CSAIL

massachusetts institute of technology — computer science and artificial intelligence laboratory

Realistic Modeling of Simple and Complex Cell Tuning in the HMAX Model, and Implications for Invariant Object Recognition in Cortex

Thomas Serre and Maximilian Riesenhuber

AI Memo 2004-017 July 2004
CBCL Memo 239

Abstract

Riesenhuber & Poggio recently proposed a model of object recognition in cortex which, beyond integrating general beliefs about the visual system in a quantitative framework, made testable predictions about visual processing. In particular, they showed that invariant object representation could be obtained with a selective pooling mechanism over properly chosen afferents through a MAX operation: For instance, at the complex cells level, pooling over a group of simple cells at the same preferred orientation and position in space but at slightly different spatial frequency would provide scale tolerance, while pooling over a group of simple cells at the same preferred orientation and spatial frequency but at slightly different position in space would provide position tolerance. Indirect support for such mechanisms in the visual system comes from the ability of the architecture at the top level to replicate shape tuning as well as shift and size invariance properties of "view-tuned cells" (VTUs) found in inferotemporal cortex (IT), the highest area in the ventral visual stream, thought to be crucial in mediating object recognition in cortex. There is also now good physiological evidence that a MAX operation is performed at various levels along the ventral stream. However, in the original paper by Riesenhuber & Poggio, tuning and pooling parameters of model units in early and intermediate areas were only qualitatively inspired by physiological data. Many studies have investigated the tuning properties of simple and complex cells in primary visual cortex, V1. We show that units in the early levels of HMAX can be tuned to produce realistic simple and complex cell-like tuning, and that the earlier findings on the invariance properties of model VTUs still hold in this more realistic version of the model.

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. MDA972-04-1-0037, 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), Daimler-Chrysler 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

Extending previous models of object recognition in cortex [1, 2], Riesenhuber & Poggio have shown that invariant object representations (similar to the ones found in inferotemporal (IT) cortex) could be explained by the combined action of two operations:

A weighted linear summation *i.e.,* units performing a weighted linear summation (followed by a Gaussian nonlinearity) over afferents tuned to different features (equivalent to template matching) would be well suited to explain the increase in complexity of the optimal stimulus driving cells en route to object recognition.

A MAX operation *i.e.,* units performing a non-linear MAX operation over afferents tuned to slightly distorted versions of the same feature (shifted and rescaled) should provide the substrate for building increasingly invariant representations.

In a benchmark simulation [3], Riesenhuber & Poggio "recorded" from the HMAX model (see Fig. 1) and showed that the range of invariances exhibited by the model VTUs (named after the view-tuned units in IT) was compatible with shift, size and depth rotation tuning properties of view-tuned cells [3, 4].

Additionally, biophysically plausible implementations of the MAX operation have been proposed [5] and neurons performing a MAX operation have been found in area V4 in the primate [6], and very recently also in complex cells in cat visual cortex [7]. The latter study showed that, consistent with the model prediction, the response of complex cells elicited by the simultaneous presentation of two bars (one optimal and one non-optimal), closely matches the response of the cells when presented with the optimal stimulus alone.

In the original paper by Riesenhuber & Poggio, tuning and pooling parameters of model units in early and intermediate areas were only qualitatively inspired by physiological data. In particular, many studies have investigated the tuning properties of simple and complex cells in primary visual cortex V1. We now take a detailed look at the compatibility of the model with population tuning at the simple and complex cells level.

We start by improving the fit between model simple (S1) units (whose tuning properties in the original model were chosen to just qualitatively resemble V1 simple cell shape) and the experimental data. In particular, we show that a better account of the simple cells population spread of tuning can be obtained with properly parameterized Gabor functions.

We further show that starting with a representative distribution of simple cell tuning properties, it is possible to adjust two of the main model parameters (spatial and frequency extent of the afferent simple cells, see Fig.1) such that the corresponding set of complex (C1) units tuning properties is compatible with the V1 complex cells. In particular, we find that the increase in receptive field size [8] and spatial frequency bandwidth [9, 10] could be well accounted by the pooling mechanisms proposed in HMAX in order to gain size and shift tolerance at the C1 level.

As a benchmark for our model units, we consider tuning properties of parafoveal cells in monkey as reported by two groups: De Valois *et al.* [9, 11] and Schiller *et al.* [10, 12, 13].[*] Focusing on this new set of S1 and C1 cells, we use a benchmark paperclip recognition task as in [3, 4] and show that the model is still able to replicate tuning properties of view-tuned cells in IT, suggesting that the model is robust to changes in the low levels.

2 Methods

2.1 Original HMAX

The precise architecture of HMAX has been described in details elsewhere [3, 14–16] and we here only highlights important features of the model (see Fig. 1). We first briefly describe the two first layers of HMAX under study, that is, simple (S1) cells and complex (C1) cells. We then highlight the other two layers of the model (S2 and C2) for further understanding on training the VTUs in the benchmark recognition task (sections 2.3.5 and 3.3).

Simple (S1) cells. Input images (160×160 gray images corresponding to 4.4° of visual angle, see [14]) are densely sampled by arrays of two-dimensional filters $G_{x,y}$ (second derivative of Gaussians) that can be expressed as:

$$G_{x,y} = \frac{(-x\cos\theta + y\sin\theta)^2}{\sigma^2(\sigma^2-1)}$$
$$\exp\left(-\frac{(x\cos\theta + y\sin\theta)^2 + (-x\cos\theta + y\sin\theta)^2}{2\sigma^2}\right).$$

Table 1 details the values of the two filters parameters: orientation θ and width σ. The response of the so-called S1 units, sensitive to bars of different orientations, thus roughly resembling properties of simple cells in striate cortex, is given by centering filters of each size and orientation at each pixel of the input image. The filters are sum-normalized to zero and square-normalized to 1 so that S1 cells activity is between -1 and 1, modeling simple cells of phase 0 and π.

While non-biological (both in its implementation and because it neglects the response saturation of V1 cells

[*] We considered parafoveal cells, as further studies of higher brain areas (V4, for instance) mostly focused on parafoveal cells population, although differences between the two groups are not always significant: Parafoveal cells tend to have slightly larger receptive fields, are slightly more broadly tuned to spatial frequency and tend to be tuned to lower-spatial frequencies.

observed at high contrast [17–19]), this simplification is convenient and does not interfere in our experiments as we work with fixed contrast. Fig. 2 shows all simple (S1) cells receptive field types used in standard HMAX.

