CBCL Paper
November 10, 2007

Derived Distance: towards a mathematical theory of visual cortex.

Steve Smale[‡], Tomaso Poggio[†], Andrea Caponnetto[◇] and Jake Bouvrie[†]

Toyota Technological Institute at Chicago and University of California, Berkeley[‡]
Department of Mathematics, City University of Hong Kong[◇]
CBCL, McGovern Institute, Artificial Intelligence Lab, BCS, MIT[†]

Abstract

We describe a "natural" metric on the space of images motivated by the neuroscience of visual cortex. We propose the notion of a hierarchical derived distance and suggest that it could be applied to the classification of imagery and text and to the analysis of genomics data.

1 Introduction

In the last few years, new models based on anatomical and physiological data about the primate visual cortex [5, 10, 6, 18, 8, 3, 6] are beginning to quantitatively account for a host of novel physiological data and to provide human-level performance on rapid categorization of complex imagery [15, 17, 16, 13, 14]. These models are the most recent examples of a family of biologically-inspired architectures [4, 21, 9, 12, 19, 1, 22], and related computer vision systems [7, 20]. The hierarchical organization of such models – and of the cortex itself – remains a challenge for learning theory, as we mentioned elsewhere [11], since classical "learning algorithms" – as described in [11] correspond to one-layer architectures.

In this note, we attempt to formalize the basic hierarchy of computations represented in the models of visual cortex. Our hope is ultimately to achieve a *theory* that may explain why such *models* work as well as they do and what the computational reasons for the hierarchical organization of the cortex are.

This report provides a simplified version of the framework introduced in [2]. In the Appendix we establish detailed connections with the Serre et al. model [16] and identify a key difference.

1.1 Notation and Preliminary Definitions

At first we consider the special case of a two stage derived distance. Consider the squares v, v' and R in $\mathbf{R}^2$ with $v \subset v' \subset R$, centered and with axis aligned (see Figure 1). For example, in our simulations (see Appendix 1) the width of v is 4 pixels, the with of v' is 8 pixels and the width of the "retina" R is 14 pixels.

Suppose $Im(R)$ is given. $Im(v)$ and $Im(v')$ are subsets of the restrictions of $f \in Im(R)$ to v and v', respectively. Let H be the set of all transformations $h : v \to v'$ with the form $h = h_\beta h_\alpha$, $h_\alpha(x) = \alpha x$ and $h_\beta(x') = x' + \beta$, where $1/2 \leq \alpha \leq 2$ and $\beta \in \mathbf{R}^2$ is such that $h_\beta h_\alpha(v) \subset v'$. We define in a similar way $h' : v' \to R$, with $h' \in H'$. The transformations h, h' are embeddings (of v in v', of v' in R,). In the following, let us assume, for simplicity, that all h, h' are restricted to translations ($\alpha = 1$) and that there is a finite number of them in H, H', respectively (thus we assume that there is a finite number of β values). h can be thought of as translating the image over "the receptive field" v or alternatively as selecting a receptive field of size and shape v that "looks" at a different regions of the image. An image f is defined on the retina R but restricted to v' or v in the following.

Let $Im(v)$, $Im(v')$, $Im(R)$ be three spaces of functions $v \to [0, 1]$, $v' \to [0, 1]$ and $R \to [0, 1]$, respectively. We also assume that $T \subset Im(v)$ and $T' \subset Im(v')$ are two different, finite sets of "templates" with probability measures ρ_T and $\rho_{T'}$ respectively, which we assume here to be uniform.

