massachusetts institute of technology — artificial intelligence laboratory

A Detailed Look at Scale and Translation Invariance in a Hierarchical Neural Model of Visual Object Recognition

Robert Schneider and Maximilian Riesenhuber

AI Memo 2002-011 August 2002
CBCL Memo 218

Abstract

The HMAX model has recently been proposed by Riesenhuber & Poggio [15] as a hierarchical model of position- and size-invariant object recognition in visual cortex. It has also turned out to model successfully a number of other properties of the ventral visual stream (the visual pathway thought to be crucial for object recognition in cortex), and particularly of (view-tuned) neurons in macaque inferotemporal cortex, the brain area at the top of the ventral stream. The original modeling study [15] only used "paperclip" stimuli, as in the corresponding physiology experiment [8], and did not explore systematically how model units' invariance properties depended on model parameters. In this study, we aimed at a deeper understanding of the inner workings of HMAX and its performance for various parameter settings and "natural" stimulus classes. We examined HMAX responses for different stimulus sizes and positions systematically and found a dependence of model units' responses on stimulus position for which a quantitative description is offered. Scale invariance properties were found to be dependent on the particular stimulus class used. Moreover, a given view-tuned unit can exhibit substantially different invariance ranges when mapped with different probe stimuli. This has potentially interesting ramifications for experimental studies in which the receptive field of a neuron and its scale invariance properties are usually only mapped with probe objects of a single type.

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. R.S. is supported by grants from the German National Scholarship Foundation, the German Academic Exchange Service and the State of Bavaria. M.R. is supported by a McDonnell-Pew Award in Cognitive Neuroscience.

1 Introduction

Models of neural information processing in which increasingly complex representation of a stimulus are gradually built up in a hierarchy have been considered ever since the seminal work of Hubel and Wiesel on receptive fields of simple and complex cells in cat striate cortex [2]. The HMAX model has recently been proposed as an application of this principle to the problem of invariant object recognition in the ventral visual stream of primates, thought to be crucial for object recognition in primates [15]. Neurons in the inferotemporal cortex (IT), the highest visual area in the ventral stream, do not only respond selectively to complex stimuli, their response to a preferred stimulus is also largely independent of the size and position of the stimulus in the visual field [5]. Similar properties are achieved in HMAX by a combination of two different computational mechanisms: a weighted linear sum for building more complex features from simpler ones, akin to a template match operation, and a highly nonlinear "MAX" operation, where a unit's output is determined by its most strongly activated input unit. "MAX" pooling over afferents tuned to the same feature, but at different sizes or positions, yields robust responses whenever this feature is present within the input image, regardless of its size or position (see Figure 1 and Methods). This model has been shown to account well for a number of crucial properties of information processing in the ventral visual stream of humans and macaques (see [6, 14, 16–18]), including view-tuned representation of three-dimensional objects [8], response to mirror images [9], recognition in clutter [11], and object categorization [1, 6].

Previous studies using HMAX have employed a fixed set of parameters, and did not examine in a systematic way the dependencies of model unit tuning properties on its parameter settings and the specific stimuli used. In this study, we examined the effects of variations of stimulus size and position on the responses of model units and its impact on object recognition performance in detail, using two different stimulus classes: paperclips (as used in the original publication [15]) and cars (as used in [17]), to gain a deeper understanding of how such IT neuron invariance properties could be influenced by properties of lower areas in the visual stream. This also provided insight into the suitability of standard HMAX feature detectors for "natural" object classes.

2 Methods

2.1 The HMAX model

The HMAX model of object recognition in the ventral visual stream of primates has been described in detail elsewhere [15]. Briefly, input images (we used 128×128 or 160×160 greyscale pixel images) are densely sampled by arrays of two-dimensional Gaussian filters, the so-called S1 units (second derivative of Gaussian, orientations 0°, 45°, 90°, and 135°, sizes from 7×7 to 29×29 pixels in two-pixel steps) sensitive to bars of different orientations, thus roughly resembling properties of simple cells in striate cortex. At each pixel of the input image, filters of each size and orientation are centered. The filters are sum-normalized to zero and square-normalized to 1, and the result of the convolution of an image patch with a filter is divided by the power (sum of squares) of the image patch. This yields an S1 activity between -1 and 1.

In the next step, filter bands are defined, *i.e.,* groups of S1 filters of a certain size range (7×7 to 9×9 pixels; 11×11 to 15×15 pixels; 17×17 to 21×21 pixels; and 23×23 to 29×29 pixels). Within each filter band, a pooling range is defined (variable *poolRange*) which determines the size of the array of neighboring S1 units of all sizes in that filter band which feed into a C1 unit (roughly corresponding to complex cells of striate cortex). Only S1 filters with the same preferred orientation feed into a given C1 unit to preserve feature specificity. As in [15], we used pooling range values from 4 for the smallest filters (meaning that 4×4 neighboring S1 filters of size 7×7 pixels and 4×4 filters of size 9×9 pixels feed into a single C1 unit of the smallest filter band) over 6 and 9 for the intermediate filter bands, respectively, to 12 for the largest filter band. The pooling operation that the C1 units use is the "MAX" operation, *i.e.,* a C1 unit's activity is determined by the strongest input it receives. That is, 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 taking the absolute value of their S1 inputs (before performing the MAX operation), modeling input from two sets of simple cell populations with opposite phase. Possible firing rates of a C1 unit thus range from 0 to 1. Furthermore, the receptive fields of the C1 units overlap by a certain amount, given by the value of the parameter *c1Overlap*. We mostly used a value of 2 (as in [15]), meaning that half the S1 units feeding into a C1 unit were also used as input for the adjacent C1 unit in each direction. Higher values of *c1Overlap* indicate a greater degree of overlap (for an illustration of these arrangements, see also Figure 5).

Within each filter band, a square of four adjacent, nonoverlapping C1 units is then grouped to provide input to a 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

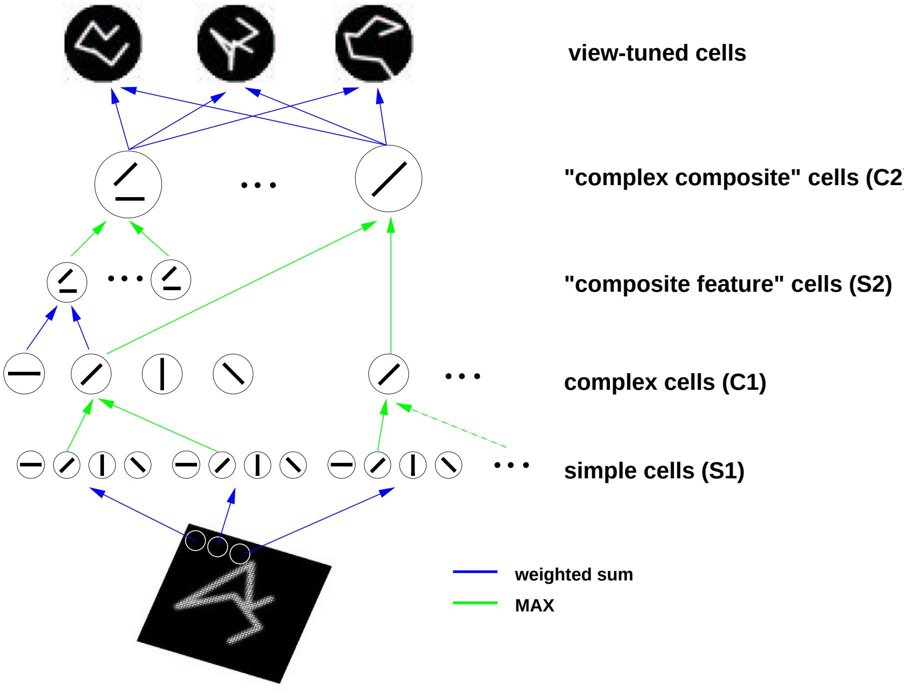

Figure 1: Schematic of the HMAX model. See Methods.

