Learning to discount transformations as the computational goal of visual cortex

Joel Z Leibo jzleibo@mit.edu
Jim Mutch jmutch@mit.edu
Tomaso Poggio tp@ai.mit.edu

Massachusetts Institute of Technology
Department of Brain and Cognitive Sciences

1. Generic transformations and invariance to them

It has been long recognized that a key obstacle to achieving human-level object recognition performance is the problem of invariance [10]. The human visual system excels at factoring out the image transformations that distort object appearance under natural conditions. Models with a cortex-inspired architecture such as HMAX [9, 13] as well as nonbiological convolutional neural networks [5] are invariant to translation (and in some cases scaling) by virtue of their wiring. The transformations to which this approach has been applied so far are *generic transformations*; a single example image of any object contains all the information needed to synthesize a new image of the tranformed object [15]. In a setting in which transformation invariance must be learned from visual experience (such as for a newborn human baby), we have shown that it is possible to learn from little visual experince how to be invariant to the translation of any object [7]. The same argument applies to all generic transformations.

Generic transformations can be "factored out" in recognition tasks (see figure 1) and this is key to good recognition performance. This is the reason underlying recent observations that random features often perform well on computer vision tasks [4, 6, 11, 12].

For simplicity consider a specific example: HMAX. In an architecture such as HMAX, if an input image is encoded in terms of similarity to a set of templates (typically via a dot product operation) and if the encoding is made invariant with respect to a transformation via appropriate pooling in C cells then recognition performance inherits the invariance built into the encoding. The actual templates themselves do not enter the argument: the set of similarities of the input image to the templates need not be high in order to be invariant. From this point of view, the good performance achieved with random features on some vision tasks can largely be attributed to the invariance properties of the architecture.

2. Class-specific transformations and invariance to them

Within the realm of fine-grained subordinate-level identification, there are several non-generic, *class-specific* transformations. For example, faces can undergo changes in expression [2] and words can undergo changes in font. Transformations of viewpoint and illumination are also non-generic since they require knowledge of the object's 3D structure and material properties which is never available in a single example. All these category-specific transformations must be taken into account by a successful within-class identification system.

3. Learning invariance to transformations

We previously showed that approximations to the hard-wired invariance in the HMAX architecture can be learned from natural videos in an unsupervised manner by employing a temporal coherence principle [3, 8, 16]. We had conjectured [7, 12] that invariance for all transformations, including class-specific transformations can be learned in an analogous manner. Since non-generic transformations are different in different object classes, the system that would result from such a learning process must pool over specific transformations of templates. For example, a viewpoint-invariant HMAX system would need to employ different C poolings of possibly the same S templates to represent the invariance to 3D rotation of faces vs. invariance to 3D rotation of chairs because these two object classes fundamentally do not rotate in the same way (knowledge of the 2D images that are evoked by rotating chairs is not any help when the task is to recognize a novel rotated face from a single training image).

We implemented several class-specific modifications of the HMAX model [9, 13]. The features we used are based on patches of images as in [13] and also similar to Bart and Ullman's extended fragments [1] but are not constrained to require similarity between all the templates to-be-pooled. Our approach is also related to Vetter and Poggio's previ-

ous work in graphics where they were able to synthesize images of a novel face at any orientation using a single example image of the novel face and a large library of other (familiar) faces seen at all orientations [2, 15]. Unlike Vetter and Poggio's previous work, the present model, with the goal of categorization rather than graphic synthesis, does not require detailed correspondence between points or regions in the library of familiar faces.

These class-specific modifications of the HMAX model achieve good viewpoint-invariant performance in a one-shot identification task (see figure 2). Performance suffers when a model that is specialized for 3D rotations of one class is tested on identification within a different class. In fact, viewpoint-pooling models employing templates from the wrong class perform worse on viewpoint invariant identification tasks than models that have no particular mechanisms for dealing with viewpoint at all (see figure 3). This is in stark contrast to the generic case where the model is invariant to all classes undergoing the transformation no matter what templates are used.

This approach to within-category identification can be extended to learn invariance to any transformation for which appropriate templates can be obtained from an object of the class undergoing the transformation.

Remarks

- It has not escaped our attention that the use of class specific tranformations by a recognition architecture implies the need for class-specific modules. This is a nice computational argument for the existence of brain modules such as the network of face patches found by Freiwald and Tsao [14].

- Based on arguments such as the ones we have sketched, we conjecture that the choice of the dictionary of S templates is not critical. The critical factor in determining recognition performace on identification and categorization tasks is the equivalence class determined by the C cells' pooling.

- We also conjecture that the hierarchical architecture of visual cortex is determined by the need to learn from experience increasingly complex transformations from translation and scaling to viewpoint, facial expression, and body pose.

