Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2004-071
AIM-2004-024
CBCL-241

November 12, 2004

Shape Representation in V4: Investigating Position-Specific Tuning for Boundary Conformation with the Standard Model of Object Recognition

Charles Cadieu, Minjoon Kouh, Maximilian Riesenhuber, and Tomaso Poggio

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

Shape Representation in V4: Investigating Position-Specific Tuning for Boundary Conformation with the Standard Model of Object Recognition

Charles Cadieu[1], **Minjoon Kouh**[1],
Maximilian Riesenhuber[2] **& Tomaso Poggio**[1]

[1]**McGovern Institute for Brain Research,**
Center for Biological and Computational Learning,
Computer Sciences & Artificial Intelligence Laboratory
and
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

[2]**Department of Neuroscience,**
Georgetown University Medical Center

Abstract

The computational processes in the intermediate stages of the ventral pathway responsible for visual object recognition are not well understood. A recent physiological study by A. Pasupathy and C. Connor in intermediate area V4 using contour stimuli, proposes that a population of V4 neurons display object-centered, position-specific curvature tuning [18]. The "standard model" of object recognition, a recently developed model [23] to account for recognition properties of IT cells (extending classical suggestions by Hubel, Wiesel and others [9, 10, 19]), is used here to model the response of the V4 cells described in [18]. Our results show that a feedforward, network level mechanism can exhibit selectivity and invariance properties that correspond to the responses of the V4 cells described in [18]. These results suggest how object-centered, position-specific curvature tuning of V4 cells may arise from combinations of complex V1 cell responses. Furthermore, the model makes predictions about the responses of the same V4 cells studied by Pasupathy and Connor to novel gray level patterns, such as gratings and natural images. These predictions suggest specific experiments to further explore shape representation in V4.

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL).

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1.

Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), DaimlerChrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sony MOU, Sumitomo Metal Industries, Toyota Motor Corporation, and WatchVision Co., Ltd..

1 Introduction

Many physiological studies have shed some light on the ventral stream in primate visual cortex [4, 11, 31], but most computational issues have yet to be resolved. Cells in the early stages of the ventral pathway have small receptive fields and are selective to simple features, such as edge orientation [1, 8], while cells far along the pathway in inferotemporal cortex (IT) have large receptive fields and are selective to complex objects like faces and hands [2, 7, 14, 20, 30]. The general selectivity at these two stages of the ventral pathway is relatively well understood. However, cells at intermediate stages, between V1 and IT, have not been fully characterized.

In one recent study of V4, Pasupathy and Connor [18] investigated the nature of shape representation in V4 cells of the primate ventral pathway. Building on a previous study [17] in which they found tuning in V4 cells to angle orientation, Pasupathy and Connor examined shape representation of V4 cells using a set of simple closed shapes formed by combining convex and concave boundary elements [18] (see Figure 2). They characterized a subpopulation of V4 cells as having selectivity for object-centered position-specific boundary conformation, such as cells that were tuned to multiple curvatures at specific angular positions from the object's center of mass.

In this paper we report that the "standard model," a recently developed computational model of object recognition [23], can reproduce the quantitative data described by Pasupathy and Connor in [18]. Model units can display the same object-centered, position-specific curvature tuning measured by Pasupathy and Connor in a subpopulation of V4 cells. These model units represent a translation invariant combination of complex V1-like subunits that may be described as "curvature filters." These results suggest that V4 selectivity is a result of the combination of complex V1 cell responses. Furthermore, the model can make quantitative predictions, yet to be verified, of how cells measured by Pasupathy and Connor will respond to novel stimuli. Such predictions can provide the basis for further experiments to explore shape representation in V4.

1.1 Physiological Results

There have been many experiments in intermediate area V4 using a variety of stimulus sets [3, 5, 6, 22], but none has been able to fully characterize V4 selectivity. In one attempt to describe V4 cell shape representation, Pasupathy and Connor systematically combined convex and concave boundary elements to produce simple closed shapes with shared boundary components [18]. They then quantified the raw responses of prescreened V4 cells to the stimulus set using Gaussian functions in a number of tuning domains: boundary conformation, edge orientation and axial orientation. Gaussian functions fit in the boundary conformation space (specifically, curvature $\times$ angular position from the object center) were found to best characterize V4 responses. Pasupathy and Connor concluded that these results, "suggest a parts-based representation of complex shape in V4, where the parts are boundary patterns defined by curvature and position relative to the rest of the object." [18]

2 Methods

2.1 The Standard Model

The standard model combines many data about the ventral stream [9, 10, 19] into a hierarchical computational model of object recognition [23–25]. The two basic cognitive requirements of object recognition, invariance and specificity, are evident at the earliest and highest stages within the ventral stream. Within the earliest stages, recordings in cat striate cortex using oriented bars show that simple cells display strong phase dependence, while complex cells display tuning that is independent of phase [9]. Hubel and Wiesel proposed that the invariance of complex cells could be created by pooling together simple cells with similar selectivities but translated receptive fields [10]. Perrett and Oram proposed a similar mechanism within IT [19], the highest layer of the ventral stream, that described invariance to any transformation as pooling over afferents tuned to transformed versions of the same stimuli. Riesenhuber and Poggio extended these proposals in a quantitative model to describe the mechanisms that achieve invariance and specificity throughout the ventral stream [23].

The resulting model is a hierarchical framework that consists of units analogous to simple and complex cells in V1, and leads to view dependent and view invariant neurons analogous to IT cells. The model layers are organized to mirror the layers from V1 to IT in the ventral stream and can be extended up to prefrontal cortex [25]. The view-based module leads to complex feature selective units that are scale and translation invariant [23]. Through the layers of the model, increasingly complex feature representations are achieved by combining intermediate features using template matching Gaussian transfer function (see Appendix A.1). Scale and translation invariance are achieved by using a max-pooling operation over similar features with different scales and translations.