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.

To finally achieve size invariance over all filter sizes in the four filter bands and position invariance over the whole visual field, 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.

C2 units then again provide input to the view-tuned units (VTUs), named after their property of responding well to a certain two-dimensional view of a three-dimensional object, thereby closely resembling the view-tuned cells found in monkey inferotemporal cortex by Logothetis *et al.* [8]. The C2 → VTU connections are so far the only stage of the HMAX model where learning occurs. A VTU is tuned to a stimulus by selecting the activities of the 256 C2 units in response to that stimulus as the center of a 256-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. To achieve greater robustness in case of cluttered stimulus displays, only those C2 units may be selected as afferents for a VTU that respond most strongly to the training stimulus [14]. We ran simulations with the 40, 100, and 256 strongest afferents to each VTU. An additional parameter specifying response properties of a VTU is its σ value, or the standard deviation of its Gaussian response function. A smaller σ value yields more specific tuning since the resultant Gaussian has a narrower half-maximum width.

2.2 Stimuli

We used the "8 car system" described in [17], created using an automatic 3D multidimensional morphing system [19]. The system consists of morphs based on 8 prototype cars. In particular, we created lines in morph space connecting each of the eight prototypes to all the other prototypes for a total of 28 lines through morph space, with each line divided into 10 intervals. This created a set of 260 unique cars, and induces a similarity metric: any two prototypes are spaced 10 morph steps apart, and a morph at morph distance, *e.g.*, 3 from a prototype is more similar to this prototype than another morph at morph distance 7 on the same morph line. Every car stimulus was viewed from the same angle (left frontal view).

In addition, we used 75 out of a set of 200 paperclip stimuli (15 targets, 60 distractors) identical to those used by Logothetis *et al.* in [8], and in [15]. Each of those was viewed from a single angle only. Unlike in the case of cars, where features change smoothly when morphed from one prototype to another, paperclips lo-

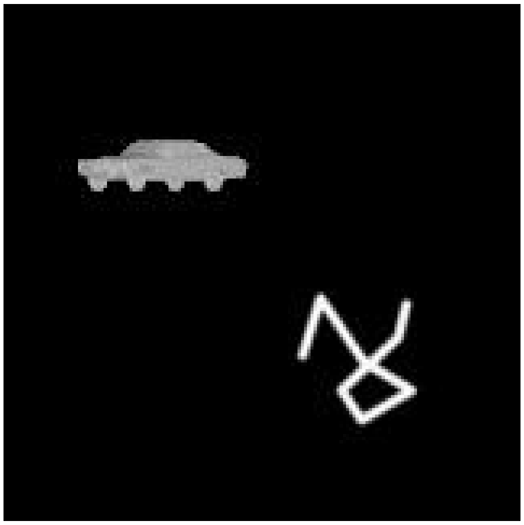

Figure 2: Examples of the car and paperclip stimuli used. Actual stimulus presentations contained only one stimulus each.

cated nearby in parameter space can appear very different perceptually, for instance, when moving a vertex causes two previously separate clip segments to cross. Thus, we did not examine the impact of parametric shape variations on recognition performance for the case of paperclips.

Examples of car and paperclip stimuli are provided in Figure 2. The background pixel value was always set to zero.

2.3 Simulations

2.3.1 Stimulus transformations

Translation. To study differences in HMAX output due to shifts in stimulus position, we trained VTUs to the 8 car prototypes and 15 target paperclips positioned along the horizontal midline and in the left half of the input image. We then calculated C2 and VTU responses to those training stimuli and all remaining 252 car and 60 paperclip stimuli for 60 other positions along the horizontal midline, spaced single pixels apart. (Results for vertical stimulus displacements are qualitatively identical, see Figure 3a.) To examine the effects of shifting a stimulus beyond the receptive field limits of C2 cells, and to find out how invariant response properties of VTUs depend on the stimulus class presented, we displayed 5 cars and 5 paperclips in isolation within a 100×100 pixel-sized image, at all positions along the horizontal and vertical midlines, including positions where the stimulus all but disappeared from the image. Responses of the 10 VTUs trained to these car and paperclip stimuli (when centered within the 100×100 image) to all these stimuli, of both stimulus classes, were then calculated.

Scaling. To examine size invariance, we trained VTUs to each of the 8 car prototypes and each of the 15 target paperclips at size 64×64 pixels, positioned at the center of the input image. We then calculated C2 and VTU responses for all cars and paperclips at different sizes, in half-octave steps (*i.e.,* squares with edge lengths of 16, 24, 32, 48, 96, and 128 pixels, and additionally 160 pixels), again positioned at the center of the 160×160 input image.

2.3.2 Assessing the impact of different filters on model unit response

To investigate the effects of different filter sizes on overall model unit activity, we performed simulations using individual filter bands (instead of the four in standard HMAX [15]). The filter band source of a C2 unit's activity (*i.e.,* which filter band was most active for a given S2 / C2 feature and thus determined the C2 unit's response) could be determined by running HMAX on a stimulus with only one filter band active at a time and comparing these responses with the response to the same stimulus when all filter bands were used.

2.3.3 Recognition tasks

To assess recognition performance, we used two different recognition paradigms, corresponding to two different behavioral tasks.

"Most Active VTU" paradigm. In the first paradigm, a target is said to be recognized if the VTU tuned to it fires more strongly to the test image than all other VTUs tuned to other members of the same stimulus set (*i.e.,* the 7 other car VTUs in the case of cars, or the 14 other paperclip VTUs in the case of paperclips). Recognition performance in a given condition (*e.g.,* for a certain stimulus size) is 100% if this holds true for all prototypes. Chance performance here is always the inverse of the number of VTUs (*i.e.,* prototypes), since for any given stimulus presentation, the probability that any VTU is the most active is 1 over the number of VTUs. This paradigm corresponds to a psychophysical task in which subjects are trained to discriminate between a fixed set of targets, and have to identify which of them appears in a given presentation. We will refer to this way of measuring recognition performance as the "Most Active VTU" paradigm.

