Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2003-019
AIM-2003-021
CBCL-231

September 8, 2003

Investigating shape representation in area V4 with HMAX: Orientation and Grating selectivities

Minjoon Kouh and Maximilian Riesenhuber

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

Abstract

The question of how shape is represented is of central interest to understanding visual processing in cortex. While tuning properties of the cells in early part of the ventral visual stream, thought to be responsible for object recognition in the primate, are comparatively well understood, several different theories have been proposed regarding tuning in higher visual areas, such as V4. We used the model of object recognition in cortex presented by Riesenhuber and Poggio (1999), where more complex shape tuning in higher layers is the result of combining afferent inputs tuned to simpler features, and compared the tuning properties of model units in intermediate layers to those of V4 neurons from the literature. In particular, we investigated the issue of shape representation in visual area V1 and V4 using oriented bars and various types of gratings (polar, hyperbolic, and Cartesian), as used in several physiology experiments. Our computational model was able to reproduce several physiological findings, such as the broadening distribution of the orientation bandwidths and the emergence of a bias toward non-Cartesian stimuli. Interestingly, the simulation results suggest that some V4 neurons receive input from afferents with spatially separated receptive fields, leading to experimentally testable predictions. However, the simulations also show that the stimulus set of Cartesian and non-Cartesian gratings is not sufficiently complex to probe shape tuning in higher areas, necessitating the use of more complex stimulus sets.

Copyright © Massachusetts Institute of Technology, 2003

This report describes research done within the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

This research was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032.

Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sumitomo Metal Industries, Toyota Motor Corporation, WatchVision Co., Ltd., and The Whitaker Foundation. M.R. is supported by a McDonnell-Pew Award in Cognitive Neuroscience.

1 Introduction

The ventral visual pathway, from primary visual cortex, V1, to inferotemporal cortex, IT, is considered to be responsible for object recognition in the primate ("what pathway"). In V1, neurons tend to respond well to oriented bars or edges. Neurons in the intermediate visual areas are no longer tuned to oriented bars only, but also show responses to other forms and shapes, at a level not found in primary visual cortex [7, 10]. Finally, in IT, neurons are responsive to complex shapes like the image of a face or a hand [4, 5, 10, 17].

Understanding how the neural population represents shape information and how such representations arise within the cortex is one of the main objectives of visual neuroscience. Many physiological studies have used different sets of visual stimuli in order to identify the underlying neural mechanisms. However, the nonlinear behaviors of the neurons in higher visual areas have made it difficult to determine the cortical computational mechanisms for increasing shape complexity: Considering the infinite number of functions that can be fitted to the limited set of data points (given by the responses of a neuron to a set of test stimuli), studies that rely on *post hoc* function fitting are doomed to fail. Rather, it is essential to have an *a priori*, biologically plausible computational hypothesis of how more complex features are built from simple features, a theory that provides testable predictions.

In this paper, we use the HMAX model of object recognition in cortex developed by Riesenhuber and Poggio [13], which has been shown to successfully model various aspects of invariant object recognition in cortex (for a recent review, see [14]), to provide a computational hypothesis of how complex features in V4 can be built from V1 cell inputs. We here focus our efforts on understanding the responses of V4 neurons to bars, Cartesian and non-Cartesian grating stimuli, as arising from a combination of V1 complex cell inputs.

2 Methods

2.1 The HMAX Model of Object Recognition in Cortex

The HMAX model, proposed by Riesenhuber and Poggio [13], is composed of four hierarchical feed-forward layers, labelled as S1, C1, S2, and C2. In the first layer, S1, a stimulus image is convolved with linear filters (*e.g.*, difference of Gaussians or Gabor filters) of various orientations and sizes. At the C1 layer, the responses from S1 units lying within certain spatial and scale ranges are pooled over, and the maximum responses are forwarded to the next S2 layer. Such maximum-based pooling increases robustness to clutter, as well as invariance to stimulus translation and scaling [13]. (Some recent physiological evidences for the maximum operation within visual cortex can be found in [6, 8].) At the S2 level, responses of C1 units are combined into more complex features.

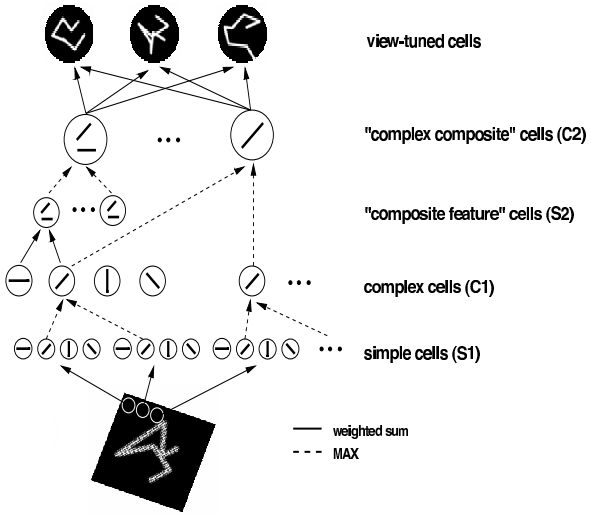


Figure 1: Schematic diagram of the HMAX model. In the standard version, S1 filters come in four different orientations (0°, 45°, 90°, 135°), and each S2 unit combines four adjacent C1 afferents in a spatial 2x2 arrangement, producing a total of 256 (4^4) different types of S2 units. At the final C2 layer, units perform another max pooling operation over all the S2 units of each type, yielding the 256 output units of the HMAX model, which can in turn provide an input to the view-tuned units with tuning properties as found in inferotemporal cortex [9, 13]. As the shape tunings of the corresponding S and C cells in the same layers are very similar, we here confine ourselves to an analysis of the shape tuning of S cells.

The receptive field of a model unit can be defined as the region of the input stimulus that produces an excitatory (positive) response of the unit. Due to the pooling operation in the C layers and the combination of afferents in the S layers, the receptive fields become progressively larger, going from the lower to the higher layers. In the current model, the receptive field of an S2 unit is about twice the size of an S1 receptive field, corresponding to the neurons in the fovea [4]. For example, an S2 unit that combines 2x2 C1 units, which in turn pool over 9x9 S1 units of 17–21 pixels with adjacent receptive field centers, will have a receptive field width of 38 pixels. The C2 units, which pool over the population of the S2 units, have the biggest receptive field size. The receptive field size of an S2 unit can be adjusted by using different pooling ranges or different feature combination schemes (other than the 2x2 spatial arrangement).

Thus, the HMAX model performs a series of weighted-sum template-matching operations in the S-layers and maximum-pooling operations in the C-layers that progressively build up feature complexity

and invariance to scaling and translation, respectively.

2.2 Approach

In our simulations, we measured the responses of the S1 and the S2 model units to different sets of stimuli (oriented bars and gratings). The baseline-subtracted responses were measured, where the baseline was defined to be the response to a null stimulus. The simulation procedures and stimuli were based directly on several physiological studies of the macaque monkeys [4, 5, 17], so that the simulation results with the HMAX model could be readily compared with the experimental data. Fig. 2 illustrates the rationale behind this study.

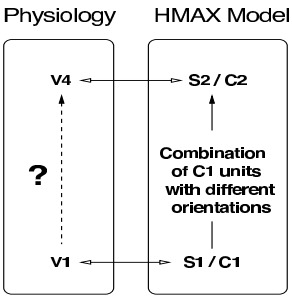


Figure 2: Experimental paradigm: The goal of the modeling study is to investigate possible computational mechanisms underlying the increase in feature complexity along the ventral pathway from V1 to V4 ("?" in the diagram). To this end, we compare the responses of the model units corresponding to the neurons in V1 and V4, with the physiological data using the same stimuli and procedures.