Tuning within the model can generally be interpreted as representing a conjunction of non-linear filters that is translation invariant within a unit's receptive field. Simulations using the methodology of Pasupathy and Connor in [18], show that units within the standard model correspond to V4 cells and exhibit object-centered, position-specific boundary conformation tuning.

2.2 Model Implementation

The model used in this paper is an extension of original, simplified version described by Riesenhuber and Poggio [23] in three ways: the combination of afferents by S2 units is less rigid, S2 units are tuned to a target stimulus (possibly by learning, see [29]), and the C2 layer pooling range is set to match the invariance properties of a V4 cell. These changes are natural (and planned, see [23]) extensions of the original model. They were made possible in a quantitative way as a consequence of the results found by Pasupathy and Connor in [18]. In this paper we will refer to S2 and C2 units that incorporate these changes. A comparison of the new units to the original model units is presented in Appendix A.2.

The present version of the model, as used here, is shown in Figure 1 and consists of five layers: S1, C1, S2, C2, and VTU. The 'S' layers perform the template matching function and the 'C' layers perform the max-pooling operation. The S1 layer units perform a convolution on regions of the raw input image using Gabor filters at different orientations and sizes. The entire population of S1 units represents a convolution map of Gabor filters of different sizes and orientations with the entire raw image.

The C1 layer performs a max-pooling operation on the S1 convolution maps of the same orientation. The max-pooling function provides some scale and translation invariance that is characteristic of complex cells in V1. S2 units perform Gaussian tuning in a multi-dimensional space created from the outputs of C1 units.

In general, we expect a variety of cells tuned in different ways within V4, possibly by a process of passive learning based on visual experience. In this study we created specific units by the following procedure. S2 units are tuned to a particular set of C1 inputs with a Gaussian function. A target stimulus is presented to the model and the outputs of the C1 layer form the center of the S2 Gaussian tuning function. Feature tuning may be considered as a type of learning in that the S2 unit learns an optimal input pattern. This method is a departure from the hard-wired tuning of the original model [23]. A similar method follows the original intents of the model and was in fact used recently to successfully recognize objects in real world settings [28, 29].

The S2 units were tuned to a target stimuli taken from the main stimulus set, (see Figure 2). The target stimulus spanned a 3×3 arrangement of C1 unit spatial locations, creating 9 spatially distinct C1 locations. 2 or 3 of these spatial locations were chosen as inputs to the S2 unit (the spatial locations in the 3×3 map will be considered as: top left, top middle, top right, left middle, middle, right middle, lower left, lower middle, and lower right). C1 unit outputs from all Gabor filter orientations (in this case 4 orientations : 0°, 45°, 90°, and 135°) are included as input from each spatial location. This results in S2 units that take either 8 or 12 C1 inputs (2 or 3 spatial locations $\times$ 4 Gabor orientations).

The units in the C2 layer then perform the max-pooling operation over a spatial region of shifted S2 units with identical tuning properties. The size of the spatial pooling region is set to match the invariance properties of the V4 cell shown in Figure 6A of [18]. C2 units are directly compared to V4 cells. To complete the model for object recognition, VTUs are tuned to object views as in the original model [23]. The VTU layer will only be used here for object recognition benchmarks (see Appendix A.4). Further details of the model implementation are described in Appendix A.1.

2.3 Response Characterization

The methodology used to characterize C2 units follows the methodology used by Pasupathy & Connor to characterize V4 cells [18]. The stimulus set is shown in Figure 2 and is reproduced using code kindly supplied by Anitha Pasupathy. The construction of the stimulus sets and the data analysis methods used to characterize responses in various tuning spaces are described in detail in [18].

2.4 V4 Cell Raw Responses

The raw responses of V4 cells described in Figures 2, 4, 5, and 6 of [18] were extracted from digital images of the Figures. Raw V4 cell responses were then scaled between 0 and 1. Correlation coefficients were computed between a cell's scaled response to the 366 stimuli and the response of the model C2 unit to the same 366 stimuli to determine how well the model response matched the cell response.

2.5 Tuning Spaces

In addition, model unit tuning was also characterized using the same shape space analysis used by Pasupathy & Connor [18]. Multi-dimensional Gaussian functions were fit for each model unit in a shape space based on the stimuli. The multi-dimensional functions used to characterize model responses are: 2-D boundary conformation, 4-D boundary conformation, edge orientation, and edge orientation + contrast polarity.

The 2-D boundary conformation domain represents the contour elements of each stimuli in a curvature $\times$ angular position space. The 4-D boundary conformation domain not only contains the same curvature $\times$ angular position space as the 2-D boundary conformation space but also includes two adjacent curvature dimensions (i.e. the central curvature is augmented by the curvatures of the contour segments that are counterclockwise and clockwise adjacent).

An edge orientation shape space analysis was used to determine if responses were selective for flat contour segments at specific orientations. For this space each contour segment of a stimulus was parameterized by the angle between the tangent line and the horizontal.