Complex (C1) cells. One prediction made by the model is that complex cells are phase invariant as well as size and position tolerant. Fig. 1 describe how size and position invariance are increased in the model. The mechanisms rely on a non-linear MAX operation (or its soft-MAX approximation, [14]) over properly chosen afferents, *i.e.,* a C1 unit's activity is determined by the strongest input it receives.

For instance, pooling over simple (S1) cells at the same preferred orientation but responding to bars of different lengths, provide invariance with respect to changes in size (see Fig. 1 B.). The amount of invariance gained is determined by the range of sizes (or equivalently spatial frequency selectivities) over which the MAX is performed. We call this *filter bands, i.e.,* groups of S1 filters of a certain size range. In standard HMAX , four filter bands are used in which filter sizes are within the range:

$$\text{ScaleRange} = \{7-9; 11-15; 17-21; 23-29\} \qquad (1)$$

Similarly, position invariance is increased by pooling over S1 cells at the same preferred orientation but whose receptive fields are centered on neighboring locations, *i.e.,* within each filter band, a pooling range is defined which determines the size of the array of neighboring S1 units of all sizes in that filter band which feed into a C1 unit (see Fig. 1 A.). It is important to mention that only S1 filters with the same preferred orientation feed into a given C1 unit to preserve feature specificity. In standard HMAX , the pooling ranges for each of the four filter bands are such that:

$$\text{PoolRange} = \{4; 6; 9; 12\} \qquad (2)$$

As a result, a C1 unit responds best to a bar of the same orientation as the S1 units that feed into it, but already with an amount of spatial and size invariance that corresponds to the spatial and filter size pooling ranges used for a C1 unit in the respective filter band. Additionally, C1 units are invariant to contrast reversal, much as complex cells in striate cortex, by pooling over on and off simple cells (before performing the MAX operation). Possible firing rates of a C1 unit thus range from 0 to 1.

S2 cells. A square of four adjacent, non-overlapping C1 units belonging to the same filter band, in a 2×2 arrangement, is grouped to provide input to each S2 unit. There are 256 different types of S2 units in each filter band, corresponding to the 4^4 possible arrangements of four C1 units of each of four types (*i.e.,* preferred bar orientation). The S2 unit response function is a Gaussian with mean 1 (*i.e.,* $\{1, 1, 1, 1\}$) and standard deviation 1, *i.e.,* an S2 unit has a maximal firing rate of 1 which is attained if each of its four afferents fires at a rate of 1 as well. S2 units provide the feature dictionary of HMAX , in this case all combinations of 2×2 arrangements of "bars" (more precisely, C1 cells) at four possible orientations.

It is worth noting that those choices of S2 units' parameters remain somewhat arbitrary. This reflects the lack of a precise characterization of the response properties of cells in intermediate layers of visual cortex. Indeed, current work is trying to improve the fit between S2 units in HMAX and biological neurons in V4 [20, 21]. We also showed in [22] that S2 units centers could be learned in order to perform robust real-world object recognition.

C2 cells. To finally achieve size invariance over all filter sizes in the four filter bands and position invariance over the whole input image, the S2 units are again pooled by a MAX operation to yield C2 units, the output units of the HMAX core system, designed to correspond to neurons in extrastriate visual area V4 or posterior IT (PIT). There are 256 C2 units, each of which pools over all S2 units of one type at all positions and scales. Consequently, a C2 unit will fire at the same rate as the most active S2 unit that is selective for the same combination of four bars, but regardless of its scale or position.

View-tuned units. C2 units in turn provide input to the view-tuned units (VTUs), named after their property of responding well to a specific two-dimensional view of a three-dimensional object, thereby closely resembling the view-tuned cells found in monkey inferotemporal cortex by Logothetis *et al.* [4]. The C2 → VTU connections are so far the only stage of the HMAX model where learning occurs (but see [22] for a method to learn S2 features with HMAX in the context of an object detection task).

A VTU is tuned to a stimulus by selecting the activities of the N C2 units (all 256 or a subset) in response to that stimulus as the center of an N-dimensional Gaussian response function, yielding a maximal response of 1 for a VTU in case the C2 activation pattern exactly matches the C2 activation pattern evoked by the training stimulus [†].

2.1.1 New HMAX

S1 cells. We here motivate the use of Gabor functions to model simple cells receptive field instead of the Gaussian derivatives as in standard HMAX. For the past decade, Gabor filters have been extensively used to

[†] We here consider the simplest way to train a set of VTUs from data as in [3]. The method is closely related to RBF networks for which a function is approximated by a weighted sum of basis functions centered on each data points (or a subset of the training data). More complex schemes include a search of the VTU centers as in generalized RBF network for instance.