2.3 Stimuli

2.3.1 Bars

Our experimental procedure for the orientation selectivity study followed that of Desimone and Schein [4] as closely as possible. The stimuli were the images of bars at varying orientations (0° to 180° at 10° intervals) and widths (1, 5, 10, 15, 20, 25, 30, 50, and 70% of the receptive field size). The bars were always long enough to cover the whole receptive field and presented at different locations across the receptive field, as shown in Fig. 3.

The orientation tuning curve of a model unit was obtained by first finding the preferred width of the bar stimulus and then measuring the maximum (baseline-subtracted) responses over different bar positions at each orientation. The orientation bandwidth was defined as a full width at half maximum with linear interpolation. Figs. 6 and 7 show examples of orientation tuning curves.

Again following the convention used in [4], the contrast of the bar image was defined as the luminance difference between the bar and the background, divided by the background luminance. Throughout the experiment, the stimulus contrast was fixed at 90%. (Orientation selectivity was invariant for a wide range of contrasts, as shown in Appendix A.1.)

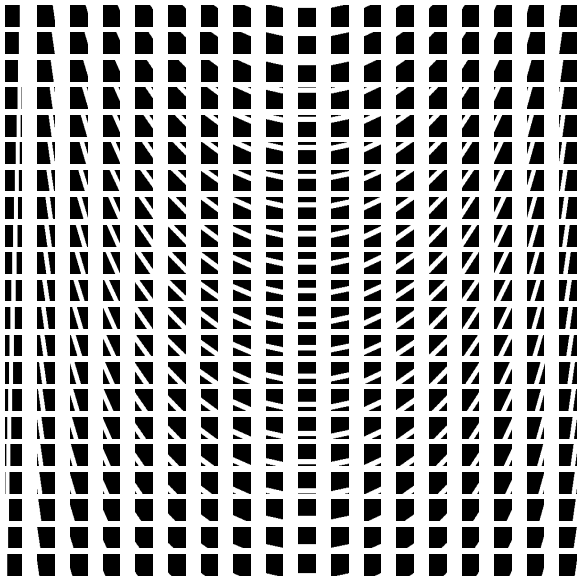


Figure 3: Examples of bar stimuli at varying orientations and positions. Each square corresponds to the receptive field of a model unit, and the width of the bars shown here is equal to 25% of the receptive field size.

2.3.2 Gratings

Three classes of gratings (Cartesian, polar, and hyperbolic) were prepared according to the following equations (as in [5]). For Cartesian gratings,

$$L_c(x,y) = A_0 + A_1 \cos(2\pi f u + \theta), \quad (1)$$
$$u(x,y) = x\cos\phi - y\sin\phi. \quad (2)$$

For polar gratings,

$$L_p(x,y) = A_0 + A_1 \cos(2\pi f_c c + 2\pi f_r r + \theta), \quad (3)$$
$$c(x,y) = \sqrt{x^2+y^2}, \quad (4)$$
$$r(x,y) = \tan^{-1}\frac{y}{x}. \quad (5)$$

For hyperbolic gratings,

$$L_h(x,y) = A_0 + A_1 \cos(2\pi f \sqrt{uv} + \theta), \quad (6)$$
$$u(x,y) = x\cos\phi - y\sin\phi, \quad (7)$$
$$v(x,y) = x\sin\phi + y\cos\phi. \quad (8)$$

The contrast of the grating stimuli was defined by

$$\text{Contrast} = \frac{L_{max} - L_{min}}{L_{max} + L_{min}}. \quad (9)$$

The mean value of the grating (A_0) was set to a nonzero constant, and its amplitude of modulation (A_1) was adjusted to fit the contrast of 90%.

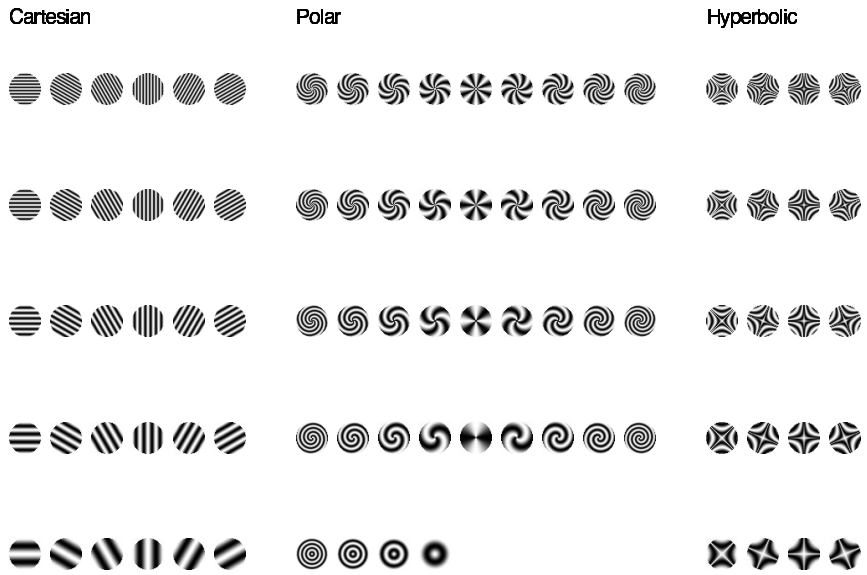


Figure 4: Grating stimuli (30 Cartesian, 40 polar, and 20 hyperbolic gratings) as used in [5].

These gratings were presented within the receptive field of a model unit at varying phases θ, in steps of 120° and 180° (as in [5]), and the baseline-subtracted maximum responses were calculated.

3 Results

3.1 Orientation Selectivity

3.1.1 (V1, S1)

Neurons in visual area V1 exhibit varying degrees of orientation selectivity. The upper left histogram in Fig. 8 shows the distribution of orientation bandwidth in V1 (from [17]). The median is 42°, while the median of the oriented cells alone (bandwidth $< 90°$) is 37°. These results are summarized in Table 1.

	Experiment		HMAX	
Bandwidths	V1	V4	S1	S2
Median:				
All neurons	42°	75°	39°	77°
Less than 90°	37°	52°	39°	59°
Percentage:				
Narrow ($< 30°$)	27%	5%	0%	0%
Wide ($> 90°$)	15%	33%	0%	11%

Table 1: Summary of the physiological data (V1, V4) and the simulation results (S1, S2). The experimental data were taken from [4, 17].

In the original HMAX model [13], each S1 feature was modeled as a difference of Gaussians. However, these features turn out to have an orientation bandwidth much broader (approximately $90°$) than found in the experiment [16], and the Gabor filters were shown to provide a good approximation to the experimental data in V1 [3, 15]. A Gabor filter is defined as

$$G(x, y) = \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \frac{\cos(kx - \phi)}{2\pi\sigma_x\sigma_y}. \qquad (10)$$

By varying σ_x, σ_y, and the wave number k, the properties of the Gabor filter can be adjusted [16]. The following parameters are used: Spatial phase $\phi = 0$, so that the peak is centered. Spatial aspect ratio (x *vs.* y), $\sigma_x/\sigma_y = 0.6$. The extent in the x direction, $\sigma_x = 1/3$ of the receptive field. The wave number $k = 2.1 \cdot 2\pi\cdot$. In this neighborhood of k, there are two inhibitory surroundings and one excitatory center as seen in Fig. 5. These parameters were chosen to produce a median bandwidth of 39°, close to the median of the V1 bandwidths.

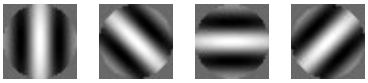


Figure 5: Gabor filters with circular receptive field at four different orientations. The circular masking was applied to reduce numerical differences between the filters at different orientations.