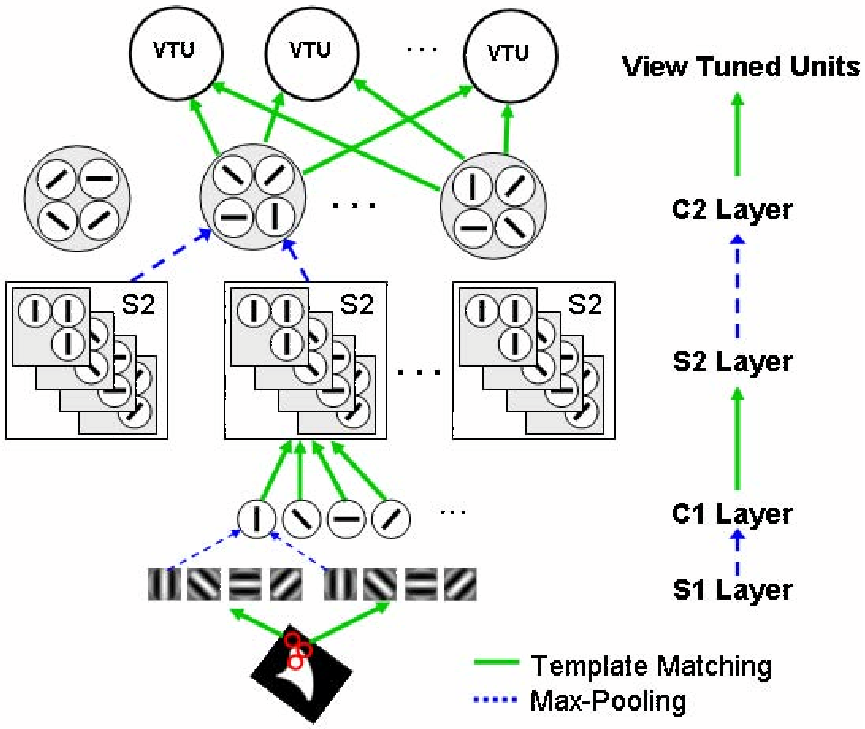

Figure 1: Schematic diagram of the implementation of the standard model of object recognition used in this paper, which is an extension of the model described in [23]. S1 filters come in four different orientations (0°, 45°, 90°, 135°), and each S2 unit is tuned to a unique combination of C1 afferents. C2 units perform max-pooling over S2 units of the same feature. C2 units of each type provide inputs to the view-tuned units with tuning properties as found in inferotemporal cortex [14, 23]. This study focuses on the units in the C2 layer that are analogous to cells in V4 recorded by Pasupathy and Connor [18].

The edge orientation space was represented by a Gaussian function over the 180 degree range of edge orientations.

2.6 Stimulus Translation

Model units were tested for invariance to stimulus translation. One optimal stimulus and one non-optimal stimulus was tested over a grid of multiple positions. The grid consisted of a 5×5 square centered on the receptive field with a translation of 0.5× S2 receptive field radius for each location on the grid.

2.7 Feature Translation Test

An additional stimulus set was used from [18] that varied the orientation and offset of a convex projection of a tear dropped stimuli. This set was used to test the dependence on angular position and orthogonal offset by fitting the response to this stimulus set to a 1-D Gaussian in either tuning space. The response to this stimulus set was used by Pasupathy and Connor as evidence for relative position tuning (a higher dependence on angular position would indicate relative position tuning). Further details are described in [18].

3 Results

3.1 Shape Tuning

Figure 3 shows the responses to the main stimulus set of a V4 cell and a C2 unit. The V4 cell response is adapted from Figure 4 of [18]. Both responses are linearly scaled between 0 and 1. The C2 unit was tuned to the stimulus shown in the lower left portion of the same figure and takes inputs from upper right and middle right spatially located C1 units. The V4 cell response is plotted against the C2 unit response in the lower right portion of Figure 3.

There is a good correspondence between the V4 response and the C2 response. Generally, both exhibit high responses to stimuli with concave curvature to the right of the object. Many subtleties of the responses match, such as the many stimuli that exhibit approximately half maximum responses. These similarities result in a high correlation coefficient of 0.77 between the V4 cell response and the C2 unit response. The C2 unit also shows shape space tuning similar to the V4 cell. Correlation coefficients of 0.67 in the 4-D boundary conformation tuning space and 0.32 in the edge orientation space both agree with the correlation coefficients found

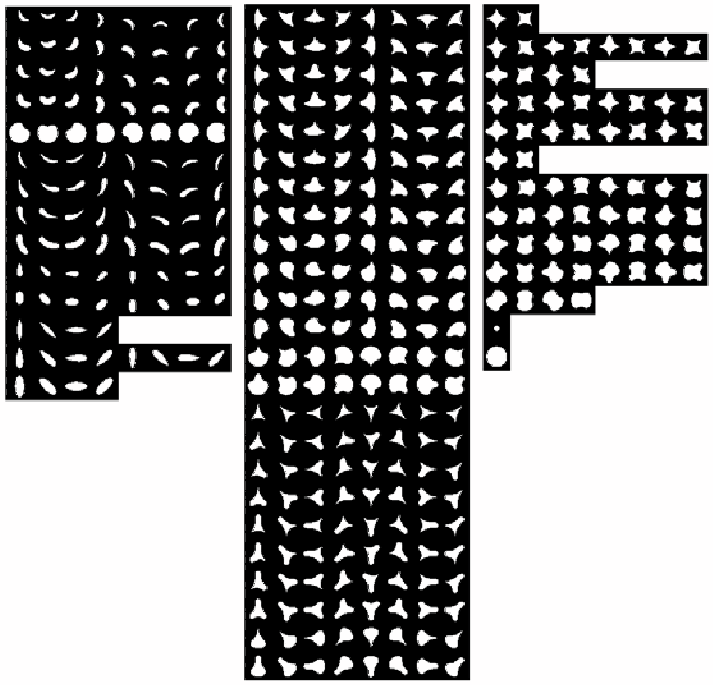

Figure 2: Stimulus set used for computational simulations based on [18]. Each closed white shape represents one individual stimulus. The background is uniform black and fills the receptive field.

for the V4 cell.

Other C2 units achieved good fits for the specific V4 cells shown in Figures 2 and 5 from [18]. A C2 unit constructed with spatial locations of the C1 subunits at middle, middle right and lower left and tuned to a stimuli with a sharp convex projection to the lower left showed a correlation coefficient of 0.67 to the V4 cell in Figure 2 of [18]. This unit also exhibited a high correlation in 4-D boundary conformation tuning space of 0.72 and a low correlation in edge orientation tuning space of 0.32. Another C2 unit, constructed with C1 spatial locations at upper right and lower left and tuned to a stimuli with a sharp convex projection to the upper right, showed a correlation coefficient of 0.70 to the V4 cell in Figure 5 of [18]. The 4-D boundary conformation correlation for this unit was 0.75 and the edge orientation correlation was 0.28.