"Target-Distractor Comparison" paradigm. Alternatively, a target stimulus can be considered recognized in a certain presentation condition if the VTU tuned to it responds more strongly to its presentation than to the presentation of a distractor stimulus. If this holds for all distractors presented, recognition performance for that condition and that VTU is 100%. Chance performance is reached at 50% in this paradigm, *i.e.,* when a VTU responds stronger or weaker to a distractor than to the target for equal numbers of distractors. This corresponds to a two-alternative forced-choice task in psychophysics

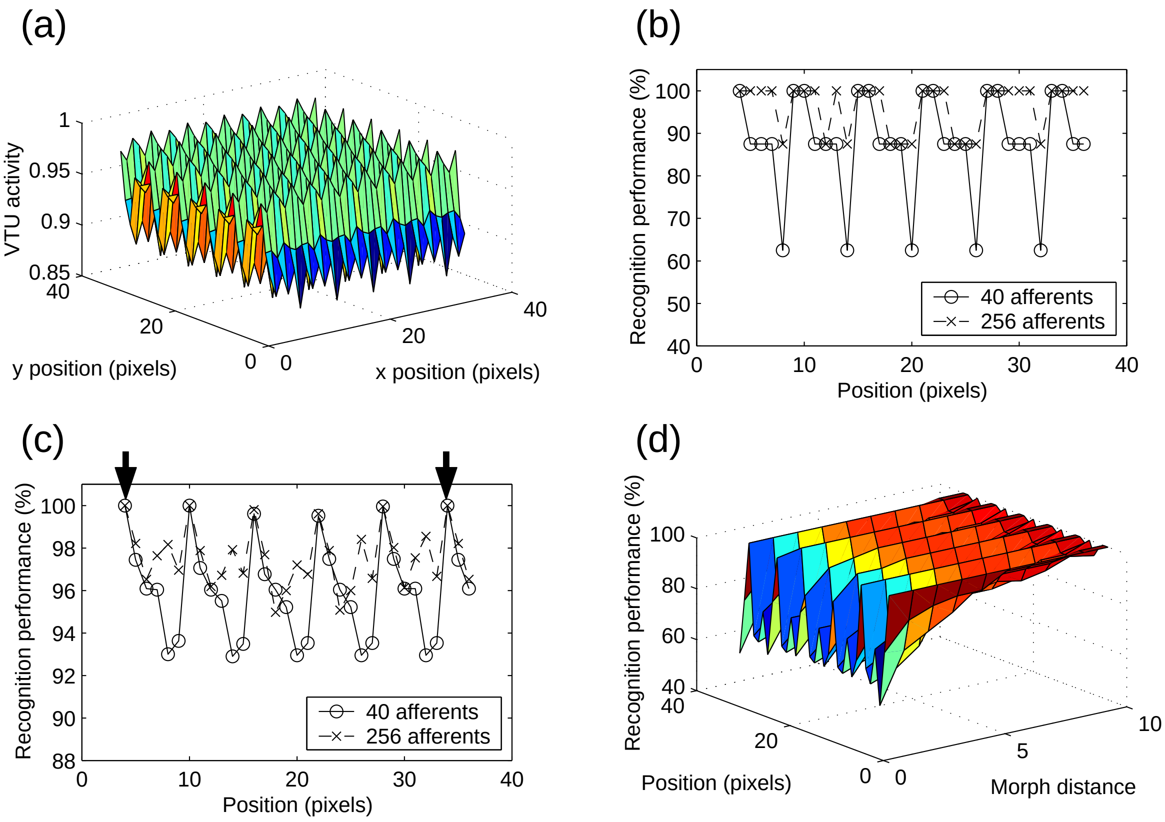

Figure 3: Effects of stimulus displacement on VTUs tuned to cars. **(a)** Response of a VTU (40 afferents, $\sigma = 0.2$) to its preferred stimulus at different positions in the image plane. **(b)** Mean recognition performance of 8 car VTUs ($\sigma = 0.2$) for different positions of their preferred stimuli and two different numbers of C2 afferents in the Most Active VTU paradigm. 100% performance is achieved if, for each prototype, the VTU tuned to it is the most active among the 8 VTUs. Chance performance would be 12.5%. **(c)** Same as (b), but for the Target-Distractor Comparison paradigm. Chance performance would be 50%. Note smaller variations for greater numbers of afferents in both paradigms. Periodicity (30 pixels) is indicated by arrows. Apparently smaller periodicity in (b) is due to the lower number of distractors, producing a coarser recognition performance measure. **(d)** Mean recognition performance in the Target-Distractor Comparison paradigm for car VTUs with 40 afferents and $\sigma = 0.2$, plotted against stimulus position and distractor similarity ("morph distance"; see Methods).

in which subjects are presented with a sample stimulus (chosen from a fixed set of targets) and two choice stimuli (one of them being the sample target, and the other being a distractor of varying similarity to the target) and have to indicate which of the two choice stimuli is identical to the sample. We will refer to this paradigm as the "Target-Distractor Comparison" paradigm.

For paperclips, distractors in the latter paradigm performance were 60 clips randomly chosen from the whole set of 200 clips. We thus used exactly the same method to assess recognition performance as in [14] for double stimuli and cluttered scenes. For cars, we either chose all 259 nontarget cars as distractors for a given target (*i.e.,* prototype) car, as in Figures 3c and 10c. Or, for a given target car, we used only those car stimuli as distractors that were morphs on any of the 7 morph lines leading away from that particular target (including the 7 other car prototypes). By grouping those distractors according to their morph distance from the target stimulus, we could assess the differential effects of using more similar or more dissimilar distractors in addition to the effects of variations in stimulus size or position. This additional dimension is plotted in Figures 3d and 10d.

3 Results

3.1 Changes in stimulus position

We find that the response of a VTU to its preferred stimulus depends on the stimulus' position within the image (Figure 3a). Moving a stimulus away from the position it occupied during training of the VTU results in decreased VTU output. This can lead to a drop in recognition performance, *i.e.,* another VTU tuned to a different stimulus might fire more strongly to this stimulus than the VTU actually tuned to it (Figure 3b), or the

