Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2004-001
AIM-2004-004
CBCL-235

January 14, 2004

On the difficulty of feature-based attentional modulations in visual object recognition: A modeling study.

Robert Schneider and Maximilian Riesenhuber

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

Abstract

Numerous psychophysical experiments have shown an important role for attentional modulations in vision. Behaviorally, allocation of attention can improve performance in object detection and recognition tasks. At the neural level, attention increases firing rates of neurons in visual cortex whose preferred stimulus is currently attended to. However, it is not yet known how these two phenomena are linked, *i.e.*, how the visual system could be "tuned" in a task-dependent fashion to improve task performance. To answer this question, we performed simulations with the HMAX model of object recognition in cortex [45]. We modulated firing rates of model neurons in accordance with experimental results about effects of feature-based attention on single neurons and measured changes in the model's performance in a variety of object recognition tasks. It turned out that recognition performance could only be improved under very limited circumstances and that attentional influences on the process of object recognition *per se* tend to display a lack of specificity or raise false alarm rates. These observations lead us to postulate a new role for the observed attention-related neural response modulations.

Copyright © Massachusetts Institute of Technology, 2004

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

At any given time, much more information enters the visual system via the retina than is actually behaviorally relevant. A first selection mechanism is already provided by the fovea, endowing stimuli at the center of the visual field with much higher acuity and disproportionately large representation. However, more sophisticated mechanisms are needed to allow an animal to focus in a more abstract way, on what is important in a given situation or for a certain task. Attention is designed to accomplish just this, to both avoid being overwhelmed by vast amounts of visual input and to find the currently relevant elements in it.

A large body of literature, both theoretical and experimental, has dealt with the phenomenon of attention in recent years and explored its effects on subjects' performance in behavioral tasks as well as on neural activity (for reviews, see [9, 25, 59]). At first, it might seem difficult to "measure" attention, and it can of course never be determined with absolute certainty what a human being or a monkey is actually attending to at any given moment. Deployment of attention can, however, be controlled by requiring a subject to perform significantly above chance in a behavioral task, *e.g.*, judging orientation differences between bar stimuli or detecting a cued picture. Then, changes in behavioral or neural response to the same stimulus when it is relevant or irrelevant to the task (*i.e.*, , as can be assumed, attended or unattended, respectively) can be attributed to attentional effects.

Such experiments indicate, for example, that human observers can increase their performance at discriminating visual stimuli according to their orientation or spatial frequency when they direct attention to the respective stimulus dimension [52] and that focusing attention on a color stimulus is equivalent to an increase in its color saliency [7]. Furthermore, experiments with rapid serial visual presentations (RSVP) have shown that subjects perform better at detecting a given target object in a rapid stream of images when they are informed about what to look for, rather than when they have to judge after the presentation whether a certain image has been shown in it or not [41]. Performance improves further with more specific cuing information, *i.e.*, knowing the basic-level category of a target object (*e.g.*, "dog") in advance facilitates target detection more than merely knowing the superordinate category it belongs to (*e.g.*, "animal") [21]. It might be asked if these results for more complex stimuli are also caused by attention directed to certain stimulus features about which the subject is informed in advance.

Single-neuron studies, on the other hand, have established that attention modulates responses of neurons in visual cortex [39, 43] such that neurons whose preferred stimulus is attended to respond more strongly while the activity of neurons coding for nonattended stimuli is attenuated. Moreover, an attended stimulus determines a neuron's response even in the presence of other stimuli. That is, a stimulus that by itself elicits only a weak response will do so even if a more optimal stimulus for the neuron under study is present in its receptive field, as long as the nonpreferred stimulus is attended to, and vice versa for preferred stimuli. These effects have been described mostly for extrastriate areas of the ventral visual stream (which is considered crucial for the processes of object recognition), namely V2, V4 [43] and inferotemporal cortex (IT) [10], but they have also been found in primary visual cortex [49] and in the dorsal stream, usually associated with processing motion information [60].

Thus far, both physiology and psychophysics suggest that attention increases the perceptual saliency of stimuli. However, it has not yet been examined systematically whether the neuronal firing rate changes observed in physiological experiments with feature attention actually influence the processes of object recognition, and whether they can explain the increases in discrimination and recognition performance observed in behavioral experiments. Modeling studies provide good opportunities to test such hypotheses about brain function. They can yield constraints for further theories and show what might work in the brain and what might not, in a rigorously defined and well-understood framework. Some modeling has already been done in the field of attention, but usually rather with a focus on the neural mechanisms alone, without regard to object recognition [62, 63]. On the other hand, the HMAX model of object recognition in visual cortex [45] (see Figure 1) has been explicitly designed to model this task, but so far has not been used to model attention. Its output model units, the view-tuned units (VTUs) at the top of the hierarchy in Figure 1, show shape tuning and invariance properties with respect to changes in stimulus size and position which are in quantitative agreement with properties of neurons found in inferotemporal cortex by Logothetis *et al.* [28]. This is achieved by a hierarchical succession of layers of model units with increasingly complex feature preferences and increasing receptive field sizes. Model units in successive layers use either one of two different mechanisms of pooling over afferent units: a "template match" mechanism generates feature specificity (by combining inputs from different simple features), while response invariance to translation and scaling is increased by a MAX-like pooling mechanism that picks out the activity of the strongest input among units tuned to the same features at different positions and scales. The model is comparatively simple in its design, and it allows quantitative predictions that can be tested experimentally.

HMAX has turned out to account, at least to some degree, for a number of crucial properties of information processing in the ventral visual stream of humans

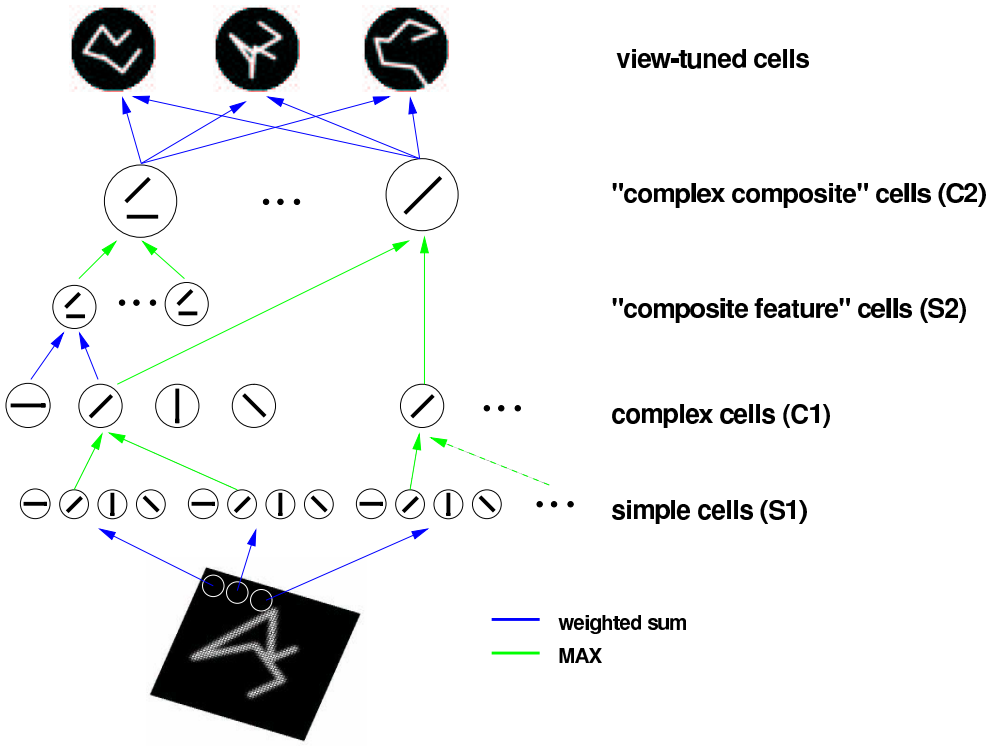

Figure 1: Schematic of the HMAX model. See Methods.

and macaques (see [27, 44, 46–48]), including view-tuned representation of three-dimensional objects [28], response to mirror images [29], performance in clutter [37], and object categorization [15]. So far, processing in HMAX is purely feedforward, without the feedback connections usually considered to be the mediators of attentional modulations, whose sources are assumed in areas such as prefrontal cortex or the frontal eye fields [25]. In this study, we investigate the effects of introducing attentional top-down modulations of model unit activity in HMAX, to learn about the possible role of such modulations in the processes of object recognition.

Before turning to the implementation of models of attentional effects in HMAX, we discuss possible mechanisms of attentional modulation, in order to situate our modeling efforts within the relevant literature and to delineate constraints for modeling attention on the basis of experimental evidence. As a foundation for the simulations employing attentional modulations, we also examine the behavior of HMAX without attentional effects for stimuli at low contrasts and in cluttered displays, the circumstances in which attention would be expected to aid object recognition most. Different models of attentional effects on neuronal responses are then investigated with respect to their potential of increasing object recognition performance in HMAX under such circumstances. Based on the results of these simulations, we finally attempt to formulate a possible role for attention in object recognition.

1.1 Spatial and featural attention

Attention can be directed to a spatial location (*spatial attention*) or to a certain object (*object* or *feature attention*), independent of where it might appear in the visual field. While the underlying neural mechanisms of these two kinds of attention are probably similar (see section 1.3), they are distinct phenomena that can be discerned, for example, by the different patterns they elicit in EEG recordings [19, 20]. In our study, we focus on modeling feature attention, without prior knowledge about the location where a target will appear, as it is employed in visual search or RSVP experiments. Spatial attention may be modeled in a relatively straightforward fashion in HMAX, for example, by only or preferentially considering visual input from an attended location which might be determined by advance cuing or based on especially salient features in the image [65].

However, it is far less clear how attention to features might be implemented in the brain or in a model. How can the visual system be "tuned" if only an abstract cue is given, *i.e.*, how can elementary visual features be selected for preferred processing if the actual visual appearance of the target object is unknown? Moreover, even in the ideal case if the target's identity is known exactly, how can this be translated into "tuning" of complex features along the processing hierarchy of the ventral visual stream? To our knowledge, no previous modeling efforts have addressed this problem; simulations usually featured populations of model units that

were defined to code for certain target objects relevant for the model task and thus were easily identifiable as recipients of attentional activity increases [26, 62, 63]. Possible solutions to this problem and their viability in the context of object recognition are the focus of this study.

1.2 Early and late selection

While the aforementioned findings of neuronal activity modulations caused by attention and of detection performance improvements for cued stimuli can be considered established [22, 25], there has long been a major controversy between two conflicting theories about how attention acts on perception, known as the early and late selection hypotheses of attention. The question here is whether attention acts *before* object recognition occurs, operating on representations of features, not of whole objects and effectively "tuning" the organism to expected stimuli in advance [8], or whether attention acts on neuronal representations of stimuli that have already been processed up to the level of recognition of individual objects [21].

An important constraint to our work comes from results about the time course of visual processing in humans and monkeys. ERP recordings suggest that recognition or categorization of objects, even in complex real-world scenes, can be accomplished in as little as 150 ms [14, 58], which is on the order of the latency of the visual signal in inferotemporal cortex (about 100 – 130 ms [50, 56]). This leaves little or no time for elaborate feedback processes. Consequently, if attention influences object recognition *per se*, it can only do so by means of pre-stimulation "tuning" processes that allow subsequent recognition in a single feedforward pass of information through the visual pathway.

Since this study focuses on the possible effects of feature attention on object recognition performance, it by definition deals with early selection mechanisms of attention. We thus model top-down influences on neuronal representations of features that occur before recognition of individual objects is accomplished or even before visual stimulation begins. As mentioned above, this entails that features have to be selected for preferred processing in advance, which poses a problem if the cue about the target stimulus is a rather general one. For this case, solutions have to be devised and tested with respect to their effectiveness and specificity for a target stimulus.

1.3 Physiology of attentional modulations

Earlier work has hypothesized that the physiological effect of attention on a neuron might be to shrink its receptive field around the attended stimulus [38] or to sharpen its tuning curve [17]. However, a receptive field remapping, possibly implemented by way of shifter circuits [1], would likely only be appropriate for spatial attention, where the locus of the object of interest is known in advance, and not for early selection mechanisms of feature attention. A sharpening of tuning curves, on the other hand, is not observed if cells' responses are corrected for baseline firing [33].

More likely mechanisms are rapid changes in synaptic weights that selectively increase input gain for the neuron population responding to the attended stimulus, as assumed in the Biased Competition model [43], or direct excitatory top-down input to cells coding for the attended stimulus, a mechanism often used in modeling studies [62, 63], causing increased activity or probably disinhibition in those cells [35]. It has been discussed to some extent whether the result of attentional modulation on a neuron's firing rate is better described as multiplicative, increasing high firing rates more than low rates [33], or by a contrast-gain model, which assumes that attention causes a leftward shift of a neuron's contrast-response function, yielding the most significant increases in firing rate when the neuron's activity is close to baseline [42]. Both have been observed experimentally—in different paradigms, however. The two viewpoints can be reconciled by assuming that a neuron's tuning curve, *i.e.*, the function describing its responses to *different* stimuli at the *same* contrast, is enhanced in a multiplicative way by attention, such that responses to more preferred stimuli increase more (in absolute terms), while the neuron's contrast response function, *i.e.*, the function describing its response to a *given* stimulus at *varying* contrast, is shifted to the left, leading to more prominent activity increases for stimuli at low and intermediate contrasts [42].