The results for the three C2 units described here are summarized in Table 1.

3.2 Invariance

Figure 4 shows the responses to an optimal and non-optimal stimuli for a V4 cell, adapted from Figure 6A of [18], and a C2 unit over a 5×5 translation grid. The C2 unit shows high responses to the optimal stimuli over a translation range that is comparable to the V4 cell. For the non-optimal stimuli, the C2 unit shows low response over all translations. This shows that stimulus selectivity is preserved over translation for the C2 unit. The degree of translation invariance is comparable to the invariance range of the V4 cell.

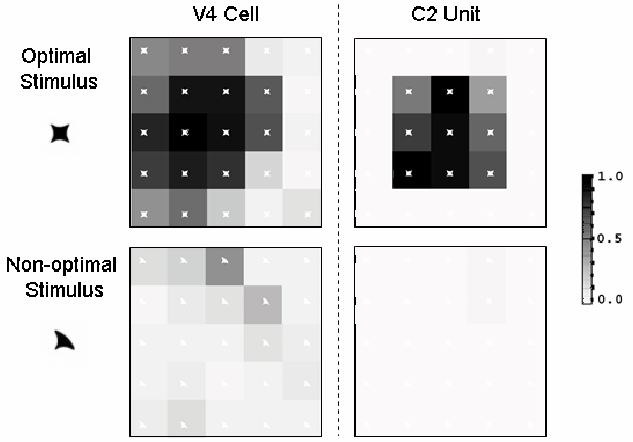

Figure 4: Translation response to an optimal and non-optimal stimulus for a V4 cell and a C2 unit. Responses are scaled between 0 and 1. (V4 cell response adapted from Figure 6A of [18])

Figure 5 shows the responses of a V4 cell and an example C2 unit to the feature translation stimulus set. The C2 unit shows a response pattern that is nearly identical to the V4 cell, adapted from Figure 6B of [18]. The C2 response is highly correlated with the angular position of the convex extremity and poorly correlated

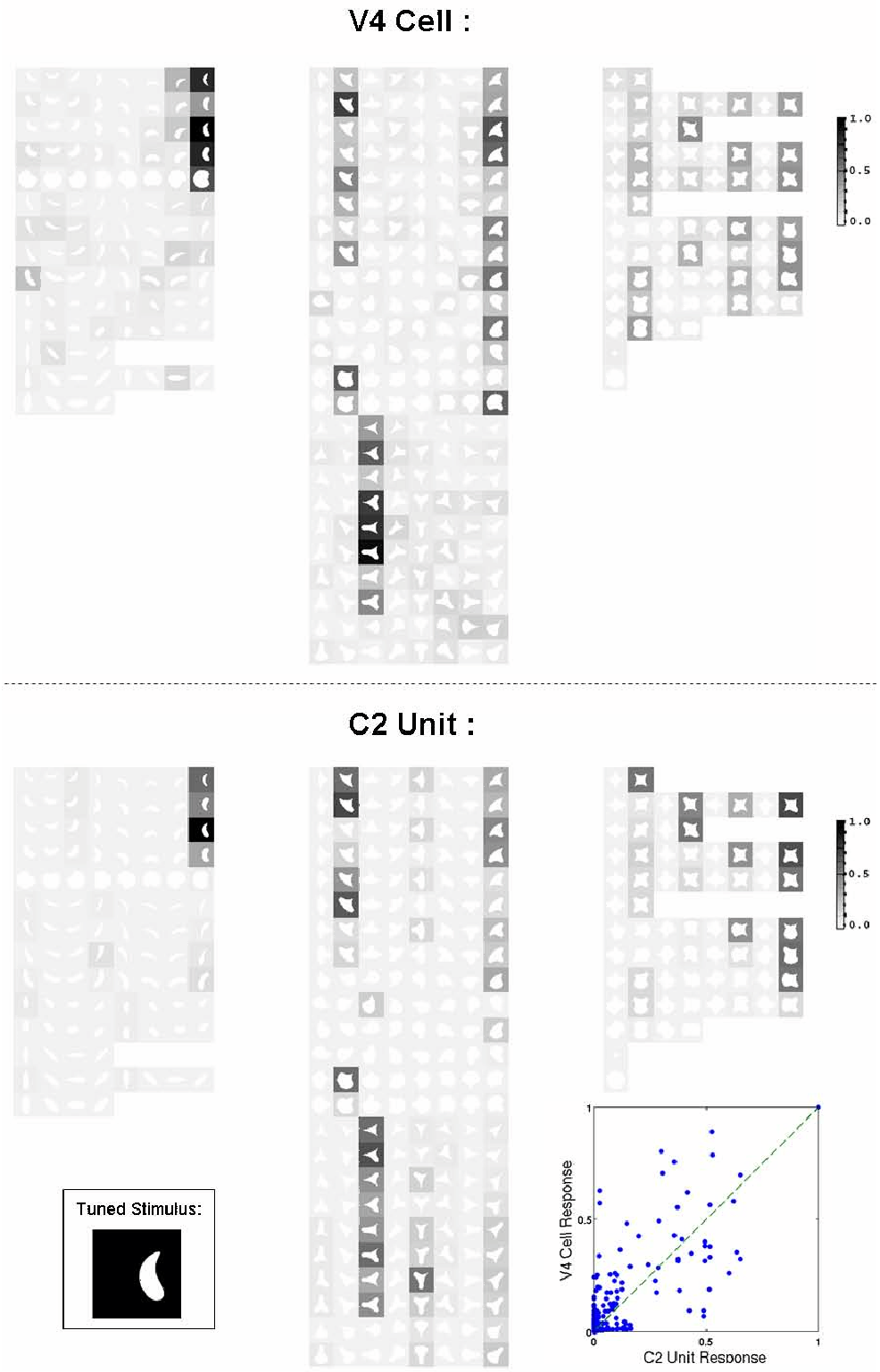

Figure 3: V4 cell (top) and C2 model unit (bottom) responses to the main stimulus set. (V4 cell response adapted from Figure 4 of [18]) The response to each stimulus is plotted in terms of the shading surrounding the stimulus and has been linearly scaled between 0 and 1. The darker the background the higher the response exhibited to that stimulus. The tuned stimulus for the S2 inputs for this C2 unit is shown in the lower left. The V4 cell response is plotted against the C2 unit response in the lower right.

