Pose-estimation and pose-invariant recognition with an extended hierarchical model of the ventral stream

McGOVERN INSTITUTE FOR BRAIN RESEARCH AT MIT

Heejung Kim, Joel Z Leibo, Tomaso Poggio

Center for Biological & Computational Learning

Introduction

When learning to recognize a novel body shape, e.g., a panda bear, we are not misled by changes in its pose. A "jumping panda bear" is readily recognized, despite having no prior visual experience with the conjunction of these concepts. Likewise, a novel pose can be estimated in an invariant way, with respect to the actor's body shape. These body and pose recognition tasks require invariance to non-generic transformations (Leibo et al. 2011) that previous models of the ventral stream do not have. We show that including biologically-plausible, class-specific mechanisms associating previously viewed actors in a range of poses enables a hierarchical model of object recognition to account for this human capability. These associations could be acquired in an unsupervised manner from past experience.

Methods

Stimuli

We used 3D graphics software (Blender, DAZ 3D Studio) to generate images of human bodies in various poses. We generated 40 bodies, each one had distinct body features such as male/femaleness, fatness, muscularity, and limb proportion.

We rendered each of the 40 bodies in 32 different poses. Each pose was defined as a specific body configuration viewed from certain angle. All the poses were natural and frequently appearing in natural vision e.g., waving, running, leaning, and clinging. With 40 bodies and 32 poses, there were 1280 total images.

Model

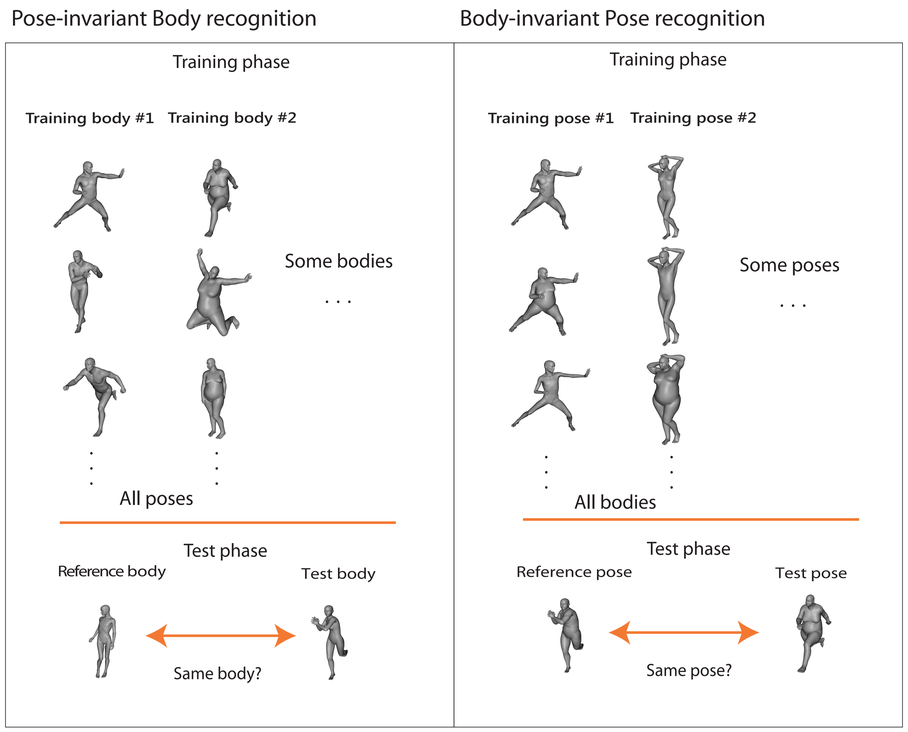
We tested a hierarchical model of object recognition (modified from Serre et al. 2007) as well as class-specific variants, which pool over pose or body transformations. We recently described a model that pools over 3D rotations of familiar faces (Leibo et al. 2011) that works analogously to the models in this study.

S cells in the HMAX model compute the normalized dot product (or a Gaussian radial basis function) with their input and a stored template. C cells pool over specific sets of S cells, such as all the S cells tuned to the same body under different poses. The final output of the model---the signature----can be regarded as a vector of similarities of the input image to the templates. If the testing objects transform similarly to the training objects, then the signature will be approximately invariant to those transformations.