A Gabor filter produces an optimal response when the bar stimulus is oriented along the same direction as the filter itself. The S1 unit shown in Fig. 6 prefers the bar oriented at 0° with an orientation bandwidth of 34°. For a given set of parameters (σ_x, σ_y, and k), the orientation tuning curves of the Gabor filters at different sizes are almost identical to one another. (See the upper right histogram in Fig. 8.) Therefore, in our model, the distribution of the orientation bandwidths in the S1 layer

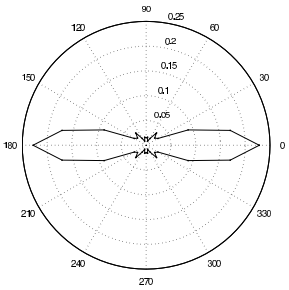


Figure 6: Tuning curve of an S1 unit with a Gabor-like receptive field. Because of the reflective symmetry of the bar stimuli, the data for 180°–360° are identical to the data for 0°–180°.

is very sharply peaked around a single value. However, as shown in the following section, even from this extremely homogeneous S1 population, the 2x2 feature combination at the S2 layer can create a wide variety of model units with different orientation bandwidths.

3.1.2 (V4, S2)

Moving from V1 to V4, the receptive field size increases, and neurons respond more to shapes of intermediate complexity [4, 5, 10, 17]. It is not clear exactly how and why neurons in V4 behave differently from those in V1.

In the standard HMAX model, because the afferents for the S2 units are systematically combined in a 2x2 arrangement of C1 afferents, every S2 unit can be categorized according to its geometric configuration of the four afferents, as shown in Table 2. Such a classification scheme turns out to be a meaningful tool for understanding the behavior of the S2 population to the bar stimuli.

For example, each orientation tuning curve in Fig. 7, typical of each class, shows that the responses to the bar stimuli depend strongly on how the afferent features are geometrically combined. Some model units, whose afferents are aligned in the same orientations, have very simple unimodal tuning curves resembling that of an S1 unit (group 8). For others (group 2–7), the tuning curves show multiple peaks at different orientations. Those in group 1, whose afferents are at orthogonal or non-parallel orientations to one another, exhibit little or no orientation tuning.

As a result, S2 units with similar feature configuration tend to have similar orientation bandwidths, as seen in Fig. 8. The S2 units in group 6, 7 and 8 have narrow bandwidths around 40°. Group 1 has an extremely broad orientation tuning profile due to its non-parallel, orthogonal afferents. The orientation bandwidths of group 3 and 4 are quite variable because of the secondary peaks: When those secondary peaks are small, only the primary peak contributes to the orientation bandwidth. Otherwise, the secondary peaks are merged with the primary peak to yield larger orientation bandwidths. Thus, by adjusting the model parameters that influence the sharpness and the relative size of the response peaks, it is possible to obtain different bandwidth distributions. In general, the distributions are upper bounded by group 1 with the flat orientation tuning profiles and lower bounded by group 6, 7, and 8.

Fig. 8 and Table 1 summarize one particular simulation result that produced a reasonable approximation to the physiological data. (See Appendix for the results using different sets of model parameters.) Note that on average, V4 neurons and S2 units tend to have wider orientation bandwidths than V1 and S1 units. With a median bandwidth of 75°, V4 neurons have wider orientation bandwidths than V1 neurons. In the model, there is a sizable increase in the population of cells with wider bandwidths. The actual percentage values are not very close to the physiological data, since the S1 population is too simple and homogeneous. (Only 11% of the S2 units in the current model are broadly tuned, whereas in V4, 33% of the neurons have wide bandwidths.) However, as seen in the next section, the model can cover a broad range of bandwidth distributions. By including a population of broadly tuned S1 units, the S2 layer will likely show more realistic distribution of orientation bandwidths.

3.1.3 (V1, S1) → (S2, V4)

The broadening of the orientation tuning from S1 to S2 layer is observed over a wide range of the model parameter values. In particular, the Gabor wave number k has a strong influence on both S1 and S2 bandwidths. Fig. 9 shows the changing shapes of the Gabor filter at different k values.

As k increases, S1 orientation bandwidth monotonically decreases. The orientation bandwidths of the S2 units also change, but rather disproportionally, as seen in Fig. 10. As explained before, for the S2 units in group 3 and 4, the secondary peaks in the orientation tuning profile can become significant enough and merge with the primary peaks to yield larger orientation bandwidths. When the S1 bandwidths get larger, the neighboring peaks in the S2 tuning profile are more likely to overlap, resulting in the sharp increase of the orientation bandwidths in Fig. 10.

Furthermore, Fig. 10 shows that with a homogeneous population of S1 units, it is possible to consistently construct a distribution of the S2 units with wider orientation bandwidths. Also note that the HMAX model can cover a wide range of orientation bandwidths in the S2 layer. Then, by incorporating a population of more broadly tuned S1 units, a larger percentage of S2 units would have a broad orientation tuning, yielding an even better fit to the experimental data. Thus, the

Group	Example	Number	Afferent Configuration
8	||||	4	All 4 in the same orientation.
7	|||/	32	3 in the same orientation, and the other at non-orthogonal orientation.
6	|||-	16	3 in the same orientation, and the other at orthogonal orientation.
5	||//	24	2 in the same orientation, and the other 2 in the same orientation that is non-orthogonal to the first 2.
4	||-/	96	2 in the same orientation, and the other 2 at different and non-orthogonal orientations to each other.
3	||/\	48	2 in the same orientation, and the other 2 at different and orthogonal orientations to each other.
2	||--	12	2 in the same orientation, and the other 2 in the same orientation that is orthogonal to the first 2.
1	-|/\	24	All 4 in different orientations.

Table 2: 8-class classification scheme for the 256 S2 units. In the *Example* column, the four characters represent the possible orientations of the afferent C1 units. The 2x2 geometric configuration was written as a 1x4 vector for notational convenience. The *Number* column shows the number of S2 units belonging to each class.

