A hierarchical model of perspective-invariant scene identification

Emily Y Ko, Joel Z Leibo, Tomaso Poggio

McGOVERN INSTITUTE FOR BRAIN RESEARCH AT MIT

Center for Biological & Computational Learning

Introduction

In the object recognition literature it is common to study within-class identification tasks---e.g., telling one person's face from another despite variations in their appearance (position, scale, viewpoint). Studies of scene perception, by contrast, typically focus on categorization---e.g., telling beaches from deserts. We conjecture that the focus on transformation invariance found in the object-recognition literature can be fruitfully applied to the study of scene ***identification*** as well.

Methods

Stimuli

We used 3D graphics software (Blender) to generate images of long hallways. The transformation that occurs when walking down such a hallway is a *perspective transformation*. We rendered images of each hallway at 31 different "depths" (perspectives). In all cases the vanishing point was at the end of the hallway. For both modeling and psychophysics experiments the task was to identify a specific hallway despite changes in perspective between the reference (training) image and query (test) image. Hallways were distinguished from one another by the locations of bumps on their walls.

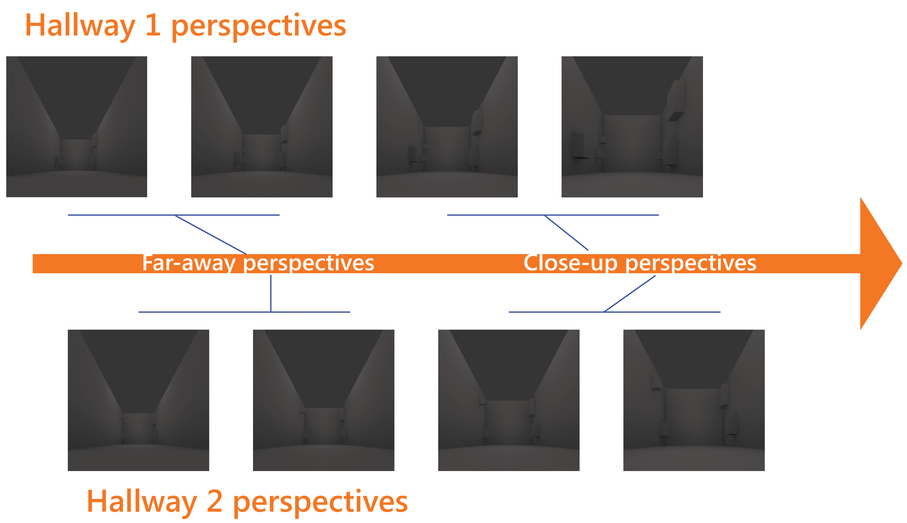
Psychophysics

Same-different task: 4 subjects were presented with a fixation cross for 1 second which then disappeared. Then the reference image was shown for 48ms and followed by a 16ms random dot mask. The query image was then presented another 800ms after the mask disappeared. Another 16ms random dot mask followed the query image. After the query image disappeared, subjects indicated whether or not the reference and query images depicted the same hallway (possibly from a different perspective). Trials were balanced so that the probability of the correct response being `same' was .5. We varied the amount of time to present the query image (24,36,48,60 and 500ms). All trial types were intermixed.

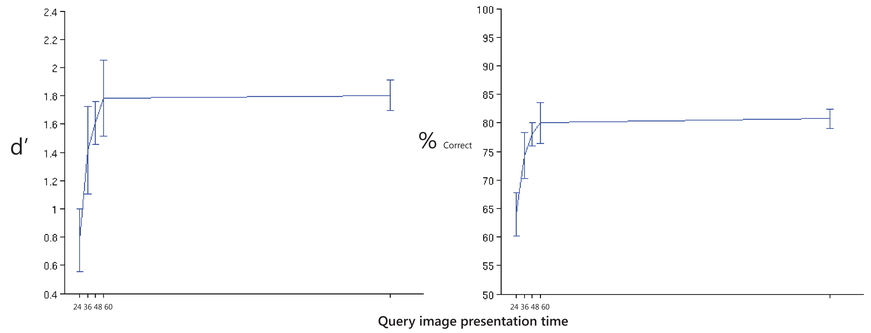
Modeling

We tested a hierarchical model of object recognition modified from Serre et al. 2007 (architecture A) as well as two variants that pool over changes in perspective of previously-viewed scenes (architectures B and C). We recently described a model that pools over 3D rotations of familiar faces (Leibo et al. 2011) that works analogously to the models in this study.