The key property we use is

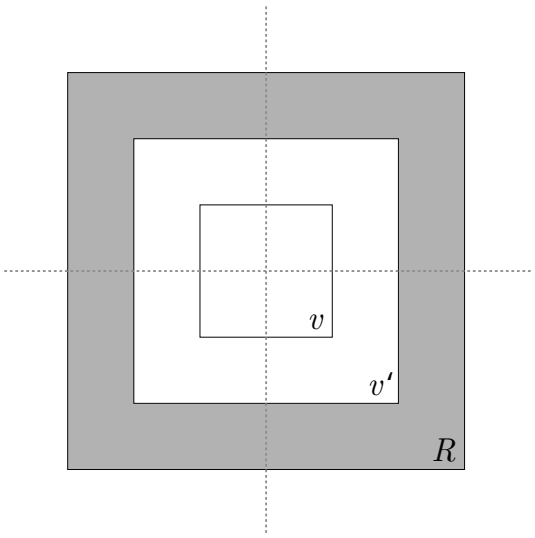


Figure 1: *Nested domains, see text.*

Axiom: $f \circ h : v \to [0,1]$ is in $Im(v)$ if $f \in Im(v')$ and $h \in H$,
that is *the restriction of an image is an image* and similarly for H'. Thus

$f \circ h : v \to [0,1] \in Im(v)$ if $f \in Im(v')$ and $h \in H$,
$f \circ h' : v' \to [0,1] \in Im(v')$ if $f \in Im(R)$ and $h' \in H'$.

We also call $\mathbf{R}_+^T$ the set of functions $T \to \mathbf{R}_+$ so that $\mathbf{R}_+^T$ can be thought of as the positive orthant of the cartesian space $\mathbf{R}^N$ with $N = |T|$ (and similarly for T'). These function spaces use ρ_T and $\rho_{T'}$, respectively, as the structure of an L_p-normed space. This $p \geq 1$ is a fixed parameter.

2 A Derived Distance

We formulate the model in the following stages:

1. The process starts with some initial distance on $Im(v)$ provided by

$$d_0'(f,g) = d(f,g) = ||f - g||_p, \tag{1}$$

where $||\cdot||_p$ is an appropriate L_p norm $\left(||f||_p = (\int_v |f(x)|^p d\mu(x))^{1/p}\right)$, for the space of functions $Im(v)$.

Then we define a first stage *Neural Similarity* as

$$N_t^1(f) = \min_{h \in H} d_0'(f \circ h, t), f \in Im(v') \tag{2}$$

Thus $N^1 : Im(v') \rightarrow \mathbf{R}_+^T$ can be defined[1] by $N^1(f)(t) = N_t^1(f)$.

We define the derived distance (with $||N^1(f)||_p = (\int_T |N_t^1|^p d\rho(t))^{1/p}$) on $Im(v')$ as

$$d_1'(f, g) = ||N^1(f) - N^1(g)||_p \tag{3}$$

Since $N^1(f)$ and $N^1(g)$ are elements in $\mathbf{R}_+^T$, this norm makes sense (we use no L_p norm on $Im(v')$).

2. We now repeat the process by defining the second stage *Neural Similarity* as

$$N_{t'}^2(f) = \min_{h' \in H'} d_1'(f \circ h', t'), f \in Im(R), t' \in T'. \tag{4}$$

The new derived distance is now on $Im(R)$

$$d_2'(f, g) = ||N^2(f) - N^2(g)||_p. \tag{5}$$

Clearly this process could continue if appropriate higher level patches were defined.

Remarks

1. The Appendix discusses relations with the model of [15] , see Figure 8.
2. Notice that the hierarchy can be considered as a hierarchy of associations. At the first level, each patch of the image is associated with a set of similarities to given templates. At the second level the process is iterated using larger patches and bigger, more complex templates – and thus establish an association with associations. At the top level the image is described as a set of similarities to the top-level templates. Clearly, translations (and scales) make the hierarchy non-trivial. If there were no translations and scales, it is not clear whether the hierarchy would offer any advantage with respect to a single stage of similarities, *apart from defining a "natural" similarity function*.
3. Notice that a supervised classifier may be defined from the derived distance at the top level. Also derived distances at lower levels may be used (this is what is done in [13, 15]).