There is broad consensus in the literature that there are not only increases in firing rates of cells whose preferred stimulus is attended, but also suppression of cells that code for nonattended stimuli, at least in areas V2, V4 and IT [10, 11, 39, 59]. This also fits the picture of attention as a means to increase stimulus salience (or, more specifically, effective contrast) selectively. However, these studies usually report attenuated firing rates in those cells whose preferred stimulus is *present* in the image but not being attended. It is not clear whether this extends to cells with other stimulus preferences – a question calling for further electrophysiological investigations. However, it seems relatively unlikely that *all* cells in a visual cortical area except those coding for attended stimuli would be actively suppressed. On the other hand, in an early selection, pre-recognition paradigm, the question arises which features should be selected, this time for suppression. This problem seems especially difficult, if not impossible to solve for distractor stimuli (*i.e.*, other stimuli appearing together with the target object). Usually, no information at all is available about their characteristic features in advance.

In our simulations, we attempt to cover a broad range of possible implementations of attentional modulation

in HMAX, examining the effects of multiplicative and contrast-gain modulations with and without concurrent suppression of other model units. Before turning to the details of implementation and possible solutions to the problem of selecting model units for attentional modulation, we first introduce the model in its basic feedforward configuration.

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 [45]. Briefly, input images (we used greyscale images 128×128 or 160×160 pixels in size) 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 [45], 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 [45]), 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.

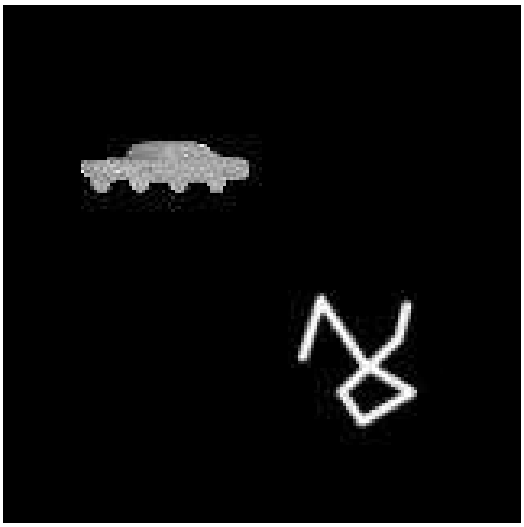

Figure 2: Examples of the car and paperclip stimuli used.

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 a mean value, called *s2Target*, of 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 responds 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 units) 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 respond 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 in turn 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.* [28]. A VTU is tuned to a stimulus by taking the activities of the 256 C2 units in response to that stimulus as the center $\vec{w}$ of a 256-dimensional Gaussian response function as given in the following equation:

$$y = \exp\left(-\frac{||\vec{x}-\vec{w}||^2}{2\sigma^2}\right) \quad (1)$$

This yields a maximal response $y = 1$ for a VTU in case the C2 activation pattern $\vec{x}$ exactly matches the C2 activation pattern evoked by the training stimulus. The σ value of a VTU can be used as an additional parameter specifying response properties of a VTU. A smaller σ value yields more specific tuning since the resultant Gaussian has a narrower half-maximum width. 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 [44]. Apart from simulations where we took all 256 C2 units as afferents to each VTU, we ran simulations using only the 40 or 100 most active C2 units as afferents to a VTU.

2.2 Stimuli

Cars We used the "8 car system" described in [47], created using a 3D multidimensional morphing system [55]. The system consisted 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 induced a similarity metric: any two prototypes were spaced 10 morph steps apart, and a morph at morph distance, *e.g.*, 3 from a prototype was 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).

Paperclips 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 [28], and in [45]. 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 located 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.

Faces For simulations with a VTU population code (see section 2.6), we used a dataset of 200 frontal-view face stimuli provided by Thomas Vetter [6]. For 10 of these faces, analogous to the car stimuli, morphed faces were available that connected any two of them by a morph line divided into 10 intervals. The remaining 190 face stimuli were unrelated to the 10 morphable faces.

Contrast The contrast measure we used was 0% when all pixels in the image had the same (background) value, and 100% when the maximum deviation of a pixel value from the background value was as large as the background value itself. This was achieved by first setting the mean pixel value $\overline{img}$ of the image matrix img to zero and then applying the following operation to each image pixel $img(i,j)$:

$$img'(i,j) = -\frac{c \cdot bg}{\min(img)} \cdot img(i,j) + bg \quad (2)$$

with c denoting the desired contrast value (from 0 to 1), bg the background pixel value, which was always set to 128, and $\min$ the minimum operation. This procedure yielded absolute pixel values ranging from 0 to a maximum of about 300 for paperclip stimuli; maximal pixel values for cars and faces were usually well below that (around 200).

Stimulus presentations For simulations examining contrast-invariant recognition, the 260 cars and 75 paperclips (each at a size of 64×64 pixels) were presented individually at the center of a 128×128 or 160×160 pixel image, at contrasts from 0% to 100% in steps of 10%. Otherwise, as a simple model of clutter, two stimuli from the same object class were presented in conjunction, each again sized 64×64 pixels, at positions $(50|50)$ and $(110|110)$ within a 160×160 pixel image, where, for example, $(50|50)$ denoted that the center of the stimulus was positioned 50 pixels to the right and 50 pixels downward from the upper left corner of the image. The positions were chosen such that the same stimulus elicited the same response in all C2 units at both positions and that the target stimulus to be recognized appeared at the same position as during VTU training, to exclude possible effects of stimulus position on recognition performance [53]. The two stimuli were also spaced far enough apart to exclude interactions within a single S1 unit receptive field. One of the two stimuli was always a target stimulus to which a VTU had been trained previously (presented in isolation at position $(50|50)$ and size 64×64 pixels, at 100% contrast), *i.e.*, one of the 8 car prototypes or one of the 15 selected target paperclips. The other stimulus could be, for a car target, any morphed car stimulus from one of the 7 morph lines leading away from that particular target, including the 7 other prototypes. For each paperclip target, any of the 60 randomly selected distractor paperclips was presented as the second stimulus in the display. Target and distractor contrast were varied independently; target contrast was usually below its training value of 100%, in accordance with the

experimental finding that attention changes neural responses only for low and intermediate contrast levels [42].

2.3 Recognition tasks

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

2.3.1 "Most Active VTU" paradigm

In the first paradigm, a target was said to be recognized if the VTU tuned to it responded more strongly in response to a display containing this target than all other VTUs tuned to other members of the same stimulus set (*i.e.*, the 7 other car VTUs in case of cars, or the 14 other paperclip VTUs in case of paperclips). This paradigm corresponded 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.

If stimuli were presented in isolation, recognition performance in the Most Active VTU paradigm reached 100% if for all presentations of prototypes the VTU tuned to the respective prototype was more active than any other VTU. If displays with a target and a distractor were used, each VTU had to be maximally active for any presentation of its preferred stimulus, regardless of the second stimulus present in the display, in order to reach 100% recognition performance. Thus, for perfect recognition of paperclips in clutter, each of the 15 VTUs tuned to paperclips had to respond more strongly than all other VTUs in response to all 60 presentations of its preferred paperclip (one presentation for each distractor paperclip). For car stimuli, the similarity metric over the stimuli could be used to group distractors according to their morph distance to the target and thus plot recognition performance as a function of target-distractor similarity. 100% recognition performance for a given morph distance between target and distractor in this paradigm was reached if each VTU was the most active VTU for all presentations of its preferred stimulus in conjunction with a distractor car at this morph distance from the target. For each prototype and any given morph distance, there were 7 such distractor cars, corresponding to the 7 morph lines leading away from each prototype to the 7 other prototypes.

Chance performance in the Most Active VTU paradigm was always the inverse of the number of VTUs (*i.e.*, prototypes), since for any given stimulus presentation, the probability for a VTU to be more active than all other VTUs is 1 over the number of VTUs. This resulted in chance levels of 12.5% for cars and 6.7% for paperclips.

2.3.2 "Stimulus Comparison" paradigm

Alternatively, a target stimulus could be considered recognized if the VTU tuned to it responded more strongly to a display that contained this target than to another display without it. This corresponded to a two-alternative forced-choice behavioral task in which subjects are presented with a sample stimulus (chosen from a fixed set of targets) and two choice stimuli, only one of which contains the sample target, and subjects then have to indicate which of the two choice stimuli is identical to the sample. We will refer to this paradigm as the "Stimulus Comparison" paradigm.

When only one stimulus was presented at a time, as in the simulations regarding the contrast-response behavior of HMAX, responses of individual VTUs to their target stimulus were compared with their responses to each of the 60 distractor paperclips if the target was a paperclip (as in [45]), or with their responses to all morphs on morph lines leading away from the target if the target was a car stimulus. 100% recognition performance in this paradigm then entailed that all VTUs always responded more strongly to their respective target than to any of these other stimuli. For two-stimulus displays of paperclips, all VTUs were required to respond more strongly to all 60 displays of their target paperclip with a distractor paperclip than to any of the other (14×60) displays not containing their respective target in order to achieve 100% recognition performance. For two-stimulus displays of cars, comparisons were only made between displays in which the morph distances between target and distractor were identical, to again enable quantification of recognition performance depending on the similarity of a distractor to a target. That is, perfect recognition performance in the Stimulus Comparison paradigm for cluttered car stimuli and a given morph distance between target and distractor was reached if all car VTUs responded more strongly to all 7 displays of their target prototype car with a distractor at the selected morph distance than to any display of another prototype car with a distractor at that morph distance.

In the Stimulus Comparison paradigm, chance level was reached at 50% performance. This value entailed that the response of a VTU to a display that did not contain its preferred stimulus was equally likely to be stronger or weaker than its response to a display containing its target stimulus. Such a VTU was thus unable to differentiate between displays that contained its preferred stimulus and displays that did not.

2.3.3 ROC analysis

Recognition performance in the Stimulus Comparison paradigm was plotted in the form of ROC curves (*Receiver Operating Characteristic*). An ROC curve evaluates the capability of a signal detection system (here, of VTUs) to differentiate between different types of signals

(here, target and other stimuli), regardless of a specific threshold an observer might use to judge which signal was detected by the system. To generate an ROC, all responses of a VTU for displays containing its preferred stimulus ("target displays") and for the other displays used for comparison in the Stimulus Comparison paradigm ("nontarget displays") were considered. The difference between the maximum and the minimum responses of the VTU to this set of stimuli was divided into a given number of intervals (we mostly used 30). For each activity value at the boundary of two intervals ("threshold"), the percentage of the VTU's responses to target displays that were above this threshold was calculated, yielding the "hit rate" for this threshold value, as well as the percentage of its responses to nontarget displays that were above the threshold value, the so-called "false alarm rate". Plotting hit rates against false alarm rates for all threshold values then yielded the ROC curve. As is true for all ROCs, it always contained the points $(0\%|0\%)$ (first figure: false alarm rate, the abscissa value; second figure: hit rate, the ordinate value) and $(100\%|100\%)$ for threshold values above the maximum or below the minimum VTU response, respectively. Perfect recognition performance in the Stimulus Comparison paradigm (*i.e.*, a VTU always responded more strongly to displays containing its preferred stimulus than to any other display) lead to an ROC curve that contained the point $(0\%|100\%)$, *i.e.*, there was at least one threshold value such that all responses of the VTU to target displays were greater and all its responses to nontarget displays were smaller than this threshold value. Chance performance, on the other hand, yielded a linear ROC through the points $(0\%|0\%)$ and $(100\%|100\%)$, *i.e.*, for any given threshold value, there was an equal chance that the firing rate of the VTU in response to a target or nontarget display was higher or lower than this threshold value. ROCs in this study were always averaged across all VTUs of a stimulus class (8 for cars, 15 for paperclips).

2.4 Attentional modulations

The units whose activities were changed by attentional effects in our simulations were either C2 units or VTUs. Since these model units received input from the whole visual field and represented complex stimuli, they were most suitable for simulation of nonspatial, object-directed attention. Furthermore, C2 units and VTUs were designed to correspond to neurons in visual areas V4 and IT, respectively, where the earliest and strongest effects of feature attention are observed [34]. Since model output was interpreted in terms of recognition of objects, any modulation of neural activities before readout by definition corresponded to an early selection mechanism of attention, *i.e.*, a form of attention that influences the processes of object recognition *per se*.