For architectures B and C, we collect templates of "familiar" hallways, previously viewed from all 31 perspectives. In the penultimate layer, each cell signals the similarity of its input to its stored template (by a normalized dot product or a Gaussian radial basis function). Each cell in the final layer pools (with a max function) over all the perspectives of a single template hallway. The output of the model's final layer is a vector of similarities of the input---a novel hallway--- to a set of template hallways. The output is invariant to perspective changes of the input insofar as the template hallways transform in the same way as the input. We used a nearest-neighbor classifier to obtain the final performance measure (AUC).

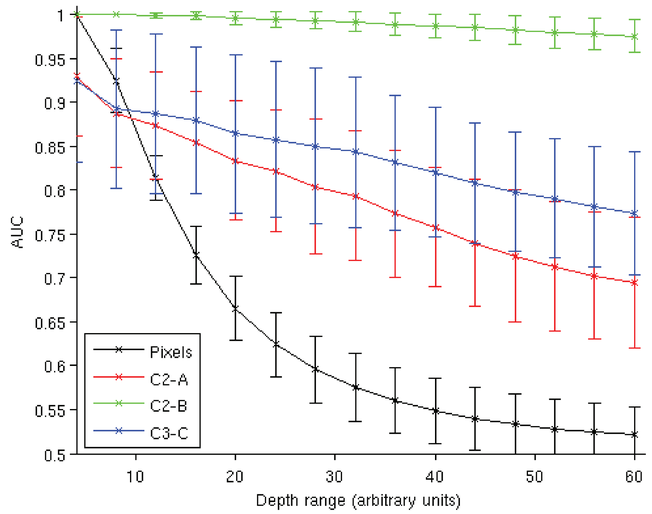


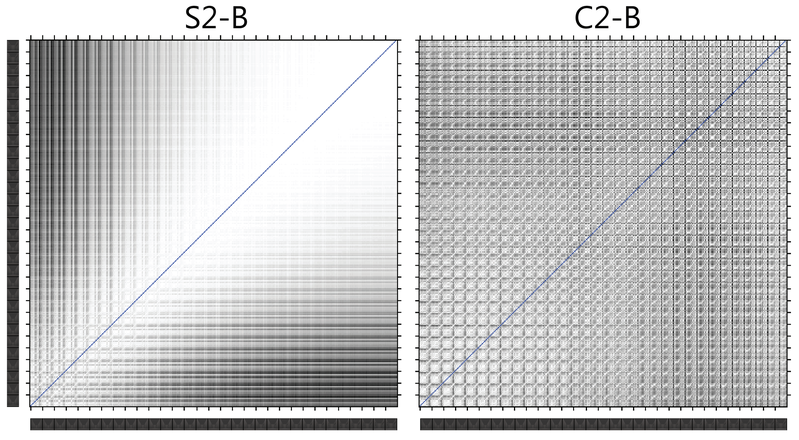
Psychophysics results



Error bars indicate the standard error of the mean (n = 4). Identification performance reaches a plateau ~60ms after query image presentation.

Modeling results





Simulation details: Architecture A used 10 templates tuned to patches of natural scenes. The final layer was translation invariant and (somewhat) scale invariant. Architecture B used 10 templates tuned to appearances of hallway images drawn from a separate set of images, different from the testing images. The final layer pooled over perspective. Architecture C used both sets of templates, the penultimate layer pooled over translation and the final layer pooled over perspective. Top figure: Performance as a function of the severity of the perspective transformations allowed in the testing set. Bottom figure: Similarity matrices (correlation). Images were ordered by perspective. The white bands off the diagonal in the matrix for C2-B indicate that this layer supports perspective-invariant scene-identification. The coarse block structure of the S2-B matrix shows that this layer could support depth-estimation from perspective cues, though not with very high accuracy.

Acknowledgments

We thank Heejung Kim and Leyla Isik for last-minute help preparing this poster. Funding for this work was provided by DARPA, NSF and IIT.

References

Serre T., Wolf L., Bileschi S., Riesenhubuer M., Poggio T. (2007). Robust object-recognition with cortex-like mechanisms. IEEE PAMI.

Leibo J. Z., Mutch J., Poggio T. (2011). Why the brain separates face recognition from object recognition. NIPS.