	V4 Cell	C2 Unit	V4 Cell	C2 Unit	V4 Cell	C2 Unit
Figure from [18]	4	-	2	-	5	-
Correlation Coefficient	0.77		0.67		0.72	
4-D Boundary	0.81	0.67	0.82	0.72	0.85	0.75
Edge Orientation	0.38	0.32	0.25	0.32	0.31	0.28

Table 1: Comparison of V4 cells to C2 units showing: correlation coefficient between V4 cell response and C2 unit response to the main stimulus set, 4-D boundary correlation coefficient for the response to the main stimulus set, and edge orientation correlation coefficient for the response to the main stimulus set.

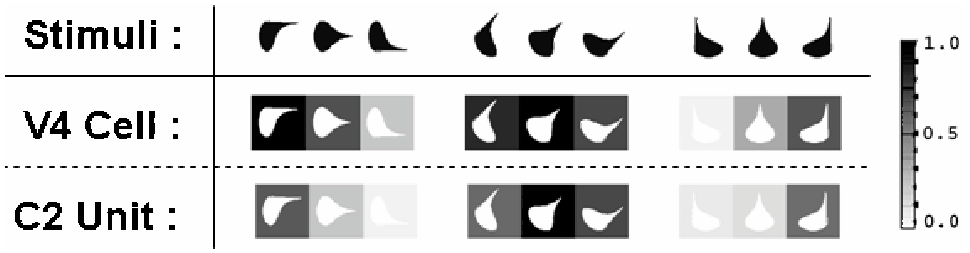

Figure 5: V4 cell and C2 unit response to the feature translation stimulus set. Responses are scaled between 0 and 1. V4 cell response is adapted from Figure 6B of [18].

with the orthogonal offset of the extremity (correlation coefficients of 0.74 and 0.04, respectively) with nearly identical results for the V4 cell (correlation coefficients of 0.72 and 0.08, respectively).

4 Discussion

The ability of C2 units to model V4 cell responses suggests an explanation of V4 cell tuning that is based on a conjunction of complex V1 cell responses. The model represents a biologically plausible mechanism that displays the same curvature and object centered tuning found by Pasupathy and Connor in a subpopulation of V4 cells. The model may be extended to predict the responses of these V4 cells to novel stimuli.

The standard model will also allow the integration and comparison of findings across physiological V4 studies. An additional set of constraints is imposed by each V4 study, limiting possible model connections and selectivities. For example, a previous study has found constraints on standard model units based on V4 findings [12]. Through such studies, a unified model of V4 connectivity and selectivity can be achieved.

C2 model units demonstrate how a local feature combination can create object-centered tuning (as defined in [18]). C2 units demonstrate selectivity that is dependent on the relative spatial locations of the C1 unit inputs combined at the S2 layer. The max-pooling operation between the S2 units and the C2 unit produces tuning that is independent of the stimulus translation and is thus independent of absolute receptive field position. As a result, C2 units demonstrate tuning for the relative spatial position of their sub-features. These results indicate that invariance mechanisms at different stages within the ventral stream may be closely associated with the tuning properties of V4 cells. For further analysis see Appendix A.3.

The specific C2 units described here display selectivity to curvature segments and therefore, may approximately be described as curvature filters. In much the same way that simple V1 cells can be thought of as filters selective for orientation [15, 16, 26, 27], some V4 cells can be thought of as filters selective for curvature. For example, C2 units show selectivity within a stimulus set of polar, hyperbolic and Cartesian gratings. The response of the C2 unit shown in Figure 3 was found for a stimulus set of 40 polar, 20 hyperbolic and 30 Cartesian gratings (similar to stimuli used in [5]). The response to the individual grating stimuli and the maximum response within each grating class is shown in Figure 6. This type of grating selectivity is consistent with reported V4 responses [5].

More generally, C2 units are tuned in a high dimensional space of C1 filters. Thus, C2 units show high selectivity to a wide range of stimuli that are not present in the closed shape stimuli sets of [18] and are not easily described as containing curvature or contours. For example, the natural image (from [13]) shown in Figure 7 elicits high responses from the C2 unit shown in Figure 3. Figure 7 shows a natural image, the response map of the C2 unit shifted over the image, and an image patch that produces a high C2 response. The C2 response to this image patch is 0.75 (the highest and lowest responses in Figure 3 are 1.0 and 0.0, respectively). Note that this image patch does not obviously contain

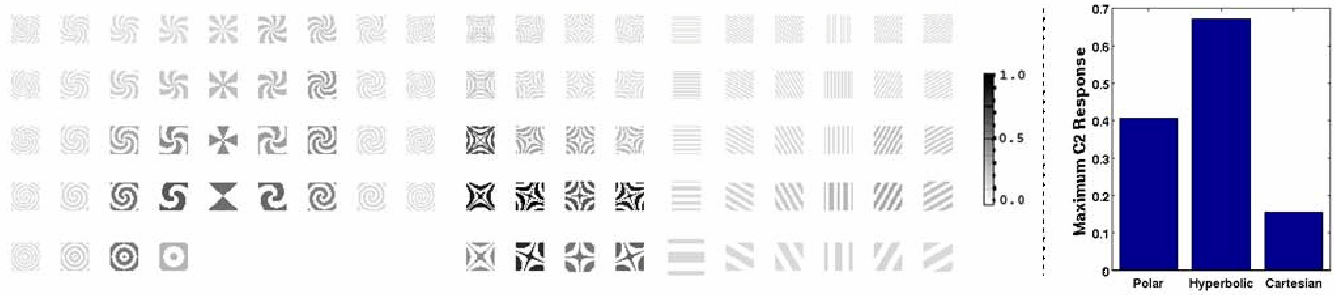

Figure 6: Response pattern of a C2 unit to polar, hyperbolic and Cartesian gratings. Response is indicated by the grey level of the stimulus and have been normalized between 0 and 1. The maximum response to each grating type is shown to the right. The C2 unit used for these measurements is the same unit shown in Figure 3.

the curvature found to produce a high response for this unit – and it is in any case not clear how to define curvature for grey level images such as this one.