[1]Notice that N_t^1 can be interpreted in terms of the model of Figure 8, see Appendix. In our simulations (see Appendix 1) the function of t, $N_t^1(f)$ is represented as a vector N_t^1 with $|T|$ components.

4. The same model can be applied to classification of text (the dictionary of templates at the first level consists of isolated letters, at the second level of frequent pairs etc.).

5. The same model can be applied to classification of genomic sequences (the dictionary of templates at the first level consists of the four isolated nucleotides, at the second level of frequent pairs etc.).

6. We suppose our probability measure ρ on $Im(R)$ is given by the real world images for the given vision problem under study. Then the templates $t_i \in T$ can be most conveniently be taken as random draws from ρ_v on $Im(v)$, where ρ_v is the probability measure induced from ρ by restriction. One may take T' in a similar fashion.

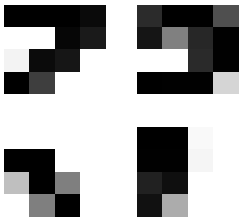


Figure 2: *Four of the 500 (4x4) templates in T.*

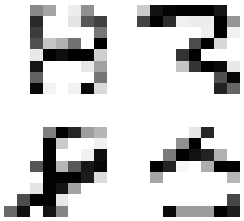
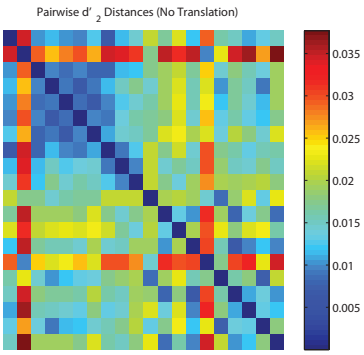


Figure 3: *Four of the 50 (8x8) templates in T'.*



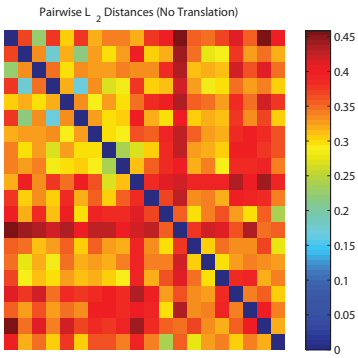


Figure 4: *Matrices of pairwise d'_2 (left) and L_2 (right) distances for the set of 20 labeled images from the database. The first 10 rows/columns correspond to images of threes, and the last 10 rows/columns correspond to images of eights. Each entry is the pairwise distance between the corresponding images.*

3 Appendix 1: Preliminary Simulations

We have conducted preliminary simulations in which derived distances are compared to the L_2 distance in the context of a simple handwritten digit classification task. Given a small labeled set of images, we use the 1-nearest neighbor (1-NN) classification rule: an unlabeled test example (restricted to 3s and 8s) is given the label of the stored, labeled instance it is closest to under the specified distance metric.

In the experiments presented here, we have used 14x14 pixel grayscale images randomly selected from the MNIST handwritten digit dataset. The digits in this dataset include a small amount of natural translation, rotation, scaling, shearing and other "transformations", as one might expect to find in a corpus containing the handwriting of many human subjects. To further explore the translation invariance of the derived distance, we subjected the labeled and unlabeled (test) sets of images to translations ranging from 0 to 6 pixels in one of 8 randomly chosen directions (the transformations h are restricted in the

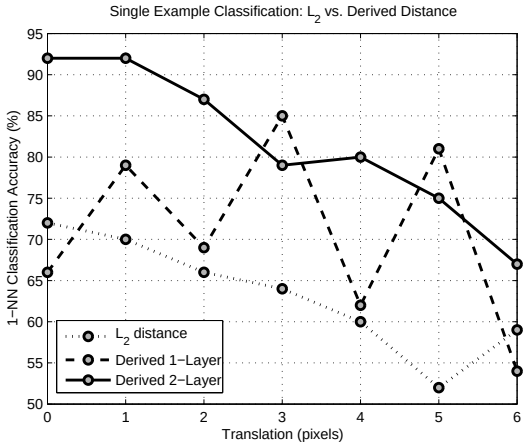


Figure 5: *Experiments comparing 1-NN classification accuracy over 100 test examples with 2 labeled examples, one per class, when using the L_2 distance vs. d'_1 and d'_2 derived distances. The experiment is repeated with images subjected to translations of 0-6 pixels (x-axis) to test robustness of the distance under translations.*