2.4.1 Facilitation

We addressed the problem of selecting appropriate features for attentional modulation, in the simplest case, by simulating attention directed to a single target object (one of the 8 cars or 15 paperclips) whose visual appearance is known, and for which a dedicated VTU (a "grandmother cell", see section 2.6 for the more general population coding case) has been trained. Activity modulations were then applied to the C2 afferents of these VTUs. This corresponded to a top-down modulation of V4 cell activities by object- or view-tuned cells in inferotemporal cortex. We used VTUs with 40 or 100 C2 afferents. For VTUs connected to all 256 C2 units, modulations of their afferents' firing rates would have affected all VTUs, which would have been at odds with the idea of specifically directing attention to a certain object. We used this situation, however, to compare the effects of nonspecific activity increases with the more specific attentional effects achieved for fewer afferents per VTU.

One method to increase activity values of model units coding for attended features was to multiply them with a factor between 1.1 and 2 (section 3.3), corresponding to findings of McAdams and Maunsell [33] and Motter [39] that attention led to an increase in neuronal response gain. Another method we used was to lower the mean value of the S2 (and, in turn, C2) units' Gaussian response function, *s2Target*, such that a given C1 input into a S2 unit yielded a greater response (section 3.4). This corresponded to the leftward shift of the contrast response function of V4 neurons that has been reported as attentional effect by Reynolds *et al.* [42], yielding higher contrast sensitivity and earlier saturation in neurons whose preferred stimulus was attended to. Instead of the *s2Target* value of 1 we used in all other simulations, we selected two slightly lower values (0.945 and 0.9275) that, on average, yielded response increases that closely matched the loss of activity encountered by C2 units upon a decrease in stimulus contrast from 100% to 60% (for car stimuli). We then applied these shifts in the S2 / C2 response function selectively to the 40 afferent C2 units of the VTU tuned to a target stimulus. Since the new mean values were optimized for car stimuli at 60% contrast, we applied this boosting method only to car stimuli with the target car at this contrast level. We also made sure that no C1 unit could possibly respond at a rate higher than the respective mean of the Gaussian response function, which would have caused a lower response from the corresponding C2 unit for a higher firing rate of the C1 unit.

In the Stimulus Comparison paradigm, for each target stimulus, responses for *all* stimulus combinations (target and distractor) of a given stimulus class (cars or paperclips) were calculated with attention directed to this target stimulus, regardless of whether it was actually present in the image. This made sure that responses

of a VTU to displays with and without its preferred stimulus were comparable and permitted the analysis of false alarms, *i.e.*, erroneous recognition of a stimulus that had been cued but was not present in an image. ROC curves were then averaged over the results for all VTUs tuned to objects of a class. An analogous control for false alarms in the Most Active VTU paradigm was to apply activity increases to the C2 afferents of a VTU whose preferred stimulus was not shown in the image under consideration. The percentage of cases in which this VTU nevertheless ended up as the most active VTU among those tuned to objects from the same class could then be considered the false alarm rate in this paradigm. This control experiment is discussed in section 3.8.

More complex situations with more than one VTU or their afferents as recipients of attentional activity modulations are discussed in section 2.6.

2.4.2 Suppression

In addition to attentional activity enhancements, suppression of neurons not coding for the attended stimulus was also modeled. In some experiments, the units selected for suppression were the C2 afferents of the most active among those VTUs that did not code for the target stimulus; in other simulations, all C2 units not connected to the VTU coding for the target were suppressed. The first mechanism could, strictly speaking, not be considered an early selection mechanism, since to determine the most active nontarget VTU, a HMAX run had to be completed first. The second mechanism, on the other hand, could be applied in advance, but suppression of *all* neurons not directly relevant for the representation of a stimulus seems to be rather effortful and probably an unrealistic assumption, as mentioned in section 1.3. However, in our model, these two were the most intuitive implementations of suppressive mechanisms, and they could also be expected to be the most specific ones. If suppression can at all enhance object recognition performance, it should do so especially if the most salient neural representation of a nontarget stimulus or all nontarget representations are suppressed. In all cases, suppressive mechanisms were modeled by multiplication of activity values with a factor smaller than 1.

2.5 Alternative coding schemes

Saturation tuning. We implemented two alternative coding schemes in HMAX that were designed to improve the model's contrast invariance properties and to allow for more robust attentional activity modifications. In the first alternative coding scheme we investigated, the VTUs were tuned such that they responded maximally (*i.e.*, at a rate of 1) if all their afferents responded at or above their activity levels during training, instead of displaying reduced activity again if any afferent C2 unit responded more strongly than during presentation of the VTU's preferred stimulus at full contrast. This kind of encoding, which we will refer to as "saturation tuning", provided for an effectively sigmoidal VTU response function and saturation of VTUs. It was achieved by setting to zero the exponent of a VTU's Gaussian response function whenever all of its afferents were either as active as or more active than during training, as can be seen from the Saturation Tuning response function:

$$y = \exp\left(-\frac{\sum_i [\min(x_i - w_i, 0)]^2}{2\sigma^2}\right) \tag{3}$$

where i runs over the VTU's afferent C2 units.

Relative rate tuning. Another alternative coding scheme was to have that VTU respond most strongly whose C2 afferents were most active, instead of the VTU whose preferred C2 activity matched the actual C2 activation pattern best. This was achieved by a VTU tuning similar to the S2 / C2 units' tuning to their C1 afferents: the same weight value w (*i.e.*, mean value of a one-dimensional Gaussian response function) was used for *all* afferent units, and it was equal to or greater than the maximum possible response of any afferent unit, such that a VTU would respond maximally if all its afferents responded at their maximum rate. This relation is given in the following formula:

$$y = \exp\left(-\frac{\sum_i (x_i - w)^2}{2\sigma^2}\right) \tag{4}$$

with the sum running over all C2 afferents again.

This means that the most active VTU was determined by which set of afferents responded most strongly, even if absolute activity levels of C2 units were very low, *e.g.*, due to low stimulus contrast. Specificity, on the other hand, was only conferred through the selection of a VTU's afferents, not through matching their activity pattern to its training value. We will refer to this coding scheme as "relative rate tuning".

For both alternative coding schemes, recognition performance was examined as in the experiments with standard HMAX encoding, with cars and paperclips as stimuli and a target and a distractor in each presentation, using both Most Active VTU paradigm and Stimulus Comparison paradigm. Multiplicative activity increases were used to model attentional effects.

2.6 Population coding

To study the more general case in which stimuli are represented in the brain by the activities of populations of neurons rather than of single neurons (see section 3.7), we performed simulations where stimuli were encoded by activation patterns over several VTUs. For these experiments, we used a dataset of face stimuli [6] that had a number of advantages over the car and paperclip stimuli in this context. For ten faces, morphed stimuli

Figure 3: Example of the face stimuli used for population code experiments.

were available that smoothly changed any one of these faces into any other of them, such that morph similarity of a distractor to a target could be added as an extra parameter in our simulations, as was the case for car stimuli. However, unlike in our car dataset, 190 more faces unrelated to the 10 morphable faces were available. We were thus able to tune a VTU to each of these 190 faces and calculate the response of this population of face-tuned VTUs to two-stimulus presentations consisting of one of the 10 morphable face prototypes as target and one morphed face as distractor. It is important here that none of the units in the VTU population was tuned to any of the morphable faces we presented as test stimuli. This allowed us to model the response of a population of neurons selective for a certain stimulus class to new members of the same object class, *i.e.*, to test generalization. Such populations have been described in temporal cortex [13, 61, 66, 67]. All face stimuli, as was the case for cars and paperclips, were 64×64 pixels in size and presented within an image of size 160×160 pixels.

We read out the response of the VTU population by means of a second level of VTUs which were tuned to the activation pattern over the VTU population when one of the target stimuli (the 10 morphable face prototypes) was presented to the model in isolation. That is, a second-level VTU responded at a maximum rate of 1 when its afferent VTUs in the population (we selected the 10, 40 or 100 population VTUs that were most active in response to the target stimulus) displayed the same activation pattern as during presentation of the target stimulus alone and at full contrast. The second-level VTUs were not designed as models of certain neurons in the brain, but rather used as a simple method to evaluate population responses. In a population code, a given stimulus is considered recognized if neural activity across the population matches the reference activity pattern elicited by this stimulus closely enough. In our model, the response of a second level of VTUs could be used as a convenient measure of the similarity of two activity patterns of the VTU population. With the second-level VTUs, we could use essentially the same means of quantifying recognition performance as for the single VTU coding scheme. Recognition was either considered accomplished if the second-level VTU tuned to the target stimulus was the most active second-level VTU overall (*i.e.*, if the VTU population response resembled the response to the target stimulus more than it resembled the response to any other face prototype; Most Active VTU paradigm) or if this second-level VTU responded more strongly to a display containing its target than to a display without its target (*i.e.*, the VTU population reliably distinguished between stimulus displays containing different target stimuli; Stimulus Comparison paradigm). As in previous sections, comparisons in this paradigm were made between responses to all displays containing a given target and responses to all other presentations, grouped according to the morph distance between the two stimuli in the displays.

A fundamental problem associated with task-dependent tuning in a processing hierarchy is how to translate modulatory signals at higher levels into modulations of units at lower levels. Attentional modulations in this population coding scheme were applied to C2 units or VTUs. Target objects were the 10 prototype faces to which the second-level VTUs had been trained. To model attention directed to one of these targets, either all population VTUs connected with the corresponding second-level VTU were enhanced in their activity (by multiplication) or a selected number of their C2 afferents. This selection of C2 units could either simply consist of *all* C2 afferents of these VTUs or only of those among them that did not at the same time project to other VTUs in the population as well. This was to test different possible solutions—with different degrees of specificity—to the problem of selecting neurons for attentional activity enhancements. Again, in the Stimulus Comparison paradigm, only responses with attention directed to the same target object were compared, regardless of whether this object actually appeared in a display, to make sure that correct values for false alarm rates were obtained.

Suppression of units not coding for the current target stimulus was also tested with population coding. Either all VTUs from the population that did not project to the second-level VTU which coded for the target were suppressed, or certain C2 units—either *all* C2 units not affected by attentional activity enhancement, or the C2 afferents of those population VTUs that were connected to the most active unit among the second-level VTUs that did not code for the target stimulus. Thus, the selection of suppressed C2 units was done in an analogous fashion as in the experiments using "grandmother cell" encoding based on a single VTU.

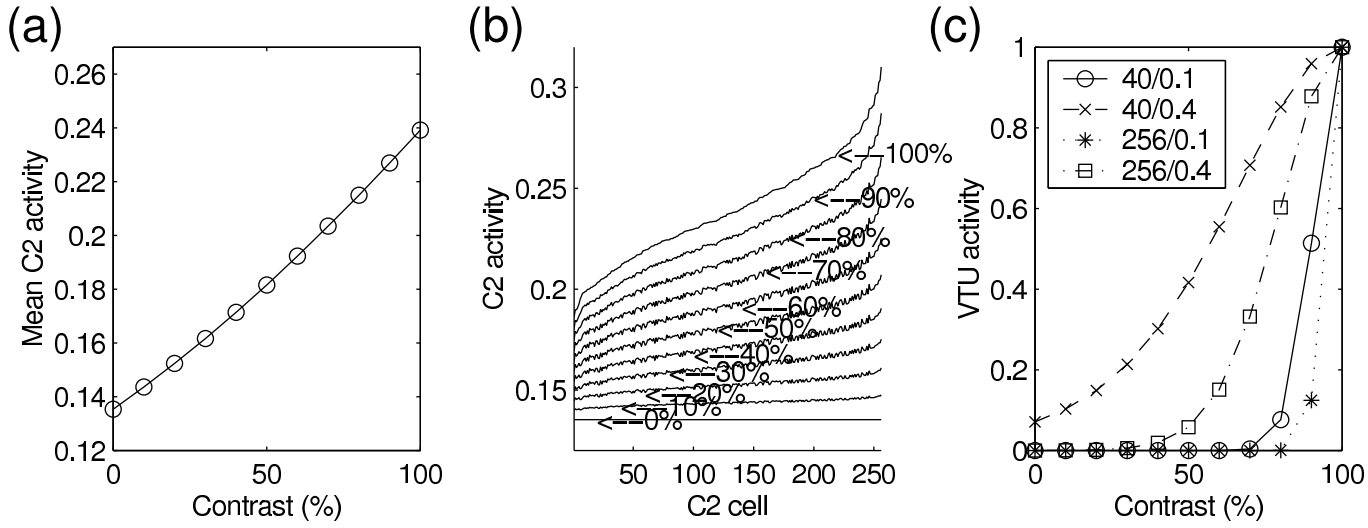

Figure 4: Contrast response behavior of C2 units and VTUs, plotted for car prototype 3 at varying levels of contrast. **(a)** Mean contrast response curve of all 256 C2 units. **(b)** Responses of all C2 units for varying contrast. Each line represents the responses of all 256 C2 units to car prototype 3 for a given contrast value, indicated by the values at the end of the arrows. Model units are sorted according to their response strength for contrast 1.0, *i.e.*, the response of any given C2 unit is always plotted at the same x-axis position for all contrasts. **(c)** Responses of the VTU tuned to car prototype 3 to its preferred stimulus at different contrasts. Values in the legend indicate the number of afferents for this VTU (40 or 256) and its σ value (0.1 or 0.4).

