Large-scale Characterization of a Universal and Compact Visual Perceptual Space

Ha Hong[1,2]*, Ethan Solomon[1,3]*, Dan Yamins[1]*, and James J. DiCarlo[1] (*equal contribution)
[1]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, MIT, Cambridge, MA. [2]Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, MA. [3]Present address: Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

Many visual psychophysics experiments hypothesize a perceptual space whose axes encode key features on which judgements are made. We characterized human perceptual space for an image set with 64,000 images of 64 objects, shown with differing positions, sizes, poses, and backgrounds. We performed online psychophysical experiments involving 703 observers, obtaining confusion matrices for 2016 two-alternative forced-choice (2AFC) pairwise object identification tasks. Generalizing Getty (1979) and Ashby (1991), we hypothesized that: (1) for each object, multiple image instances sample a Gaussian pointcloud in perceptual space; and (2) identity decisions could be modeled with distance-based classifiers applied to these Gaussian clouds. The dimension, locations, and spreads of the Gaussians were then chosen to be consistent with experimentally observed confusions.

The resulting representation almost perfectly predicts confusions on held-out images and is stable to the addition of new objects. It also generalizes to visual tasks well beyond the original 2AFC task, predicting human responses for: (1) 8-way AFC recognition tasks, (2) ratings of objects with adjectives (e.g. “rectangular”, "cuddly"), and (3) subjective similarity judgements between objects. The representation scales efficiently with object number, requiring ~47 dimensions to encode 10,000 distinct objects (Biederman, 1987).

Given the scale and precision of the dataset, we were able to make direct comparisons to neural data. We found that the object layout in the inferred human perceptual space correlated highly with those from the neural population representation measured in Inferior Temporal (IT) cortex. Taken together, these results suggest that the human brain produces a visual perceptual space that is both universal (underlies behavior for many different tasks) and compact (requires few dimensions to represent many entities). We anticipate extensions of this method will further bridge neural and perceptual observations, and help characterize how interventions (e.g., learning and attention) modify perceptual representations.

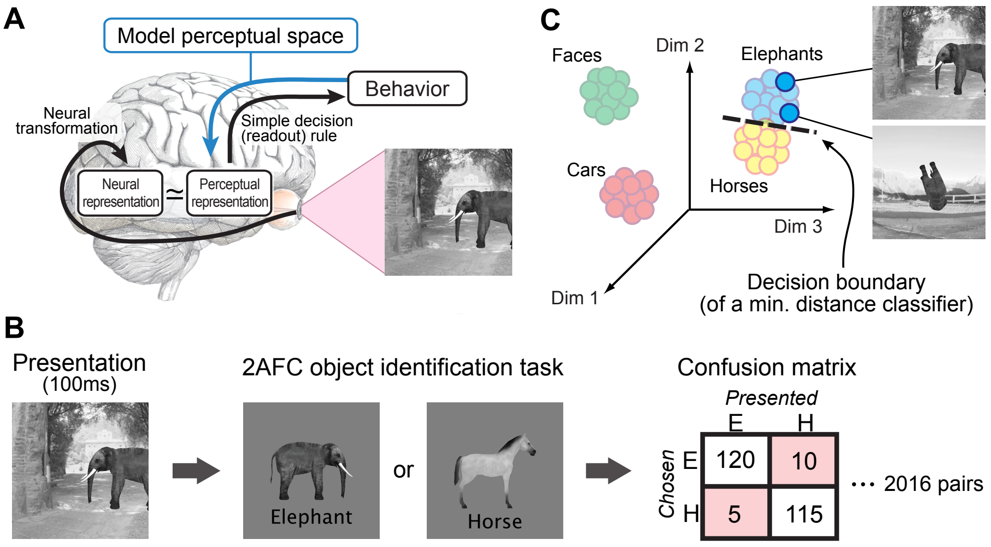

Fig. 1. Construction of the perceptual space from large-scale psychophysics. (**A**) Inference of perceptual space (blue arrow) attempts to invert observed behavioral results to extract the underlying perceptual representation. (**B**) Human observers (*N* = 703) were tested on binary recognition problems arising from pairwise comparisons of 64 objects in a 2AFC paradigm. 2016 binary confusion matrices were obtained. (**C**) The perceptual space inversion process relies on two basic assumptions: (1) that the various images corresponding to a single object form a Gaussian cloud in perceptual space and (2) behavioral choices are generated by applying a simple distance-based classifier to the collection of Gaussian clouds. The overall dimension of, and the 64 centers and 64 spreads of the Gaussian clouds are chosen so that they simultaneously generate the 2016 confusion matrices that are consistent with the observed human confusion matrices. Asymmetries in binary confusion matrix correspond to the situation in which the Gaussian cloud for one of the classes has a larger spread in perceptual space than the other.