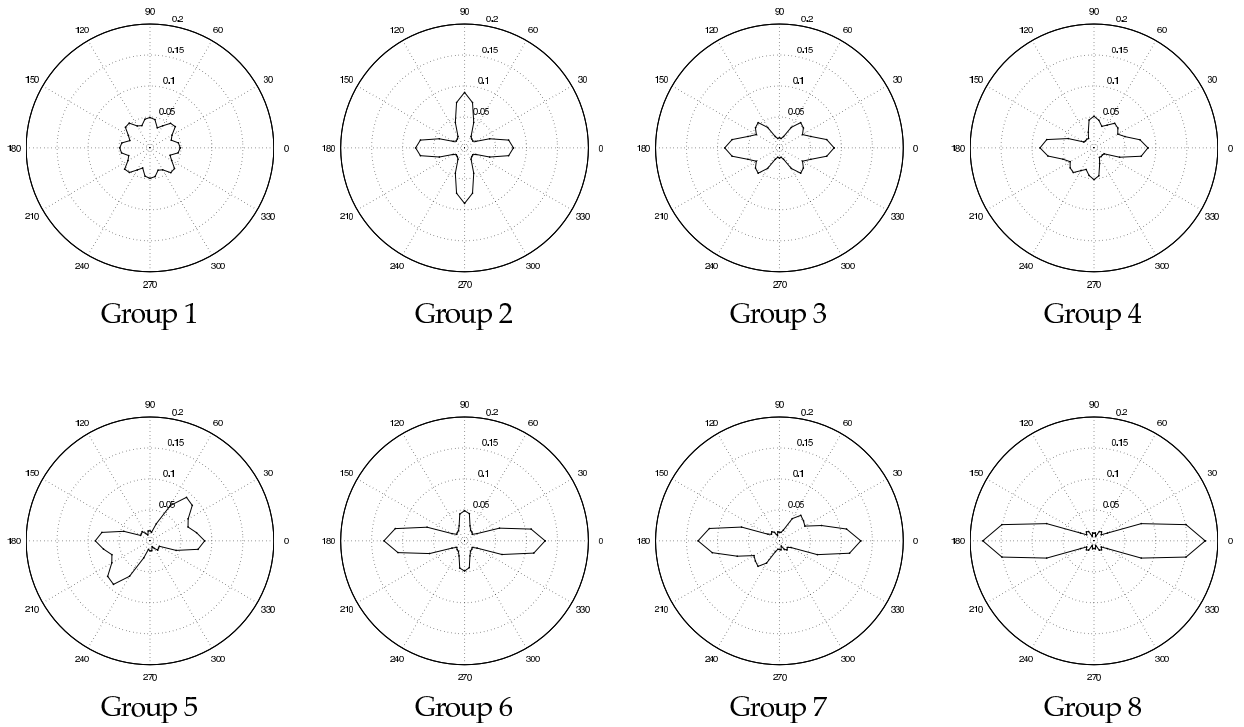


Figure 7: Sample tuning curves of S2 units: The S2 units in group 1 (cf. Table 2) do not respond much to the bar stimuli, yielding a flat tuning curve. Group 2 shows a sharp bimodal tuning, whereas in group 5, two peaks are merged to give a larger orientation bandwidth. Groups 3 and 4 have a large node and two small nodes, while group 6 and 7 have one large node and one small node, according to the geometric configuration of the afferents. Group 8 has a sharp, unimodal tuning curve. These tuning curves represent typical results for each group.

increase in orientation bandwidth from V1 to V4 found in the experiments can be explained as a byproduct of cells in higher areas combining complex cell afferents.

3.2 Grating Selectivity

3.2.1 (V1, S1)

Neurons in visual area V1 are known to be most responsive to bar-like or Cartesian stimuli, even though

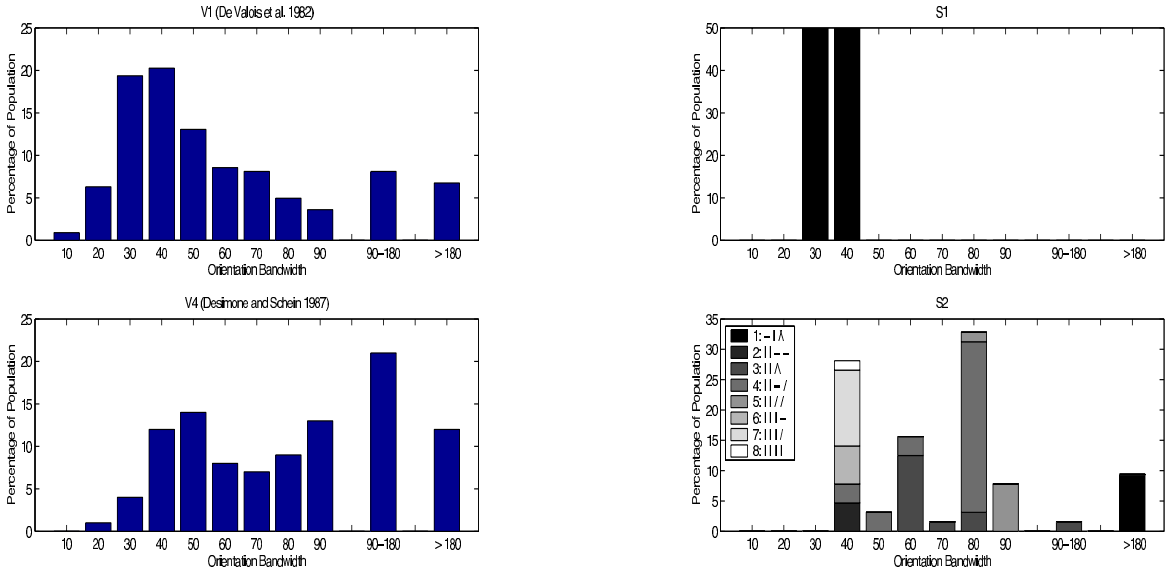


Figure 8: Distributions of the orientation bandwidths from the physiological data (V1 and V4, taken from [4, 17]) and from the simulation results (S1 and S2). The legend in the lower right histogram shows the 8-class classification scheme given in Table 2.

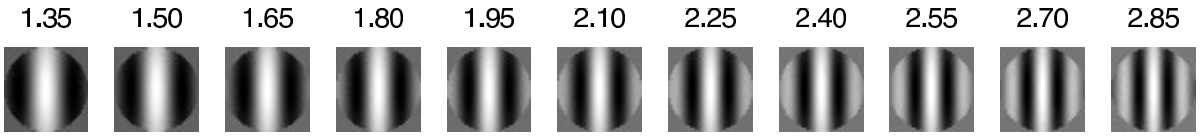


Figure 9: Gabor filters with varying wave numbers. From left to right, wave number k is increased from 1.35 to 2.85 in units of 2π. The central excitatory region becomes narrower, and the orientation bandwidth decreases, going from left to right.

there appears to be a small population of V1 cells more responsive to non-Cartesian stimuli [10]. In our model, the S1 population is quite homogeneous and clearly shows a bias toward Cartesian stimuli, as shown in Fig. 11.

3.2.2 (V4, S2)

Using three different classes of gratings as shown in Fig. 4, Gallant *et al.* [5] reported that the majority of neurons in visual area V4 gave comparable responses (within a factor of 2) to the most effective member of each class, while the mean responses to the polar, hyperbolic, and Cartesian gratings were 11.1, 10.0, and 8.7 spikes/second respectively, as summarized in Table 3. Furthermore, there was a population of neurons highly selective to non-Cartesian gratings. Out of 103 neurons, there were 20 that gave more than twice the peak responses to one stimulus class than to another: 10 showed a preference for the polar, 8 for the hyperbolic, and 2 for Cartesian gratings, as shown in Table 4.

When the HMAX model (with the same set of parameters used in the orientation selectivity studies) is presented with the same set of gratings, the S2 population exhibits a similar bias toward non-Cartesian gratings, as summarized in Tables 3 and 4.

	Gallant *et al.*	HMAX
Polar	11.1	0.14 ± 0.07
Hyperbolic	10.0	0.15 ± 0.06
Cartesian	8.7	0.05 ± 0.04

Table 3: Mean responses to three different classes of gratings. Physiological data are in units of spikes/second, whereas the model responses (baseline-subtracted) lie between 0 and 1. Even though the literal comparison of the numerical value is meaningless, the model units and the neurons both show a clear bias toward non-Cartesian gratings.