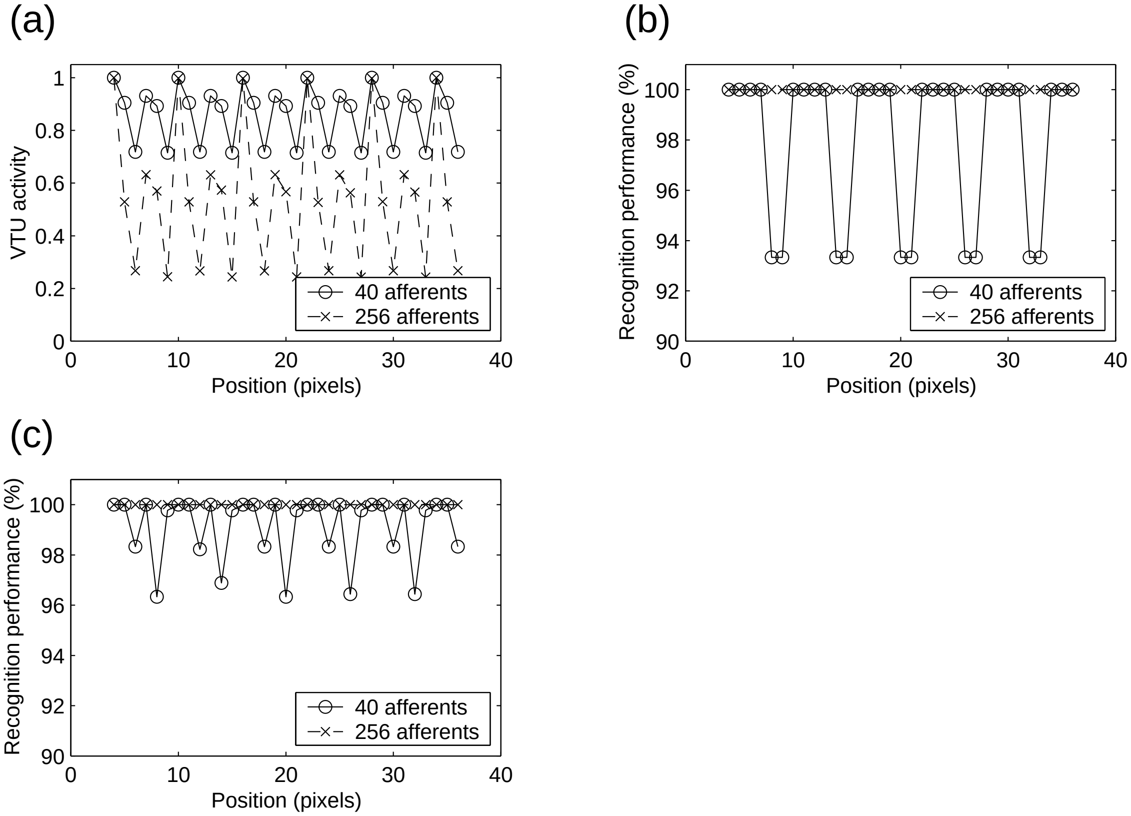

Figure 4: Effects of stimulus displacement on VTUs tuned to paperclips. **(a)** Responses of VTUs with 40 or 256 afferents and $\sigma = 0.2$ to their preferred stimuli at different positions. Note larger variations in VTU activity for more afferents, as opposed to variations in recognition performance, which are smaller for more afferents. **(b)** Mean recognition performance of 15 VTUs tuned to paperclips ($\sigma = 0.2$) in the Most Active VTU paradigm, depending on position of their preferred stimuli, for two numbers of C2 afferents. Chance performance would be 6.7%. **(c)** Same as (b), but for the Target-Distractor Comparison paradigm. Chance performance would be 50%.

VTU might respond more strongly to a nontarget stimulus at that position (Figure 3c). This is more likely for a similar than for a dissimilar distractor (Figure 3d). The changes in VTU output and recognition performance depend on position in a periodic fashion, but they are qualitatively identical across stimulus classes (see Figure 4 for paperclips). Larger variations in recognition performance for cars are due to the fact that, on average, any two car stimuli are more similar to each other (in terms of C2 activation patterns) than any two paperclips.

The "wavelength" λ of these "oscillations" in output and recognition performance is found to be a function of both the spatial pooling range of the C1 units (*poolRange*) and their spatial overlap (*c1Overlap*) in the different filter bands, in a manner described as follows:

$$\lambda = lcm_i \left[ceil \left(\frac{poolRange_i}{c1Overlap_i} \right) \right] \qquad (1)$$

with i running from 1 to the number of filter bands, and lcm being the lowest common multiple. Thus, standard HMAX parameters (pooling range 4, 6, 9, or 12 S1 units for C1 units in the four filter bands, respectively; C1 overlap 2) yield a λ value which is the least common multiple of 2, 3, 5, and 6, *i.e.,* 30. This means that changing the position of a stimulus by multiples of 30 pixels in x- or y-direction does *not* alter C2 or VTU responses or recognition performance. (Smaller λ values for paperclips apparent in Figure 4 derive from the dominance of small filter bands activated by the paperclip stimuli, as discussed later and in Figure 11).

These modulations can be explained by a "loss of features" occurring due to the way C1 and S2 units sample the input image and pool over their afferents. This is depicted in Figure 5 for only one filter size (*i.e.,* a single spatial pooling range). A feature of the stimulus in its original position (symbolized by two solid bars) is detected by adjacent C1 units that feed into the same S2 unit. Moving the stimulus to the right can position the right part of the feature beyond the limits of the right C1 unit's receptive field, while the left feature part is still detected by the left C1 unit. Consequently, this feature is "lost" for the S2 unit these C1 units feed into. However, the feature is not detected by the next S2 unit

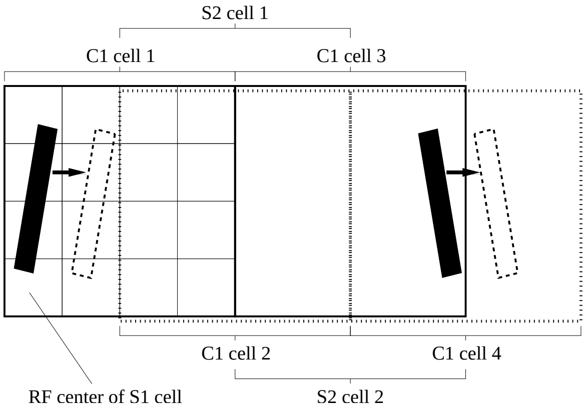

Figure 5: Schematic illustration of feature loss due to a change in stimulus position. In this example, the C1 units have a spatial pooling range of 4 (*i.e.,* they pool over an array of 4×4 S1 units) and an overlap of 2 (*i.e.,* they overlap by half their pooling range). Only 2 of 4 C1 afferents to an S2 unit are shown. See text.

to the right, either (whose afferent C1 units are depicted by dotted outlines in Figure 5), since the left feature part does not yet fall within the limits of this S2 unit's left C1 afferent. Not until the stimulus has been shifted far enough so that all features can again be detected by the next set of S2 units to the right will the output of HMAX again be identical to the original value. As can be seen in Figure 5, this is precisely the case when the stimulus is shifted by a distance equal to the offset of the C1 units with respect to each other, which in turn equals the quotient *poolRange/c1Overlap*. If multiple filter bands are used, each of which contributes to the overall response, the more general formula given in Eq. 1 applies.

Note that stimulus position during VTU training is in no way special, and shifting the stimulus to a different position might just as well cause "emergence" of features not detected at the original position. However, due to the VTUs' Gaussian tuning, *any* deviation of the C2 activation pattern from the training pattern will cause a decrease in VTU response, regardless of whether individual C2 units display a stronger or weaker response.

To improve performance during recognition in clutter, only a subset of the 256 C2 units – those which respond best to the original stimulus – may be used as inputs to a VTU, as described in [14]. This results in a less pronounced variation of a VTU's response when its preferred stimulus is presented at a nonoptimal position, simply because fewer terms appear in the exponent of the VTU's Gaussian response function so that fewer deviations from the optimal C2 unit activity are summed up (see Figure 4a). On the other hand, choosing a smaller value for the σ parameter of a VTU's Gaussian response function – corresponding to a sharpening of its tuning – leads to larger variations in its response for shifted stimuli, since the response will drop more sharply already for a small change of the C2 activation pattern. In any case, it should be noted that even sharp drops in VTU activity need not entail a corresponding decrease in recognition performance. In fact, while using a greater number of afferents to a VTU increases the magnitude of activity fluctuations due to changing stimulus position, the variations in recognition performance are actually *smaller* than for fewer afferents (see, for example, Figures 3b, c, and 4b, and c). This is likely due to the increasing separation of stimuli as the dimensionality of the feature space increases.

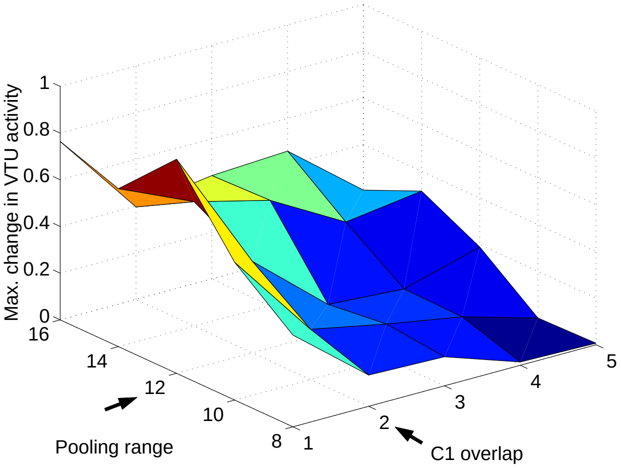

Figure 6: Dependence of maximum stimulus shift-related differences in VTU activity on pooling range and C1 overlap. Results shown are for a typical VTU with 256 afferents and $\sigma = 0.2$ tuned to a car stimulus using only the largest S1 filter band (23×23 to 29×29 pixels) employed. Arrows indicate standard HMAX parameter settings.