3 Results

3.1 Effects of contrast changes on recognition performance

Figure 4 shows the contrast-response behavior of C2 units and of a typical VTU for a car stimulus. Response of C2 model units, the output units of the HMAX core module, is obviously dependent on stimulus contrast, in accordance with properties of V4 neurons, which the C2 units are designed to model [42]. The average C2 contrast response for this stimulus (Figure 4 a) was nearly, although not perfectly, linear within the range of contrasts used in our experiments. In physiological terms, we were thus operating within the dynamic range of the C2 model units. However, clearly there were different slopes of the different C2 units' contrast response curves, corresponding to cells displaying stronger or weaker activation for a given stimulus at full contrast (see Figure 4 b). Of course, all C2 units had the same response function, a Gaussian centered at 1 and with a standard deviation of 1, as mentioned in section 2.1. In response to a given stimulus, however, different units exhibited different response strengths, and since all C2 units had the same baseline response for zero contrast, the slopes of their response curves drawn as a function of this particular stimulus' contrast varied, depending on the stimulus. To avoid confusion, we will call the function describing a C2 unit's response in relation to the contrast of a certain stimulus the "contrast response curve", while the basic and for all C2 units identical function describing C2 output in relation to input from afferent C1 units will be called the "response function".

Since the view-tuned units were tuned to a certain activity pattern of all or some C2 units, their activities also changed with changing stimulus contrast. The roughly linear C2 contrast response curve gave rise to a "sigmoidal" VTU activity profile for different contrast levels (see Figure 4 c). Strictly speaking, the VTU response curve was a Gaussian, not sigmoidal; however, since we were not interested in the saturating regime the Gaussian response was a good model for the sigmoidal response found in the experiment (see section 3.6.1). VTU response curves were steeper for greater numbers of afferent C2 units and smaller VTU σ values, since these parameter settings provided for a more specific VTU tuning to a certain C2 activity pattern.

Figure 5 shows recognition performance in the Most Active VTU paradigm and ROCs for cars and paperclips at different contrasts. Obviously, object recognition in HMAX is not contrast-invariant; most notably for cars, performance for contrasts below the training value quickly dropped to very low levels in both paradigms (a, b). Even limiting the number of a VTU's afferents to the 40 C2 units that responded best to its preferred stimulus did not improve performance here. However, for paperclip stimuli, recognition performance in HMAX at low contrasts was significantly better than for car stimuli, at least for 40 C2 afferents per VTU (see Figure 5 a, c). Thus, the representations of

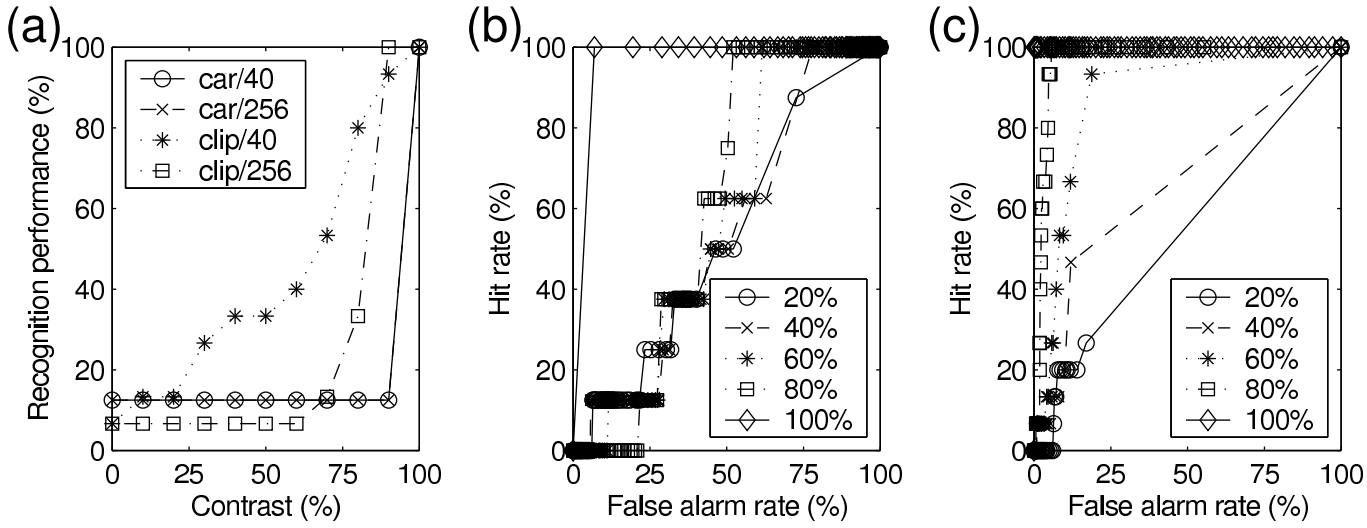

Figure 5: Recognition of car and paperclip stimuli for varying contrasts. **(a)** Recognition performance in the Most Active VTU paradigm for single cars and paperclips and contrasts from 0% to 100%. Legend indicates stimulus class and number of VTU afferents (40 or 256). Chance performance is 12.5% for cars and 6.7% for paperclips. **(b)** ROC curves for recognition of single cars in the Stimulus Comparison paradigm at different contrasts, as indicated in the legend, for 40 VTU afferents. **(c)** Same as (b), but for single paperclips.

paperclips in HMAX were more "well-behaved" than those of cars, *i.e.*, they scaled more regularly with contrast. The reason for this difference can be found in the set of features used in HMAX. The features detected by S2 / C2 units consist of four adjacent bars, an arrangement which appears well matched to the actual features displayed by paperclips. While this also caused S2 and C2 units to exhibit a stronger overall response to paperclip stimuli, higher firing rates were not so much the reason for better contrast-invariant recognition of this stimulus class in HMAX. It was much more critical that different paperclips elicited C2 activation patterns that were more different from each other than were the patterns elicited by different car stimuli (see also section 3.2). Consequently, even for small overall C2 activity levels due to low stimulus contrast, different paperclips could still be distinguished quite well by the model. This makes clear that invariant recognition for different contrasts is aided by a suitable set of features for the respective stimulus class. A suitable feature in this case need not be a specialized feature for a certain object class, but it should reflect stimulus properties better than the current S2 / C2 features do in case of cars. Such features can be extracted from natural images by learning algorithms (see [54]).

Of course, there are methods by which HMAX responses can be made invariant to changes in contrast as we have defined it. For example, by normalizing the mean of each image patch that is processed by a S1 unit to zero, all changes in stimulus contrast effectively become multiplicative changes to pixel values, which are compensated for by the sum-normalization the S1 units perform. However, the biological plausibility of such input normalization is questionable, and it would rid C2 unit responses of *any* contrast dependence, in contrast to data from physiology [42] and recent fMRI results from V4 cortex [2]. Furthermore, attentional modulations of neural activity are usually observed with low or intermediate stimulus contrasts and, consequently, firing rates well below saturation [31, 42]. Since C2 units responded nearly linearly within the range of contrasts used in our simulations, *i.e.*, in a similar fashion as real, *e.g.*, V4 neurons when their firing rate can be modulated by attention, we retained the contrast dependence of C2 units. Their response linearity also allowed for straightforward multiplicative firing rate increases to be used as models for the increases in effective contrast which are commonly associated with attentional effects [42] (see section 3.3).

3.2 Addition of a distractor stimulus

As described in the Methods section, clutter in our experiments was modeled by the presence of a distractor stimulus of the same object class. Ideally, adding a distractor would not interfere with the recognition of the target stimulus. In the Most Active VTU paradigm this would mean that the VTU tuned to the target still responded most strongly among the set of VTUs tuned to individual members of the stimulus class, excluding the VTU tuned to the distractor. For successful recognition of a stimulus in the Stimulus Comparison paradigm, on the other hand, we demanded that a VTU responded more strongly to a two-stimulus display that contained its preferred stimulus than to any of the two-stimulus

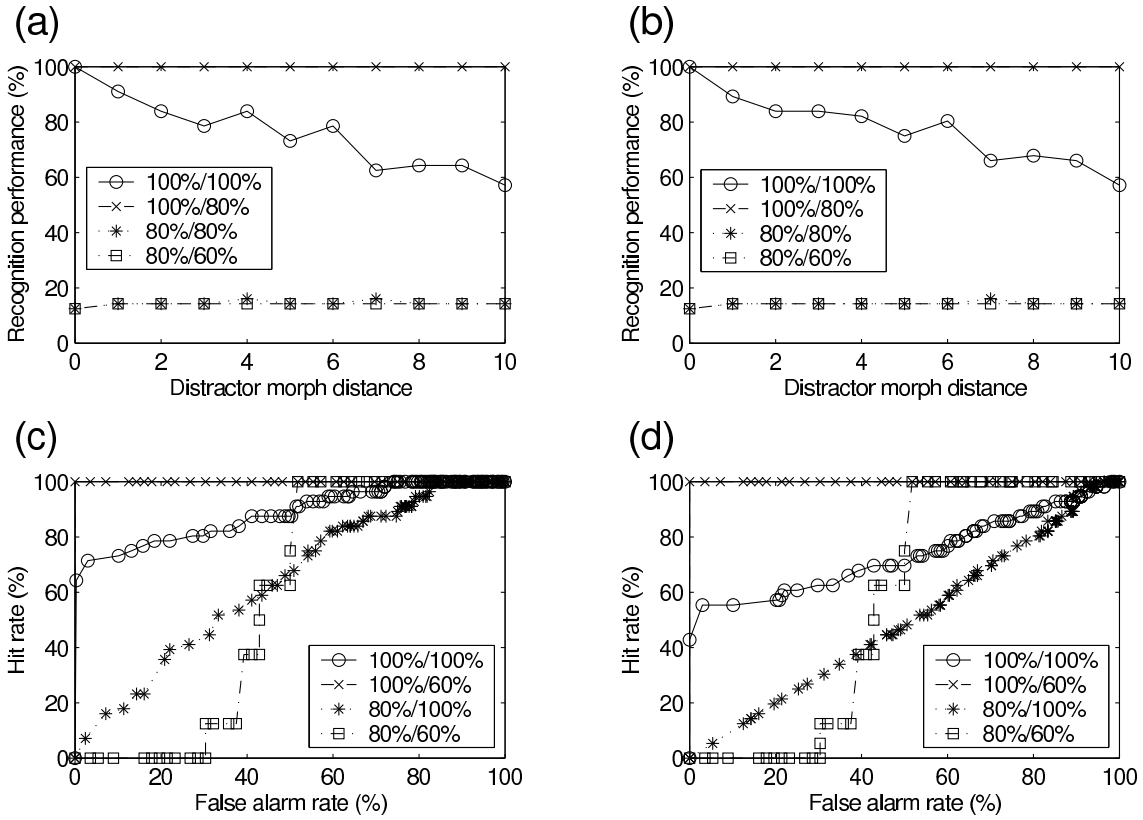

Figure 6: Recognition of cars in the presence of a distractor car. **(a)** Recognition performance in the Most Active VTU paradigm at varying target and distractor contrasts for 40 C2 afferents to each VTU. Legend indicates target (first value) and distractor contrast (second value). **(b)** Same as (a), but for 100 afferents. **(c)** ROC curves for car stimulus recognition in clutter according to the Stimulus Comparison paradigm, 40 C2 afferents to each VTU. The distractor was always at morph distance 5 from the target. Legend indicates target (first value) and distractor contrast (second value). **(d)** Same as (c), but for distractors at maximum morph distance (10).