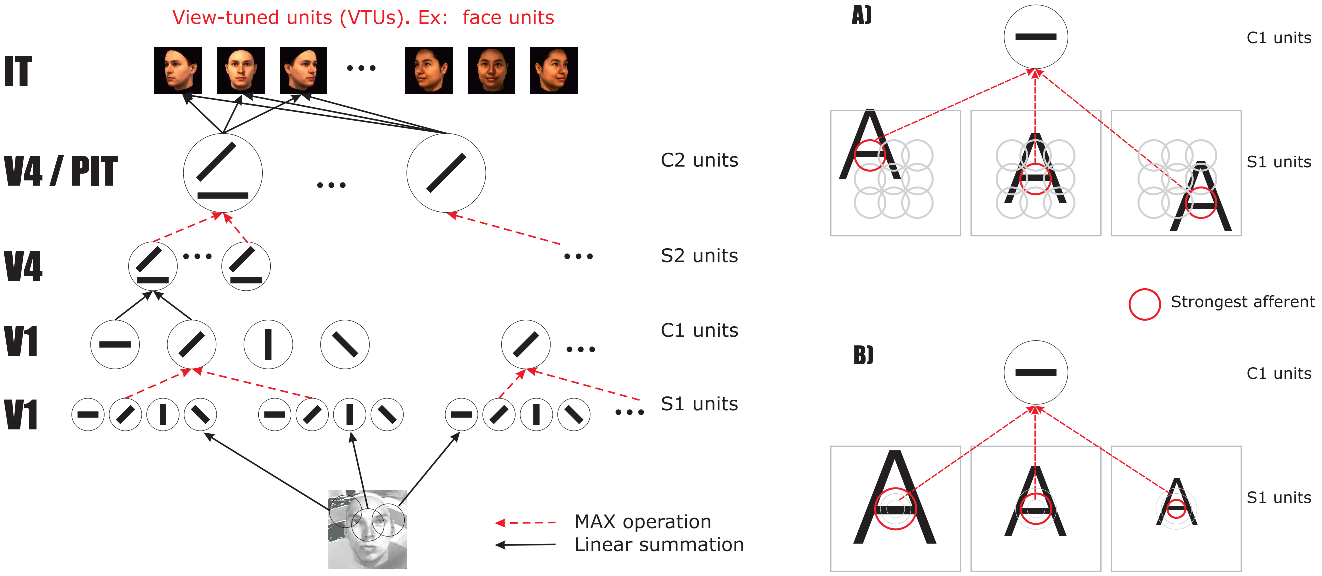

Figure 1: **Left**: Schematic of the model. Two types of computations *i.e.,* linear summation and non-linear MAX operation alternate between layers. Input images are first densely sampled by arrays of two-dimensional filters at four different orientations, the simple (S1) units. Within a pooling band, S1 cells (*i.e.,* a group of cells at the same preferred orientation but at slightly different scales and positions, see text) feed into complex (C1) cells through a MAX operation (see right figure for illustration). In the next (S2) level, and within each filter band, a square of four adjacent, non overlapping C1 units in a 2×2 arrangement is grouped to provide input to an S2 unit. To finally achieve size invariance over all filter sizes in the four filter bands and position invariance over the whole input image, the S2 units are again pooled by a MAX operation to yield C2 units that again provide input to the view-tuned units (VTUs). **Right**: Schematic of how size and shift tolerances are increased at the (C1) level: A complex (C1) cell pools over S1 cells (within a pooling band, see text) at the same orientation but **A)** centered at different location thus providing some translation invariance and **B)** at different scales providing some scale invariance to the complex cell.

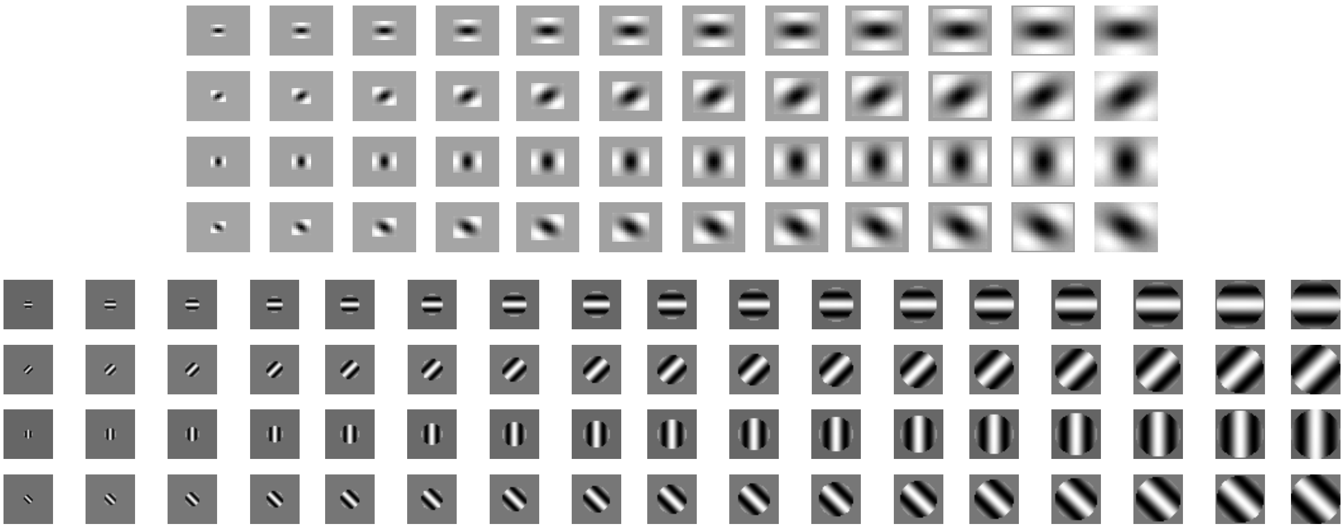

Figure 2: Top: Model simple cells receptive field used in standard HMAX [14]. Receptive field sizes range from 0.19^o to 0.8^o at four different orientations. Bottom: Modeling simple cells receptive field with Gabor functions. Receptive field sizes range from 0.19^o to 1.07^o at four different orientations. In order to obtain receptive field sizes within the bulk of the simple cell receptive fields (0.1° -1°) reported in [8, 12], we cropped the Gabor receptive fields and applied a circular mask so that, for a given parameter set (λ, σ), cell tuning properties are independent of their orientations. Note that receptive fields were set on a gray background for display only and so that relative sizes were preserved.

model the receptive fields of simple cells. Gabor functions have been shown to be solutions of an optimization problem that is minimizing simultaneously uncertainty in both position and spatial frequency [23] and to fit well with physiological data recorded from cat striate cortex [24]. We here motivate the use of Gabor functions to model cortical simple cell receptive fields because they have more free parameters and allow more accurate tuning than their homologue (Gaussian derivatives) used in standard HMAX (see section 3 for a comparison between the two).

Placing the origin of the x and y axis coordinates at the center of the receptive field, the filter response is given by:

$$G_{x,y} = \exp\left(-\frac{(x\cos\theta + y\sin\theta)^2 + \gamma^2(-x\sin\theta + y\cos\theta)^2}{2\sigma^2}\right) \times \cos\left(2\pi\frac{1}{\lambda}(x\cos\theta + y\sin\theta) + \phi\right).$$

The five parameters, *i.e.*, orientation θ, aspect ratio γ, effective width σ, phase ϕ and wavelength λ determine the properties of the cells spatial receptive fields. The tuning of simple cells in cortex along these dimensions varies substantially. Rather than attempting to replicate the precise distribution (which differs between the different studies), our aim is to show that model S1 unit tuning can capture more robust statistics (such as sample mean or median) and the range of experimental neurons.

