Texture Statistics: The Mechanism Behind Ensemble Perception in Human Vision

Sasen S Cain (sasen@ucsd.edu)
Department of Psychology, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093 USA

Matthew S. Cain (matthew.s.cain6.civ@mail.mil)
U.S. Army NSRDEC, 15 General Green Ave., Natick, MA 01760 USA
Center for Applied Brain and Cognitive Sciences, Tufts University, 200 Boston Ave., Medford, MA 02155 USA

Abstract

How do ensemble representations subvert well-established perceptual capacity limits, such as attention and visual working memory? Here we ask if an off-the-shelf texture statistics representation can explain ensemble judgments, without explicitly representing and measuring objects. We found that an ideal observer using only texture statistics was able to perform an ensemble mean size comparison task as well as humans, and further, that this model replicated previously unexplained human perceptual biases. This means that the virtues of ensemble representations could actually be due to the compressive power of texture statistics. We thus present the first generalizable computational account of ensemble perception, while also explaining a long-standing mystery about limitations on human performance in ensemble tasks.

Keywords: vision; ensemble perception; texture statistics

Introduction

How do people make judgments about scenes with many items? On one hand, visual cognition seems severely limited by precious attention and working memory resources. On the other hand, most of us can navigate a grocery store, despite each aisle containing a multitude of goods for sale. Potter (1976) and others have demonstrated that we can meaningfully understand detailed real-world scene information in less than 100ms. What representations could support such fast scene perception?

Ensemble Representations

Ariely (2001) proposed that sets of similar items could be stored compactly by summary statistics—for example, the items' mean and variance—of properties like size, position, and color. These *ensemble representations* are thought to support scene gist perception and to guide eye movements and attention deployment. Following a series of experiments on mean circle size judgments, as shown in Figure 1 (Chong & Treisman, 2003), others have found that ensemble averaging can apply to high-level judgments like mean emotion of a set of faces (see Whitney & Yamanashi Leib, 2018, for a review). Thus far, the only mechanisms proposed for ensemble representations have implied multiple parallel pre-attentive pathways operating on the level of objects (Alvarez, 2011). In this light, ensemble representations can seem a bit too good to be true, even if the phenomenon of ensemble perception (i.e., human performance on ensemble tasks) is agreed upon. What mechanisms explain ensemble task performance?

Modeling the Set-size Effect

It has been found repeatedly that human performance suffers somewhat when comparing sets of unequal size, manifesting as a bias to choose the set with more items as having larger mean size (Chong & Treisman, 2005; Sweeny, Wurnitsch, Gopnik, & Whitney, 2015). We have previously shown that this bias is systematic (Cain, Dobkins, & Vul, 2016), affecting participants' point of subjective equality (Figure 2A). This set-size effect casts doubt on object-based accounts of ensemble perception, but no model yet accounts for it.

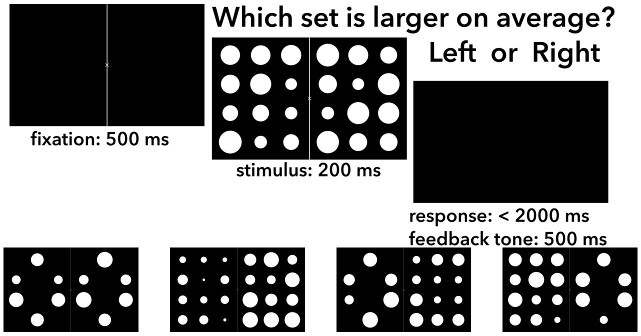

Figure 1: (Top) Classic two-alternative forced choice ensemble perception experiment. (Bottom) Cain, Dobkins, and Vul (2016) compared equal (6&6 and 12&12) versus unequal (6&12 and 12&6) set-sizes.

What representations could account for this set-size bias? Is individuating or registering objects truly necessary? Since high-level object-based accounts cannot explain the set-size bias, we consider low-level and mid-level representations that can be computed on any image (i.e., "image-computable"). In particular, we consider a powerful texture statistics representation (Portilla & Simoncelli, 2000) within an ideal observer framework for modeling the set-size bias. A texture representation makes sense as the mechanism behind ensemble performance, as capacity limitations cannot apply to any representation that does not individuate or register objects.

We found that this texture representation neatly accounts for human performance in the classic larger-mean-size task. This mid-level representation provides the first principled explanation of the set-size bias. Overall, we present the first image-computable approach to modeling ensemble perception, and provide hints that these mid-level statistical representations—rather than object-based accounts—underlie high-level ensemble judgments and support rapid scene perception.