Fig. 11 shows that there is a general trend away from the Cartesian sector, confirming the bias toward non-Cartesian stimuli. A small population of the S2 units responds significantly more to one class of stimuli than to another, as illustrated by Table 4 and by the data points

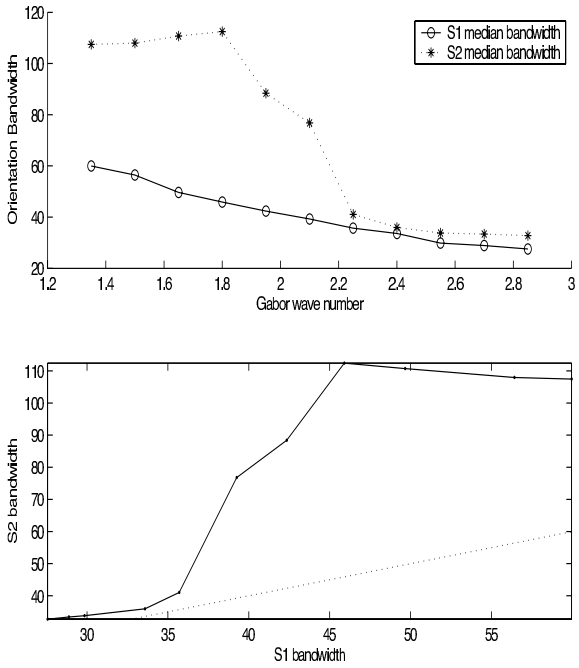


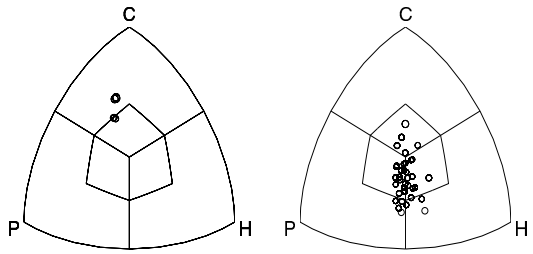
Figure 10: Top: Median orientation bandwidths of S1 and S2 units *vs.* the Gabor wave number k, plotted in units of 2π. Bottom: Same data, plotted as S1 bandwidth *vs.* S2 bandwidth. The dashed line represents the condition where S1 bandwidth = S2 bandwidth.

	Gallant *et al.*	HMAX
Polar	10%	10%
Hyperbolic	8%	5%
Cartesian	2%	0%

Table 4: Percentage of cells that gave more than twice the peak responses to one stimulus class than to another.

lying outside of the inner region in Fig. 11. Note that the proportions of the cells preferring non-Cartesian gratings in model and in experiment agree surprisingly well. While there is no S2 cell preferring Cartesian gratings in the standard version of HMAX, this is not a fundamental shortcoming of the model — S2 units that receive input from a single C1 unit would show the required preference for Cartesian gratings.

The ratio of the maximum and the minimum responses to three grating classes shows that most S2 units (82%, very close to the estimate of 80% in [5]) respond to all three types of gratings comparably (within a factor of two) as seen in Fig. 12. However, for a small fraction of cells, this maximum-over-minimum ratio exceeds 2, indicating an enhanced selectivity toward one class of stimuli. In particular, the S2 units in group 1 (Table 2) stand out in the distribution, since they respond weakly to Cartesian stimuli, but strongly enough



(a) S1 Responses (b) S2 Responses

Figure 11: Responses to the three grating classes (polar, hyperbolic, and Cartesian gratings), drawn in the same convention as in Fig. 4 of [5]. For each model unit, the maximum responses within each grating class are treated as a 3-dimensional vector, normalized and plotted in the positive orthant. This 3-dimensional plot is viewed from the $(1,1,1)$-direction, so that the origin will correspond to a neuron whose maximum responses to three grating classes are identical. Cartesian-preferring units will lie in the upper sector, polar in the lower left, and hyperbolic in the lower right sector. The symbols outside of the inner region correspond to the model units that gave significantly greater (by a factor of 2) responses to one stimulus class than to another. The size of each symbol reflects the maximum response obtained across the entire stimuli. Note that all S1 units prefer Cartesian over polar and hyperbolic gratings, whereas most S2 data points lie in the lower part of the plot, indicating a general bias toward non-Cartesian gratings.

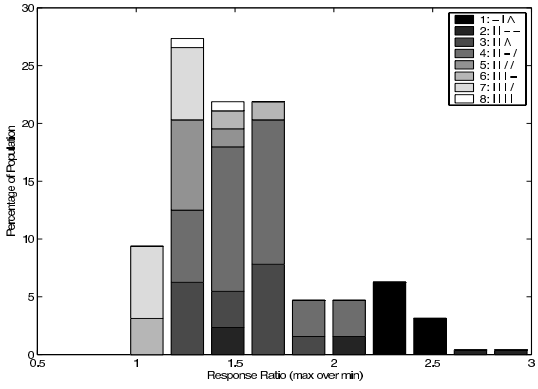


Figure 12: Distribution of the response ratio (maximum over minimum) to three grating classes. The ratio of 1 indicates that the cell gave the same maximum responses to all three grating classes.

to non-Cartesian stimuli.

Fig. 13 and 14 show the distribution of the S2 unit responses, along with the 8-class classification scheme (Table 2). They illustrate that the S2 units in group

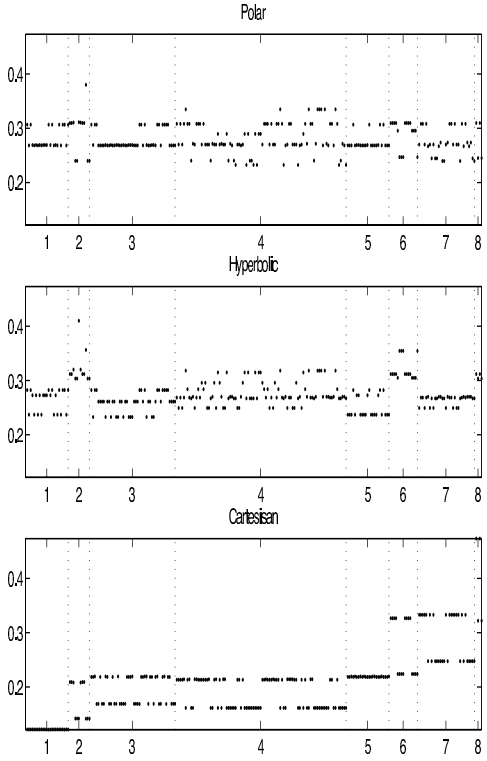


Figure 13: Distribution of the maximum responses to three grating classes: Each dot corresponds to one of the 256 S2 units, as categorized according to the 8-class classification scheme along the x-axis. Note that the distribution for Cartesian grating is significantly different from the other two distributions. Some S2 units that do not respond much to Cartesian gratings respond well to non-Cartesian gratings (group 1 and 2), and vice versa (group 8). Thus, in visual cortex, the Cartesian-selective cells may receive afferent inputs from the cells with similar orientation selectivities, while non-Cartesian cell's afferents would be composed of cells with different orientation selectivities.

8, whose afferents are pointing in parallel orientations, produce large responses to Cartesian gratings, as expected. On the other end of the spectrum, the S2 units in group 1, whose pooled afferents are selective to different orientations, show higher responses to non-Cartesian gratings.

The average response of the population to each grating is plotted in Fig. 15(a), where the bias in favor of non-Cartesian stimuli is again apparent. In a good qualitative agreement with Figure 3-D of [5], the average population responses are high for polar and hyperbolic gratings of low/intermediate frequencies. Within the Cartesian stimulus space, the average response is also peaked around the low/intermediate frequency region. The concentric grating of low frequency (marked with *) shows the maximum average response. For reference, Fig. 15(b,c,d) show the tuning curves of three individual S2 units that are most selective to each grating type.

One of the major differences between the physiological data in [5] and the aforementioned simulation results is the lack of the S2 units highly selective to one stimulus class only. (In the scatter plot, those units would lie along the direction of $(1,0,0)$, $(0,1,0)$, or $(0,0,1)$.) In fact, as seen in Fig. 11(b), most of the S2 units lie near the boundary between the polar and the hyperbolic sectors, meaning they respond quite similarly to these gratings, but differently to Cartesian gratings. Fig. 13 indeed shows that the response distributions for the polar and the hyperbolic gratings are quite similar.