As in standard HMAX, we considered four orientations only ($\theta = 0^\circ$, 45° , 90° , and 135°). This is an over-simplification but this has been shown to be sufficient to provide rotation and size invariance at the VTU level in good agreement with recordings in IT [3]. ϕ was set to 0° while different phases are crudely approximated by centering receptive fields at all locations.

In order to obtain receptive field sizes consistent with values reported for parafoveal simple cells [12], we increased the number of filter sizes covered with standard HMAX leading to 17 filters sizes from 7×7 (0.19° visual angle) to 39×39 (1.07° visual angle) obtained by steps of two pixels instead of the 12 filters sizes ranging between 7×7 (0.19° visual angle) and 29×29 (0.80° visual angle) as in standard HMAX .

When fixing the values of the remaining 3 parameters (γ, λ and σ), we tried to account for general cortical cell properties, that is: (i) Cortical cells' peak frequency selectivities are negatively correlated with their receptive field sizes [10]. (ii) Cortical cells' spatial frequency selectivity bandwidths are positively correlated with their receptive field sizes [10]. (iii) Cortical cells orientation bandwidths are positively correlated with their receptive field sizes [13].

We empirically found that one way to account for all three properties was to include fewer cycles in the units' receptive fields as their sizes (*R*F size) increase. We found that the two following (ad hoc) formulas gave good agreement with the tuning properties of cortical cells:

$$\sigma = 0.0036 * \mathit{RF\ size}^2 + 0.35 * \mathit{RF\ size} + 0.18 \quad (3)$$

$$\lambda = \frac{\sigma}{0.8} \quad (4)$$

Table 1 gives the values of parameters that determine Gabor filter tuning properties and how they differ from those in standard HMAX (Gaussian derivatives).

For all cells with a given set of parameters (λ_0, σ_0) to share similar tuning properties at all orientations, we applied a circular mask to the Gabor filters (see Fig. 2 bottom) which was not done in standard HMAX . Cropping Gabor filters to a smaller size than their effective length and width, we found that the aspect ratio γ had only a limited effect on the cells tuning properties and was fixed to 0.3 for all filters.

C1 cells. In order to better account for complex cells tuning properties, we assigned new values to the two parameters *ScaleRange* and *PoolRange* that control the filter bands in HMAX (see section 2.1). The number of filter bands was increased from 4 to 8 while the number of filters within each filter bands was decreased (from 3 to 2 in each band) thus providing less scale tolerance (therefore narrower spatial frequency bandwidth) to complex cells. Values for the *PoolRange* variables varied from 8 to 22 and new values were assigned to *ScaleRange*:

$$\text{PoolRange} = \{8; 10; 12; 14; 16; 18; 20; 22\} \quad (5)$$

$$\text{ScaleRange} = \{7-9; 11-13; 15-17; 19-21; 23-25; 27-29; 31-33; 35-39\} \quad (6)$$

	standard HMAX	Gabor filters
RF size (receptive field size)	$7 \times 7 \rightarrow 29 \times 29$ 12 filters in steps of 2	$7 \times 7 \rightarrow 39 \times 39$ 17 filters in steps of 2
θ (orientation)	0, $\frac{\pi}{4}$, $\frac{\pi}{2}$, $\frac{3\pi}{4}$	0, $\frac{\pi}{4}$, $\frac{\pi}{2}$, $\frac{3\pi}{4}$
σ (effective width)	RF size/4 1.8-7.3	$a\mathit{RF\ size}^2 + b\mathit{RF\ size} + c$ a = 0.0036; b = 0.35; c = 0.18 2.8-19.5
γ (aspect ratio)	1	0.3
λ (wavelength)	N/A N/A	$\sigma/0.8$ 3.5-24.4

Table 1: Comparison between parameters used in standard HMAX to model simple (S1) cells with Gaussian derivatives and the ones used to model simple (S1) cells with Gabor filters to better account for properties of parafoveal simple cells.

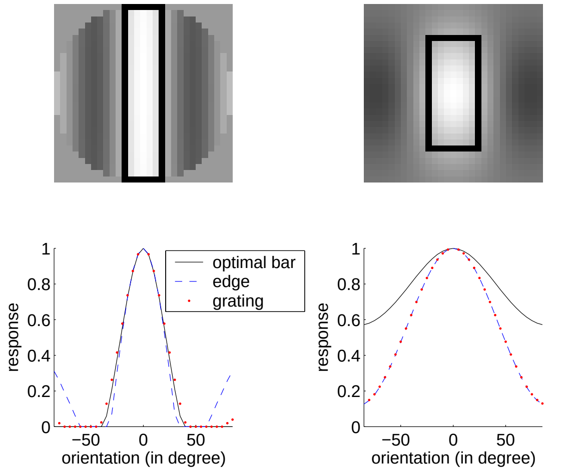

Figure 3: Top: Filters (Gabor (left) and Gaussian derivatives (right)) and preferred bar stimulus superimposed. Bottom: Corresponding orientation tuning curves obtained with optimal bars, gratings and edges. The three stimuli produced similar curves with Gabor filters but not with Gaussian derivatives as simple (S1) units tend to select shorter and wider bars.

2.2 Assessing model unit tuning properties

2.2.1 Orientation tuning

Orientation tuning was assessed in two ways: First, following [11], we swept sine wave gratings of optimal frequency over the receptive field of a model unit at thirty-six different orientations (spanning 180^o of the visual field in steps of 5^o). For each cell tested, the maximum response elicited for each orientation was recorded to fit a tuning curve and the orientation bandwidth at half-amplitude was calculated. For comparison with [13], we also swept edges and bars of optimal dimensions: For each cell the orientation bandwidth at 71% of the maximal response was calculated as in [13].

