How can cells in the anterior medial face patch be viewpoint invariant?

Joel Z Leibo, Jim Mutch, Tomaso Poggio

McGOVERN INSTITUTE FOR BRAIN RESEARCH AT MIT

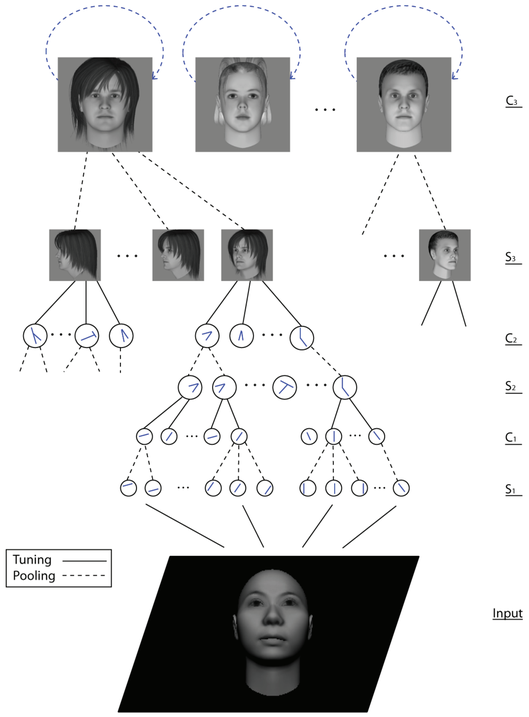
Center for Biological & Computational Learning

Introduction

In a recent paper, Freiwald and Tsao [1] found evidence that the responses of cells in the macaque anterior medial (AM) face patch are invariant to significant changes in viewpoint. The monkey subjects had no prior experience with the individuals depicted in the stimuli and were never given an opportunity to view the same individual from different viewpoints sequentially. These results cannot be explained by a mechanism based only on the temporal association of experienced views.

Affine transformations such as translation and scaling can be computed independently of the particular object and are thus generic. On the other hand, non-generic transformations including changes in viewpoint and illumination depend on the 3D structure or material properties of the objects [2].

Within a restricted class of objects having a similar 3D structure - like faces - it may be possible to have model cells which are identity-specific and invariant to non-generic transformations using class-specific features [3]. Freiwald and Tsao's finding of a population of cells that appear to be invariant to changes in viewpoint could constitute evidence that the brain is employing face-specific features for this purpose.



We implemented a biologically-plausible model of the visual system (modified from [4]). This model converts images into a feature representation via a series of processing stages referred to as layers. In order, the layers of the basic model were: S1 --> C1 --> S2 --> C2. For some simulations we added two additional layers: S3 --> C3. In our model, an object presented at a position A will evoke a particular pattern of activity in layer S2. When the object is moved to a new position B, the pattern of activity in layer S2 will change accordingly. However, this translation will leave the pattern in the C2 layer unaffected.

At the first stage of processing, the S1 units compute the responses of Gabor filters (at 4 orientations) with the image's (greyscale) pixel representation. The S1 units model the response of Hubel and Wiesel's V1 simple cells. At the next step of processing, C1 units pool over a set of S1 units in a local spatial region and output the single maximum response over their inputs. Thus a C1 unit will have a preferred Gabor orientation but will respond invariantly over some changes in the stimulus' position. We regard the C1 units as modeling Hubel and Wiesel's V1 complex cells. Each layer labeled S is computing a selectivity-increasing operation while the C layers perform invariance-increasing operations.