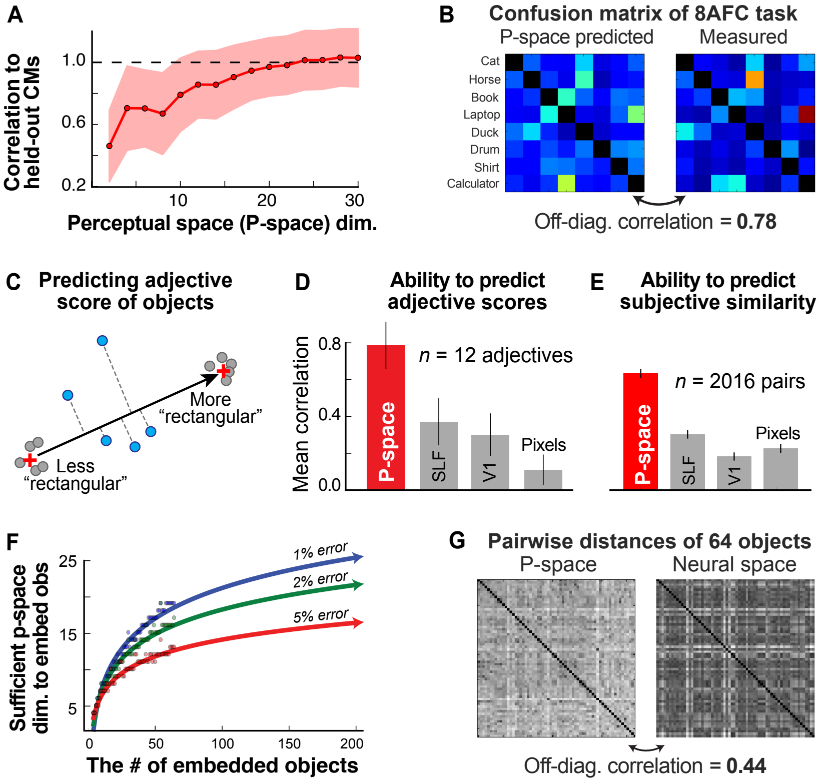

Fig. 2. Validation and generalization of the perceptual space model. (**A**) At sufficiently high dimensionality, the inferred perceptual space achieved nearly perfect results for confusions of held-out object pairs. (**B**) The perceptual space inferred from 2AFC data generalized to 8-way AFC tasks. Shown is one example of an 8-way object identification task for which human data was collected. Over 20 such 8-way tasks, the mean actual/predicted correlation was 0.80. (**C**) The 2AFC-based perceptual space also generalized to predicting adjectival ratings for the objects. For a variety of adjectives (including shape adjectives, such as "rectangular" or "globular", texture adjectives such as "striped", and more semantic adjectives like "cuddly"), we measured human subjective ratings for each object using a 0-100 sliding scale bar (0 for "least rectangular" to 100 for "most rectangular"). To produce model predictions for each adjective, we identified two "anchor points" corresponding to the most and least highly rated objects for that adjective. For the remaining objects, we projected the object center positions in the 2AFC-based perceptual space to the line connecting the two anchor points, using the position along that line as the predicted adjectival rating. We then measured the correlation between the actual human ratings and the predicted ratings for these remaining objects. (**D**) Bar height represents the Pearson correlation between predicted and actual adjective scores (for *N*=140 human subjects), averaged over the 12 tested adjectives. Error-bars correspond to the standard error due to adjective variation. "P-space" bar is the result for 2AFC-based perceptual space model; "SLF" and "V1" bars are control models with visual features extracted from images by either the HMAX-SLF model (Mutch, 2008) or an optimized V1-like model (Pinto, 2008); the "Pixels" bar represents the trivial 90000-pixel "model" from 300x300 images. (**E**) We also tested generalization to pairwise subjective similarity. Human subjects (*N*=157) rated 2016 object pairs on a scale from 0 to 100, with 0 as "least similar" and 100 as "most similar". Similarity predictions were generated from the 2AFC-based perceptual space by using distance between object centers. Bar height represents Pearson correlation between actual and predicted similarities. (**F**) For each level of desired accuracy, we characterized the sufficient number of dimensions necessary to embed objects, as a function of the number of objects to be embedded. We project that ~47 dimensions would be required to achieve 1% error for 10,000 objects. (**G**) Comparison of the 2-AFC perceptual space to neural data collected on the same images. Left panel: pairwise distances between object centers predicted by the perceptual space. Right panel: for each image, neural responses of 141 Inferior Temporal cortex sites were obtained using array electrophysiology methods. Neural responses were averaged over images of each of the 64 objects, and 2016 pairwise distances were computed between these. The Spearman rank correlation between the off diagonal elements in the neural and perceptual space distances was 0.44.

References

Ashby, F. G. and Lee, W. W. Predicting Similarity and Categorization From Identification. *J Exp Psychol Gen*, 1991.
Getty, D. J., Swets, J. A., Swets, J. B., and Green, D. M. On the prediction of confusion matrices from similarity judgments. *Percept Psychophys*, 1979.
Mutch, J. and Lowe, D. G. Object class recognition and localization using sparse features with limited receptive fieds. *IJCV*, 2008.
Pinto, N., Cox, D. D., and DiCarlo, J. J. Why is Real-World Visual Object Recognition Hard. PLoS Comp. Bio., 2008.