalit, Biederman, Herald, Yue, & von der Malsburg, 2016). Smaller Euclidean distances between images indicate similar simple cell responses, which is used as a measure of physical similarity. These similarity measures are depicted in Fig. 6; note the striking similarity between this matrix and the confusability matrix from Fig. 4. The more similar an image was to another image, as determined by the Gabor filters, the more likely observers were to confuse them for one another ($r = .80, p < .005$). This strong correlation suggests that a physical similarity analysis of this nature might serve as a reasonable proxy for the two-alternative forced choice discrimination task carried out in Experiment 1.

Even with this analysis, one is still faced with the difficult task of choosing the level at which two images may be considered “confusable.” In our behavioral task, we labeled any two images that were incorrectly identified as “the same” 50% of the time as confusable—a rather conservative criterion. With the physical similarity analysis, however, the units are arbitrary, and thus the same principled approach is not available. For our purposes, we looked for a similarity value that resulted in a rate of confusability comparable to the one in the behavioral task. Specifically, the model implementation for our behavioral data identified 21% of the image pairs as confusable. To match this rate of confusability in the physical similarity analysis, values below the 60th percentile of the maximum dissimilarity value were labeled as confusable.

That is, if the maximum dissimilarity between two morphs was 340 units, every pair with a dissimilarity score less than 204 units would be considered confusable.

This approach proved effective. Correcting the same data described in Experiment 2 using the physical similarity analysis yielded an improvement in ensemble performance estimation comparable to the correction based on the behavioral data (the differences between the corrected and uncorrected data from the behavioral analysis and the physical similarity analysis were 3.7° and 3.8° , respectively). Overall, these results suggest that this particular image similarity analysis (Gabor filters simulating tiled simple cells) was a reasonable proxy for our behavioral analysis. Of course, the criterion for confusability was guided by the behavioral results, without which deciding what image pairs to label as confusable would have been challenging. One possible solution to this would be to derive rough discrimination thresholds for a given morph sequence and to use these to guide the selection of what level of dissimilarity to label as confusable in the physical similarity analysis.

Choosing a model

The choice of how to implement the correction is not as important as implementing *some* correction. Detection of potentially small effects, such as ones that might emerge using attentional manipulations (e.g., Attarha, Moore, & Vecera, 2014; Emmanouil & Treisman, 2008), becomes difficult if stimulus confusability is not accounted for. As we noted above, however, the choice of how to implement a correction is flexible. There are multiple possible decision points in

Fig. 6 The “dissimilarity matrix,” depicting the Euclidean distance between morphs in terms of simulated V1 simple cell responses. See the text for details

accounting for confusability, ranging from what level of performance to call “confusable” (we chose 50%), to the size of the error to correct for (we chose greater than 30°), to the distribution of noise to apply to the correction (we chose a Gaussian with a standard deviation of approximately 20°). These particular decision points are fairly conservative, and other choices would be justifiable. For example, rather than correcting for regions of confusability, one could simply remove trials in which responses fell within a confusable region.

Although there are many decision points in the implementation of this model, one should be careful to make such decisions on principled grounds, choosing model parameters before examining the end result. To do so, it is important to understand ahead of time how such decisions might impact a given dataset. For example, in our implementation we chose to label as “confusable” any morph pair that 50% of our observers incorrectly identified as the same. This resulted in 21% of the total morph pairs being potentially confusable. When we relaxed the “confusable” criterion to 30% (i.e., observers incorrectly identified morphs as identical 30% of the time), the total percentage of confusable morph pairs increased to 25%, which increased the likelihood that a given trial would be corrected.

All of the parameters available to tweak will affect the likelihood that a given trial will undergo correction. How it will impact this likelihood is fairly straightforward (i.e., less conservative criteria will result in more model corrections). However, the model will not ever reverse the direction of an effect (since it is applied to all conditions), nor will it turn typically unusable data into something usable. This last point was verified by simulating a randomly responding observer and subjecting these data to our model. Although the ensemble performance improved slightly after correction, it still did not pass our criterion for inclusion, which tested whether an observer’s response distribution differed significantly from uniform (see Fig. 5 for a reference—note that this participant’s performance is visibly centered around the mean; a random guesser would have responses uniformly distributed about the mean).