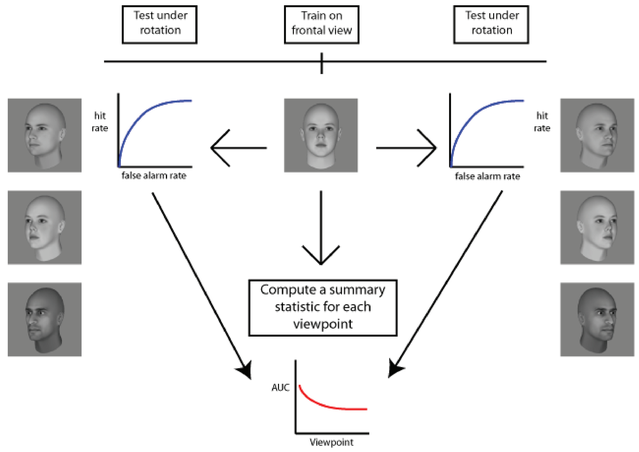
The S2 units employ a template-matching operation (Gaussian radial basis function) to detect features of intermediate complexity. The preferred features of S2 units are preprocessed versions small patches extracted from natural images. Here preprocessed means that the template-matching operation is performed on the output of the previous layer and so is encoded in the pattern of activity of a set of C1 units.

We chose the preferred features of S2 units by randomly sampling patches from a set of natural images and storing them (C1-encoded) in the S2 weights. So the response of an S2 unit to a new image can be thought of as the similarity of the input to a previously encountered template image. In the final layer, a C2 unit pools over all S2 units with the same preferred feature. Thus the pattern of activity over all C2 units is a vector of distances to previously encountered template images.

The S3 and C3 layers implement class-specific features for non-generic transformations. For faces and viewpoint invariance, each S3 cell matches the preprocessed (in C2 units) input to a stored face template at a particular viewpoint. The response of an S3 unit can be thought of as encoding the similarity of the input to a previously encountered face at a particular viewpoint. The S3 units are analogous to the "view-tuned units" in [5,6]. The cells in the C3 layer pool over the S3 cells preferring each viewpoint of the same face. Thus the pattern of activity over all the C3 units is a vector of similarities to previously encountered template faces invariantly of viewpoint. We can use any class of objects and any transformation to produce S3/C3 layers specialized for that object class and transformation.

Simulation Procedure

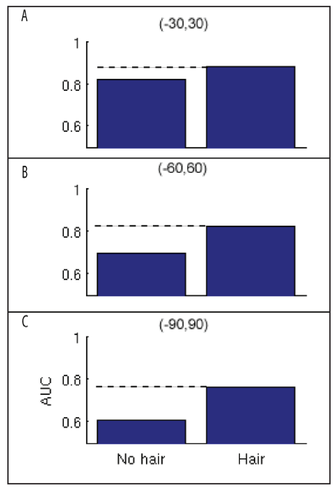
We simulate a same-different task by presenting images that contain either the target object or a distractor object on a neutral background. The object may appear at a range of viewpoints. We trained a simple correlation-based classifier on a single view of the target object presented frontally (rotation of 0). The task was for the classifier to rank transformed versions of the target object as being more similar to the trained view than to distractors and rotated distractors. We tested the model with increasing ranges over which we allowed targets and distractors to rotate. The range of viewpoints tested in each condition is plotted as the abscissa in the following figures. The classifier only used the output of the model (C2, S3 or C3) and no other information.



Can non-shape cues account for viewpoint-invariant face identification?

Under natural conditions we are rarely met with the task of recognizing faces by shape information alone. Usually faces come with a host of extra identifying attributes that all usefully contribute to recognition. Skin color and hair style both provide information that is largely preserved by 3D rotation. In order to mimic natural experience as closely as possible, Freiwald and Tsao presented stimuli that included all these extra identifying features. We demonstrate that a computational model of object recognition employing no specialized viewpoint-invariant mechanisms performs very well under these circumstances and accuracy is substantially boosted when tested on faces with hair.

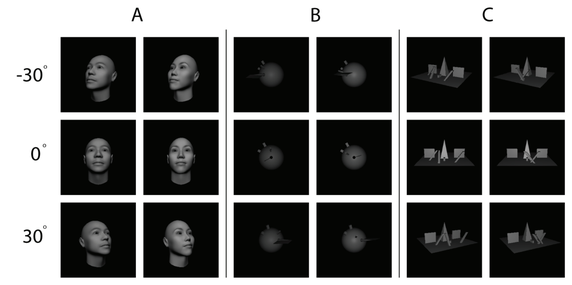
It is likely that at least some of the effect observed in AM can be attributed to these cues.