The above result therefore suggests that the 2x2 arrangement of the Gabor-like features may be too simplistic, possibly because the sampling of the adjacent afferents is too correlated to contain enough distinguishing features across the polar-hyperbolic dimension. The HMAX model can be extended to investigate these issues. The feature complexity can be increased by using different combination schemes (*e.g.*, 3x3) or by sampling the afferents from non-adjacent regions. Fig. 16 illustrates that by introducing such modifications, it is possible to obtain more uniformly distributed responses in the polar-hyperbolic-Cartesian space, while maintaining a general bias toward non-Cartesian stimuli. This result suggests that combining non-local, less-correlated features would be important in building features that can distinguish object classes better (in this case, polar *vs.* hyperbolic gratings). Interestingly, preliminary data [2], indicating that some V2 receptive fields appear to show separate directional subfields, are compatible with this hypothesis of separated C1 afferents to an S2 receptive field.

Using more C1 afferents, it is also possible to introduces other variants of S2 units with different grating selectivities. Using 3x3 feature combination with 4 different orientations yields $4^9 = 262144$ possibilities. However, by increasing the number of the afferents, the bias toward non-Cartesian grating is also increased, since it is less likely to have most of the afferents with the same orientation selectivities.

3.2.3 (V1, S1) $\rightarrow$ (S2, V4)

Physiological data show that along the ventral pathway, the selectivity for non-Cartesian stimuli increases. Mahon and De Valois [10] reported that there were more neurons responsive to non-Cartesian gratings in V2 than in V1. Gallant *et al.* [5] reported that the selectivity for non-Cartesian gratings was quite enhanced in the visual area V4 and that there were very few neurons highly responsive to Cartesian gratings only.

A similar trend is apparent in our model, or rather it has been implicitly built into it, by combining oriented filters (naturally responsive to Cartesian gratings)

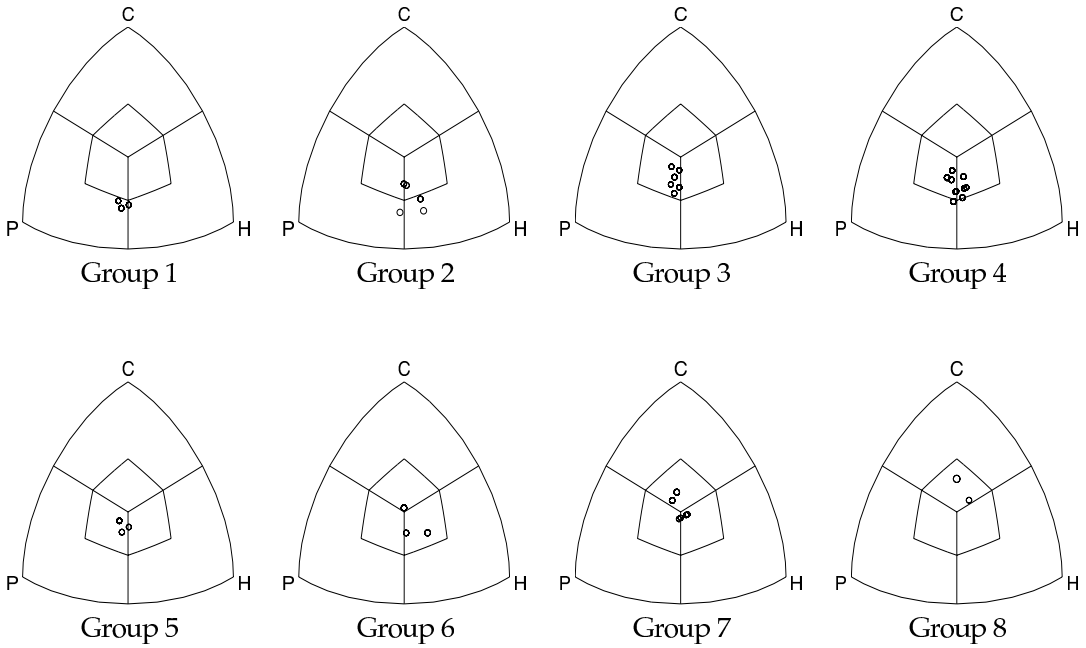


Figure 14: When all 256 S2 units are plotted in the same format as Fig. 11, it is apparent that group 1 and 2 are composed of highly non-Cartesian units, while the preference for Cartesian stimuli slowly increases toward group 8.

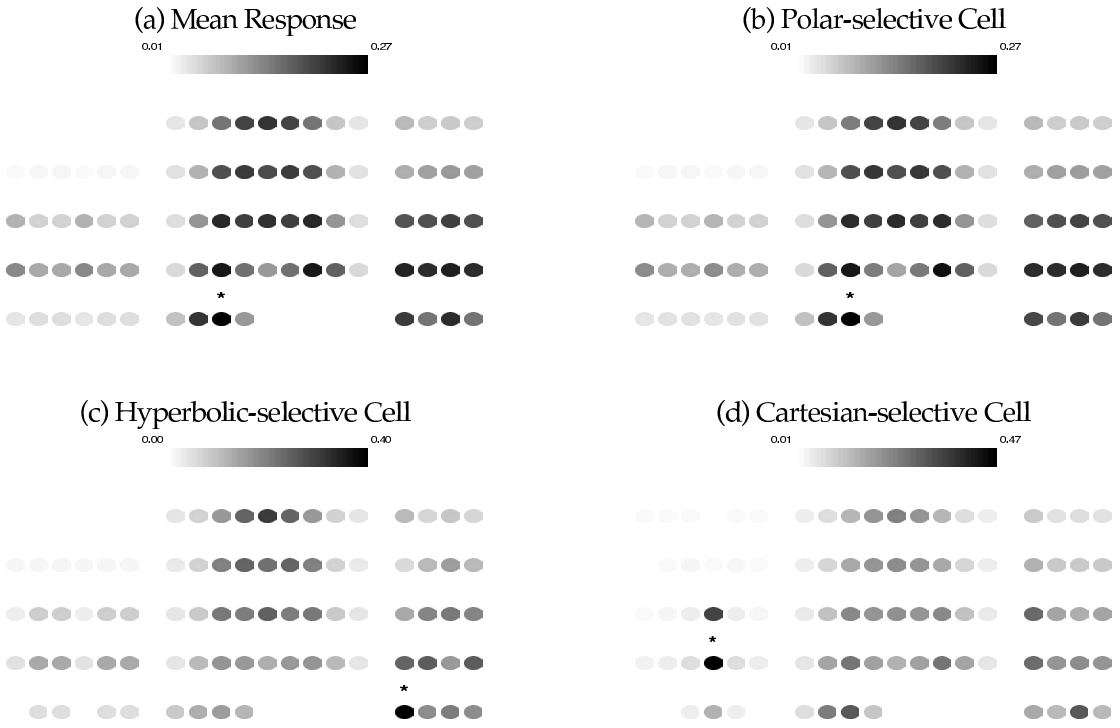


Figure 15: (a) Average population responses and (b,c,d) three sample tuning curves most selective to each of the three grating classes. The responses are arranged in the same layout as in Fig. 4. The most effective stimulus is marked with an asterisk (*) on top.