Figure 6 shows the maximum differences in VTU activity encountered due to variations in stimulus position for different pooling ranges and C1 overlaps, using only the largest S1 filter band (from 23×23 pixels to 29×29 pixels). Activity modulations are generally greater for larger pooling ranges and smaller C1 overlaps, since the input image is sampled more coarsely for such parameter settings, and thus the chance for a given feature to become "lost" is greater. Note that for a pooling range of 16 S1 units and a C1 overlap of 4, VTU response is subject to larger variations at different stimulus positions than for a pooling range of 8 and C1 overlap of 2, even though both cases, in accordance with equation 1, share the same λ value. This can be explained by the fact that large S1 filters only detect large-scale variations in the image. If nevertheless small C1 pooling ranges are used, the resulting small receptive fields of S2 units will experience only minor differences in their input when the stimulus is shifted

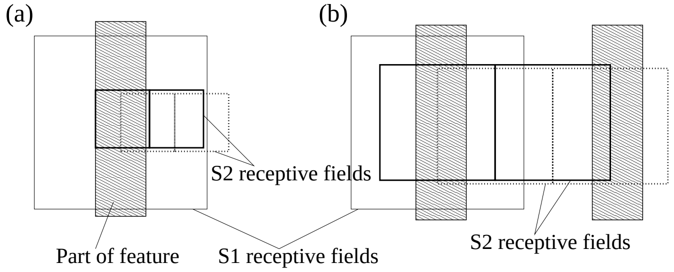

Figure 7: Illustration of different effects of stimulus displacement for different C1 pooling ranges and large S1 filters. As in Figure 5, boundaries of C1 / S2 receptive fields are drawn around *centers* of S1 receptive fields. Actual S1 receptive fields, however, are of course much larger, especially for large filter sizes. **(a)** Large S1 filters yield large features. For small C1 pooling ranges, only few neighboring S1 cells feed into a C1 cell, and they largely detect the same feature due to coarse filtering and since they are only spaced single pixels apart. Hence, moving a stimulus does not change input to a C1 cell much. **(b)** C1 / S2 units with larger pooling ranges are more likely to detect complex features (indicated by two shaded bars instead of only one in (a)) even if the S1 filters used are large. Thus, in this situation, C1 / S2 cells are more susceptible to variations in stimulus position.

by a few pixels (see Figure 7). Conversely, when small S1 filters are used, the filtered image varies most on a small scale, and smaller pooling ranges will yield larger variations in HMAX output than larger pooling ranges (not shown).

The position-dependent modulations of C2 unit activity (Figure 8), one level below the VTUs, are in agreement with the observations made earlier. The additional drop in activity for the leftmost stimulus positions is due to an edge effect at the corner of the image. Since care has been taken in HMAX to ensure that all S1 filters, even those centered on pixels at the image boundary, receive an equal amount of input (to achieve this, the input image is internally padded with a "frame" of zero-value pixels), even a feature situated at an image boundary cannot slip through the array of C1 units. This is different, however, for S2 units; an S2 unit that detects such a feature via its top right and/or bottom right C1 afferent might not be able to detect this feature any more if it is positioned at the far left of the input image.

3.2 Changes in stimulus size

Size-invariance in object recognition is achieved in HMAX by pooling over units that respond best to the same feature at different sizes. As has already been shown in [15], this works well over a wide range of stimulus sizes for paperclip stimuli: VTUs tuned to paperclips of size 64×64 pixels display very robust recognition performance for enlarged presentations of the clips and also good performance for reduced-size presentations (see Figure 9).

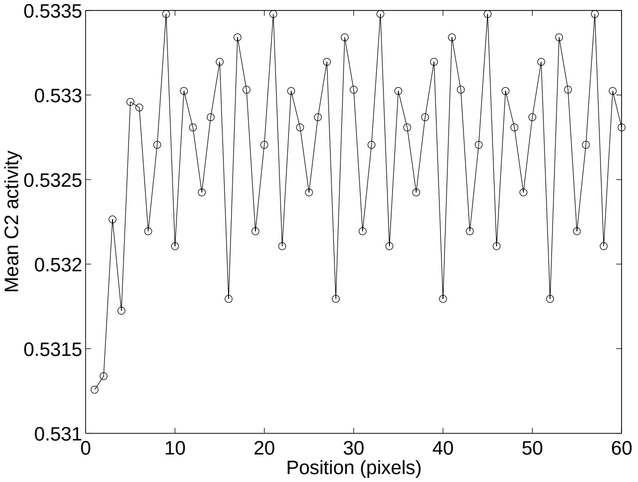

Figure 8: Mean activity of all 256 C2 units plotted against position of a car stimulus. Note that in the simulation run to generate this graph, spatial pooling ranges of the four filter bands used were still 4, 6, 9, and 12 S1 units, respectively, as in previous figures, while the C1 overlap value was changed to 3. Thus, in agreement with equation 1, the periodicity of activity changes was reduced to 12 pixels.

Interestingly, with cars a different picture emerges. As can be seen in Figure 10b and c, recognition performance in both paradigms drops considerably both for enlarged and downsized presentations of a car stimulus. Figure 10d shows that this low performance does not even increase if rather dissimilar distractors are used (corresponding to higher values on the morph axis). Shrinking a stimulus understandably decreases recognition performance since its characteristic features quickly disappear due to limited resolution. Figure 11 suggests a reason why performance for car stimuli drops for increases in size as well. While paperclip stimuli (size 64×64 pixels) elicit a response mostly in filter bands 1 and 2 (containing the smaller filters from 7×7 to 15×15 pixels in size), car stimuli of the same size mostly activate the large filters (23×23 to 29×29 pixels) in filter band 4. This is most probably due to the fact that our "clay model-like" rendered cars contain very little internal structure so that their most conspicuous features are their outlines. Consequently, enlarging a car stimulus will blow up its characteristic features (to which, after all, the VTUs are trained) beyond the scale that can effectively be detected by the standard S1 filters of the model, reducing recognition performance considerably.