implementation of this paper to translations, mainly for computational simplicity). For the classification experiments, we explored two different labeled image sets: one in which we included only one labeled image per class, and another in which the labeled set was chosen to include 10 threes and 10 eights, giving 20 labeled images total. In all experiments reported here, classification accuracies are averaged over 100 test examples.

The sets T of templates consists of 500 randomly extracted 4x4 image patches (corresponding to v). The set T' consists of 50 randomly extracted 8x8 image patches[2] (corresponding to v'). Examples of the patches in T and T' can be seen in Figures 2 and 3, respectively. The smaller templates in T are large enough to include semi-circles and distinct stroke intersections as extracted from the digits. At 8x8 pixels, the templates in T' are seen to include nearly full digits where more discriminative structure is present. Here, R corresponds to the full domain of the image, eg 14x14 pixels. Patches are extracted from images of threes and eights which are not used in the classification experiments. In all experiments, we have used the same images to build the labeled, unlabeled, and template sets, respectively.

The p-norms invoked in the definitions of d'_0, d'_1 and d'_2 (equations (1), (3) and (5)) are as given in Section 2. For the simulations in this section, the norm involving images takes the form

$$\|f\|_p = \left(\int_v |f|^p d\mu(x) \right)^{1/p} \implies \left(\frac{1}{M} \sum_{i=1}^{M} |f_i|^p \right)^{1/p} \tag{6}$$

[2]This latter number, imposed by computational limitations, is probably too small.

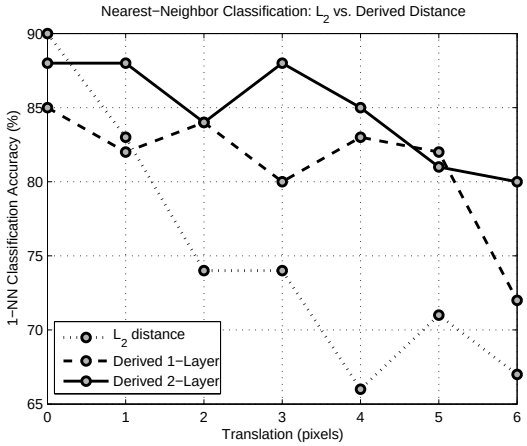


Figure 6: *Experiments similar to Figure 5 comparing 1-NN classification accuracy over 100 test examples, but using here 10 labeled examples per class.*

where the image patch has M pixels. The neural similarities N^i require the following norm (we drop the index i of the layer for simplicity)

$$\|N(f)\|_p = \left(\int_T |N_t|^p d\rho(t) \right)^{1/p} \implies \left(\frac{1}{|T|} \sum_{i=1}^{|T|} |N_{t_i}|^p \right)^{1/p} \tag{7}$$

where we have assumed a uniform measure ρ over the finite space T of templates. We denote the number of templates in the set by $|T|$. For purposes of the experiments discussed in this note, we set $p = 2$.

Before discussing classification, we pause to illustrate graphically in Figure 4 the derived distances when applied to pairs of digits. On the left we show d'_2 distances, while the L_2 distances are provided for comparison on the right. In this example, we have used the labeled set of 20 images described above, and with no artificial translations. The first 10 rows/columns correspond to images of threes, and the last 10 rows/columns correspond to images of eights. Each entry is the pairwise distance between the corresponding images. The images themselves are shown (with some translation) in Figure 7 (small, right), where the index written to the left of each digit corresponds to rows/columns of the distance matrices shown in Figure 4.