References

[1] E. Bart and S. Ullman. Class-based feature matching across unrestricted transformations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(9):1618–1631, 2008. 1

[2] D. Beymer and T. Poggio. Image Representations for Visual Learning. *Science*, 272(5250):1905–1909, 1996. 1, 2

[3] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991. 1

[4] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? *IEEE International Conference on Computer Vision*, pages 2146–2153, 2009. 1

[5] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, pages 255–258, 1995. 1

[6] J. Leibo, J. Mutch, L. Rosasco, S. Ullman, and T. Poggio. Learning Generic Invariances in Object Recognition: Translation and Scale. *MIT-CSAIL-TR-2010-061, CBCL-294*, 2010. 1

[7] J. Leibo, J. Mutch, S. Ullman, and T. Poggio. From primal templates to invariant recognition. *MIT-CSAIL-TR-2010-057, CBCL-293*, 2010. 1

[8] T. Masquelier, T. Serre, S. Thorpe, and T. Poggio. Learning complex cell invariance from natural videos: A plausibility proof. *AI Technical Report #2007-060 CBCL Paper #269*, 2007. 1

[9] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, Nov. 1999. 1

[10] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, 12(2):162–168, 2002. 1

[11] A. Saxe, M. Bhand, Z. Chen, P. W. Koh, B. Suresh, and A. Y. Ng. On random weights and unsupervised feature learning. *NIPS: Workshop on deep learning and unsupervised feature learning*, 2010. 1

[12] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *CBCL Paper #259/AI Memo #2005-036*, 2005. 1

[13] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):411–426, 2007. 1

[14] D. Tsao, W. Freiwald, and R. Tootell. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670, 2006. 2

[15] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):733–742, 2002. 1, 2

[16] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. 1

Learning to discount transformations as the computational goal of visual cortex

Joel Z Leibo, Jim Mutch, and Tomaso Poggio

jzleibo@mit.edu, jmutch@mit.edu, tp@ai.mit.edu

McGovern Institute for Brain Research at MIT

Center for Biological & Computational Learning

Conjecture: Transformation invariance is key

There are generic transformations – which are the same for all object classes– and class-specific transformations. Invariance to both can be acquired by unsupervised temporal association-based learning [1, 2, 3].

Invariance is the hard part

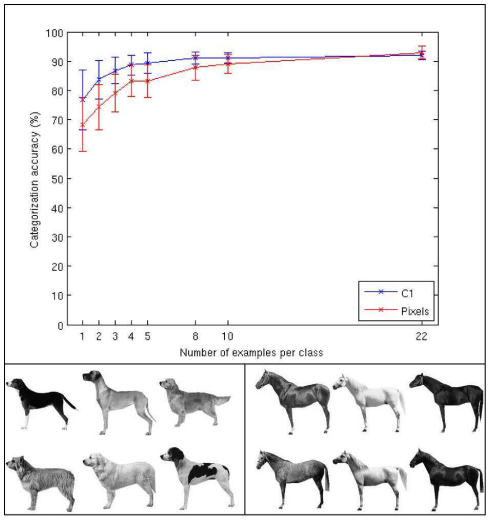
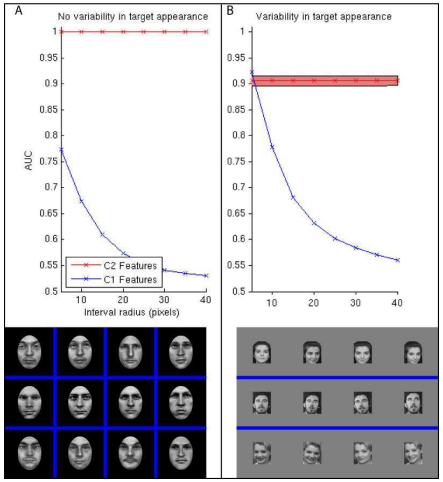
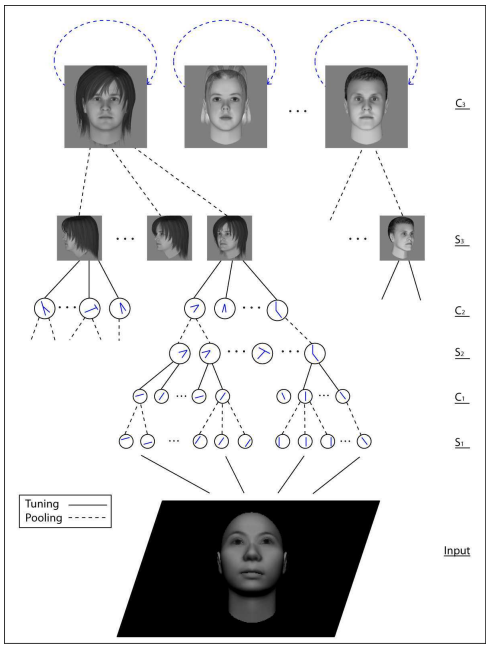


Image dataset thanks to Krista Ehinger and Aude Oliva.

