Visual Cortex and Deep Networks

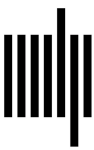
Learning Invariant Representations

By: Tomaso A. Poggio, Fabio Anselmi

Citation:

Visual Cortex and Deep Networks: Learning Invariant Representations
By: Tomaso A. Poggio, Fabio Anselmi
DOI: 10.7551/mitpress/10177.001.0001
ISBN (electronic): 9780262336710
Publisher: The MIT Press
Published: 2016

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

4 Class-Specific Approximate Invariance in Inferior Temporal Cortex

4.1 From Generic Templates to Class-Specific Tuning

As discussed in chapter 1, approximate invariance for transformations beyond the affine group requires highly tuned templates and therefore highly tuned simple cells. This is expected to take place in higher, nonretinotopic visual areas of the hierarchy. This is consistent with the architecture of the ventral stream and the the existence of class-specific modules in the primate cortex such as a face module and a body module [91–94]. We saw that areas in the hierarchy up to V4 or PIT provide signatures for larger parts or full objects. Thus we expect the following:

- Inputs to the class-specific modules are more scale and shift invariant than in earlier areas such as V1.
- Class-specific templates are large. For instance, in the case of faces, templates should cover significant regions of the face. Note that only large templates support pose invariance: the image of an isolated eye does not change much under rotations in depth of the face (and at most it undergoes scaling and shift).

4.2 Development of Class-Specific and Object-Specific Modules

A conjecture emerging from i-theory offers an interesting perspective on anterior inferior temporal cortex (AIT) [13]. For transformations that are not affine transformations in 2-D (we assume 3-D information is not available to the visual system or used by it, which may not always be true), an invariant representation cannot be computed from a single view of a novel object because the information available is not sufficient: lacking are the 3-D structure and the material properties of the object. Thus exact invariance to rotations

in depth or to changes in the direction or spectrum of the illuminant cannot be obtained. However, as shown in chapter 1, approximate invariance to smooth nongroup transformations can still be achieved in several cases (but not always) using a standard Hubel-Wiesel (HW) module. The reason this will often approximately work is because it effectively exploits prior knowledge of how similar objects transform. The image-to-image transformations caused by a rotation in depth are not the same for two objects with different 3-D structures. However, objects that belong to an object class where all the objects have similar 3-D structures transform their 2-D appearance in (approximately) the same way. This commonality is exploited by an HW module to transfer the invariance learned from (implicitily supervised) experience with template objects to novel objects seen only from a single example view. This is effectively our definition of an object class: *a class of objects such that the transformation for a specific object can be approximately inferred from how other objects in the class transform*. The necessary condition for this to hold is that the 3-D shapes be similar between any two objects in the class. The simulation in figure 4.1 shows that HW modules tuned to templates from the same class of (always novel) test objects provide a signature that tolerates substantial viewpoint changes (plots on the diagonal). It also shows the deleterious effect of using templates from the wrong class (plots off the diagonal). There are, of course, several other class-specific transformations besides rotation in depth, such as face expression and body pose transformations, to which the same arguments apply.

An interesting conjecture is that the visual system is continuously and automatically clustering objects and their transformations, observed in an unsupervised or implicitely supervised way, into class-specific modules [13]. Images of an object and of its transformations correspond to a new orbit Λ_k. New images are added to an existing module only if their transformations are well predicted by it. If no module can be found with this property, the new orbit will be the seed of a new object cluster/module.

For the special case of rotation in depth, Leibo et al. [13] ran a simulation using 3-D modeling, rendering software to obtain the orbits of objects for which there exist 3-D models. Faces had the highest degree of clustering of any natural category, unsurprising since recognizability likely influenced face evolution. A set of chair objects had broad clustering, implying that little invariance would be obtained from a chair-specific region. A set of synthetic wire objects, very similar to the paperclip objects used in several classic experiments on view-based recognition (e.g., [95–97]) were found to have the smallest index of compatibility [13]. Experience with familiar wire objects does not