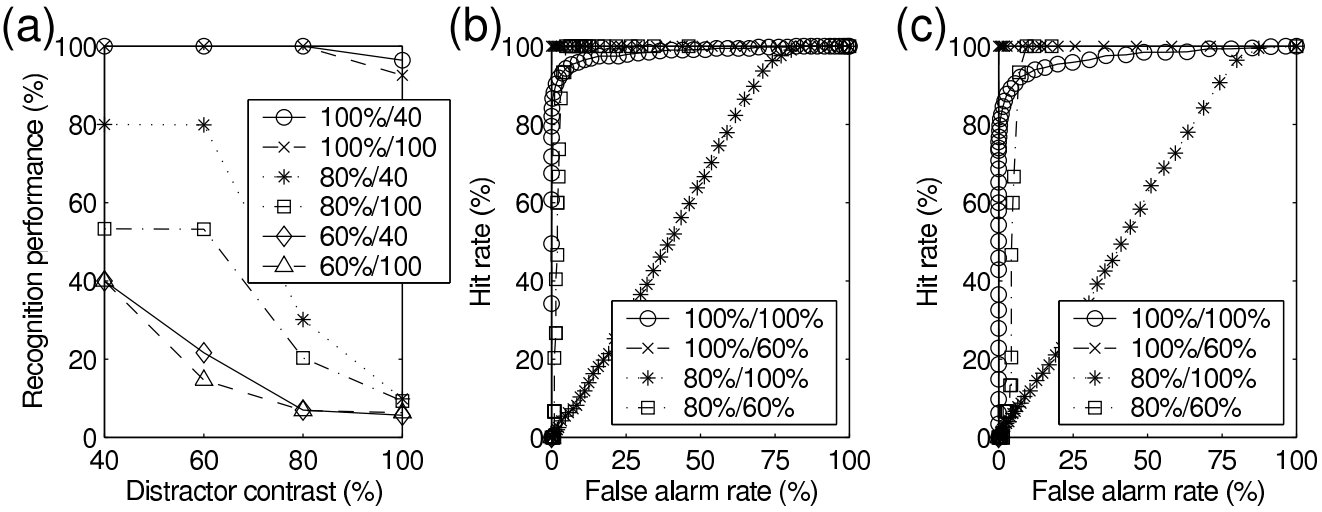

Figure 7: Recognition of paperclips in the presence of a distractor paperclip. **(a)** Recognition performance in the Most Active VTU paradigm at varying distractor contrast levels (abscissa values). Legend indicates target contrast (first value) and number of VTUs' afferents (second value). **(b)** ROC curves for paperclip recognition in clutter (Stimulus Comparison paradigm), 40 C2 afferents to each VTU. Legend indicates target (first value) and distractor contrast (second value). **(c)** Same as (b), but for 100 afferents.

displays containing distractor objects only.

From Figures 6 and 7, it can be seen that a distractor did in fact interfere with recognition of a target stimulus, if its contrast was equal to or greater than that of the target stimulus. For cars, due to the low contrast invariance exhibited by HMAX for this stimulus class, performance quickly reached chance level at contrasts lower than the training value, and a distractor did not change that result any more. For target stimuli at 100% contrast, interference was greater, and performance lower, if the distractor car was more dissimilar with respect to the target (*i.e.*, if the distractor's morph distance to the target was greater), but no dependence of performance on the number of afferents was apparent (compare Figures 6 a and b), as opposed to paperclips (see Figure 7 a). This might seem surprising; one would expect a dissimilar distractor to interfere less with the neuronal representation of a target stimulus than a more similar one, and for paperclips, it has already been shown in [44] that HMAX recognition performance in clutter is more robust if fewer C2 afferents to the VTUs are used. However, a close look at the C2 representations of the eight car prototypes revealed a considerable degree of overlap between them. If 40 afferents to a VTU were used, only 63 out of 256 C2 units constituted all the afferents to the eight car prototype VTUs, *i.e.*, different VTUs mostly shared the same afferent C2 units. In short, rather than activating *different* sets of units most strongly, different car stimuli activated the *same* units *differently*, and thus, for cars, interference was greater for more dissimilar distractors. This high degree of overlap was also the reason why recognition performance for cars hardly at all depended on the number of a VTU's afferents. Already for small numbers of afferents, the sets of afferents of different VTUs overlapped to a great extent. Thus, adding a distractor car to the display affected firing rates of even the most robust target VTU afferents, giving smaller sets of afferents no advantage over larger ones in terms of recognition performance.

For paperclips, on the other hand, the relation between number of afferents and robustness of recognition in clutter was as expected: the more afferents a VTU used, the higher the probability that a distractor stimulus significantly affected the firing rates of at least some of them, and the more recognition performance dropped (see Figure 7 a). This relation is in agreement with previous findings [44]. It held for paperclips because the sets of afferents of VTUs tuned to paperclips overlapped much less than those of cars. If again each VTU had 40 afferents, the combined set of afferents of eight paperclip VTUs (units 1 to 8 in this case) consisted of 142 C2 units, as opposed to only 63 for eight car VTUs. Thus, for paperclips, recognition performance was indeed more robust if smaller sets of afferents were used. It is also evident that recognition performance for paperclips, even when a distractor was present, still dropped a lot more slowly with decreasing target contrast than was the case for cars (compare Figures 6 a and b with Figure 7 a), just as we found for single stimuli in section 3.1. Interestingly, ROC analysis showed little dependence of recognition performance for paperclips on either target contrast (if distractor contrast was lower than target contrast) or number of VTU afferents (see Figure 7 b, c). This demonstrates that performance in different tasks can depend on parameters such as contrast or number of afferents in different ways. As discussed in Methods, in the Stimulus Comparison paradigm, we measured performance in a simulated two-alternative forced choice task, while the Most Active VTU paradigm performance value measured the ability of the system to distinguish between a certain number of trained stimuli. Our data indicate that, even if, at low stimulus contrasts, a neuron does not fire more strongly than others any more upon presentation of its preferred stimulus, it might still respond selectively by firing more strongly when its preferred stimulus is present than when it is not.

These results confirm that, as has already been discussed in [44], robust recognition performance in clutter can be achieved in HMAX, even without attentional mechanisms, provided (a) only a subset—the most strongly activated—of the C2 units are used as afferents for the VTUs, (b) target contrast and, consequently, model unit activation caused by the target are high enough to avoid interference by a distractor, and (c) the neuronal representations of different stimuli are sufficiently distinct, as is the case for paperclips. However, stimuli that share a common shape structure, like the car images in our experiments, can have less distinct representations in HMAX feature space, leading to a loss of recognition performance in clutter. Like HMAX's weaker performance at contrast-invariant recognition of car stimuli, this is a consequence of the feature dictionary used in the standard version of HMAX. As discussed above and in [53], the standard features appear well-suited for paperclip stimuli, but not necessarily for real-world images. While this performance can be improved by learning object class-specific feature detectors [54] in a non-attentive paradigm, we can also expect selective attention to the features of a target stimulus to increase performance "on the fly", *i.e.*, without requiring learning of new features. This will be the focus of the following sections.

3.3 Introducing attentional effects: Multiplicative attentional boosts

We first modeled attentional enhancement of neural activity by multiplying firing rates of model units with a factor greater than 1, corresponding to the hypothesis of McAdams and Maunsell that attention boosts firing rates of cells coding for attended features in a mul-

tiplicative fashion [33]. In our experiments, attention was directed to a stimulus by increasing activity values of those C2 units that projected to the VTU which had been trained on this stimulus (see Methods). This corresponded to an attentionally induced activation of a target stimulus' stored representation, presumably located in IT [36], that sensitizes upstream feature detectors in V4 which are critical for recognition of the target.

Figures 8 to 11 show the results of this kind of attentional priming in terms of recognition performance after presentation of the stimulus display with target and distractor and its feedforward processing by the primed model. Results are shown for both cars and paperclips, evaluated in the Most Active VTU and Stimulus Comparison paradigms. For comparison, results for two situations were included here that were then excluded from further study: target stimuli at 100% contrast and VTUs with 256 C2 afferents. Due to the MAX operation performed by C2 units, the addition a second stimulus could, if anything, only increase their firing rates as long as the two stimuli were spaced far enough apart to exclude interactions at the S1 unit level. With a target stimulus at 100% contrast, a further increase in C2 activity by attentional effects thus changed the C2 activity pattern even more from its value for 100% target contrast without a distractor—to which, after all, the VTU was trained. Consequently, Figures 8 to 11 always show declining recognition performance due to attention effects if the target stimulus was presented at 100% contrast. This effect is independent of the specific boosting method used, and thus we will not further discuss attentional effects for targets presented at full contrast (except in simulations using alternative coding schemes or population coding). This is also in accordance with experiments that do not observe attention effects for high-contrast stimuli [42]. Thus, our experiments were performed within the dynamic range of the C2 units; effects of response saturation due to higher contrast levels and attentional firing rate boosts will be considered later in section 3.6. On the other hand, attentional effects applied to C2 units could not be specifically directed to a certain stimulus if all VTUs were connected with all 256 C2 units, as discussed in the Methods section. This situation was only considered in control simulations to find out whether nonspecific effects could also affect recognition performance (see below).

The figures show that, in HMAX, recognition performance could in fact be increased by attentional activity modulation, both for cars and paperclips. For certain boost and stimulus contrast values (*e.g.*, for cars, 1.3 at 60% target contrast and 1.1 at 80% target contrast), the target's VTU was more often the most active VTU when the target was in the stimulus display, and its responses were more often selective for the target in the sense that it responded more strongly when the target was present than when it was not. However, success of this boosting method was highly dependent on exact boost value in relation to image contrast. For any given target contrast, only a small range of boost values actually improved performance; others were either too small to influence relative VTU activities or actually boosted the afferents' firing rates beyond their levels during training, again reducing absolute and possibly relative activity of the target's VTU. In our model, with the VTUs tuned to certain levels of activity of their C2 afferents, whose firing rates are again contrast-dependent, it is clear that a single attentional boost value cannot improve recognition performance for all stimulus contrast levels. If, however, it is assumed that units firing below saturation participate in stimulus encoding, and if the firing rate carries information about the stimulus—both of which are realistic assumptions, as also mentioned in the Methods section—, then the problem is a general one. (The case of using saturated units for stimulus encoding will be considered in section 3.6.) In a pre-recognition attention paradigm, it also remains unanswered how the system should determine in advance how much attentional activity enhancement it must apply.

Moreover, this boosting method was not very efficient at resolving the effects of a distractor whose contrast was higher than that of the target. Since a high-contrast distractor could, if anything, only increase firing rates of some of the target's C2 afferents due to the MAX pooling mechanism, as discussed previously, an attentional boost of all target afferents could not compensate for this perturbation. In the Most Active VTU paradigm, performance improvements were observed even if distractor contrast was higher than target contrast (see, for example, the plot for target contrast 60% and distractor contrast 80% in Figure 8 or the plots for 60% and 80% target contrast in Figure 10). However, it is important to note that this method of measuring recognition performance did not account for false alarms (*i.e.*, "hallucinations"), as opposed to ROC curves. In calculation of ROC curves, events of erroneous detection of a stimulus that had been cued (*i.e.*, the C2 afferents of the VTU coding for it had been increased in their activity) but not presented in the image were explicitly counted as false alarms. There was no such false alarm measure incorporated in the recognition performance value of the Most Active VTU paradigm. We will discuss a way to account for false alarms in the Most Active VTU paradigm later in section 3.8. The ROC curves, however, with their correction for false alarms, notably do not show any significant increases in performance for distractor contrast values above target contrast (see Figures 9 and 11).

Finally, Figure 12 shows that occasionally a boost of *all* C2 units yielded better recognition performance than a selective boost of the C2 units that were most strongly activated by the target, even if VTUs with all 256 C2

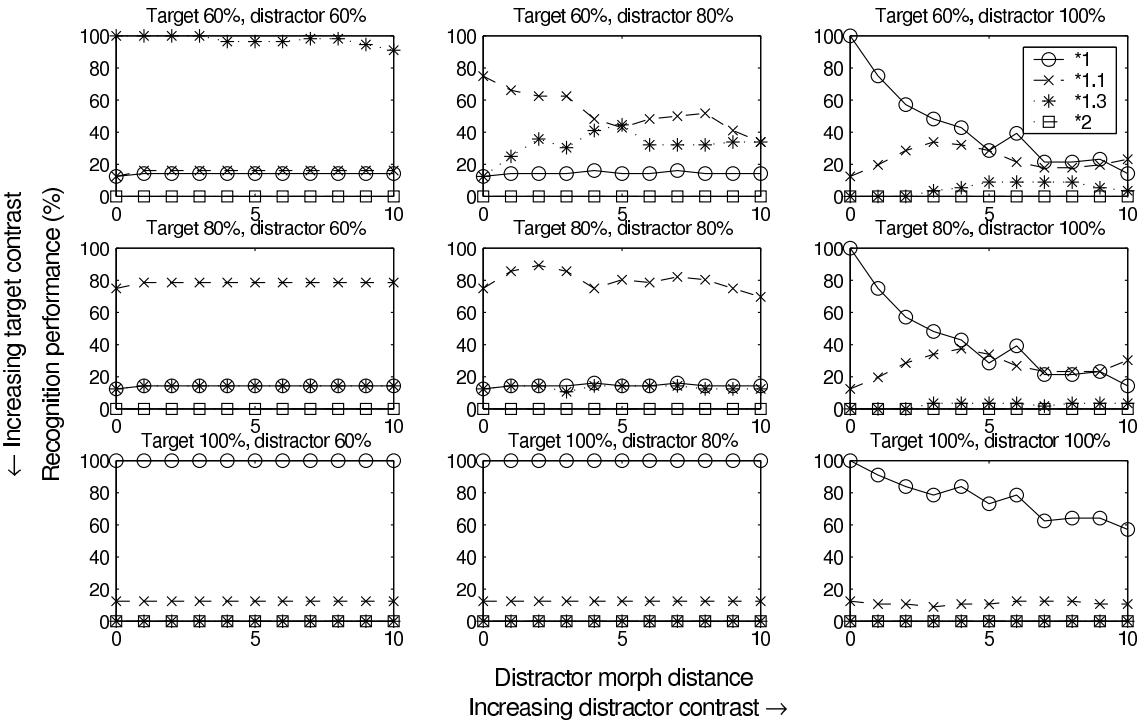

Figure 8: Multiplicative attentional boosts of C2 units. Recognition performance (Most Active VTU paradigm) for target cars in presence of a distractor car (contrasts indicated above each plot), averaged over all target cars, for different distractor morph distances. Legend in top right plot indicates (for all plots) values of multiplicative boosts applied to target VTU's C2 afferents in generation of each graph. All results shown are for 40 C2 afferents to each VTU.