In Figures 5 and 6 we turn to the classification experiments, and show accuracies using the 1-NN rule for each of the image translations ranging from 0-6 pixels (as discussed above) in the case of the L_2, d'_2 and d'_1 distances. For the experiments summarized in Figure 5 we have used one labeled example per class, while in Figure 6 we used the labeled image set with 10 examples per class. As might be expected, the derived distances are better able to accommodate image translations than L_2 on the whole, and classification accuracy decays more gracefully in the derived distance cases as we increase the

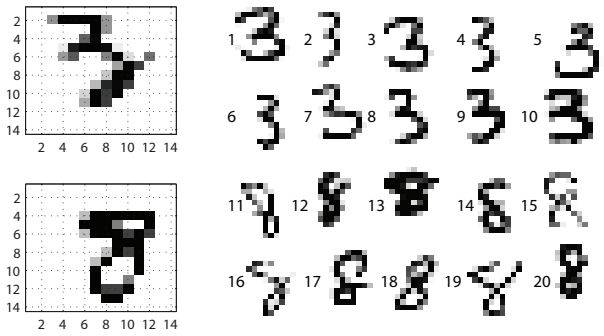


Figure 7: *Two instances from the set of test examples (large, left), and the full labeled set (small, right, with* 10 *examples for each of the two digits). All images in the experiments are fixed at 14x14 pixels. Images here have been randomly translated by 3 pixels.*

Test Digit	Train 3s (d'_2)	Train 8s (d'_2)	Train 3s (L_2)	Train 8s (L_2)
3	0.0096	0.0116	0.3028	0.3339
8	0.0111	0.0082	0.3832	0.3360

Table 1: *Minimum distances from the test examples shown in Figure 7 (large digits, left) to the labeled examples in Figure 7 (small digits, right). We have divided the labeled set up into threes and eights, and show the minimum d'_2 and L_2 distances to each category separately.*

size of the translation. In addition, the 2-layer d'_2 derived distance is seen to generally outperform the 1-layer d'_1 derived distance. Whether there exists an optimal number of layers and template set sizes for a given problem, and how to choose them, is an open topic.

In Figure 7 we show one instance each from the test set of threes and eights on the left (large boxed digits), and on the right we show the complete labeled image set (small digits). Both the labeled and test images in this case have been artificially translated by 3 pixels in one of 8 random directions. In Table 1 we give the minimum distances, in both the L_2 and d'_2 derived distance cases, from the test instances shown in Figure 7 to the labeled set shown in the same figure. While both L_2 and d'_2 give the correct 1-NN classification, the difference between minimum distances across the two classes of training examples is more pronounced (relative to the distances) for d'_2.

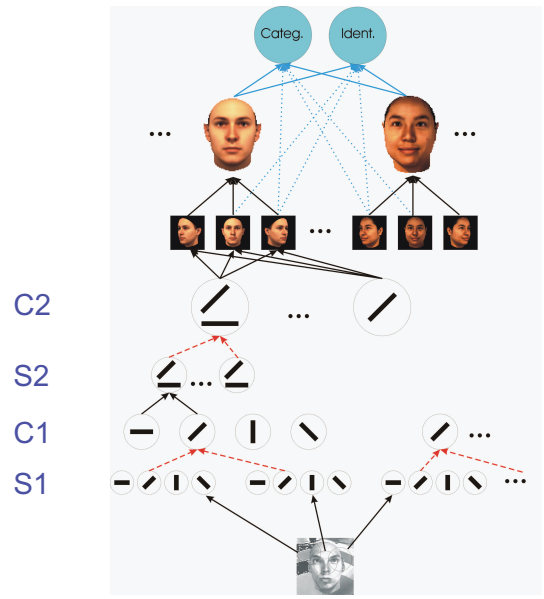


Figure 8: *The model of Serre et al [16]. We consider here the layers up to C2.*

4 Appendix 2: Derived distances and Visual Cortex

In this Appendix, we extend the neural distance definition in Section 2 towards establishing an exact connection with the model of Serre et al. We consider a two-stage model (comprising S1, C1, S2, C2 as in [16], as illustrated in the first 4 layers of Figure 8.