Results

Object-based, high-level accounts predict that no human perceptual bias is induced by unequal set-size comparisons on the larger-mean-circle-size task. However, participants perceived sets with more items as having larger mean, which shifted their psychometric curves. We measure this bias using the point of subjective equality (PSE), the log mean-ratio at which the sets are perceived to have equal mean diameter. Can low-level or mid-level representations better account for human PSE shifts in the unequal condition? We consider a low-level luminance representation and a mid-level texture statistics representation. For each of these candidates, we construct ideal observer models, generate trial-by-trial predictions, and fit psychometric functions separately for equal and unequal conditions to extract the model's PSE shift. We evaluated models by comparing their PSE shifts to those of 20 participants. This allowed us to determine which of these non-object-based representations yielded PSE shifts consistent with human set-size bias.

Low-level Luminance Representation

Could luminance in each visual field predict mean size judgments? We computed luminance in each visual field (i.e., the total area of the items in each ensemble). Taking the simplest possible model, we selected as larger the side with greater luminance. Of course, for equal set-size trials, this luminance heuristic is perfectly correct. For our unequal set-size trials, the heuristic induces a PSE shift toward the side with more items. However, the size of this bias is too extreme: an ensemble with six items would need to have circles that are twice as large on average than an ensemble with twelve items. Humans PSE shifts are significantly less shifted than this estimate ($t(19) = -10.553, p < .001$). Since this one-parameter luminance model overpredicts the PSE shift induced by unequal set-sizes, we conclude that this is not a good model of human ensemble perception.

Mid-level Texture Representation

Next, we asked if an ideal observer trained on Portilla and Simoncelli (2000) texture statistics can replicate human PSE shifts. This representation compresses an image into about 2500 image statistics which were inspired by physiological findings in primary visual cortex. Since we believe that ensemble perception relies on texture representations, we expect stimuli with discriminably different means to have discriminably different texture statistics. For simplicity, we assumed a single pooling region per visual field, covering each ensemble. This is akin to a visual system with two receptive fields: one computing texture statistics in the left periphery, and one in the right periphery. For each trial, we computed texture statistics in each pooling region, and then subtracted to construct the trial's feature vector: $T_\Delta = T_\mathrm{L} - T_\mathrm{R}$. Then we trained a linear support vector machine (SVM) to perform the mean size task, using the difference-of-statistics features, on the 48 easiest trials (Equal condition only, 2-to-1 ratio of mean circle size). We then used the linear classifer to generate predictions for the remaining 768 held-out trials (Equal and Unequal). Using the same analysis workflow as for the human data, we fit a cumulative gaussian to obtain the ideal observer model's own psychometric curves. As seen in Figure 2B, this ideal observer's PSE shift is similar to the human group-level PSE shift ($t(19) = -1.285, p > .05$). Therefore, the texture representation—without any parameter tuning or fitting—contains sufficient information to do ensemble tasks. Crucially, our ideal observer reproduces the set-size bias using a linear readout of texture features.

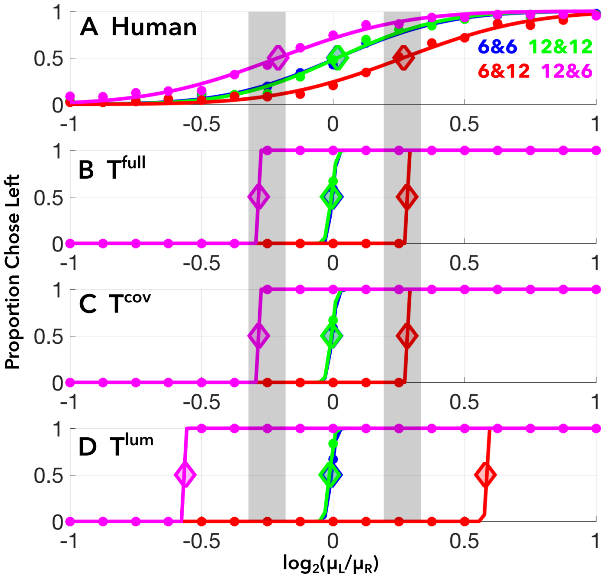

Figure 2: Psychometric curves for each trial type, with the point of subjective equality (PSE) marked by large open diamonds. (A) group-level human data (20 participants × 1600 trials), (B,C) texture-based ideal observers, and (D) a luminance-based ideal observer. (A) Set-size effect: unequal set-sizes conditions have non-zero PSE shifts, gray bars. (B) T^{full} uses the full Portilla and Simoncelli (2000) texture model. (C) T^{cov} retains second-order texture statistics. (D) T^{lum} uses only multiscale luminance image statistics.