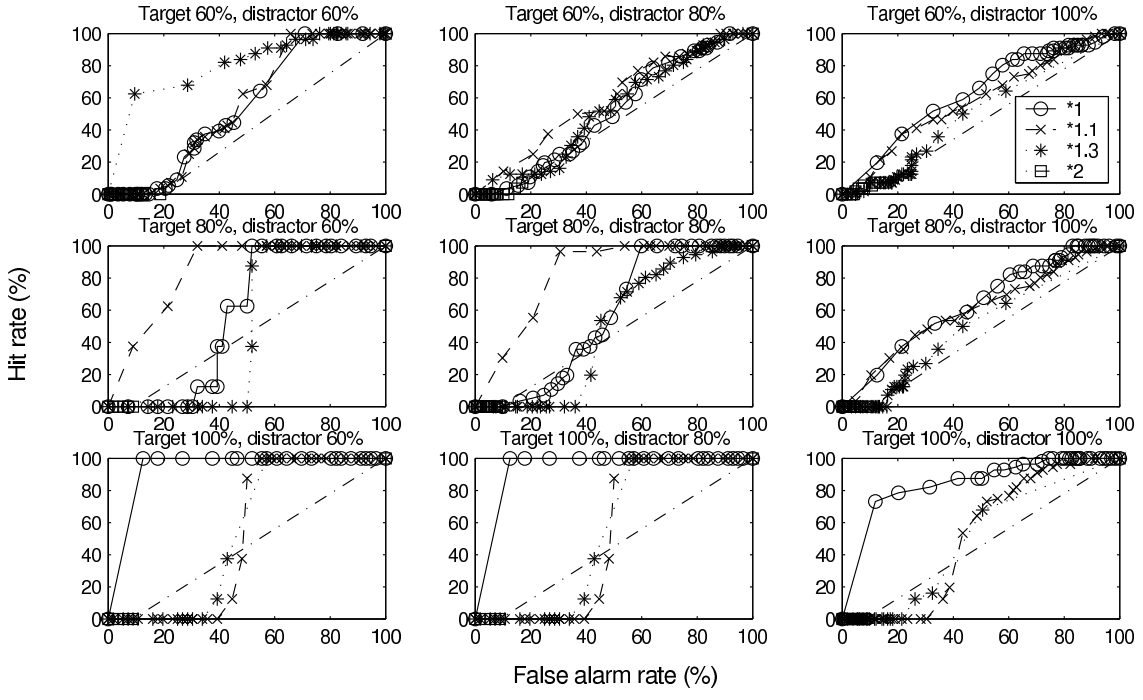

Figure 9: Multiplicative attentional boosts of C2 units. ROCs for recognition of target cars in presence of a distractor car (Stimulus Comparison paradigm), at different contrasts and multiplicative attentional boosts, averaged over all target cars, for 40 afferents per VTU. Distractors were always at morph distance 5 from the target stimulus. Legend as in Figure 8.

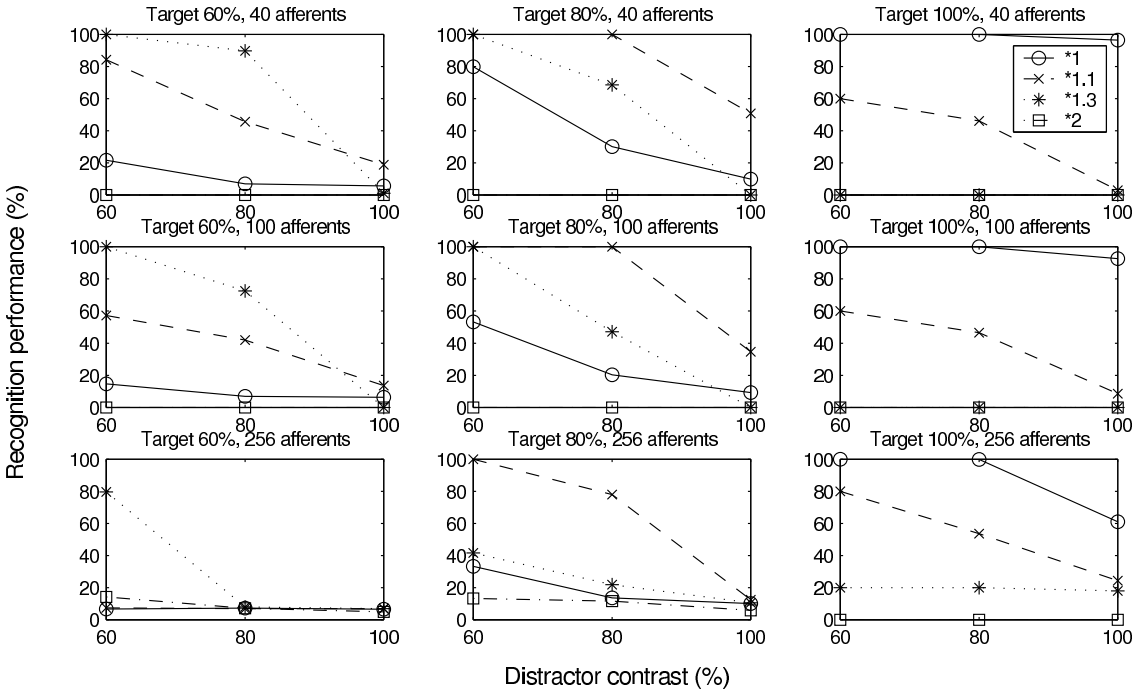

Figure 10: Multiplicative attentional boosts of C2 units. Recognition performance (Most Active VTU paradigm) for target paperclips in presence of a distractor paperclip, averaged over all target paperclips, for different distractor contrasts (abscissa values). Target contrast and number of afferents per VTU indicated above each plot. Legend in top right plot indicates (for all plots) values of multiplicative boosts applied to target VTU's C2 afferents in generation of each graph.

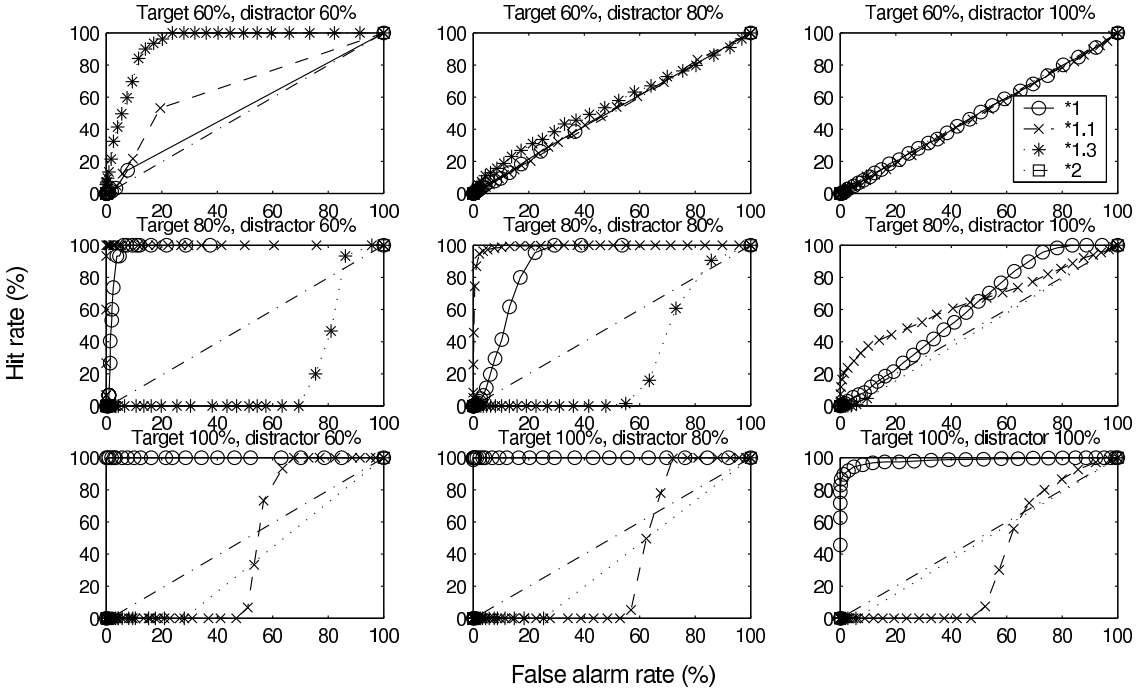

Figure 11: Multiplicative attentional boosts of C2 units. ROCs for recognition of target paperclips in presence of a distractor paperclip (Stimulus Comparison paradigm), at different contrasts and multiplicative attentional boosts, averaged over all target paperclips, for 40 afferents per VTU. Legend as in Figure 10.

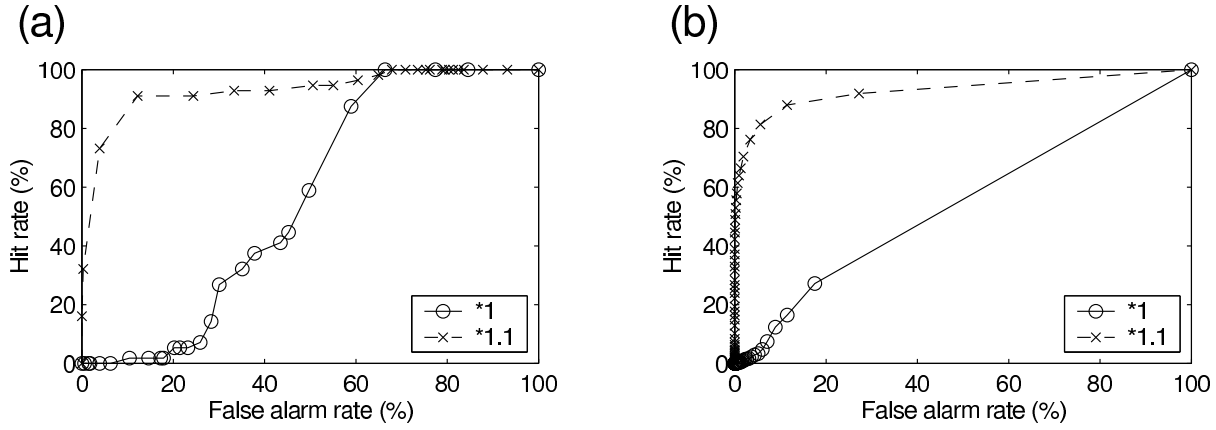

Figure 12: Multiplicative attentional boosts applied to *all 256* C2 units when all 256 cells were used as afferents to each VTU. **(a)** ROCs for recognition of target cars in the presence of a distractor car in the Stimulus Comparison paradigm, averaged over all target cars, with and without a multiplicative boost, as indicated in the legend. Target and distractor contrast 80%. **(b)** Same as (a), but for paperclips. Target and distractor contrast 80%.

units as afferents—which are usually less resistant to clutter—were used. While Figure 10 confirms the intuitive notion that boosting more afferents is less specific and yields smaller performance gains, it again does not account for possible false alarms. For enhancing the response of the target stimulus' VTU over that of others, and thus increasing recognition performance in the Most Active VTU paradigm, boosting only its afferents is of course more effective than boosting all C2 units. However, selectivity of VTU responses to different stimuli might in some cases be improved more by a simple general increase in effective contrast, arguing against the need for selective feature attention in these instances.

All in all, results for a multiplicative boost of C2 units coding for attended features are rather mixed. Our data show that, even though significant performance increases are possible, best results are obtained only if the distractor is presented at a contrast level equal to or lower than that of the target and if the appropriate boost value for the target's contrast level is known. This problem of the mechanism of attention discussed here is not unique to our model if it is assumed that units firing below saturation are used for stimulus encoding and that exact firing rates of neurons carry information. Moreover, selectivity of model unit responses seems to be just as well improved by a nonspecific general C2 activity increase. This might, however, be partly due to the multiplicative boosting method used here. It increases firing rates of model units already displaying high levels of activation more than those of units firing less strongly, by absolute measures. As discussed in section 1.3, a more realistic assumption might be that attention causes a leftward shift in the response function of neurons that code for attended features, resulting in larger activity increases for neurons firing just above baseline [42]. This will be explored in the following section.

3.4 Shifting the response function as attentional boost

To account for the findings of Reynolds *et al.* that attention to a stimulus might cause a leftward shift of the contrast response function of neurons that participate in the representation of this stimulus [42], we modeled this behavior in HMAX by selecting a smaller value for the mean value *s2Target* of the S2 / C2 units' Gaussian response function, thus shifting it to the left and allowing for greater C2 firing rates at lower C1 input strengths. We hypothesized that this method would modulate C2 unit firing rates in a more "natural" way than multiplication with a factor did.

