Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2010-057
CBCL-293

December 4, 2010

From primal templates to invariant recognition

Joel Z Leibo, Jim Mutch, Shimon Ullman, and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

From primal templates to invariant recognition

Joel Z Leibo, Jim Mutch, Shimon Ullman, Tomaso Poggio

Department of Brain and Cognitive Sciences, McGovern Institute,

Massachusetts Institute of Technology

We can immediately recognize novel objects – seen only once before -- in different positions on the retina and at different scales (distances). Is this ability hardwired by our genes or learned during development -- and if so how? We present a computational proof that developmental learning of invariance in recognition is possible and can emerge rapidly. This computational work sets the stage for experiments on the development of object invariance while suggesting a specific mechanism that may be critically tested.

Position- and scale-invariant recognition in primates depends on representations in inferotemporal visual (IT) cortex. Individual cell responses in IT are themselves invariant to translations of up to ±2º of eccentricity (equivalent to ±240 cones) and ±1 octave of scale (2). Standard models of visual cortex by Fukushima and others (1,3,4,5) attempt to replicate this phenomenon. In these models, units in layers corresponding to IT achieve invariance due to the wiring of the model, sometimes described as a *convolutional* architecture (3).

Networks with this architecture achieve invariance using receptive fields that pool together the responses to the same set of templates at many different locations. This operation is repeated in each successively higher layer until, in the top layer, unit responses are invariant across a large region (see figure 1). The templates themselves are derived from previously-seen image fragments; a unit's response is a measure of the similarity between the image patch it's seeing now and its optimal stimulus -- the remembered image fragment. In the top layer, each unit will respond invariantly to its optimal stimulus no matter where it appears. The combined activities of the top-layer units are the network's representation of an object. This representation is like a fingerprint for the object which can be used by a classifier for recognition; it inherits its invariance from that of its component units.

The problem of developing invariance via this scheme is thus equivalent to the problem of associating units representing the same template at different positions and scales in the visual field. While it is straightforward to achieve this in a computational model, for visual cortex it is one of a class of notoriously difficult correspondence problems. One way the brain might solve it is via the trace rule suggested by Foldiak (6). The trace rule explains how many simple cells having the same selectivity at different spatial positions could be wired to the same complex cell by exploiting continuity of motion: cells that fire to the same stimulus in close temporal contiguity are all presumably selective to the same moving stimulus. The biological plausibility of using the trace rule -- or a similar mechanism -- for learning of invariant object recognition by a developing organism depends on the number of templates needed to achieve a useful level of performance, as learning each invariant template requires seeing the same object patch moving continuously across all positions and scales. For at least some templates to be similar enough to parts of future novel objects, it would seem that many templates would be required. Indeed, the full computational model in figure 1 uses thousands of top-layer templates.

We investigated these issues using a reduced version of this class of models, for which simulation software is openly available (8). Surprisingly, it seems that *any* object can be recognized fairly well by using a set of templates which can be very different from it: for instance, templates learned from images of random dots can be used to recognize a large number of different objects, such as faces. Even more surprisingly, using a small number of templates is sufficient for good recognition performance. Figure 1 shows good invariant recognition performance for identification of different faces using a network with just 10 top-layer templates derived from patches of a face. Figure 1 shows invariance for translation; similar results were obtained for scale (7).

We conjecture that a small number of "primal" templates could be easily imprinted during early development in an invariant way by a mechanism implementing a trace rule. Patches of a mother's face or patterns derived from spontaneous retinal waves, for instance, could serve as primal templates from which the network will inherit invariance for any novel object. The number of invariant templates could later be increased by any of several bootstrapping mechanisms (7), anchored by the initial invariance provided by the primal templates.