Mathematically, both scaling and translation are ex-

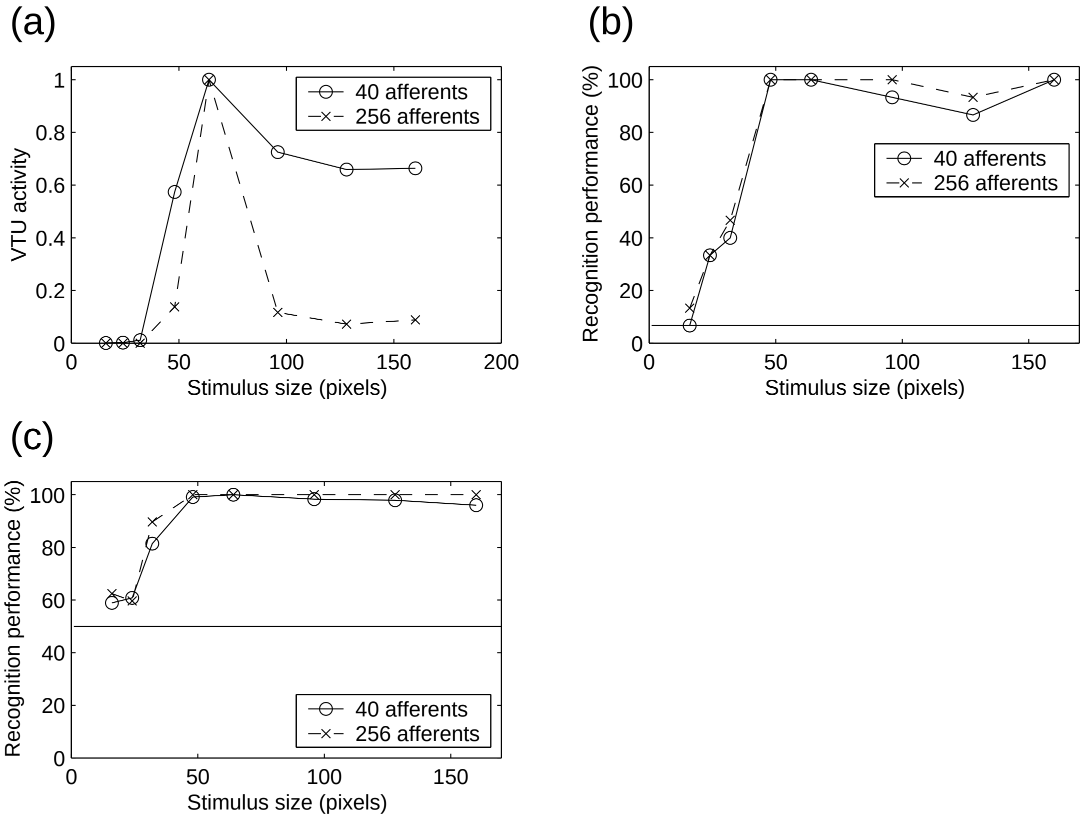

Figure 9: Effects of stimulus size on VTUs tuned to paperclips of size 64×64 pixels. **(a)** Responses of VTUs with 40 or 256 afferents and $\sigma = 0.2$ to their preferred stimuli at different sizes. **(b)** Mean recognition performance of 15 VTUs tuned to paperclips ($\sigma = 0.2$) in the Most Active VTU paradigm, depending on the size of their preferred stimuli, for two different numbers of C2 afferents. Horizontal line indicates chance performance (6.7%). **(c)** Same as (b), but for the Target-Distractor Comparison paradigm. Horizontal line indicates chance performance (50%).

amples of 2D affine transformations whose effects on an object can be estimated exactly from just one object view. The two transformations are also treated in the same fashion in HMAX, by MAX-pooling over afferents tuned to the same feature, but at different positions or scales, respectively. However, it appears that, while the behavior of the model for stimulus translation is similar for the two object classes we used, the scale invariance ranges differ substantially. This is, however, most likely not due to a fundamental difference in the representation of these two stimulus transformations in a hierarchical neural system. It has to be taken into consideration that, in HMAX, there are more units at different positions for a given receptive field size than there are units with different receptive field sizes for a given position. Moreover, stimulus position was changed in a linear fashion in our experiments, pixel by pixel, and only within the receptive fields of the C2 units, while stimulus size was changed exponentially, making it more likely that critical features appear at a scale beyond detectability. Conversely, with a broader range of S1 filter sizes and smaller, linear steps of stimulus size variation, similar periodical changes of VTU activity and recognition performance as with stimulus translation might be observed. Or, recognition performance for translation of stimuli could depend on stimulus class in an analogous manner as found here for scaling if, for example, for a certain stimulus class the critical features were located only at a single position or a small number of nearby positions within the stimulus, which would cause the response to change drastically when the corresponding part of the object was moved out of the receptive field (see also next section).

However, control experiments with larger S1 filter sizes (up to 59×59 pixels) failed to improve recognition performance for scaled car stimuli over what was observed in Figure 10, because in this case, again the largest filters were most active already in response to a car stimulus of size 64×64 pixels (not shown). This makes clear that, especially if neural processing resources are limited, recognition performance for transformed stimuli depends on how well the feature set is matched to the object class. Indeed, paperclips are composed of features that the standard HMAX S2/C2 units apparently capture quite well, namely combinations of bars of various orientations, and the receptive field sizes

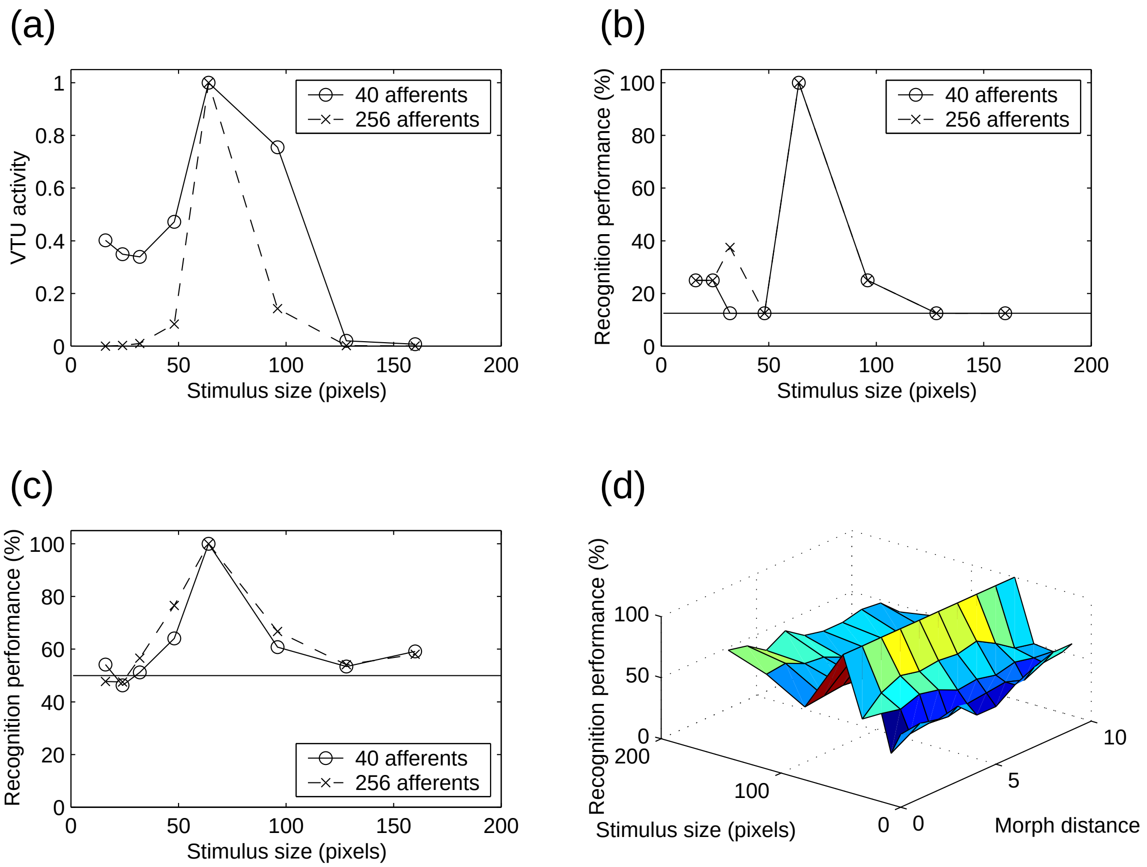

Figure 10: Effects of stimulus size on VTUs tuned to cars of size 64×64 pixels. **(a)** Responses of VTUs with 40 or 256 afferents and $\sigma = 0.2$ to their preferred stimuli at different sizes. **(b)** Mean recognition performance of 8 car VTUs ($\sigma = 0.2$) for different sizes of their preferred stimuli and two different numbers of C2 afferents in the Most Active VTU paradigm. Chance performance (12.5%) indicated by horizontal line. **(c)** Same as (b), but for the Target-Distractor Comparison paradigm. Chance performance (50%) indicated by horizontal line. **(d)** Mean recognition performance in the Target-Distractor Comparison paradigm for car VTUs with 40 afferents and $\sigma = 0.2$, plotted against stimulus size and distractor similarity ("morph distance").

of the S2 filters they activate most are in good correspondence with their own size. Consequently, the different filter bands present in the model can actually be employed appropriately to detect the critical features of paperclip stimuli at different sizes, leading to a high degree of scale invariance for this object class in HMAX.