Sweeping edges, bars and gratings gave similar tuning curves for Gabor filters, suggesting that if simple cells can be well modeled by Gabor filters, measurements made by groups with different stimuli (bars, grating and edges) are indeed consistent. Bar stimuli with Gaussian derivatives as in standard HMAX , however lead to inconsistent tuning curves compared with edges and gratings, indicating that Gaussian derivatives are a poor model of simple cell processing.

2.2.2 Spatial frequency tuning

Spatial frequency selectivity was assessed by sweeping sine wave gratings of various spatial frequencies over a model unit's receptive field. For each grating frequency, the maximal cell response was recorded to fit a tuning curve and the spatial frequency selectivity bandwidth was calculated as in [9] by dividing the frequency score at the high crossover of the curve at half-amplitude by the low crossover at the same level.

Taking the $\log_2$ of this ratio gives the bandwidth value (in octaves):

$$\text{bandwidth } = \log_2 \frac{\text{high cut}}{\text{low cut}} \tag{7}$$

For comparison with [10], we also calculated the *selectivity index* as defined in [10], by dividing the frequency score at the high crossover of the curve at 71% of the maximal amplitude by the low crossover at the same level and multiplying this value by 100 (a value of 50 representing a specificity of 1 octave):

$$\text{selectivity index} = \frac{\text{high cut}}{\text{low cut}} \times 100 \tag{8}$$

2.3 Benchmark paperclip recognition task

2.3.1 Stimuli

To test translation, size and rotation invariance properties of the VTUs, we used 80 out of a set of 200 "paperclip" stimuli (20 targets, 60 distracters) similar to those used previously in [3, 4]. Examples of paperclip stimuli are shown in Fig. 4. The background pixel value was always set to zero (contrast 100%), as in [3, 4].

2.3.2 Shift

To examine shift invariance, we trained VTUs to each of the 20 target paperclips at size 64×64 pixels, positioned at the center of the 160×160 pixel input image. We then calculated C2 and VTU responses for all paperclips at eight random positions around the reference position. An example of tested positions for one paper clip (positions varied from one paperclip to another) is shown Fig. 4a.

2.3.3 Scaling

To examine size invariance, we trained VTUs to each of the 20 target paperclips at size 64×64 pixels, positioned at the center of the 160×160 pixel input image. We then calculated C2 and VTU responses for all paperclips at different sizes, in quarter-octave steps (*i.e.,* squares with edge lengths of 27, 32, 38, 45, 54, 64, 76, 91 108, 129 and 154 pixels), again positioned at the center of the 160×160 input image. Examples of three paperclips rescaled by ± 1 octave from reference (center) are shown in Fig. 4b.

2.3.4 Rotation

To examine invariance to rotation in depth, we trained VTUs to each of the 20 target paperclips at 0° rotation and size 64×64 pixels, positioned at the center of the input image (160×160). We then calculated C2 and VTU responses for all paperclips at different rotations from the origin (± 50° by steps of 4°). Examples of three paperclips at -20° , 0° and +20° are shown in Fig. 4c.

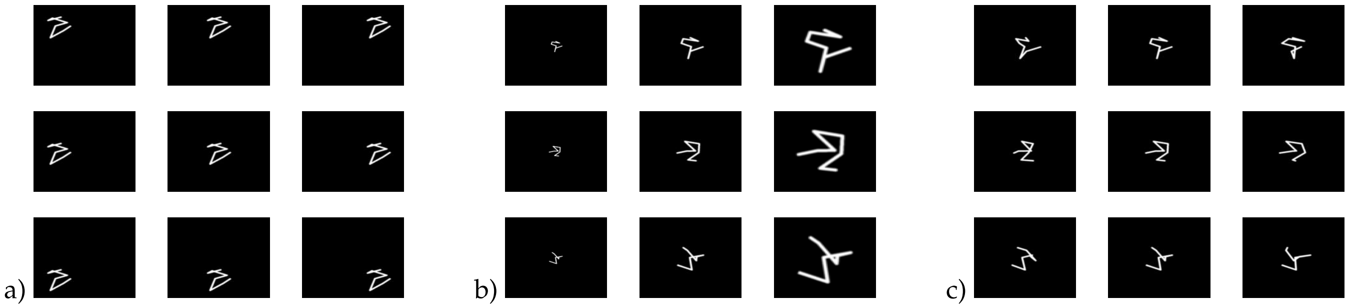

Figure 4: Stimulus transformations to test **a)** shift invariance (all tested positions), **b)** scaling invariance (each row shows a reference paperclip rescaled by ± 1 octave -left and right-) and **c)** 3-D rotation invariance (reference paperclip rotated by ± 20° -left and right-). For all invariance tests, the reference is the 64×64 pixel center paperclip.

2.3.5 Task

To assess the degree of invariance to stimulus transformations, we used a paradigm similar to the one used in [3, 4], in which a transformed (rescaled or rotated in depth) target stimulus is considered recognized in a certain presentation condition if the VTU tuned to the original target (default size and view), responds more strongly to its presentation than to the presentation of any distracter stimulus. This measures the hit rate at zero false positives.

3 Results

3.1 Original HMAX

3.1.1 Spatial frequency tuning

S1 units. We found that simple cells in original HMAX were too broadly tuned to spatial frequency: Spatial frequency bandwidth measured at half-amplitude was about 1.7 octaves for all units. De Valois *et al.* report a median value of 1.32 [9]) for parafoveal simple cells, with most cells lying around 1-1.5 octaves. We found a similar discrepancy between model units and cortical cells from data collected by Schiller *et al.* who report spatial-frequency selectivity index values in the range of 40-80. (HMAX cells index values vary between 34 and 41).

Because Gaussian derivatives only have one free parameter, we found it impossible to have them match both simple cells spatial frequency distribution and bandwidth. Setting σ so that spatial frequency selectivities of the two populations match [9] (1-5.6 for parafoveal cells *vs.* 1.4-5.8 cycles/degree as in standard HMAX) lead to overly broad spatial frequencies tuning profiles while setting σ so that spatial frequencies bandwidth match lead to peak frequencies too high. This motivates the use of functions with more degrees of freedom such as Gabor functions.

C1 units. Similarly, we found that complex cells were too broadly tuned to spatial frequency with a median spatial frequency bandwidth measured at half-amplitude around 2.1 octaves (range: 2.0-2.2 octaves) which is high compared to a value of 1.6 for Y cells parafoveal reported in [9]. Similarly, the spatial frequency index was around 30 and therefore lay outside the bulk (30-70) reported in [10].