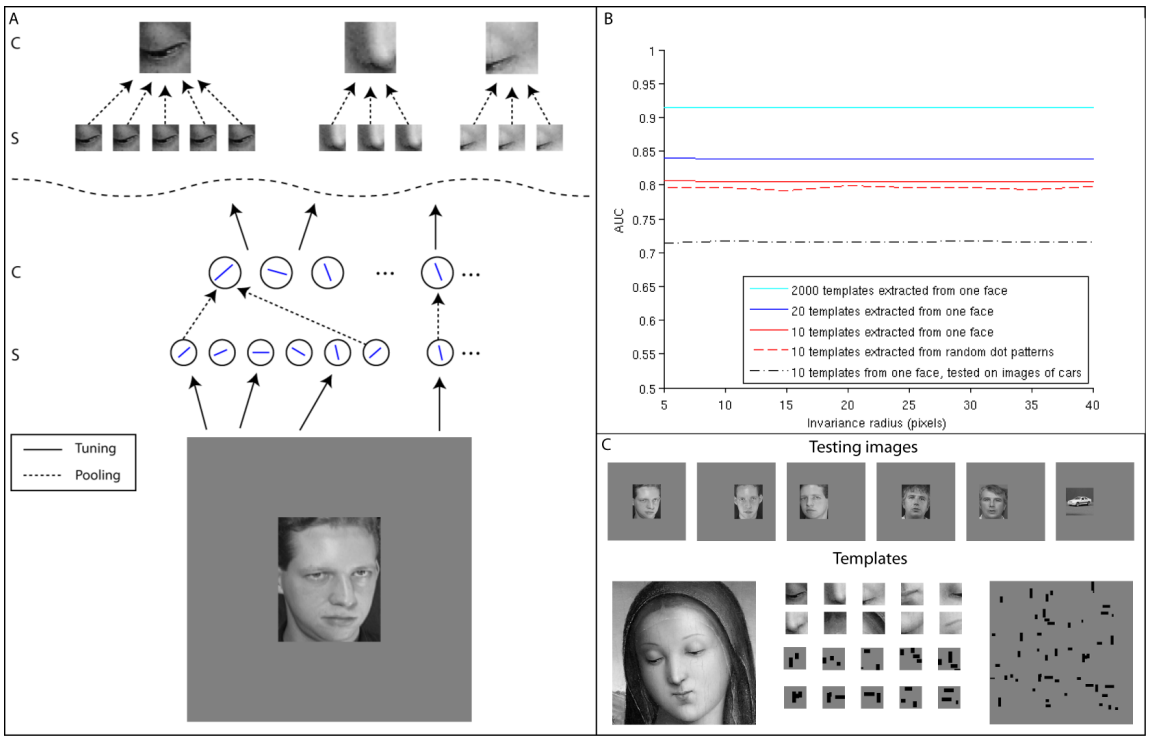

Fig. 1. Invariant recognition via a small number of invariant cells **(A)** Illustration of a generic hierarchical model of object recognition in the spirit of Hubel and Wiesel. In the first layer (S), the "simple" units are tuned to oriented edges. Each "complex" unit in the second (C) layer pools the first layer units with the same preferred orientation but from different locations in the visual field. In the penultimate layer, cells are tuned to patches of natural images. Each high level C unit pools S cells tuned to the same template replicated at different locations. The image fingerprint computed by the top level C units is then fed into a classifier. **(B)** Model accuracy: We plot a summary statistic of the ROC curve for classifying test images as either containing the same person's face or a different person's face (see (7)). AUC ranges from 0.5, indicating chance performance, to 1, indicating perfect performance. We repeat this classification allowing objects to translate different distances (pixels). We plot AUC as a function of the range over which objects could appear. The receptive field of the top-level C units was 256x256 pixels; the faces were approximately 100 pixels horizontally. In order to introduce variability into the positive class, we used 10 test images of the target face under slightly variable pose and lighting conditions; each was replicated at every position in a radius of 40 pixels (see panel c). For distractors, we used 390 different faces presented at the same locations. The simulations were done with a linear correlation classifier using only a single training view presented at the center of the receptive field. One can view the goal of this task as being to determine that all 10 views of the target face at all positions are more similar to the trained view of the target face presented at the center than they are to any of the distractor images. We repeated the experiment 40 times using different target faces or cars. In panel b we plot the mean AUC over all 40. *Results*: Just 10 top-level C units were sufficient for good performance. Each C cell pooled from ~3000 S cells optimally tuned to the same stimulus at each location. See (7). The black trace shows the results from testing on images of cars. In this case there were 5 views of the target car to be associated at each position and 45 distractor cars replicated at each position. Variability of results: We repeatedly restarted each simulation using a different set of randomly chosen templates. The standard deviation of the AUC (averaged across invariance range) was as follows: Testing on faces: 2000 face templates: stdev=.0074. 20 face templates: stdev=.069. 10 face templates: stdev=.106. 10 random dot pattern templates: stdev=.125. Testing on cars: 10 face templates: stdev=.130 **(C)** Examples of test images and top level templates. If the entire 256x256 image represents a 12 degree IT receptive field then the (~100x100 pixels) targets and distractors would occupy around 5 degrees of visual angle and the maximum translations considered here (+/- 40 pixels) would be about 2 degrees away from the example location in either direction

References

1. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*. **36**, 193-202 (1980)

2. Logothetis, N., Pauls, J., Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, **5**(5):552–563 (1995)

3. LeCun, Y., Bengio, Y. Convolutional networks for images, speech, and time series, *The Handbook of Brain Theory and Neural Networks*. , 255–258 (1995)

4. Riesenhuber, M., Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neuroscience*. **2**(11):1019–1025 (1999)

5. Serre, T., Oliva, A., Poggio, T. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*. **104**(15):6424–6429 (2007)

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

7. Leibo, J. Z., Mutch J., Rosasco L., Ullman S., Poggio., Invariant recognition of objects by vision. *CBCL*-291 (2010)

8. http://cbcl.mit.edu/software-datasets/index.html