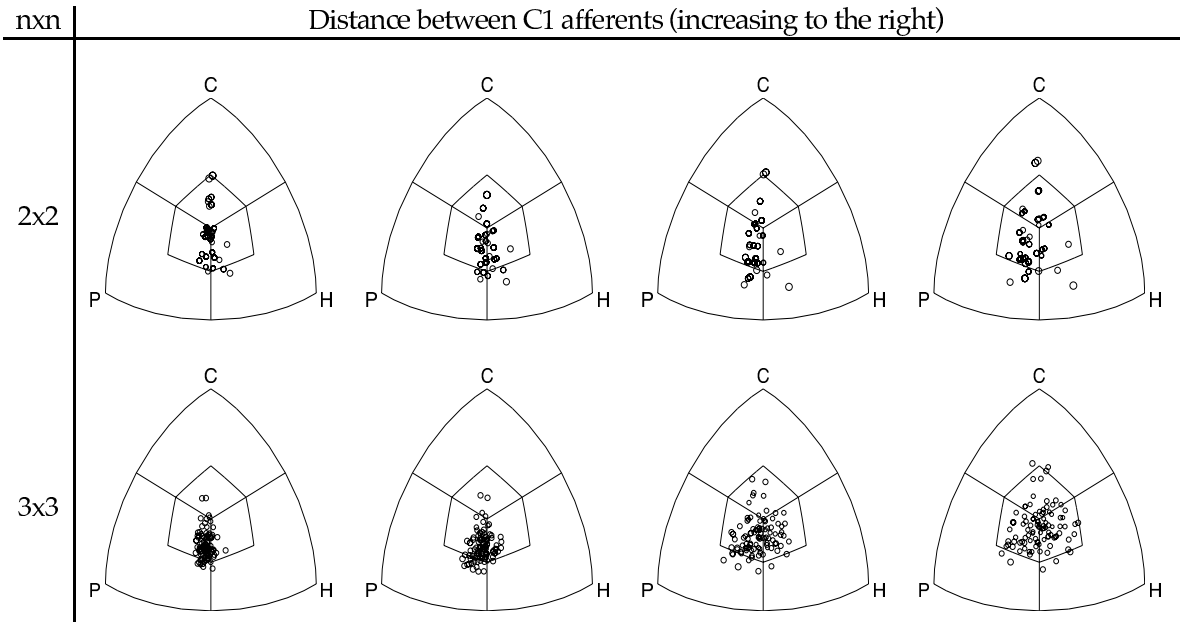


Figure 16: Population responses to three grating classes, from 100 S2 units that are chosen randomly from all possible feature combinations. Top row shows the results with the 2x2 feature combination scheme, and the bottom row shows the 3x3 scheme. Going from left to right, the distance between the C1 afferents is increased. In the first column, the C1 afferents are partially (1/2) overlapping. In the second column, the C1 afferents are adjacent. (Thus, the plot in the top row of the second column represents the result using the standard HMAX parameters.) In the third and the fourth columns, the C1 afferents are even farther apart (1 or 2 times the C1 pooling range). As the distance between the C1 afferents increases, the features combined in one S2 receptive field are sampled from farther regions of the stimulus image.

into non-parallel, non-Cartesian features. In the S1 layer, there is no model unit more responsive to non-Cartesian gratings, whereas in the S2 layer the majority prefers non-Cartesian gratings.

4 Discussion

In this paper, a computational model of the ventral visual stream was used to provide hypotheses regarding possible mechanisms underlying the observed change in neuronal feature tuning from V1 to V4. The model posits that the increase in complexity results from a simple combination of complex cell afferents. Despite its simplicity, the model turned out to approximate several physiological data along the ventral pathway of the primate visual cortex. In particular, the model exhibited the broadening of the orientation bandwidth and the bias toward non-Cartesian stimuli, while successfully reproducing some of the population statistics. Interestingly, even a simple 2x2 combination of the afferents could yield a fairly complex behavior from the population. Furthermore, it was noted that the model units whose afferents were non-parallel and orthogonal served to yield a wide orientation bandwidth and a high selectivity for non-Cartesian stimuli.

The gratings provided a richer set of stimuli and showed some discrepancies between the standard HMAX model and the physiological data, in particular the lack of model units strongly selective for either polar or hyperbolic gratings. Such model units could only be obtained by increasing the spatial separation of the C1 afferents. This provides an interesting prediction for experiments regarding the receptive field substructure of neurons in higher visual areas, for which there are some preliminary experimental evidences in V2 [2]. Interestingly, features based on spatially-separated, complex cell-like afferents have been previously postulated based on computational grounds [1]. An alternative, more trivial way to obtain cells strongly selective for non-Cartesian gratings, even though not explored here, would be to assume more complex, non-Cartesian S1 features that are more selective toward the features found in the stimuli set. Physiological data indicate that V1 does contain neurons responsive to radial, concentric, or hyperbolic gratings [10].

Finally, it appears that the bar and grating stimuli are too limited as a stimulus set to provide strong constraints for the model, as the standard HMAX model seemed to have enough degrees of freedom to cover

various bandwidth distributions and grating selectivities. For example, the present data did not require any significant modification of feature combination schemes or the inclusion of more complex features in the lower layer of the hierarchical architecture. It will be interesting to test model unit responses to more complex stimuli, such as the contour features used by Pasupathy and Connor [11, 12]: While the tuning of model unit is currently based on shape only, Pasupathy and Connor have postulated that V4 neurons show evidence for an object-centered reference frame.

A Further Orientation Selectivity Studies

Most of the main results in this paper were obtained with the following standard parameters (adopted from [16]).

Parameter	Value
Stimulus Contrast	90%
σ_{S2}	1.25
Gabor wave number	$2\pi \cdot 2.1$
S1 receptive field size	17, 19, 21 pixels
S2 receptive field size	38 pixels

In this appendix, we study the effects of these parameters in more detail.

A.1 Stimulus Contrast

The luminance of a stimulus can be varied in several different ways. The total luminance, the sum or squared sum of all pixel values, can be set to a constant. Alternatively, the background luminance or the minimum luminance can be set to a constant, and the maximum luminance can be adjusted according to the definition of contrast (Eqn. 9). In our study of orientation selectivity, the background of the stimulus image was set to a constant value of 1, and the luminance of the bar was adjusted.

As shown in the following table, the mean orientation bandwidths were invariant over a wide range of stimulus contrasts. However, the high contrast stimuli produced higher responses from the S1 and, thus, the S2 units.

	Bandwidths		Responses	
Contrast	S1	S2	S1	S2
10%	39.8°	78.0°	0.032	0.013
30%	39.1°	77.5°	0.090	0.037
50%	39.1°	77.2°	0.140	0.059
70%	39.1°	77.2°	0.184	0.078
90%	39.3°	76.8°	0.223	0.095

A.2 Orientation Bandwidth

The distribution of the orientation bandwidth is in general lower bounded by group 8 and upper bounded by group 1. Between these bounds, the S2 units in group 3 and 4 have the most variable range of orientation bandwidths, because of their secondary response peaks in the orientation tuning curve. Therefore, by manipulating the model parameters that affect the tuning profiles, it is possible to obtain different bandwidth distributions. Some of such parameters are the feature sensitivity (σ_{S2}), receptive field geometry, and the scale. The sharpness of tuning at the S1 level (Gabor wave number k) was treated in section 3.1.3.

A.2.1 σ_{S2}

The response of an S2 unit is determined by the afferent C1 units, which are combined as a product of Gaussians.

$$S2 = e^{-\left(\sum_i (C1_i - 1)^2\right)/2\cdot(\sigma_{S2})^2}. \quad (11)$$

Each Gaussian is centered at 1, since the C1 responses lie between 0 and 1.

The response of an S2 unit, or the sensitivity to a feature, is affected by σ_{S2}. For example, for a large value of σ_{S2}, the S2 unit will produce a fairly large response (close to 1) regardless of the stimulus. If σ_{S2} is small, the S2 unit will only respond to a very specific feature, determined by the afferent C1 units. Then, as σ_{S2} is varied from a small, to a medium, and to a large value, the baseline-subtracted response of an S2 unit will go from 0, to an intermediate value, and to 0 again. (In the limit of $\sigma_{S2} << 1$, the S2 response will be 0, unless the stimulus is the optimal feature. If $\sigma_{S2} >> 1$, the S2 unit will yield the maximum response 1, for any stimulus. It will also respond well to a blank stimulus, and therefore, the baseline-subtracted response will be 0 again.)

The orientation selectivity at various σ_{S2} can be understood similarly. For a large σ_{S2}, all orientations of the bar stimulus will produce similar responses, yielding a flat tuning curve. This will be especially true for the S2 units in group 3 and 4, whose primary and secondary peaks can then easily merge into one wide peak. Therefore, increasing σ_{S2} will have a broadening effect on the orientation bandwidths, as seen in the following Table and Fig. 17.