Generic transformations

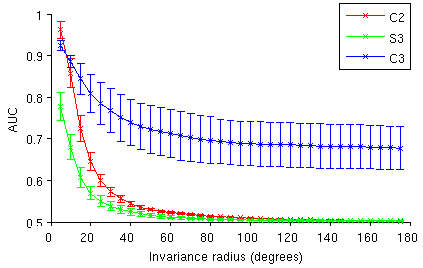


Hierarchies for invariance

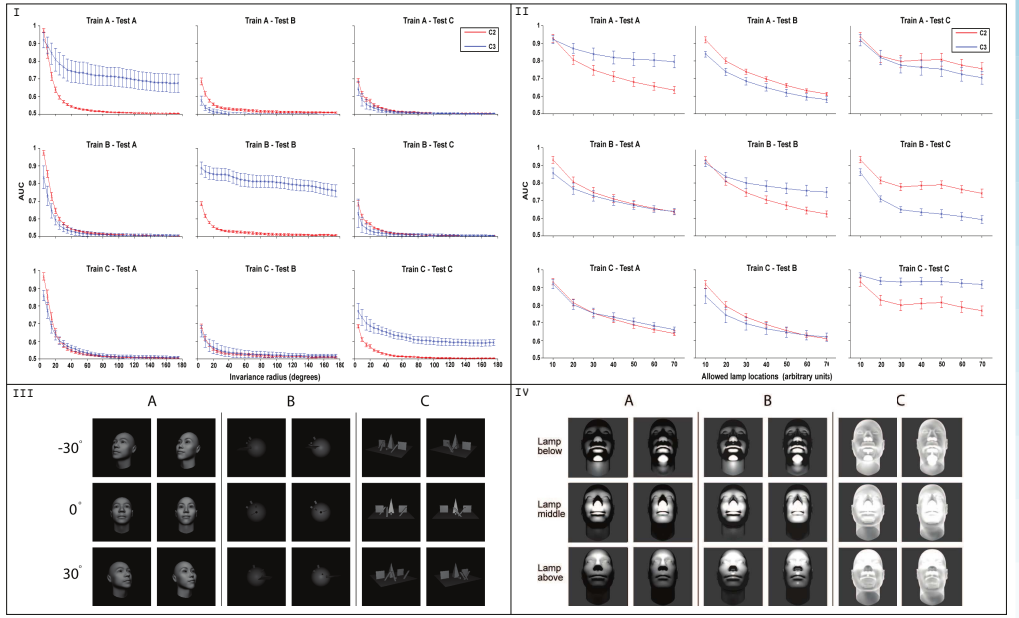


Following [4], S1 –> C2 layers discount translation and scaling. We added two additional layers: S3 –> C3 for viewpoint or illumination.

Pose invariant recognition



Invariance to class-specific transformations



Summary

Viewpoint and illumination transformations depend on the object's 3D structure and material properties. These are normally consistent within, but not between, classes. Class-specific modifications of the HMAX model achieve good viewpoint and illumination tolerant performance in a one-shot identification task. Performance suffers when a model that is specialized for transformations of one class is tested on identification within a different class. In fact, viewpoint-pooling models employing templates from the wrong class perform worse on viewpoint invariant identification tasks than models that have no particular mechanisms for dealing with viewpoint at all. The same situation arises for illumination invariance. This is in stark contrast to the generic case where the model is invariant to all classes undergoing the transformation no matter what templates are used.

The need to acquire invariance to class-specific transformations provides a nice computational argument for the existence of specialized face-processing patches in visual cortex [5, 6].

References

[1] T. Masquelier, T. Serre, S. Thorpe, and T. Poggio, "Learning complex cell invariance from natural videos: A plausibility proof," *AI Technical Report #2007-060 CBCL Paper #269*, 2007.

[2] P. Földiák, "Learning invariance from transformation sequences," *Neural Computation*, vol. 3, no. 2, pp. 194–200, 1991.

[3] L. Wiskott and T. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, no. 4, pp. 715–770, 2002.

[4] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 15, pp. 6424–6429, 2007.

[5] D. Tsao, W. Freiwald, R. Tootell, and M. Livingstone, "A cortical region consisting entirely of face-selective cells," *Science*, vol. 311, no. 5761, p. 670, 2006.

[6] N. Kanwisher, J. McDermott, and M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *The Journal of Neuroscience*, vol. 17, no. 11, p. 4302, 1997.