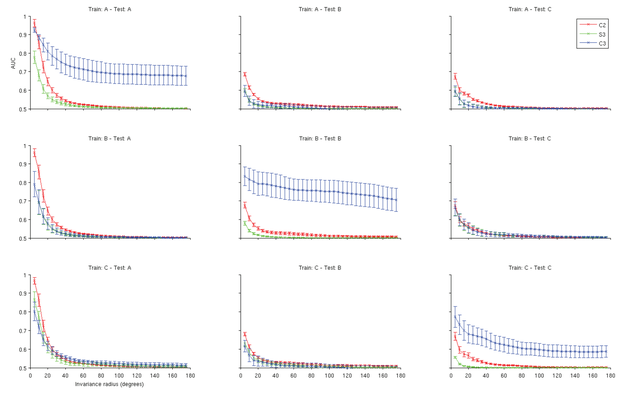


For these simulations we compute the area under the ROC curve for a simple classifier that computes the correlation of C2 model unit responses evoked by the target presented in the frontal view to responses evoked by targets and distractors presented at other views. In panel A: views between -30 and 30 degrees were used. B: -60 and 60 degrees. C: Targets and distractors were presented at views between -90 and 90 degrees away from the frontal view. Panel C is the full 180 degrees of variation from a leftward facing profile to a rightward facing profile. In all these simulations there were 25 position/scale (but not viewpoint) invariant templates. Stimuli were produced using FaceGen (face modeling software by Singular Inversions).

How can class-specific mechanisms contribute to viewpoint invariant identification?

Faces, as a class transform under 3D rotation in similar enough ways that it is possible to use previously viewed example faces to learn a general model of how all faces rotate. Novel faces can be encoded relative to these previously encountered "template" faces and thus recognized despite substantial changes in viewpoint. Different object classes transform differently under 3D rotation, thus these features must be class-specific.





Stimuli above: Class A consists of faces produced using FaceGen (Singular Inversions). Class B is a set of synthetic objects produced using Blender (Stichting Blender Foundation). Each object in this class has a central spike protruding from a sphere and two bumps always in the same location on top of the sphere. Individual objects differ from one another by the direction in which another protusion comes off of the central spike and the location/direction of an additional protrusion. As with faces, there is very little information available to disambiguate individuals available from views of the objects' back (180 degree rotation away from the front view). Class C is another set of synthetic objects produced using Blender. Each object in this class has a central pyramid on a flat plane and two walls on either side. Additionally, all objects in this class have a small bump in front of the central pyramid. Individual objects differ in the location and slant of three additional bumps. As with the other two classes, there is very little information to disambiguate individuals from views of the back of the object.

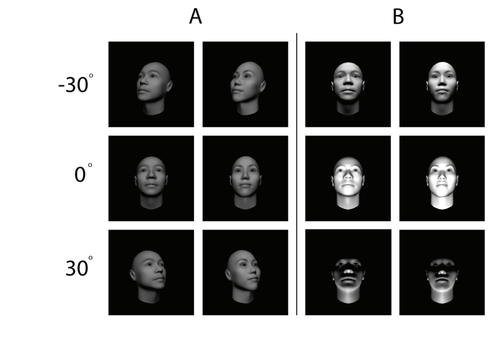
Simulation above: S3 cells were tuned to specific views of training objects. C3 cells pool over all views of the same template object. The above figure was produced using templates tuned to objects in one class at a time. Each panel shows the results of a simulation using templates tuned to objects of one stimulus class and tested on another class. The panels along the diagonal show that good viewpoint tolerance is achieved when encoding novel objects relative to templates from the same class. The panels off the diagonal show that these templates really are class-specific. They improve performance only when testing with the class of objects for which they are specialized.