σ_{S2}	Median Bandwidth
0.5	34.7°
1.0	52.2°
1.5	79.3°
4.0	82.8°

A.2.2 Receptive Field Geometry

For all the simulations described in this paper, the S1 units were given a circular receptive field, in order to reduce the numerical differences between the principal (0°, 90°) and the oblique (45°, 135°) orientations and thus to emphasize only the effects coming from the inherent architecture (*e.g.*, 2x2 feature combination) of the model.

For a comparison, the circular mask was lifted from the Gabor filters, thereby giving a square receptive field

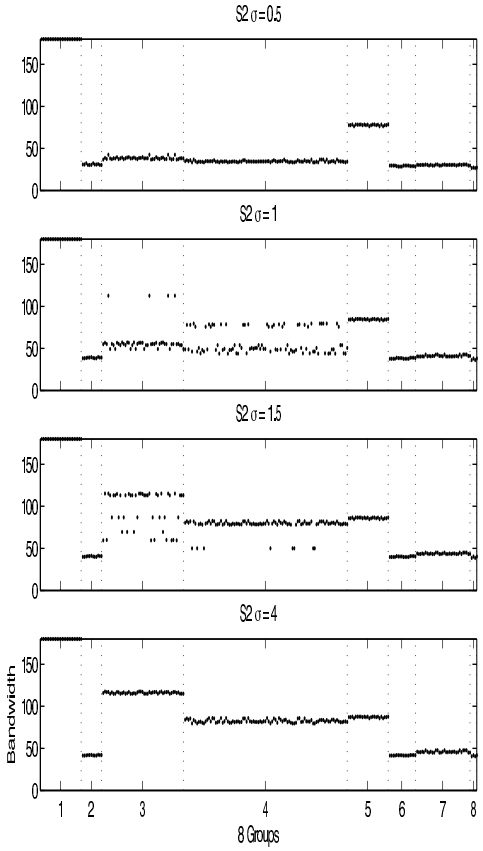


Figure 17: Distribution of the orientation bandwidth at various σ_{S2}. Note that the bandwidths are most variable for group 3 and 4. Bandwidths $> 180°$ are shown as 180°.

to the S1 units. Then, the bandwidth distribution becomes smoother with less sharp changes from one histogram bin to another, resembling the physiological data more. This effect seems to arise from the asymmetry of the S1 filters, and, therefore, differences between individual V1 neurons may play a role in producing a broad, smooth distribution of the orientation bandwidths in V4.

The Gabor filters with a square receptive field have a slightly sharper orientation tuning, since they have more elongated excitatory and inhibitory regions. However, regardless of the circularity of the receptive field, the overall shapes of the tuning profiles are equal to what is shown in Fig. 7, resulting in the similar broadening trend from the S1 to the S2 layer.

A.2.3 Scaling

The receptive field sizes of V1 and V4 neurons are widely distributed. In general, they are positively correlated with eccentricity. At the fovea, the receptive fields of V4 neurons are on average twice as large as those of V1 neurons [4]. When the orientation selectivity experiments are performed at different scales, the orientation bandwidth again increases from the S1 to the S2 layer. The following table summarizes the simulation results, where the receptive field sizes are given in pixels.

Scale	Receptive Field		Bandwidth	
	S1	S2	S1	S2
1	7, 9	16	36.0°	127.7°
2	11, 13, 15	26	39.3°	81.4°
3	17, 19, 21	38	39.3°	76.8°
4	23, 25, 27, 29	52	38.6°	50.0°

Fig. 18 explains this broadening of the bandwidths at higher scales (= resolutions) as a discretization effect: As the resolution becomes finer (going from top to bottom), the primary and the secondary peaks in the orientation tuning profiles of group 3 and 4 are better distinguished, resulting in lower median bandwidths.

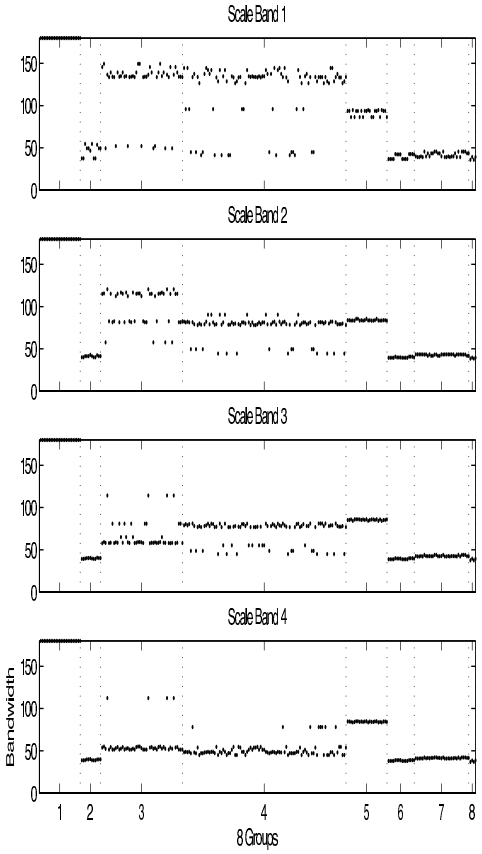


Figure 18: Distribution of the orientation bandwidth for the S2 units. Going from top to bottom, the receptive field size (and correspondingly the receptive field's resolution) increases.

B Further Grating Selectivity Studies

B.1 Scaling

Here, the behavior of the model at four different scales was studied as in Section A.2.3, using the same set

of polar, hyperbolic, and Cartesian gratings. The result shows that at all scales, the overall response distributions to each class of gratings are quite similar, with an apparent bias toward non-Cartesian stimuli. The following table summarizes the average baseline-subtracted responses and the standard deviations, which are almost identical across all four scales.

	Responses		
Scale	Polar	Hyperbolic	Cartesian
1	0.15 ± 0.05	0.18 ± 0.04	0.06 ± 0.03
2	0.15 ± 0.06	0.16 ± 0.05	0.05 ± 0.04
3	0.14 ± 0.07	0.15 ± 0.06	0.05 ± 0.04
4	0.15 ± 0.07	0.16 ± 0.06	0.05 ± 0.03

Finally, the following table shows the breakdown of the population within three grating sectors. The values inside the parentheses represent the percentage of S2 units whose peak response to one grating class was twice the response to another. Even though the breakdowns for the polar and the hyperbolic gratings are quite variable since most S2 units lie near the boundary, an overall preference for non-Cartesian stimuli is again apparent.

	Population Statistics		
Scale	Polar	Hyperbolic	Cartesian
1	36% (7%)	53% (8%)	11% (0%)
2	50% (7%)	39% (8%)	11% (0%)
3	57% (10%)	32% (5%)	11% (0%)
4	52% (10%)	37% (5%)	11% (0%)

References

[1] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[2] A. Anzai et al. Receptive field structure of monkey V2 neurons for encoding orientation contrast [abstract]. *Journal of Vision*, 2(7), 2002.

[3] P. Dayan and L. Abbott. *Theoretical Neuroscience: Computational and mathematical modeling of neural systems*. MIT Press, 2001.

[4] R. Desimone and S. Schein. Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *Journal of Neurophysiology*, 57:835–868, 1987.

[5] J. Gallant et al. 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] 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.

[8] I. Lampl et al. The MAX operation in cells in the cat visual cortex [abstract]. *Society for Neuroscience Abstracts*, 2001.

[9] N. Logothetis et al. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5:552–563, 1995.

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

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

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

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

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

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

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

[17] R. De Valois et al. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22:531–544, 1982.