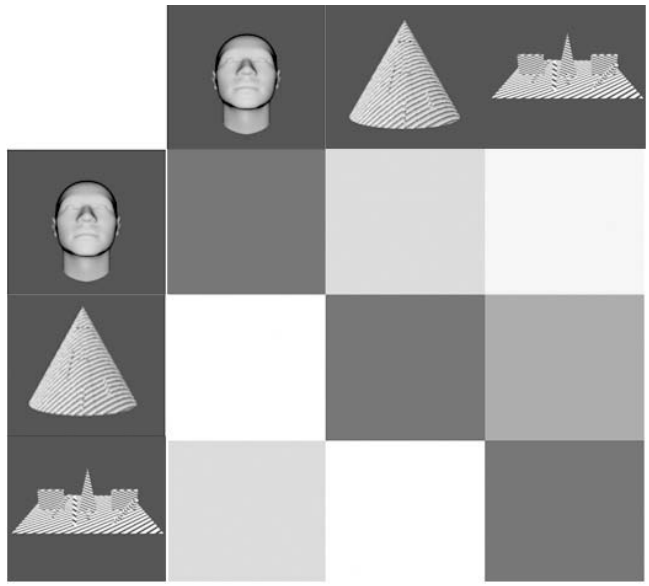


Figure 4.1

Class-specific transfer of depth rotation invariance for images from three classes: faces (A), cylinders (B), and composed (C). The left column of the matrix shows the results of the test for invariance for a random image of a face (class A) in different poses with respect to 3-D rotation using 3-D rotated templates from (classes A, B, and C). Similarly, the middle and the right columns show the invariance results for classes B and C, tested on rotated templates of A, B, and C, respectively. The different gray shades in the matrix show the maximum invariance range (degrees of rotation away from the frontal view). Only the diagonal values of the matrix (train A–test A; train B–test B; train C–test C) show an improvement of the view-based model over the pixel representation. That is, only when the test images transform similarly to the templates is there any benefit from pooling [13].

transfer to new wire objects because the 3-D structure is different for each individual paperclip object.

It is instructive to consider the limit case of object classes that consist of single objects like individual paperclips. If the object is observed under rotation, several frames are memorized as transformations of a single template (identity is implicitly assumed to be conserved by a Földiák-like rule, as long as there is continuity in time of the transformation; this is a case of implicitly supervised learning). The usual HW module pooling over them will allow view-independent recognition of the specific object. A few comments:

- Remarkably, the HW module for class-specific transformations, when restricted to multiple-views, single-object, is equivalent to the Poggio-Edelman model [98] for view invariance.

- The class-specific module is also effectively a gate. In addition to providing a degree of invariance it performs a template-matching operation with templates that can effectively block images of other object classes. This gating effect may be important for the system of face patches discovered by Freiwald and Tsao [105, 106], and it is especially obvious in the case of a single-object module.
- From the point of view of evolution, the use of the HW module for class-specific invariance can be seen as a natural extension from its role in single-object view invariance. The latter case is computationally less interesting, since it effectively implements a look-up table, albeit with interpolation power. The earlier case is more interesting, since it allows generalization from a single view of a novel object. It also represents a clear case of *transfer of learning*.
- Some of the best evidence for modules in IT for individual objects is provided by the Logothetis data from monkeys who were trained intensively to specific unfamiliar paperclip objects [96–98]. Unfortunately, fMRI data about spatial clustering in cortex was not available at the time.

4.3 Domain-Specific Regions in the Ventral Stream

There are other domain-specific regions in the ventral stream besides faces and bodies [13]. We think that additional regions for less common or less transformation-compatible object classes will appear with higher resolution imaging techniques. One example may be the fruit area, discovered in macaques with high-field fMRI [99]. Others include the body area and the lateral occipital complex (LOC), which is not really a dedicated region for general object processing but a heterogeneous area of cortex containing many domain-specific regions too small to be detected with the resolution of fMRI [100]. The visual word form area (VWFA) [101] seems to represent printed words. In addition to the generic transformations that apply to all objects, printed words undergo several nongeneric transformations that never occur with other objects. For instance, reading is rather invariant to font transformations and can deal with handwritten text. Thus, VWFA is well accounted for by the invariance hypothesis. Words are frequently viewed stimuli that undergo class-specific transformations. i-theory suggests that the limit case is object classes that consist of single objects. We expect that IT will include many such small modules in addition to large ones such as the face and the body modules. This prediction is consistent with evidence (e.g., [102]) of clustering of tuning in IT by category. Our prediction does not depend on supervised learning.

The justification—actually a prediction—by i-theory for domain-specific regions in cortex is different from other proposals. However, it is in complementary with respect to some of them, rather than exclusive. For instance, it would make sense that the clustering depends not only on the index of compatibility but also on the relative frequency of each object class. The conjecture claims that transformation compatibility is the critical factor driving the development of domain-specific regions and that there are separate modules for object classes that transform differently from one another.

4.4 Tuning in the Inferior Temporal Cortex

Considering the limit case of object classes that consist of single objects is important in order to understand the functional architecture of most of IT. If a specific object is observed under a set of transformations, several images of it can be memorized and linked together by continuity at the time of the transformation. As we mentioned, the usual HW module pooling over them will allow view-independent recognition of the specific object. Since this is equivalent to the Poggio-Edelman model for view invariance [98], there is physiological support for this proposal (see [97, 103, 104]).

4.5 Mirror-Symmetric Tuning in the Face Patches and Pooling over Principal Components

The theory can explain wether its framework the Freiwald-Tsao data [105, 106] on the face patch system. The most posterior patches, middle lateral and fundus (ML/MF), provide view- and identity-specific input to the anterior lateral patch (AL), where most neurons show tuning that is an even function of the rotation angle around the vertical axis. The anterior medial patch (AM), which receives inputs from AL, is identity specific and view invariant. The puzzling aspect of these data is the mirror-symmetric tuning in AL. Why does this appear in the course of a computation that leads to view invariance?