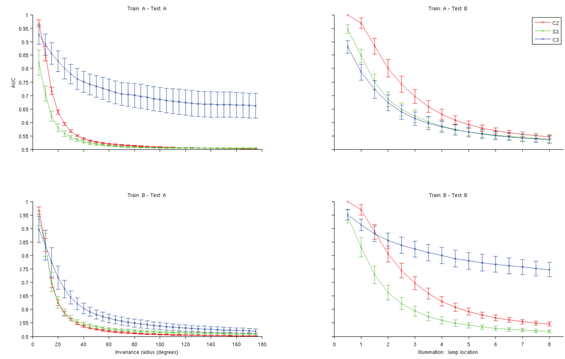
Simulation above details: These simulations used 2000 translation and scaling invariant C2 units tuned to patches of natural images. Error bars (+/- one standard deviation) show the results of cross validation by randomly choosing a set of example images to use for producing S3/C3 templates and testing on the rest of the images. The above simulations used 710 S3 units (10 exemplar objects and 71 views) and 10 C3 units.

Stimuli right: Class A consists of faces produced using FaceGen and rendered at each viewpoint (increments of 5 degrees). Class B consists of the frontal views of the same faces. In this class of stimuli, the location of the lamp illuminating the object was varied. The lamp only moved vertically relative to the face.

Simulation right: Consistent with psychophysical results [7], the extreme illumination changes produced faces that were difficult to recognize. S3 cells were tuned to specific lighting conditions (lamp locations). C3 cells pool over all lighting conditions for the same template face. The figure was produced using templates for either viewpoint or illumination. These templates are transform-specific. C3 cells that pool over illumination changes are not useful when testing on viewpoint and cells that pool over viewpoints are not useful when testing on illumination changes.

Simulation right details: These simulations also used 2000 translation and scaling invariant C2 units tuned to patches of natural images. Error bars (+/- one standard deviation) show the results of cross validation by randomly choosing a set of example images to use in producing S3/C3 templates and testing on the rest. The above simulations used 710 S3 templates for rotation and 320 templates for illumination. The final C3 vector had 10 elements in both cases.





Summary and Discussion

The observation of a viewpoint tolerant population of cells in face patch AM [1] can be partially attributed to the use of stimuli with low-level viewpoint independent diagnostic features such as skin color and hair style. We also described a method by which class-specific features could be used to identify novel objects invariantly to viewpoint without resorting to new learning for each object.

It is unlikely that the physiology result could be entirely accounted for by low-level image features in the stimulus set. Notably, Freiwald and Tsao showed that an index of population viewpoint tolerance increases to its maximum value roughly 200 ms after the first wave of spikes reaches patch AM. It is possible that this long delay is necessary to implement a class-specific mechanism for viewpoint tolerant identification. Such mechanisms seem to require a class detection step to occur prior to identification using class-specific features. One way to implement such an architecture would be to use feedback from downstream categorization-related areas in prefrontal cortex or elsewhere. Such a scheme could account for the observed delay of viewpoint tolerant information in patch AM.

Acknowledgments

Support:
DARPA (IPTO and DSO), NSF and IIT.

Affiliations:
Center for Biological and Computational Learning, Cambridge MA 02139
McGovern Institute for Brain Research, Cambridge MA 02139
MIT Department of Brain and Cognitive Science, Cambridge MA 02139

References

1. Freiwald, W., Tsao, D., Science. 330, 845 (2010)
2. Leibo, J.Z., Mutch, J., Rosasco, L., Ullman, S., Poggio, T., MIT-CSAIL-TR-2010-061, CBCL-294 (2010)
3. Vetter, T., ICASSP-97. vol. 1, p. 143-146 (1997)
4. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., IEEE Trans. Pattern Anal. Mach. Intell. 29, 411-426 (2007)
5. Poggio, T., Edelman, S., Nature. 343, 6255 (1990)
6. Riesenhuber, M., Poggio, T., Nature Neuroscience. 1097-6256 (1999)
7. C. H. Liu, C. A. Collin, A. M. Burton, and A. Chaurdhuri, B., Vision Res., vol. 39, pp. 4003–4009 (1999)