3.3 Influence of stimulus class on invariance properties

Results in the previous section demonstrated that model VTU responses and recognition performance can depend on the particular stimulus class used, and on how well the features preferentially detected by the model match the characteristic features of stimuli from that class. This was shown to be an effect of the S2 level features. Therefore, we would expect to see different invariance properties not only for VTUs tuned to different objects, but also for a *single* VTU when probed with different stimuli, depending on how well the shape of the different probe stimuli is matched by the S2 features.

Figure 12 shows that this is indeed possible. Panel (a) displays responses of a car VTU to its preferred stimulus and a paperclip stimulus, respectively, at varying stimulus positions including positions beyond the receptive fields of the VTU's C2 afferents. While different strengths of the responses to the two stimuli, different periodicities of response variations — corresponding to the filter bands activated most by the two stimuli, as discussed in section 3.1 — and edge effects are observed, invariance ranges, *i.e.*, the range of positions that can influence the VTU's responses, are approximately equal. This indicates that the stimulus regions that maximally activate the different S2 features do not cluster at certain positions within the stimulus, but are fairly evenly distributed, as mentioned above in section 3.2. Panel (b), however, shows differing invariance properties of another VTU tuned to a car stimulus when presented with its preferred stimulus or a paperclip, for

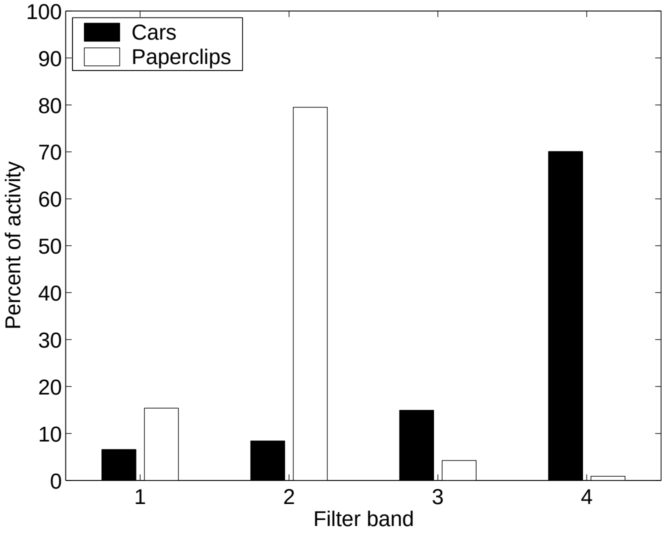

Figure 11: HMAX activities in different filter bands. Plot shows filter band source of C2 activity for 8 car prototypes and 8 randomly selected paperclips, all of size 64×64 pixels. Filter band 1: filters 7×7 and 9×9 pixels; filter band 2: from 11×11 to 15×15 pixels; filter band 3: from 17×17 to 21×21 pixels; filter band 4: from 23×23 to 29×29 pixels. Each bar indicates the mean percentage of C2 units that derived their activity from the corresponding filter band when a car or paperclip stimulus was shown. The filter band that determines the response of a C2 unit contains the most active units among those selective to that C2 unit's preferred feature. It indicates the size at which this feature occurred in the image.

stimuli varying in size. While this VTU's maximum response level is reached with its preferred stimulus, its response is more invariant when probed with a paperclip stimulus, due to the choice of filter sizes in HMAX and the filter bands activated most by the different stimuli. (Smaller invariance ranges of VTU responses for nonpreferred stimuli are also observed, but usually invariance ranges for preferred and nonpreferred stimuli are not the same.) This suggests that data from a physiology experiment about response invariances of a neuron, which are usually not collected with the stimulus the neuron is actually tuned to [7], might give misleading information about its actual response invariances, since these can depend on the stimulus used to map receptive field properties.

4 Discussion

The ability to recognize objects with a high degree of accuracy despite variations of their particular position and scale on the retina is one of the major accomplishments of the visual system. Key to this achievement may be the hierarchical structure of the visual system, in which neurons with more complex response properties (*i.e.,* responding to more complex features or showing a higher degree of invariance) result from the combination of outputs of neurons with simpler response properties [15]. It is an open question how the parameters of the hierarchy influence the invariance and shape-tuning properties of neurons in IT. In this paper, we have studied the performance of a hierarchical model of object recognition in cortex, the HMAX model, on tasks involving changes in stimulus position and size using abstract (paperclips) and more "natural" stimuli (cars). Invariant recognition is achieved in HMAX by pooling over model units that are sensitive to the same feature, but at different sizes and positions — an approach which is generally considered key to the construction of complex receptive fields in visual cortex [2, 3, 10, 13, 20]. Pooling in HMAX is done by the MAX operation, which preserves feature specificity while increasing invariance range [15].

A simple yet instructive solution to the problem of invariant recognition consists of a detector for each object, at each scale and each position. While appealing in its simplicity, such a model suffers from a combinatorial explosion of the number of cells — for each additional object to be recognized, another set of cells would be required — and from its lack of generalizing power: If an object had only been learned at one scale and position in this system, recognition would not transfer to other scales and positions.

The observed invariance ranges of IT cells after training with one view are reflected in the architecture used in HMAX (see [14]): One of its underlying ideas is that invariance and feature specificity have to grow in a hierarchy so that view-tuned cells at higher levels show sizeable invariance ranges even after training with only one view, as a result of the invariance properties of the afferent units. The key concept is to start with simple localized features — since the discriminatory power of simple features is low, the invariance range has to be kept correspondingly low to avoid the cells being activated indiscriminately. As feature complexity and thus discriminatory power grow, the invariance range, *i.e.,* the size of the receptive field, can be increased as well. Thus, loosely speaking, feature specificity and invariance range are inversely related, which is one of the reasons the model avoids a combinatorial explosion in the number of cells — while there are more different features in higher layers, there do not have to be as many units responding to these features as in lower layers since higher-layer units have bigger receptive fields and respond to a greater range of scales.