In conclusion, the standard model can be used to quantitatively predict how the V4 cells studied by Pasupathy and Connor will respond to novel stimuli, such as natural and arbitrary gray level images. These predictions, which may be incorrect, suggest a possible empirical test, driven by the standard model, to further explore shape selectivity in area V4.

5 Acknowledgements

The authors thank Anitha Pasupathy and Charles Connor for sharing the stimulus sets, granting the use of V4 recording data, and providing useful comments. Their study of V4 shape representation was instrumental in developing the version of the standard model presented in this paper.

Appendix

A.1 Model Units

The hierarchical structure of the standard model is based on two computations to produce invariance and selectivity. Invariance is achieved by max-pooling over afferents with identical tuning properties but transformed over the invariant dimension, i.e. scale or translation. For example, the output of a C1 cell taking inputs from $S1_j$ afferents with different spatial locations is given as

$$C1 = \max_j[S1_j]$$

and produces translation invariance. Selectivity is achieved using a template matching Gaussian transfer function. For example, if each S2 unit combines n C1 afferents, the output of the S2 unit is

$$S2 = \prod_{i=1}^{n} e^{-(C1_i - \mu_i)^2/2\sigma^2}.$$

The parameter μ_i is the C1 target and σ is the standard deviation of the Gaussian. The C1 targets and the standard deviation are set for each modeled V4 cell using the following procedure.

For each given experimental V4 neuron, the stimulus that elicits the highest response is selected as the target stimulus for the model S2 unit intended to replicate the experimental neuron's tuning. A 3×3 grid of half overlapped C1 units at all 4 Gabor orientations is centered on the target stimulus. To simplify the combinations of C1 units considered at the S2 level, C1 units of different orientation but the same spatial location are grouped together. From this set of 9 unique spatial locations (a 3×3 grid with all 4 Gabor orientations at each location) S2 units are created that combine 2 or 3 spatial locations at a time. This creates a total of 120 ($\binom{9}{2} + \binom{9}{3}$) possible C1 afferent combinations. The standard deviation for each S2 unit Gaussian is determined by minimizing the mean squared error between the V4 cell response and the C2 unit response.

Each C2 unit takes inputs from a 7×7 grid of S2 units with the same tuning properties but shifted by half the receptive field of the C1 units. This produces C2 units with receptive fields of about 1.5 × S2 receptive field size. The pooling range for the C2 units was set to compare with the translation invariance of the V4 cell shown in Figure 6A of [18].

In summary, the variable parameters involved in this methodology are: the C2 pooling range (1), the tuned stimulus (1), the standard deviation of the S2 Gaussian (1), and the combination of C1 spatial locations (3). This gives a total of 6 'degrees of freedom.' The C2 pooling range is determined by the example V4 cell shown in Figure 6A of [18]. The stimulus resulting in the highest response for a given V4 cell is the tuned stimulus. An optimal standard deviation of the S2 Gaussian is determined based on the given V4 response to the main stimulus set for each of the C1 spatial locations considered (2 or 3 locations from the 3×3 spatial grid). From this population, the unit with the highest correlation to the given V4 cell is selected.

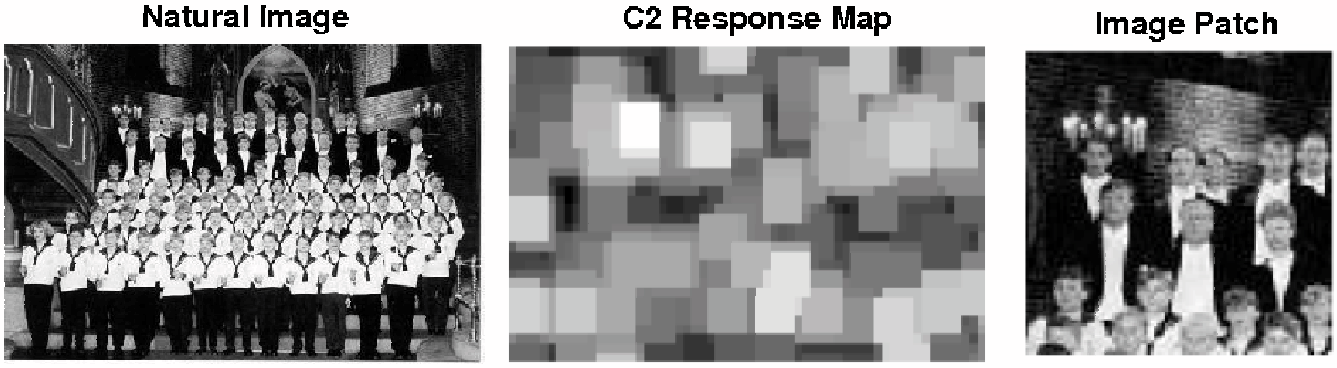

Figure 7: A natural image from [13] (left) and the response pattern of a C2 unit (middle) for the C2 unit shown in Figure 3. The C2 unit's receptive field is shifted over the image and its response is shown for each shift as a gray level with high response represented as white pixels. An image patch (right) that elicits a C2 response of 0.75 is magnified.