Conclusion

We note that our results, although limited to the specific face morph used here, highlight the importance of accounting for perceptual confusability when estimating ensemble representation abilities—even for paradigms that do not utilize continuous report (e.g., the method of constant stimuli). We acknowledge that the amount of data required in order to implement such corrections is daunting (each of our observers performed psychophysics for approximately 12 h), and as such have made our data publicly available at <https://jasonmarchaberman.wordpress.com/zeeabrahamson-and-haberman-2017-vss-abstract/>. Additionally, our physical

similarity analysis (Yue et al., 2012) revealed a strong correlation with psychophysical performance, providing an alternative and efficient means of assessing morph confusability.

Despite the potential challenges associated with accounting for perceptual errors, we contend that future research should endeavor to do so, particularly when attempting to make comparisons across stimulus domains (e.g., average orientation vs. average expression) or exploring questions that typically yield small effect sizes. Critically, these considerations do not apply exclusively to research on ensemble perception—any domain using nonlinearized morph sequences would benefit from characterizing their perceptual space, unless noise is not a serious consideration.

References

- Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, 19, 392–398. <https://doi.org/10.1111/j.1467-9280.2008.02098.x>
- Andrews, D. (1967). Perception of contour orientation in the central fovea part I: Short lines. *Vision Research*, 7, 975–997.
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12, 157–162.
- Attarha, M., Moore, C. M., & Vecera, S. P. (2014). Summary statistics of size: Fixed processing capacity for multiple ensembles but unlimited processing capacity for single ensembles. *Journal of Experimental Psychology: Human Perception and Performance*, 40, 1440–1449. <https://doi.org/10.1037/a0036206>
- Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, 31(10), 1–21. <https://doi.org/10.18637/jss.v031.i10>
- Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22, 384–392. <https://doi.org/10.1177/0956797610397956>
- Cant, J. S., & Xu, Y. (2012). Object ensemble processing in human anterior-medial ventral visual cortex. *Journal of Neuroscience*, 32, 7685–7700.
- Demeyere, N., Rzeskiewicz, A., Humphreys, K. A., & Humphreys, G. W. (2008). Automatic statistical processing of visual properties in simultanagnosia. *Neuropsychologia*, 46, 2861–2864.
- Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Perception & Psychophysics*, 70, 946–954. <https://doi.org/10.3758/PP.70.6.946>
- Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, 17, 738–743.
- Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: Not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 807–820. <https://doi.org/10.1037/a0025836>
- Haberman, J., Brady, T. F., & Alvarez, G. A. (2015a). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, 144, 432–446. <https://doi.org/10.1037/xge0000053>
- Haberman, J., Lee, P., & Whitney, D. (2015b). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, 15(4), 16. <https://doi.org/10.1167/15.4.16>

- Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17, R751–R753.
- Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, 72, 1825–1838. <https://doi.org/10.3758/APP.72.7.1825>
- Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, 18, 855–859.
- Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia*, 50, 1698–1707.
- Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces (KDEF) (CD ROM). Stockholm: Karolinska Institutet, Department of Clinical Neuroscience, Psychology section.
- Manassi, M., Liberman, A., Chaney, W., & Whitney, D. (2017). The perceived stability of scenes: Serial dependence in ensemble representations. *Scientific Reports*, 7, 1971.
- Margalit, E., Biederman, I., Herald, S. B., Yue, X., & von der Malsburg, C. (2016). An applet for the Gabor similarity scaling of the differences between complex stimuli. *Attention, Perception, & Psychophysics*, 78, 2298–2306. <https://doi.org/10.3758/s13414-016-1191-7>
- Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, 5, 644–649. <https://doi.org/10.3758/BF03208840>
- Suchow, J. W., Brady, T. F., Fougner, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, 13(10), 9. <https://doi.org/10.1167/13.10.9>
- Whitney, D., Haberman, J., & Sweeny, T. D. (2014). From textures to crowds: Multiple levels of summary statistical perception. In J. S. Werner & L. M. Chalupa (Eds.), *The new visual neurosciences* (pp. 685–709). Cambridge, MA: MIT Press.
- Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, 15, 160–168. <https://doi.org/10.1016/j.tics.2011.02.005>
- Yue, X., Biederman, I., Mangini, M. C., von der Malsburg, C., & Amir, O. (2012). Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision Research*, 55, 41–46.