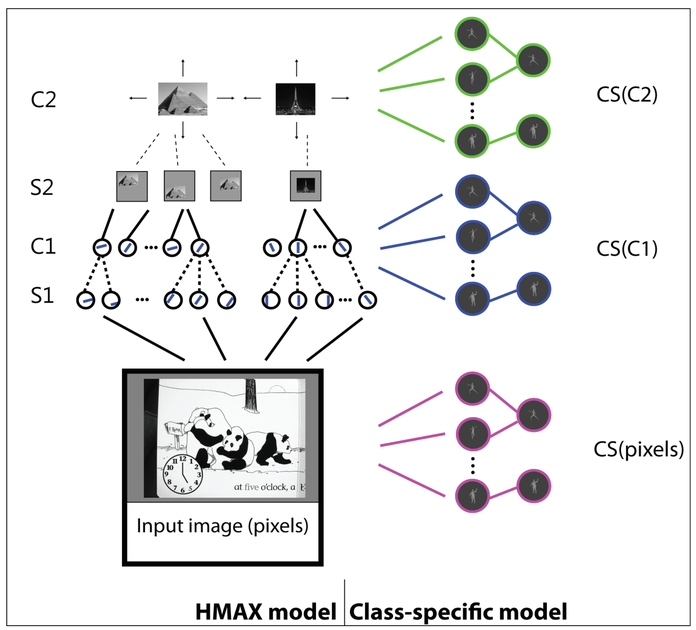
Three class-specific models were built based on the raw pixel values of the input image, the HMAX C1 vector (modeling complex cells in primary visual cortex), and the HMAX C2 vector (modeling later visual areas, e.g., V4 or IT).

We modeled a same-different psychophysical test of initial invariance. A nearest-neighbor classifier ranked the similarity of the signatures of a reference image to the signatures of all the images in a testing set containing both the reference object under various transformations, and distractor objects under the same transformations. None of the images used in the testing phase ever appear in the training phase. The pose-invariant body-recognition task required the classifier to rank images of the same body as most similar to one another despite variation in its pose. The body-invariant pose-recognition task required the classifier to rank images of the same pose as most similar to one another despite variation in the body.

Recognition tasks for the models

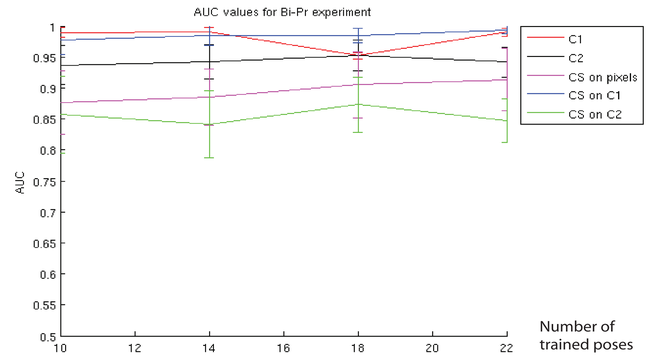


Model architecture

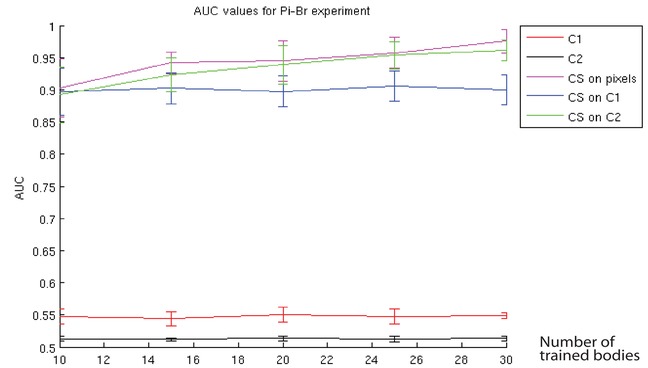


Modeling results

Bi-Pr task



Pi-Br task



For all simulations, we resampled training and test sets 10 times. The reported results are the means and standard deviations across these cross-validation splits. Each test set contained 10 objects. We showed that simple methods (e.g., HMAX's C1 layer) are sufficient to achieve good pose-recognition (at least, in our simplified setting). However, on the pose-invariant body-recognition task, the HMAX models we tested perform almost at chance. The addition of our class-specific mechanism significantly improves performance on this difficult task.

We previously conjectured that modularity in the ventral stream arises because of the need to discount class-specific transformations (Leibo et al. 2011). On this poster we showed that the task of pose-invariant body-recognition is not accounted for by standard models of the ventral stream, and that the addition of cells that pool over pose transformations makes the task relatively easy. This observation suggests that the underlying computational reason that the brain separates the processing of images of bodies from the processing of other images is the need to recognize specific people invariantly to their pose (cf. Poggio 2011, Downing et al. 2001).

Acknowledgments

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.

Poggio, T. (sections with J. Mutch, J.Z. Leibo and L. Rosasco), The Computational Magic of the Ventral Stream: Towards a Theory, Nature Precedings, doi:10.1038/npre.2011.6117.1 July 16, 2011

Downing, Paul E. Jiang, Yuhong, Shuman, Miles, Kanwisher, Nancy (2001) A cortical area selective for visual processing of the human body. Science 28 September 2001