The theoretical framework implies that neurons in patch AL [105, 107], when exposed to several different faces, each one generating several images corresponding to different rotations in depth, may become tuned to the eigenvectors corresponding to the set of views. Since faces are bilaterally symmetric and since for each view there is a mirror-symmetric view, the associated covariance function has eigenvectors (PCs) that are either even or odd functions (see figure 4.2 and appendix section A.5).

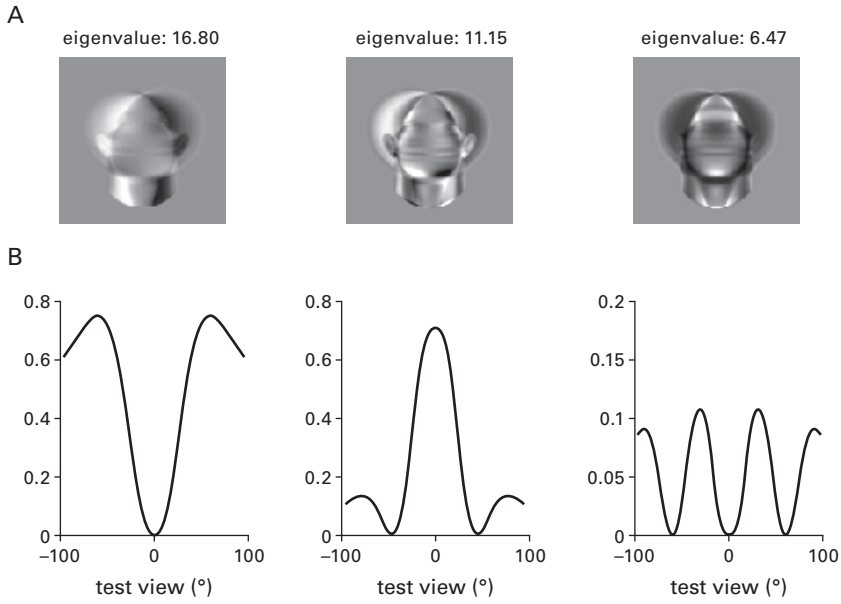


Figure 4.2

Face identity is represented in the macaque face patches [105]. Neurons in the middle areas of the ventral stream face patch (ML, MF) are view specific, while those in the most anterior area (AM) are view invariant. Neurons in an intermediate area (AL) respond similarly to mirror-symmetric views. In i-theory view invariance is obtained by pooling over simple neurons whose tuning corresponds to the PCA of a set of faces previously experienced each under a range of poses. Because of the bilateral symmetry of faces, the eigenvectors of the associated covariance matrix are even or odd. This is shown in (A), where the first three PCs of set of grey-level faces under different poses are plotted. The same symmetry arguments apply to neural images of faces. (B) Response of three model AL units to a face stimulus as a function of pose under different poses [13].

The observed AL properties could then be explained in different ways.

- The class-specific theorem and the spectral pooling proposition (see chapter 2) suggest that square pooling (i.e., energy pooling) over these face PCs can provide approximate invariance to rotations in depth. The full argument goes as follows. Rotations in depth of a face around a certain viewpoint, say, $\theta = \theta^0$, can be approximated well by linear transformations (by transformations $g \in GL(2)$). The HW algorithm can then provide invariance around $\theta = \theta^0$. Finally, if different sets of simple cells are plastic at somewhat different times, exposure to a partly different set of faces yields different eigenvectors summarizing different sets of faces. The different sets of faces play the role of different object templates in the standard theory. According to the theory, the result should be expected if AL contains

simple cells that are tuned by a synaptic Hebb-like Oja rule and the output of the cells is roughly a squaring nonlinearity as required by the spectral pooling proposition. In this interpretation, cells in AM pool over several of the squared eigenvector filters to obtain invariant second moments (see figure 4.2). Detailed models from V1 to AM show properties that are consistent with the data. It is doubtful, however, that such models can have a sufficiently good performance in face recognition [7, 15, 35, 108].

- An alternative explanation is that the network leading to each AL cell is balanced [108] and the cell has a mirror-symmetric tuning (induced by a set of template faces averaged over transformations).

Background and Bibliography

This part of the book describes work in our group. Its main sources are the following technical reports: [9, 11, 13, 15].

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.

Library of Congress Cataloging-in-Publication Data

Names: Poggio, Tomaso, author. | Anselmi, Fabio, author.
Title: Visual cortex and deep networks : learning invariant representations / Tomaso A. Poggio and Fabio Anselmi.
Description: Cambridge, MA : MIT Press, [2016] | Series: Computational neuroscience | Includes bibliographical references and index.
Identifiers: LCCN 2016005774 | ISBN 9780262034722 (hardcover : alk. paper)
Subjects: LCSH: Visual cortex. | Vision. | Neural networks (Neurobiology) | Perceptual learning. | Computational neuroscience.
Classification: LCC QP383.15 .P64 2016 | DDC 612.8—dc23 LC record available at http://lccn.loc.gov/2016005774

10 9 8 7 6 5 4 3 2 1