3.1.2 Orientation tuning

S1 units. As in section 3.1.1 for spatial frequency, we found that Gaussian derivatives could not account for simple cell orientation tuning properties. Measured at half-amplitude, we found an orientation tuning bandwidth of 97° for all cells while De Valois *et al.* report a median value of 34° (range 20° - 90°). Even though the value reported is surprisingly low (parafoveal simple cells would thus be more narrowly tuned than foveal simple and complex cells), the discrepancy is still large when compared to data collected by Schiller *et al.* who report a bulk in the range 20° -50° [13] (measured at 71% of the maximal response with edges and bars) whereas HMAX unit orientation bandwidth calculated in this way was about 69° .

C1 units. Consistent with the fact that all model simple cells share similar orientation tuning properties and since complex cells pool over simple cells at the same preferred orientation, we found that HMAX C1 orientation tuning was identical to those of S1 units (97° at half amplitude and 69° at 71% max amplitude).

3.2 New HMAX with Gabor filter sets

3.2.1 Spatial frequency tuning

S1 units. As described in section 2.1.1, Gabor filter peak frequencies are parameterized by the inverse of their wavelength $\nu = \frac{1}{\lambda}$ (*i.e.*, the wavelength of the modulating sinusoid). We found that the values measured experimentally by sweeping optimally oriented gratings were indeed close to ν. As expected (see section 2.1.1), we also found a positive correlation between receptive field size and frequency bandwidth, as well as a negative correlation with peak frequency selectivities, which is consistent with recordings made in primate striate cortex [9, 10].

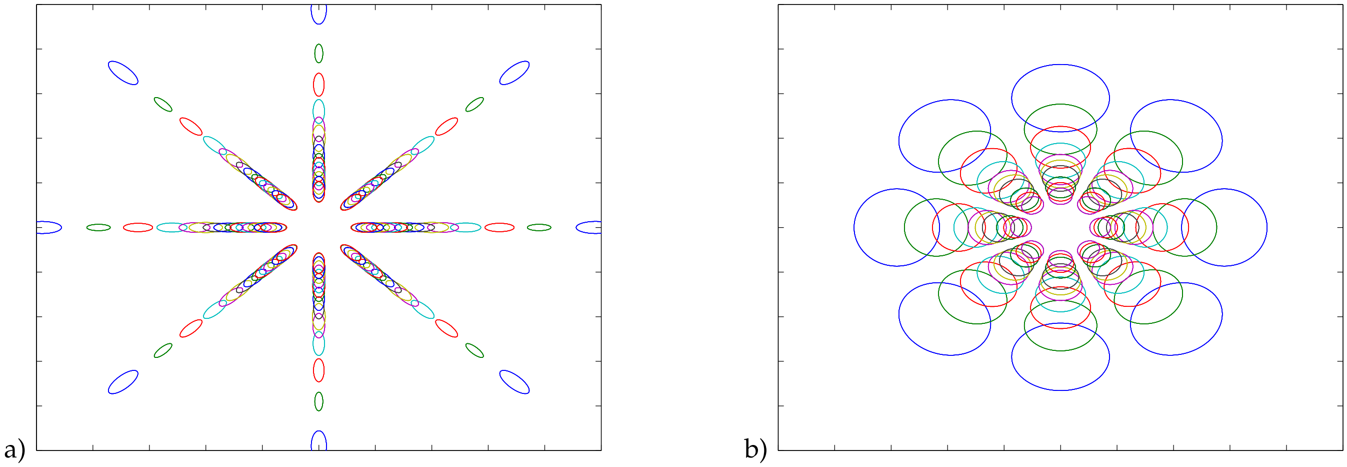

Figure 5: Coverage of the spatial frequency plane by Gabor function (a) and Gaussian derivatives as in standard HMAX (b). The length of the ellipses along the 4 axes of orientation ($\frac{\pi}{4}$, $\frac{\pi}{2}$, $\frac{3\pi}{4}$, π) indicate the filter frequency bandwidth and their widths, the filter orientation bandwidth (both measured at half-amplitude). The new S1 cells are more tightly tuned for both orientation and frequency but cover a wider range of spatial frequencies.

Model units' peak frequencies were in the range 1.6-9.8 cycles/degree (mean and median value of 3.7 and 2.8 cycles/degree respectively). This provides a reasonable fit with cortical simple cells peak frequencies lying between values as extreme as 0.5 and 8.0 degree/cycles but a bulk around 1.0-4.0 cycles/degree (mean value of 2.2 cycles/degree) [9]. Indeed, using our formula (3) to parameterize Gabor filters (see section 2.1.1), a cell with a peak frequency around 0.5 cycles/degree would have a receptive field size of about 2° which is very large compared to values reported in [8, 12] for simple cells.

Spatial frequency bandwidths measured at half-amplitude were all in the range 1.1-1.8 octaves, which corresponds to a subset of the range exhibited by cortical simple cells (values reported as extreme as 0.4 to values above 2.6 octaves). For the sake of simplicity, we tried to capture the range of "bulk frequency bandwidths" (1-1.5 octaves for parafoveal cells) and focused on population median values (1.45 for both cortical [9]) and model cells). For comparison with Schiller *et al.* , we measured the spatial frequency index and found values in the range 44-58 (median 55) which lies right in the bulk (40-70) reported in [10].

C1 units. Peak frequencies ranged from 1.8 to 7.8 cycles/degree (mean value and median values of 3.9 and 3.2 respectively) for our model complex cells. In [9], peak frequencies range between values as extreme as 0.5 and 8 cycles/degree with a bulk of cells lying between 2-5.6 cycles/degree (mean around 3.2).

We found spatial frequency bandwidths at half-amplitude in the range 1.5-2.0 octaves. Parafoveal complex cells lie between values as extreme as 0.4 to values above 2.6 octaves. Again, we tried to capture the bulk frequency bandwidths ranging between 1.0 and 2.0 octaves and matched the median values for the populations of model and cortical cells [9] (1.6 octaves for both). The spatial frequency bandwidth at 71% maximal response were in the range 40-50 (median 48) which lies within the bulk (40-60) reported in [10]. Fig 6 shows the complex *vs.* simple cells spatial frequency bandwidths.