This hierarchical buildup of invariance and feature specificity greatly reduces the overall number of cells required to represent additional objects in the model: The first layer contains a little more than one millions cells (160×160 pixels, at four orientations and 12 scales

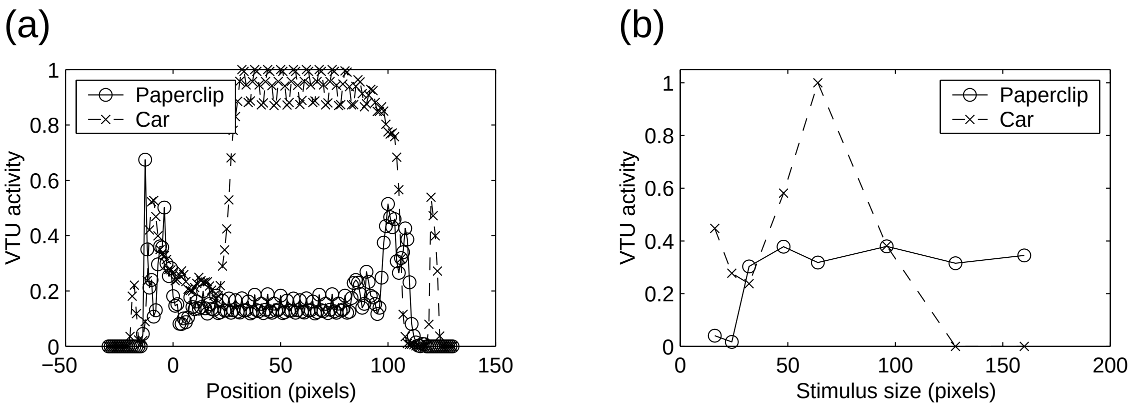

Figure 12: Responses of VTUs to stimuli of different classes at varying positions and sizes. **(a)** Responses of a VTU tuned to a centered car prototype to its preferred stimulus and a paperclip stimulus at varying positions on the horizontal midline within the receptive field of its C2 afferents, including positions partly or completely beyond the receptive field (smallest and largest position values). Size of stimuli 64×64 pixels, image size 100×100 pixels, 40 C2 afferents to the VTU, $\sigma = 0.2$. **(b)** Responses of a VTU tuned to a car prototype (size 64×64 pixels) to its preferred stimulus and a paperclip stimulus at varying sizes. Stimuli centered within a 160×160 pixel image, 40 C2 afferents to the VTU, $\sigma = 0.2$.

each). The crucial observation is that if additional objects are to be recognized irrespective of scale and position, the addition of only one unit, in the top layer, with connections to the (256) C2 units, is required.

However, we have shown that discretization of the input image and hierarchical buildup of features yield a model response that is not completely independent of stimulus position: A "feature loss" may occur if a stimulus is shifted away from its original position. This is not too surprising since there are more possible stimulus positions than model units. We presented a quantitative analysis of the changes occurring in HMAX output and recognition performance in terms of stimulus position and parameter settings. Most significantly, there is no feature loss if a stimulus is moved to an equivalent position with respect to the discrete organization of the C1 and S2 model units. Equivalent positions are separated by a distance that is the least common multiple of the *(poolRange/c1Overlap)* ratio for the different filter bands used.

It should be noted that the features affected by feature loss are the *composite* features generated at the level of S2 units, not the simple C1 features – if only four output (S2/C2) unit types with the same feature sensitivity as C1 units are used, no modulations of activity with stimulus position are observed (as in the "10 feature" version of HMAX in [15]). As opposed to a composite feature, a simple feature will of course never go undetected by the C1 units as long as there is some overlap between them (Figure 5).

The basic mechanisms responsible for variations in HMAX output due to changes in stimulus position are independent of the particular stimuli used. We showed that feature loss occurs for cars as well as for paperclips, and that it follows the same principles in both cases. However, we found that while HMAX performs very well at size-invariant recognition of paperclips, its performance is much worse for cars. This discrepancy relates to the more limited availability of model units with different receptive field sizes, as compared to units at different positions, in the current HMAX model, as well as to the particular feature dictionary used in HMAX. The model's high performance for recognition of paperclips derives from the fact that its feature detectors are well-matched to this object class — they closely resemble actual paperclip features, and the size of the detectors activated most by a stimulus is in good agreement with stimulus size. Especially the latter is important for the model to take advantage of its different filter sizes for detection of features regardless of their size. This correspondence between stimulus size and size of the most active feature detectors is not given for cars; hence the model's low performance at size-invariant recognition for this object class.

These findings show that invariant recognition performance can differ for different stimuli depending on how well the object recognition system's features match the stimuli. Our simulation results suggest that invariance ranges of a particular neuron might depend on the shape of the stimuli used to probe it. This is especially relevant as most experimental studies only test scale or position invariance using a single object, which in general is not identical to the object the neuron is actually tuned to [7] (the "preferred" object). Thus, invariance ranges calculated based on the responses to just one object are possibly different from the actual values that would be obtained with the preferred object (assuming that the neuron receives input from neurons in lower areas tuned to the features relevant to the preferred object [14]).

It is important to note that for any changes in HMAX response due to changes in stimulus position or size, drops in VTU output are not necessarily accompanied by drops in recognition performance. Recognition performance depends on relative VTU activity and thus on number and characteristic features of distractors, and it can remain high even for drastically reduced absolute VTU activity (see [14]). Thus, size- and position-invariant recognition of objects does not require a model response that is independent of stimulus size and position. Furthermore, as Figure 6 shows, the magnitude of fluctuations can be controlled by varying the parameters that control pooling range and receptive field overlap in the hierarchy. It will be interesting to examine whether one can derive constraints for these variables from the physiology literature. For instance, recent results on the receptive field profiles of IT neurons [12] suggest that the majority of IT neurons have Gaussian, *i.e.,* unimodal, profiles. In the framework of our model this corresponds to a periodicity of VTU response with a wavelength which is greater than the size of the receptive field. This would argue (Eq. 1) for either low values of C1 overlap or high values of the pooling range. The observations that the average linear extent of a complex cell receptive field is 1.5-2 times that of simple cells [4] (greater than the pooling ranges in the standard version of the model) is compatible with this requirement. Clearly, more detailed data on the shape tuning of neurons in intermediate visual areas, such as V4, are needed to quantitatively test this hypothesis.

Acknowledgements

We thank Christian Shelton for providing the car morphing software, and Tommy Poggio and Jim DiCarlo for inspiring discussions.

References

[1] Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312–316.

[2] Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154.

[3] Hubel, D. and Wiesel, T. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophys.* **28**, 229–289.

[4] Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Phys.* **195**, 215–243.

[5] Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophys.* **73**, 218–26.

[6] Knoblich, U., Freedman, D., and Riesenhuber, M. (2002). Categorization in IT and PFC: Model and Experiments. AI Memo 2002-007, CBCL Memo 216, MIT AI Lab and CBCL, Cambridge, MA.

[7] Knoblich, U. and Riesenhuber, M. (2002). Stimulus simplification and object representation: A modeling study. AI Memo 2002-004, CBCL Memo 215, MIT AI Lab and CBCL, Cambridge, MA.

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

[9] Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Ann. Rev. Neurosci* **19**, 577–621.

[10] Martinez, L. M. and Alonso, J.-M. (2001). Construction of complex receptive fields in cat primary visual cortex. *Neuron* **32**, 515–525.

[11] Missal, M., Vogels, R., and Orban, G. (1997). Responses of macaque inferior temporal neurons to overlapping shapes. *Cereb. Cortex* **7**, 758–767.

[12] op de Beeck, H. and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* **426**, 505–518.

[13] Perret, D. and Oram, M. (1993). Neurophysiology of shape processing. *Image Vision Comput.* **11**, 317–333.

[14] Riesenhuber, M. and Poggio, T. (1999). Are cortical models really bound by the "Binding Problem"? *Neuron* **24**, 87–93.

[15] Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**(11), 1019–1025.

[16] Riesenhuber, M. and Poggio, T. (1999). A note on object class representation and categorical perception. AI Memo 1679, CBCL Paper 183, MIT AI Lab and CBCL, Cambridge, MA.

[17] Riesenhuber, M. and Poggio, T. (2000). The individual is nothing, the class everything: Psychophysics and modeling of recognition in object classes. AI Memo 1682, CBCL Paper 185, MIT AI Lab and CBCL, Cambridge, MA.

[18] Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci. Supp.* **3**, 1199–1204.

[19] Shelton, C. (1996). *Three-Dimensional Correspondence*. Master's thesis, MIT, Cambridge, MA.

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