Our results (Figures 13 and 14) were similar to those for multiplicative boosts. Since we carefully chose the boost value to fit the stimulus contrast we used (see Methods), recognition performance increased considerably. Even a slight ROC performance gain for a distractor of higher contrast (80%) than the target (60%) was observed. This is no contradiction with our above claim that boosting methods cannot resolve the effects of a distractor of higher contrast. As long as a distractor was presented at a contrast level lower than that used during training, C2 units activated by both the target and the distractor did not reach their training activity level, and further increases in their firing rates by attention had a chance of bringing their activity closer to the training value, thus increasing probability of target recognition. This effect could, however, also be observed with a multiplicative gain of appropriate value (not shown). Thus, it is not a unique characteristic of the response function shift boosting method to enable better recognition of a target in the presence of a high-

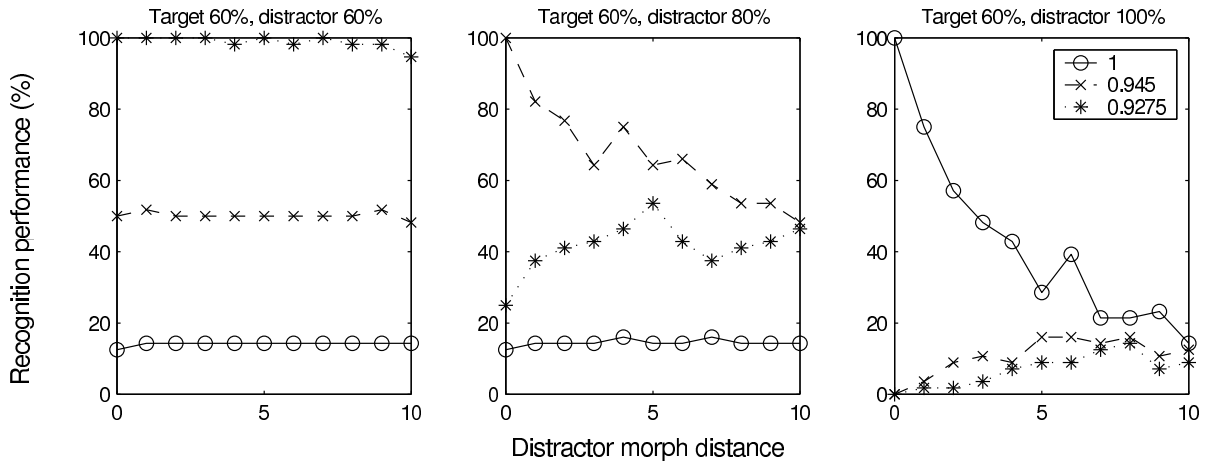

Figure 13: Effects of shifts in the C2 response function. Recognition performance (Most Active VTU paradigm) for target cars in presence of a distractor car (contrasts indicated above each plot), averaged over all target cars, for different distractor morph distances. Legend in top right plot indicates (for all plots) values of the mean of the Gaussian response function of the target VTU's 40 afferent S2 / C2 units used in generation of each graph, with 1 being the normal value used in all other simulations.

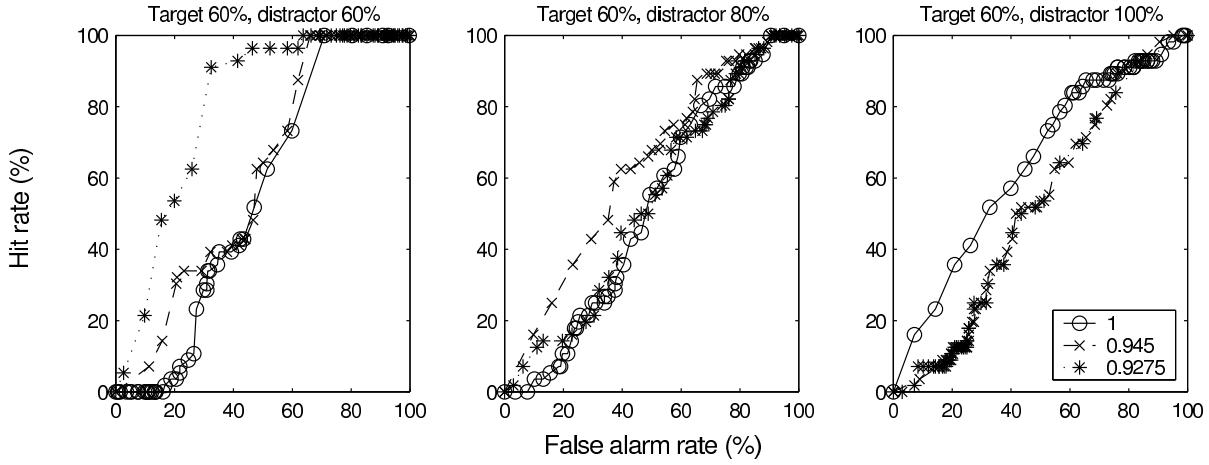

Figure 14: Effects of shifts in the C2 response function. ROCs for recognition of target cars in presence of a distractor car (Stimulus Comparison paradigm), averaged over all target cars, at varying contrasts, using different values for the mean of the Gaussian response function of those 40 C2 units that fed into the target stimulus' VTU, as indicated in the legend. Distractors were always at morph distance 5 from the target stimulus.

contrast distractor. Most noteworthy, however, and in analogy to multiplicative boosting, a shift in the response function also had to be exactly appropriate for the target's contrast level in order to improve recognition performance.

Furthermore, as was the case for multiplicative boosts, simulating uniform attentional activity enhancement of all 256 C2 units and using all of them as afferents to each VTU resulted in a performance gain similar to that for selective feature attention when an ROC measure was applied (not shown). Again, no directed attentional effect seemed to be necessary to improve performance. However, the slight performance increase for target contrast 60% and distractor contrast 80% mentioned above could not be replicated with this general boost. Thus, if improvement of recognition performance for a target with a higher-contrast distractor is at all achievable, it is most probably limited to situations where only a subset of C2 units is used as afferents to any given VTU.

Our results make clear that a leftward shift in the response function of neurons, as reported by Reynolds *et al.*, can be used in HMAX as a model of attentional activity enhancement and to improve recognition performance. However, again, the value of the shift has to be exactly matched to the target's contrast level and would have to be known in advance in an early selection paradigm. There is no qualitative difference with respect to the problem of neutralizing the effects of a high-contrast distractor, and, as with other boosting methods, a nonspecific general increase in effective stimulus contrast by equally boosting all C2 units has effects very similar to our model of selective feature attention. In fact, within the framework of the HMAX model, both boosting methods we discussed here, as well as a third method, where we experimented with constant additive activity boosts, behave very similarly. Since a leftward shift of the response function is computationally more expensive in our model, and since an additive constant boost accounts less well for the response characteristics of C2 units, we used multiplicative boosts as models of attentional activity modulations in our further simulations. This emulated a shift in the response function very well since, for the stimuli and contrast levels we used, firing rates of all C2 units were relatively closely spaced and within the C2 units' approximately linear operating range.

3.5 Suppression

So far, we only described activity enhancements of C2 units coding for features of the attended stimulus. We also performed simulations where we added suppression of other units, as described in Methods, to account for the numerous experimental results mentioned in section 1.3 that find firing rate reductions in cells whose preferred stimulus is not attended. Typical results of suppressing the afferents of the most active nontarget VTU or all C2 units that did not project to the target VTU are shown in Figures 15 to 18. First of all, it is obvious that boosting alone or in conjunction with suppression of all C2 units that had not been boosted yielded exactly the same ROC curves. After all, calculation of ROC curves was based solely on the activity of the VTU tuned to the target stimulus, for different stimulus presentations, and was thus not influenced by whatever modification was applied to C2 units that did not feed into this VTU. This just reflects the fact that the discrimination performance of a neuron can of course not be changed by modulating the activities of other neurons that are not connected with it. On the other hand, if, in addition to an attentional boost of the target's features, suppression of the afferents of the most active nontarget VTU was added, ROC curves in most cases actually show performance deterioration, for both paperclips and cars. The reason for this is that the most active nontarget VTU very likely had a number of afferent C2 units in common with the target VTU—especially for cars, where the sets of afferents overlapped to a large degree anyway, as discussed in section 3.2, but also for paperclips. Thus, improvements in discrimination performance of a single VTU achieved by boosting were diminished by suppression of some of its afferents that had originally been boosted, and performance was lower than with the corresponding attentional boost alone, or at least not better than without any attentional modulation.

On the other hand, in the Most Active VTU paradigm of measuring recognition performance, where activity of a VTU with respect to *other* VTUs in response to the *same* stimulus counted, the combination of boost and suppression was very effective. For paperclips, especially suppression of all C2 units that were not boosted yielded near-perfect performance in all circumstances tested. Firing rate attenuation of those C2 units that fed into the most active nontarget VTU also led to performance levels equal or superior to that reached with boosting alone. This means that, even though the sets of afferents of VTUs tuned to different paperclips overlapped enough so that the firing rate of the target's VTU was affected by suppression of the afferents of the most active nontarget VTU, they were still sufficiently distinct to give the VTU tuned to the target a net advantage over the other VTUs, despite some of its afferents were both boosted and suppressed. The effects of boosting the afferents of a VTU and suppressing those of others then added and set the firing rate of the boosted VTU further apart from that of the others.

The situation was different for car stimuli, however. Still, attenuating all those C2 units that did not feed into the target's VTU, in conjunction with boosting the afferents of this VTU, yielded superior performance in the Most Active VTU paradigm, just as was seen for pa-

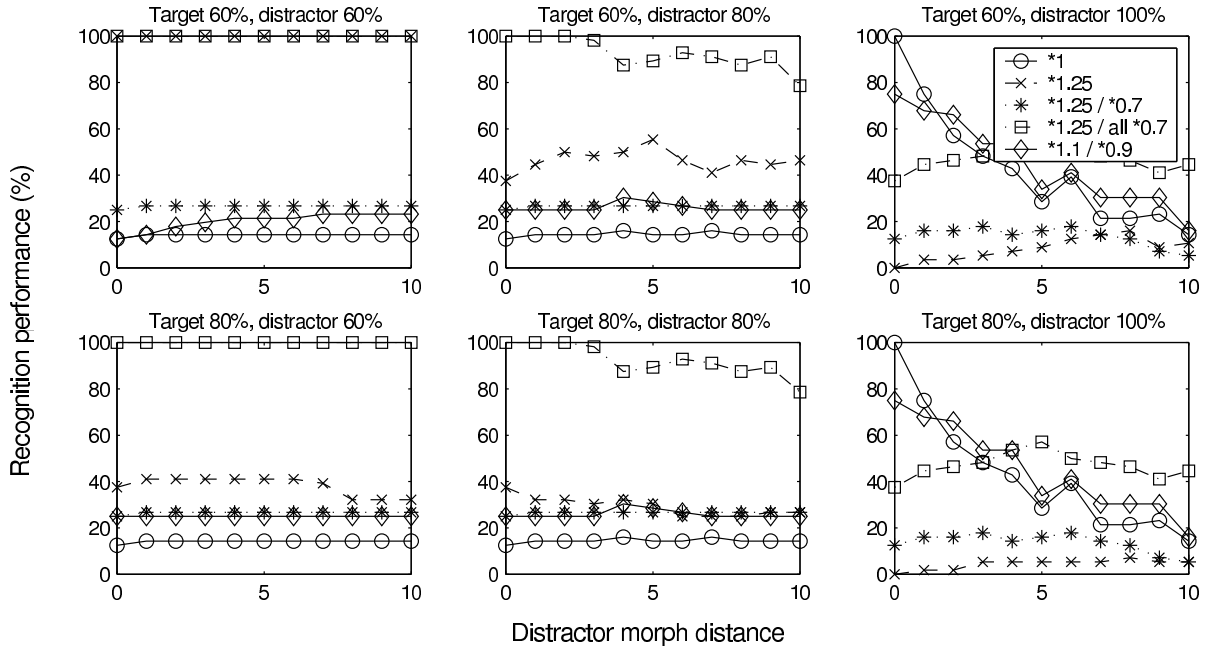

Figure 15: Multiplicative attentional boosts and suppression of C2 units. Recognition performance (Most Active VTU paradigm) for target cars in presence of a distractor car (contrasts indicated above each plot), averaged over all target cars, for different distractor morph distances. Legend in top right plot indicates (for all plots) values of multiplicative boosts and suppression values applied to C2 units in generation of each graph (first value: boost applied to target VTU's afferents; second value, if applicable: suppression applied to all other C2 units ("all") or to afferents of the most active nontarget VTU (otherwise)). All results shown are for 40 C2 afferents to each VTU.