We would expect a wide variety of connections and selectivities throughout the ventral stream leading to V4. Because of computational limitations we are only able to model a restricted subset of the possible connections. For example, S2 units in this paper rigidly combine all 4 orientation filters from each spatial subfield. If a slightly less rigid combination of C1 inputs is considered, the correspondence of C2 units to V4 cells can be improved further. For example, a C2 unit that does not combine all 4 orientation filters at each spatial location can achieve a correlation coefficient of 0.84 over the main stimulus set with the V4 cell shown in Figure 3.

A.2 Comparison to the Original Version of the Model

The parameters and connections of the units described in A.1 and used throughout this paper were made as a direct result of the work by Pasupathy and Connor in [18]. These units differ from the units in the original model [23] in three ways.

The first difference is that the original S2 units take inputs from a 2×2 grid of C1 spatial locations with only 1 Gabor orientation at each spatial location. The units used here take a subset of inputs from a 3×3 grid of C1 spatial locations with all 4 Gabor orientations at each spatial location. The resulting S2 units are less rigid in the combination of spatial subunits and have a more descriptive representation of the input at each spatial location.

The second difference is that the S2 unit Gaussian transfer function [21] in the original model has a fixed center (all C1 targets, μ_i, are set to 1). The S2 units used here learn the center of the Gaussian transfer function from the tuned stimulus. As a result, these S2 units are tuned to a specific pattern of C1 activation, while the original S2 units are tuned to maximum firing of all C1 afferents.

The final difference is that the original C2 units have a larger pooling range. The original pooling range was simply set to match data from a population of V4 cells, while the C2 pooling range used here is set to model one individual V4 cell invariance. In general we expect a spectrum of pooling ranges in the population of V4 cells.

A.3 Relative vs. Absolute Position Tuning

A simple simulation shows how relative position tuning may arise from the max-pooling characteristics of C2 units. A simple 3 layer network that takes inputs from a 1-D space with 2 unique features was created and its output was fit in either an absolute position space (absolute position of each feature) or a relative position space (distance between the two features) using multi-dimensional Gaussians. The 3 layer network consists of a C1-like layer that performs feature detection, an S2-like layer that combines the two features at specific C1 spatial locations and a C2-like layer that pools S2-like units over a max-pooling range.

Figure 8 shows the correlation coefficients to the absolute position tuning and the relative position tuning for the C2-like unit as the max-pooling range of the C2-like unit varies from 1 to 21 S2-like units. For small pooling ranges the absolute position tuning is more effective at describing the response of the network, while at larger pooling ranges the relative position tuning becomes more effective at describing the response. These findings can naturally be extended to the 2-D case of vision.

This simulation indicates that relative position tuning can be an effect of mechanisms that produce invariance, such as max-pooling. For example, the relevance of relative position tuning for describing V4 cell responses points to an invariance mechanism that maintains relative feature properties at the expense of absolute position properties. As the translation invariance range

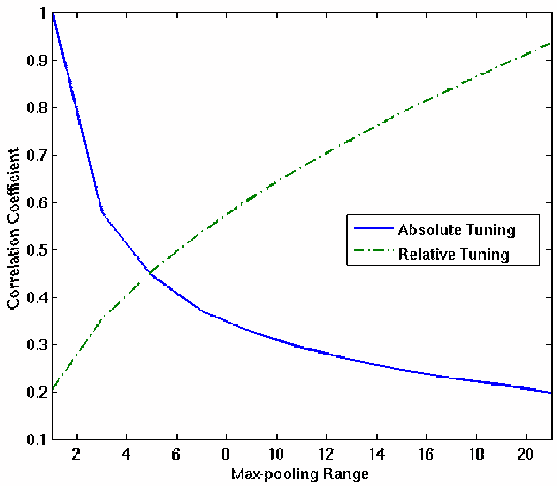

Figure 8: Correlation coefficients of relative and absolute position tuning as a function of max-pooling range for a simplified network. As the max-pooling range increases, relative position tuning becomes more effective at describing the network response.

increases for a V4 cell, relative position tuning should become a better predictor of neuron response.

A.4 Paperclip Benchmark

A test is necessary to show that the extension of the original model [23] used here produces quantitatively similar selectivity and invariance at the VTU level. A benchmark borrowed from [23] compares the single VTU object recognition performance in both models on a set of paperclip stimuli under rotation, scale and translation transformations.

For the benchmark, a population of S2 units was constructed by selecting 3 C1 spatial locations from a 3×3 grid at a time (ignoring redundancies) and then randomly sampling the center of the S2 Gaussian function for each unit five times to achieve an S2 population of 250 units (50 spatial location groupings × 5 random samples). The C2 pooling range was set to match the range of the original model. Therefore, this benchmark compares the original S2 units to the new S2 units presented in this paper. The results for the paperclip benchmark, shown in Figure 9, indicate that the model with the new S2 units achieves a similar degree of rotation, scale and translation invariance and selectivity against distractors as the original model.

References

[1] J. Baizer, D. Robinson, and B. Dow. Visual responses of Area 18 neurons in awake, behaving monkey. *Journal of Neurophysiology*, 40:1024–1037, 1977.

[2] R. Desimone, T. Albright, C. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the Macaque. *Journal of Neuroscience*, 4:2051–2062, 1984.

[3] R. Desimone and S. Schein. Visual properties of neurons in Area V4 of the Macaque: Sensitivity to simulus form. *Journal of Neurophysiology*, 57:835–868, 1987.

[4] D. Felleman and D. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex*, 1:1–47, 1991.

[5] J. Gallant, C. Connor, S. Rakshit, J. Lewis, and D. Van Essen. Neural responses to polar, hyperbolic, and Cartesian gratings in Area V4 of the Macaque monkey. *Journal of Neurophysiology*, 76:2718–2739, 1996.

[6] T. Gawne and J. Martin. Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *Journal of Neurophysiology*, 88:1128–1135, 2002.