Texture Statistics in "Higher-Order" Ensembles

Could texture statistics plausibly explain high-level ensemble perception? Since it is image-computable, a texture representation could subserve both "low-level" and "high-level" ensemble tasks. If synthetic images that are generated from the statistics of an ensemble of faces do not allow emotion extraction, then texture statistics may not underlie processing of high-level ensembles.

To demonstrate the rich information captured by texture statistics, we extracted Portilla & Simoncelli statistics from two grids of faces used by Haberman, Lee, and Whitney (2015). One ensemble is more happy (and has lower emotion variance) than the other ensemble. Figure 3 shows texture syntheses based on those original images.[1] Upon inspection, the synthesis based on the happier ensemble appears happier. This proof-of-concept demonstrates that texture statistics, even in a single pooling region, are sufficient to support ensemble emotion perception.

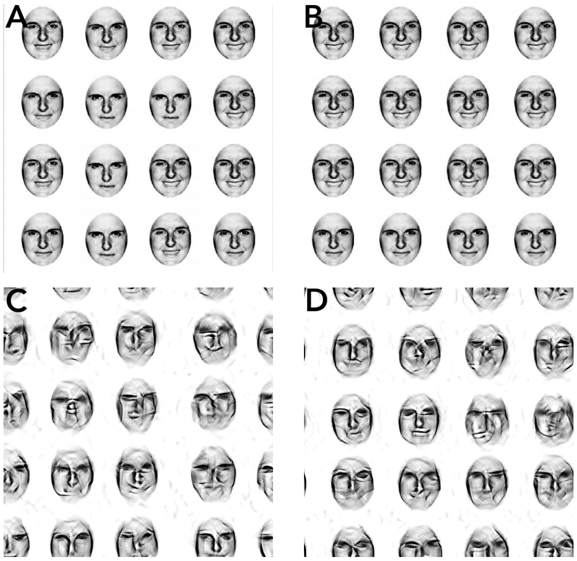

Figure 3: Texture representations of face ensembles preserve emotion information. The crowd in (A) is less happy on average than the crowd in (B). This is also apparent in corresponding texture syntheses: (C) was constrained to match the Portilla and Simoncelli (2000) statistics of (A), and (D) was constrained to match (B). [(A) was taken from Figure 1 in Haberman, Lee, and Whitney (2015), used with permission. (B) was created by repeating the last column of (A) four times.]

Reduced Texture Models

Returning to the mean circle size task, would lower-order statistical representations suffice for the ideal observer model? We trained two new SVMs: one with first- and second-order statistics only, T^{cov} (about 400 features), and one with multiscale luminance only, T^{lum} (16 features). We generated predictions from both models, and extracted each model's PSE shift. To our surprise, T^{cov} gave identical predictions and PSE shift as the full higher-order texture statistics (see Figure 2C). T^{lum}'s PSE shift was too big $(t(19) = -13.554, p < .001)$, overestimating human bias: compare Figures 2A and 2D. Therefore, second-order texture statistics are sufficient to explain human performance on our task, and multiscale luminance measurements are insufficient. To be clear, we believe that the true encoding mechanism—full-field texture statistics in many pooling regions that scale with eccentricity, like in Balas, Nakano, and Rosenholtz (2009) and Freeman and Simoncelli (2011)—includes the higher-order statistics for each of many pooling regions. But the dimensionality of the decoding mechanism necessary to perform our particular task happens to be at most second-order.

Discussion

We found that a texture statistics representation provided sufficient information for a linear classifier ideal observer to replicate human biases on a mean size comparison task. Our results indicate that the puzzling set-size bias is a natural consequence of using a texture statistics representation. Because texture statistics can capture human successes and failures in a classic ensemble task, we conclude that a mid-level texture representation underlies ensemble perception phenomena. This is why performance on ensemble tasks is so much better than object-based representations and their capacity-limited accounts would allow. Our evidence supports the idea that fast scene perception is made possible by texture representations (Rosenholtz, 2015).