As shown in the figure, we consider one additional stage and thus a v'' square. Similarly to the above, we assume that the squares v, v', v'' and R in $\mathbf{R}^2$ with $v \subset v' \subset v'' \subset R$, centered and with axis aligned. For example, the width of v is $1/2$ of the with of v' which is also $1/2$ of the with of v'' which is also $1/2$ of the width of the "retina" R. Let H be the set of all transformations $h : v \to v'$ with the form $h = h_\beta h_\alpha$, $h_\alpha(x) = \alpha x$ and $h_\beta(x') = x' + \beta$, where $1/2 \leq \alpha \leq 2$ and $\beta \in \mathbf{R}^2$ is such that $h_\beta h_\alpha(v) \subset v'$. We define in a similar way $h' : v' \to v''$ and $h'' : v'' \to R$, with $h', h'' \in H', H''$, respectively.

As before, we assume, for simplicity, that all h, h', h'' are restricted to translations and that there is a finite number of them in H, H', H'', respectively (thus we assume that there is a finite number of β values).

Let $Im(v)$, $Im(v')$ and $Im(v'')$ the three spaces of functions $v \to [0,1]$, $v' \to [0,1]$ and $v'' \to [0,1]$, respectively. We also assume that $T \subset Im(v)$ and $T' \subset Im(v'')$ are two different, finite sets of "templates", each with a probability measure ρ_T and $\rho_{T'}$, respectively, which we again assume here to be uniform.

Analogous to the assumptions made earlier, we assume that

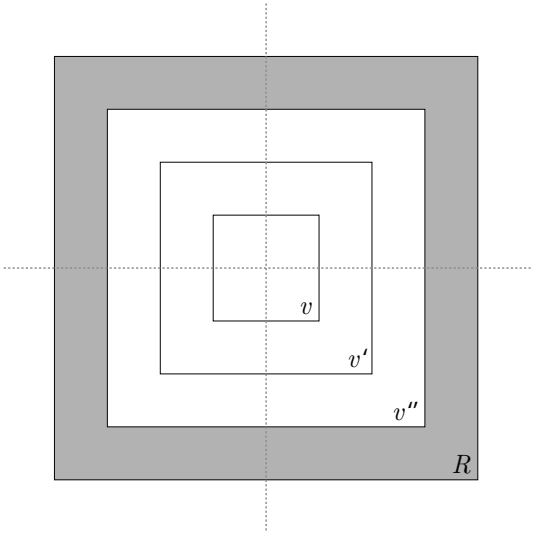


Figure 9: *Nested domains, with an additional v''.*

$$f \circ h : v \to [0,1] \in Im(v)$$

if $f \in Im(v')$ and $h \in H$,

$$f \circ h' : v' \to [0,1] \in Im(v')$$

if $f \in Im(v'')$ and $h' \in H'$,

$$f \circ h'' : v'' \to [0,1] \in Im(v'')$$

if $f \in Im(R)$ and $h'' \in H''$,

The definition of Neural Similarity can be broken down into two steps:

- First we have
$$N^{1,S}_{h,t}(f) = d(f \circ h, t)$$
where $N^{1,S}_{h,t}(f)$ corresponds to the response of a S1 cell with template t with receptive field $h \circ v$.

- The Neural Similarity is now
$$N^{1,C}_{t}(f) = \min_{h \in H} d(f \circ h, t)$$
where $N^{1,C}_{t} : Im(v') \to \mathbf{R}^{T}_{+}$. $N^{1,C}_{t}(f)$ corresponds to the response of a C1 cell with template t and with receptive field – the region over which the min is taken – corresponding to v'.