3.2.2 Orientation tuning

S1 units. We found a median orientation bandwidth at half amplitude of 44° (range 38° -49°). In [11], a median value of 34° is reported. Again, as already mentioned earlier, this value seems surprising (it would imply that parafoveal cells are more tightly tuned than their foveal homologue, both simple (median value 42°) and complex (45°). When we used instead a measure of the bandwidth at 71% of the maximal response for comparison with Schiller *et al.* , the fit was better with a median value of 30° (range: 27° -33°) compared with a bulk of cortical simple cells within 20° -70° [13].

C1 units. We found a median orientation bandwidth at half amplitude of 43° which is in excellent agreement with the 44° reported in [11]. The bulk of cells reported in [13] is within 20° -90° and our values range between 27° -33° (median 31°), therefore placing our model units as part of the most narrowly tuned subpopulation of cortical complex cells. As in both experimental data sets, the orientation tuning bandwidth of the model complex units is very similar to that of simple units.

3.2.3 Summary

We found that model simple S1 cells in the original HMAX were too broadly tuned to both orientation and spatial frequency compared to cortical simple cells (see section 3.1). We motivated the use of Gabor filters for simple S1 cells and empirically determined a set of pa-

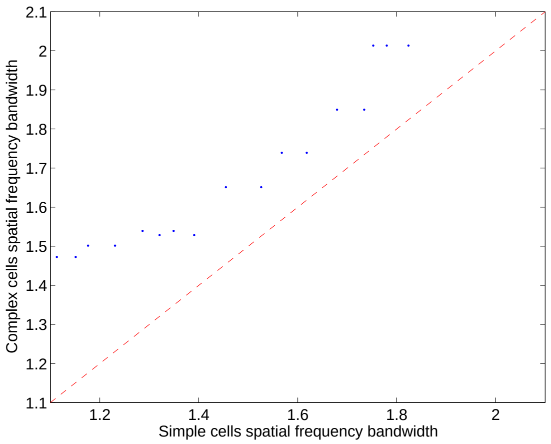

Figure 6: Complex cells spatial frequency bandwidth *vs.* simple cells spatial frequency bandwidth. There is an increase of about 20% from simple to complex cells spatial frequency bandwidth, consistent with parafoveal cortical cells [9, 10].

rameters so that model simple cells tuning properties match those of cortical simple cells (see section 3.2).

The new set of S1 cells differ from S1 cells in standard HMAX with respect to their orientation bandwidth (median 46° *vs.* 97° in standard HMAX), their peak frequencies selectivity (1.6-9.8 cycles/degree *vs.* 1.4-5.8 cycles/degree for standard HMAX), frequency selectivity bandwidth (median 1.47 *vs.* 1.7 in standard HMAX) and receptive field sizes (0.2° -1.1° *vs.* 0.2° -0.8° in standard HMAX). The new set of S1 cells is more narrowly tuned to both spatial frequency and orientation, span a larger range of frequencies and receptive field sizes (see Fig. 5) and match more closely parafoveal simple cells tuning properties.

It also appeared from our study that the pooling mechanisms inferred in the model for building complex cells from simple cells are indeed consistent with complex cells tuning properties. A comparison between HMAX and parafoveal complex cells, showed that position invariance (parameterized by the variable *PoolRange* (see section 2.1) is actually larger than in standard HMAX, while scale invariance (parameterized by the variable *ScaleRange* (see section 2.1) is actually smaller than in standard HMAX.

The new set of C1 cells differ from C1 cells in standard HMAX in terms of their orientation bandwidth (median 43° *vs.* 97° in standard HMAX), their peak frequencies selectivity (1.8-7.8 cycles/degree *vs.* 1.6-5.6 cycles/degree for standard HMAX), frequency selectivity bandwidth (median 1.6 *vs.* 2.1 in standard HMAX) and receptive field sizes (0.4° -1.7° *vs.* 0.3° -1.1° in standard HMAX). The new set of C1 cells is more narrowly tuned to both spatial frequency and orientation, span a larger range of frequencies (see Fig. 5) and match more closely parafoveal complex cells tuning properties.

3.3 Performance on a benchmark recognition task

To investigate the impact of this new representation on the VTUs' shape specificity as well as invariance to shift, size and rotation, we performed a benchmark recognition task with paperclip stimuli (see section 2.3.5) similar to the one used in [3, 4] and found that invariance properties were maintained. This suggests that the architecture in HMAX is robust to changes in the tuning properties of cells at the entry-level.

VTUs with the new sets of S1 and C1 units had a mean invariance to rotation in depth of about 34° (reference being on the same task 33° for HMAX and 29° reported by [4] for IT cells). For size invariance, we found a bandwidth of about 2.8 octaves (reference being at least 2.4 octaves for standard HMAX and about 2 octaves for IT cells [3]). Translation invariance was maintained with respect to all positions tested across the units receptive field compared to distracters at the center of the receptive field.

We also quantified the effect of the complex (C1) units receptive field size (controlled by the variable *P*oolRange, see section 2.1) on VTU scale invariance properties. We found that larger complex cells receptive field sizes lead to larger scale invariance at the VTUs level (see Fig. 7).

4 Discussion

4.1 Impact of the new S1 and C1 cells population on HMAX architecture

We proposed a new set of receptive field shapes and parameters for cells in the S1 and C1 layers of HMAX . We increased position invariance (parameterized by the variable *PoolRange*, see section 2.1) in model C1 cells while scale invariance (parameterized by the variable *ScaleRange*) was decreased compared to standard HMAX.

We showed that invariance properties at the VTU level were not substantially affected by these changes, indicating that the model appears to be robust to changes in the lower level of the hierarchy. Thus there exists a number of different pooling schemes between S1 and C2 cells that still account for VTU invariance properties.

We showed that a mechanism in which cells pool over afferents tuned to the same preferred features but at slightly different positions and scales is well suited to explain the increase in receptive field size and spatial frequency bandwidth from simple to complex cells. As we mentioned in section 2.1, the following S2 layer in the model (equivalent to V4 in primate cortex) was only qualitatively inspired by physiological data. Further studies should focus on intermediate visual areas such as V4 in which it was shown that the increase in receptive field sizes and spatial frequency bandwidth are even more pronounced.