Thus far, we have only verified the superiority of texture representations for one ensemble computation, mean circle size. We do not yet know if it will generalize to face stimuli, or to other ensemble computations such as variance, but this work is in progress. Furthermore, our approach, while productive in some contexts, cannot represent arbitrary scenes with large variation in local texture statistics. For this, one would need a richer texture representation in which statistics are computed within many pooling regions (Balas et al., 2009; Freeman & Simoncelli, 2011). Our purpose was to demonstrate the surprising representational power of texture statistics, so we chose a minimal two-pooling-region surrogate. As a consequence, some ensemble phenomena may not be captured well, for instance, outlier rejection. Outlier items produce more local variation in texture statistics, so it remains to be seen how their contributions would be incorporated.

Our approach is image-computable, meaning that this representation can, in principle, be computed on any image. Furthermore, we combine these fixed mid-level statistical encodings with task-dependent decoding. This two stage encoding-decoding approach can be applied flexibly to different tasks, just by training a new classifier, enabling new approaches to rapid scene understanding. For instance, our two-stage approach gives us a different perspective on the question of how domain-general ensemble processing might be. Haberman, Brady, and Alvarez (2015) examined correlations between participants' performance on low- and high-level ensemble tasks, concluding that there must be multiple separate levels of ensemble representation. However, we have dissociated encoding (representation) from decoding (task-specific readout). In our view, there is a single domain-general encoding mechanism: texture representation. The finding of Haberman and colleagues of uncorrelated task performance could

[1] Different hyperparameters were used here than in the mean-circle-size analysis. A suitable size of the steerable pyramid was determined for the less-happy image, and these same hyperparameters were applied to the more-happy image.

be due to independent readout mechanisms for color, orientation, face identity, etc. Because the encoding is not hierarchical, the readout mechanisms need not be correlated.

We have presented an account of ensemble perception that already has the properties needed for a parallel preattentive pathway. Texture computation is not subject to cognitive capacity limits—it is merely a biologically-plausible (Portilla & Simoncelli, 2000; Balas et al., 2009) image compression scheme. A texture-based mechanism behind ensemble perception may seem surprising, since these computations do not operate on objects or even parts of objects (Whitney & Yamanashi Leib, 2018). But texture statistics can be computed for any portion of a scene, even if it happens to contain objects (Rosenholtz, 2015). This information represents the visual input well under the same conditions as required for ensemble perception (e.g., items are indistinct; see Ariely (2001)). Now we have evidence that ensemble phenomena could be readily explained by rapidly-computed mid-level statistics, clarifying how ensemble perception works.

Furthermore, we have provided a parsimonious explanation for a long-standing mystery. The set-size effect has been ignored in the literature, despite being inescapable in experimental settings. To our knowledge, it has only been reported for mean circle size comparisons. But, by definition, it must contaminate other response modes (e.g., Method of Adjustment), and it could in principle arise in high-level ensemble tasks as well. Crucially, we were able to address this problem without relying on the concept of objects, and without denying the existence of rapid perceptual faculties. While others debate capacity limits in order to place all ensemble representation findings within the realm of object representation, we instead interpret them as special cases of texture representation. Based on our success in modeling the set-size bias via texture representations, we suggest that the broader goal of understanding scene perception may also require a different paradigm than the traditional object-centric view. Indeed, low- (Schyns & Oliva, 1994) and mid-level (Renninger & Malik, 2004) image statistics support sub-100ms scene categorization. They are also thought to be computed relatively early in cortical processing (Okazawa, Tajima, & Komatsu, 2015; Ziemba, Freeman, Movshon, & Simoncelli, 2016) and perhaps in the retina (Gollisch & Meister, 2008). Rather than seeing these as nuisance features to be experimentally controlled, we seek a more detailed mapping: *how* do these image features contribute to perceptual decisions? We see this as one of the frontiers of cognitive computational neuroscience, where the time is ripe for integrating theory, physiology, and behavior to gain a new understanding of scene perception.

References

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

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

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13.

Cain, S. S., Dobkins, K., & Vul, E. (2016). Texture properties bias ensemble size judgments. *Journal of Vision*, *16*(12), 54–54.

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

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

Freeman, J. & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201.

Gollisch, T. & Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, *319*(5866), 1108–1111.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446.

Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, *15*(4), 1–11.

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences*, *112*(4), E351–E360.

Portilla, J. & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

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

Renninger, L. W. & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*(19), 2301–2311.

Rosenholtz, R. (2015). Texture perception. In J. Wagemans (Ed.), *The Oxford Handbook of Perceptual Organization*. Oxford, U.K.: Oxford University Press.

Schyns, P. G. & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195–200.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental Science*, *18*(4), 556–568.

Whitney, D. & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*(1), 105–129.

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences*, *113*(22), E3140–E3149.