We now extend $N^{1,C}$ to the region v'' as

$$N^{1,C}_{t,h'}(f) = N^{1,C}_{t}(f \circ h')$$

where $N^{1,C}_{t,h'}(f)$ can be regarded as a vector with components indexed by t, h'; $N^{1,C}_{t,h'}(f)$ is a function of f restricted to v'. $N^{1,C}_{t,h'}(f)$ is the response of a C1 cell with receptive field $h' \circ v'$ in v''.

The two steps at the second stage – corresponding to S2 and C2 cells, respectively – can be now defined as

- $N^{2,S}_{t',h''}(f) = d(N^{1,C}_{t,h'}(f \circ h''), N^{1,C}_{t,h'}(t'))$ where the distance is computed averaging over t, h' and f is restricted to v''. $N^{2,S}_{t',h''}$ is the response of a S2 cell with receptive field $h'' \circ v''$. Furthermore

- $N^{2,C}_{t'}(f) = \min_{h'' \in H''} d(N^{1,C}_{t,h'}(f \circ h''), N^{1,C}_{t,h'}(t'))$
  is the response of a C2 cell with receptive field R.

What is described above gives a detailed correspondence between the model of Figure 8 and the mathematics (apart from the use of the Gaussian instead of the derived distance and the presence of scale invariance in the Serre model). The special case considered in the main text corresponds to the case of $v' = v''$ eg H' contains only the identity transformation. The special case corresponds to the model of [2]. It is not completely faithful to the Serre model [16, 15] and to the commonly accepted view of physiology. However, S2 cells could have the same receptive field of C1 cells and C2 cells could be the equivalent of V4 cells. Thus the known physiology may not be inconsistent.

References

[1] Y. Amit and M. Mascaro. An integrated network for invariant visual detection and recognition. *Vis. Res.*, 43(19):2073–2088, 2003.

[2] A. Caponnetto, Tomaso Poggio, and Steve Smale. On a model of visual cortex. CBCL Paper, MIT, 2007.

[3] R. Desimone. Face-selective cells in the temporal cortex of monkeys. *J. Cogn. Neurosci.*, 3:1–8, 1991.

[4] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cyb.*, 36:193–202, 1980.

[5] D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J. Phys.*, 195:215–243, 1968.

[6] E. Kobatake and K. Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophys.*, 71:856–867, 1994.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, November 1998.

[8] N.K. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.*, 5:552–563, 1995.

[9] B.W. Mel. SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comp.*, 9:777–804, 1997.

[10] D.I. Perrett and M. Oram. Neurophysiology of shape processing. *Img. Vis. Comput.*, 11:317–333, 1993.

[11] T. Poggio and S. Smale. The mathematics of learning: Dealing with data. *Notices of the american Mathematical Society (AMS)*, 50(5), 2003.

[12] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2:1019–1025, 1999.

[13] 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. AI Memo 2005-036 / CBCL Memo 259, MIT, Cambridge, MA, 2005.

[14] T. Serre, M. Kouh., C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A quantitative theory of immediate visual recognition. *Progress in Brain Research*, 165:33–56, 2007.

[15] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Science*, 104:6424–6429, 2007.

[16] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29:411–426, 2007.

[17] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In IEEE Computer Society Press, editor, *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, San Diego, 2005.

[18] K. Tanaka. Inferotemporal cortex and object vision. *Ann. Rev. Neurosci.*, 19:109–139, 1996.

[19] S.J. Thorpe. Ultra-rapid scene categorisation with a wave of spikes. In *BMCV*, 2002.

[20] S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermdediate complexity and their use in classification. *Nat. Neurosci.*, 5(7):682–687, 2002.

[21] G. Wallis and E. T. Rolls. A model of invariant object recognition in the visual system. *Prog. Neurobiol.*, 51:167–194, 1997.

[22] H. Wersing and E. Koerner. Learning optimized features for hierarchical models of invariant recognition. *Neural Comp.*, 15(7):1559–1588, 2003.