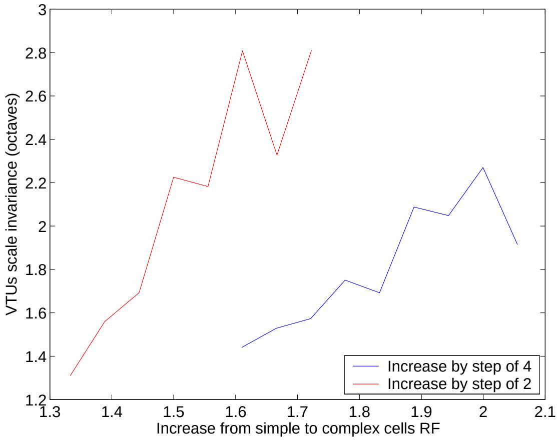

Figure 7: Effect of the variable $PoolRange$ on VTU scale invariance properties. Plotted are the mean VTU scale invariance *vs.* the increase in receptive field size from simple to complex cells when increased by step of 2 and 4 between pooling bands.

Schein & Desimone showed that the spatial frequency bandwidth median value at the V4 level was about 2.2 octaves [25] (which represents an increase of about 40% from the complex cells population). It is not clear whether this remains consistent with the four layer architecture of HMAX and further investigation on C2 units tuning properties should be performed. Also it has been shown in [25] that V4 contains cells covering a wide range of tuning properties (from 0.5 to > 4.0 octaves spatial frequency bandwidths). Although this could be an artifact of their methods (probing cells with the wrong stimuli), it is possible that direct pooling from C1 to C2 should be added.

We have thus shown that the physiological data on simple and complex cell receptive field size, spatial frequency and orientation bandwidth are in good agreement with the model hypothesis of complex cells performing a MAX pooling over simple cell afferents, a key step in the model towards invariant object recognition.

Invariance, *i.e.*, the ability to recognize a pattern under various transformations, is one goal of object recognition, another one being specificity, *i.e.*, the ability to discriminate between different patterns. The next challenge is to understand how shape complexity is increased along the ventral visual stream, from the Gabor-like preferred stimuli in V1 to neurons tuned to complex real-world stimuli such as faces and hands in IT.

References

[1] K. Fukushima. Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, 36:193–201, 1980.

[2] D.I. Perrett and M.W. Oram. Neurophsyiology of shape processing. *Imaging Vis. Comp.*, 11:317–33, 1993.

[3] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–25, 1999.

[4] N. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.*, 5:552–63, 1995.

[5] A. Yu, M. Giese, and T. Poggio. Biophysiologically plausible implementations of the maximum operation. *Neur. Comp.*, 14(12):2857–81, 2002.

[6] T. J. Gawne and J. M. Martin. Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J. Neurophysiol.*, 88:1128–35, 2002.

[7] I. Lampl, D. Ferster, T. Poggio, and M. Riesenhuber. Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *J. Neurophys*, page In press, 2004.

[8] D. Hubel and T. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.*, 28:229–89, 1965.

[9] R. L. De Valois, D. G. Albrecht, and L. G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vis. Res.*, 22:545–59, 1982.

[10] P.H. Schiller, B.L. Finlay, and S. F. Volman. Quantitative studies of single-cell properties in monkey striate cortex III. Spatial frequency. *J. Neurophysiol.*, 39(6):1334–51, 1976.

[11] R. L. De Valois, E. W. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vis. Res.*, 22:531–44, 1982.

[12] P.H. Schiller, B.L. Finlay, and S. F. Volman. Quantitative studies of single-cell properties in monkey striate cortex I. Spatiotemporal organization of receptive fields. *J. Neurophysiol.*, 39(6):1288–1319, 1976.

[13] P.H. Schiller, B.L. Finlay, and S. F. Volman. Quantitative studies of single-cell properties in monkey striate cortex II. Orientation specificity and ocular dominance. *J. Neurophysiol.*, 39(6):1334–51, 1976.

[14] M. Riesenhuber and T. Poggio. Models of object recognition. *Nat. Neurosci.*, 3 supp.:1199–1204, 2000.

[15] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Curr. Op. Neurobiol.*, 12:162–68, 2002.

[16] R. Schneider and M. Riesenhuber. A detailed look at scale and translation invariance in a hierarchical neural model of visual object recognition. Technical Report CBCL Paper 218 / AI Memo 2002-011, Massachusetts Institute of Technology, 2002.

[17] BG. Sclar, J. R. Maunsell, and P. Lennie. Coding of image contrast in central visual pathways of the macaque monkey. *Vis. Res.*, 30(1):1–10, 1990.

[18] B. C. Skottun, A. Bradley, G. Sclar, I. Ohzawa, and R. D. Freeman. The effect of contrast on visual orientation and spatial frequency discrimination: A comparison of single cells and behavior. *J. Neurophysiol.*, 57:773–86, 1987.

[19] D. G. Albrecht and D. B. Hamilton. Striate cortex of monkey and cat: Contrast response function. *J. Neurophysiol.*, 48:217–37, 1982.

[20] M. Kouh and M. Riesenhuber. Investigating shape representation in area V4 with hmax: Orientation and grating selectivities. Technical Report CBCL Paper 231 / AIM 2003-021, Massachusetts Institute of Technology, 2003.

[21] C. Cadieu, M. Kouh, and T. Poggio. Investigating position-specific tuning for boundary conformation in v4 with the standard model of object recognition. Technical Report *In prep.*, Massachusetts Institute of Technology, 2004.

[22] T. Serre, J. Louie, M. Riesenhuber, and T. Poggio. On the role of object-specific features for real world recognition in biological vision. In *BMCV*, pages 387–97, Tuebingen, Germany., 2002.

[23] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimization by two-dimensional visual cortical filters. *J. Opt. Soc. Am.*, 2:1160–69, 1985.

[24] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58:1233–58, 1987.

[25] R. Desimone and S. J. Schein. Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *J. Neurophysiol.*, 57:835–68, 1987.