[7] C. Gross, C. Rocha-Miranda, and D. Bender. Visual properties of neurons in inferotemporal cortex of the Macaque. *Journal of Neurophysiology*, 35:96–111, 1972.

[8] D. Hubel and M. Livingstone. Segregation of form, color, and steropsis in primate Area 18. *Journal of Neuroscience*, 7:3378–3415, 1987.

[9] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, 160:106–154, 1962.

[10] D. Hubel and T. Wiesel. Receptive fields and functional architecture in two nonstriate visual ares (18 and 19) of the cat. *Journal of Neurophysiology*, 28:229–289, 1965.

[11] E. Kobatake and K. Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the Macaque cerebral cortex. *Journal of Neurophysiology*, 71:856–867, 1994.

[12] M. Kouh and M. Riesenhuber. Investigating shape representation in Area V4 with HMAX: Orientation and grating selectivities. *MIT AI Memo*, 021, 2003.

[13] F.F. Li. 100 natural photograph dataset, July 2002. Available online: www.vision.caltech.edu/feifeili/Datasets.htm.

[14] N. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5:552–563, 1995.

[15] L. Mahon and R. De Valois. Cartesian and non-Cartesian responses in LGN, V1, and V2 cells. *Visual Neuroscience*, 18:973–981, 2001.

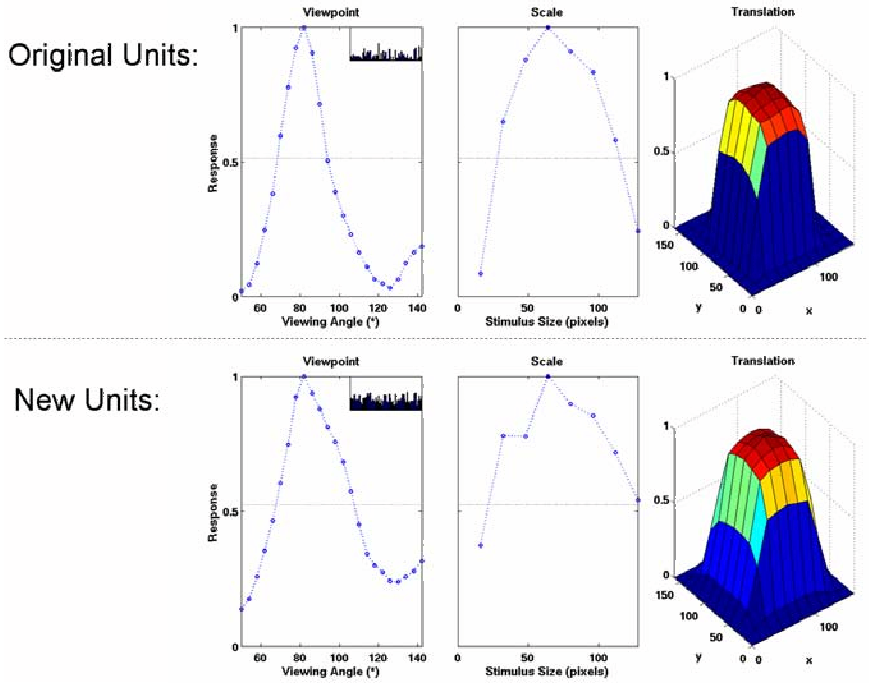

Figure 9: Paperclip benchmark comparing the model using original units to the model using a novel S2 unit population. Note that the original and new units show similar degree of invariance (view point rotation, scale, and translation) and selectivity against distractors (indicated by horizontal line). See [23] for more details.

[16] F. Mooser, W. Bosking, and D. Fitzpatrick. A morphological basis for orientation tuning in primary visual cortex. *Nature Neuroscience*, 7:872–879, 2004.

[17] A. Pasupathy and C. Connor. Responses to contour features in Macaque Area V4. *Journal of Neurophysiology*, 82:2490–2502, 1999.

[18] A. Pasupathy and C. Connor. Shape representation in Area V4: Position–specific tuning for boundary conformation. *Journal of Neurophysiology*, 86:2505–2519, 2001.

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

[20] D. Perrett, E. Rolls, and W. Caan. Visual neurons responsive to faces in the monkey temporal cortex. *Exp. Brain Res.*, 47:329–342, 1982.

[21] T. Poggio and E. Bizzi. Generalization in vision and motor control. *Nature*, 431:768–774, 2004.

[22] D. Pollen, A. Przybyszewski, M. Rubin, and W. Foote. Spatial receptive field organization of Macaque V4 neurons. *Cerebral Cortex*, 12(6):601–616, 2002.

[23] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

[24] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Current Opinions in Neurobiology*, 12:162–168, 2002.

[25] M. Riesenhuber and T. Poggio. How the visual cortex recognizes objects: The tale of the standard model. *Visual Neuroscience*, 2:1640–1653, 2003.

[26] D. L. Ringach. Spatial structure and symmetry of simple-cell receptive fields in Macaque V1. *Journal of Neurophysiology*, 88:455–463, 2002.

[27] T. Serre and M. Riesenhuber. Realistic modeling of cortical cells for simulations with a model of object recognition in cortex. *MIT AI Memo*, 2003.

[28] T. Serre, M. Riesenhuber, J. Louie, and T. Poggio. On the role of object-specific features for real world object recognition in biological vision. *Biologically Motivated Computer Vision*, Second International Workshop (BMCV 2002):387–397, 2002.

[29] T. Serre, L. Wolf, and T. Poggio. High-performance object recognition with features inspired by visual cortex. *Currently in submission*, 2004.

[30] K. Tanaka, H. Saito, Y. Fukada, and M. Moriya. Coding visual images of objects in the inferotemporal cortex of the Macaque monkey. *Journal of Neurophysiology*, 66:170–189, 1991.

[31] L. Ungerleider and M. Mishkin. *Two cortical visual systems*. MIT Press, 1982.