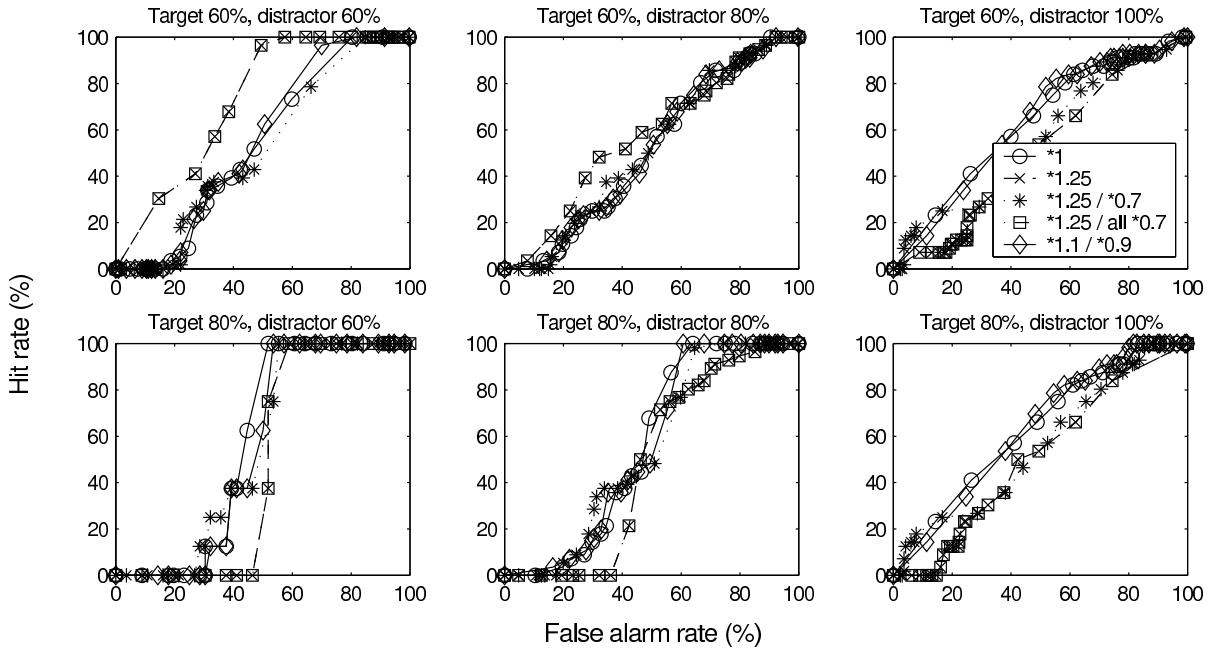

Figure 16: Multiplicative attentional boosts and suppression of C2 units. ROCs for recognition of target cars in presence of a distractor car (Stimulus Comparison paradigm) at different contrasts, averaged over all target cars, for 40 afferents per VTU. Distractors were always at morph distance 5 from the target stimulus. Legend as in Figure 15.

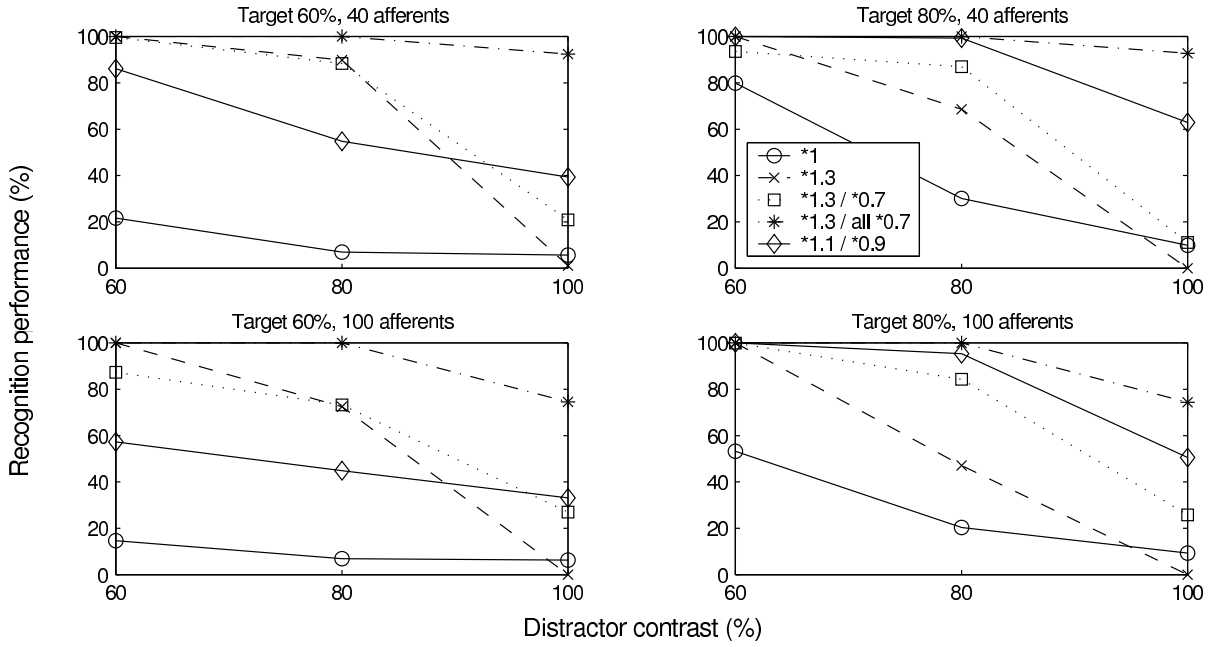

Figure 17: Multiplicative attentional boosts and suppression of C2 units. Recognition performance (Most Active VTU paradigm) for target paperclips in presence of a distractor paperclip, averaged over all target paperclips, for different distractor contrasts. Target contrast and number of afferents per VTU indicated above each plot. Legend in top right plot indicates (for all plots) values of boosts and suppression applied to C2 units in generation of each graph, as in Figure 15.

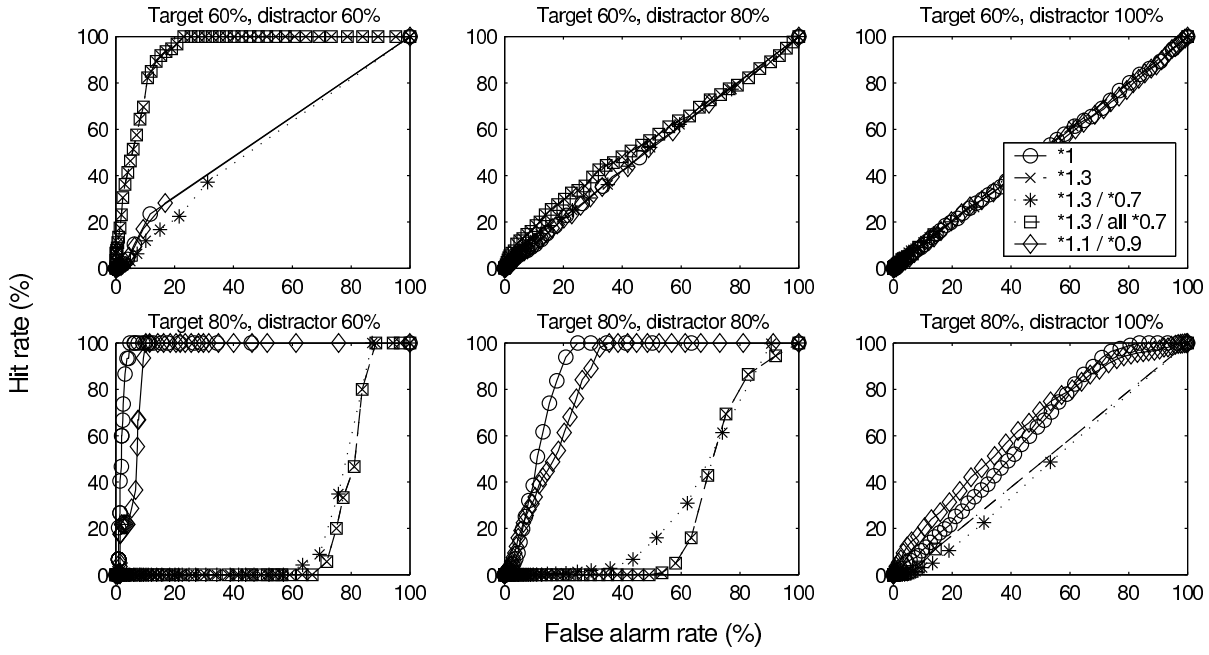

Figure 18: Multiplicative attentional boosts and suppression of C2 units. ROCs for recognition of target paperclips in presence of a distractor paperclip (Stimulus Comparison paradigm), at different contrasts, averaged over all target paperclips. 40 afferents per VTU. Legend as in Figure 17.

perclips. However, since the sets of afferents of VTUs tuned to cars overlapped much more than was the case for paperclips, suppression of the most active nontarget VTU's afferents strongly affected the firing rate of the VTU tuned to the target. Thus, for cars, performance with this kind of suppression was actually lower than when only an attentional boost was applied to the target VTU's afferents, except when a distractor at 100% contrast was presented (see Figure 15). Here, however, the reason why performance was not lower than for boosting alone was that, at best, boost and suppression more or less canceled out and performance for no modulation at all was restored, which in this case was higher than with an attentional boost alone.

All in all, our results indicate that suppression of model units coding for nonattended features can greatly improve recognition performance when it comes to deciding which of a set of stimuli appears in a cluttered scene. In this task, which we modeled in the Most Active VTU paradigm, a combination of boost and suppression yielded the most promising results, even if distractor contrast was higher than target contrast (compare, for example, Figures 15 and 17 with Figures 8 and 10). However, apart from the prevailing problem of having to know in advance which amounts of boost and suppression should be applied, it is also difficult to envision a strategy to select suitable features for suppression. Attenuating the firing rates of neurons that code for features of a distractor might affect neurons critical for target recognition as well, and advance knowledge about the possible identity of the distractor would be needed, which is usually not available. On the other hand, simply suppressing all but those neurons that participate in the neuronal representation of the target stimulus is very effective in improving recognition of this target. However, assuming such a large-scale suppression is probably, if at all, only realistic if stimuli are drawn from a rather limited set and knowledge about this limitation is provided in advance to the subject, such that suppressive mechanisms can be applied in a more directed fashion. Otherwise, one would have to consider attenuating the activity of practically all neurons of a cortical processing stage whose firing is not critical in encoding of the target, which seems to be a quite effortful mechanism. Moreover, as in previous sections, our Most Active VTU paradigm of measuring recognition performance does not account for possible false alarms, which are likely an issue, especially if the representation of one stimulus is enhanced as much over others as in the case of combined boost and suppression. We will return to this question in section 3.8.

3.6 Alternative coding schemes

Firing rates of neurons in the ventral visual stream are not insensitive to stimulus contrast, as is true for model units in HMAX (see section 3.1) [2]. In HMAX, however, the exact response levels of individual units are important for encoding visual stimuli. The details of how the brain represents sensory information in neural responses are of course still unknown, but if exact activity values carry as much information as in HMAX, one might not be able to expect the high degree of contrast invariance of object recognition that is observed experimentally [2]. Furthermore, in this case, an early selection mechanism of attention would not only have to select appropriate neurons for attentional modulation in advance, but also determine modulation strength before stimulus presentation. Otherwise, an attentional modulation can even excessively increase firing rates of neurons that respond to the target stimulus—for example, if the stimulus is shown at a higher level of contrast than expected—which can actually have a diminishing effect on recognition performance (see section 3.3). This could be avoided, for example, if stimuli were encoded only by neurons firing at their maximum firing rate. However, while it is perfectly reasonable to assume that neurons firing at high rates are important for the representation of a stimulus, it would most likely be unrealistic to expect that only neurons firing at saturation participate in stimulus encoding.

To address these problems, we examined alternatives to standard HMAX encoding of stimuli that relied less on exact firing rates of C2 units, in order to try to increase both the model's contrast invariance properties and the effectiveness of boosting and suppression mechanisms. Since, so far, our model units did not exhibit saturation, it was also of special interest whether introducing saturation of firing rates would influence recognition performance or alter the effects of attentional modulation. We hypothesized that, with less dependence of the model's response on exact C2 unit activity patterns, recognition performance might drop less sharply with stimulus contrast than seen in section 3.1, and that the beneficial effects of attentional activity modulations on recognition performance would be more robust since it would not be necessary to restore a certain model unit activity pattern as closely as possible.

3.6.1 Saturation tuning

The first alternative coding scheme we devised, the so-called "saturation tuning" scheme, avoided declines in activities of VTUs if their C2 afferents responded more strongly than during VTU training with the target stimulus, as described in Methods. This kind of encoding is actually very plausible biologically, since it provides for an effectively sigmoidal VTU response function and saturation of VTUs, with different possible saturation levels for different units. However, one also has to bear in mind that the Gaussian response function of VTUs was originally designed to perform template matching in an abstract feature space. Permitting an

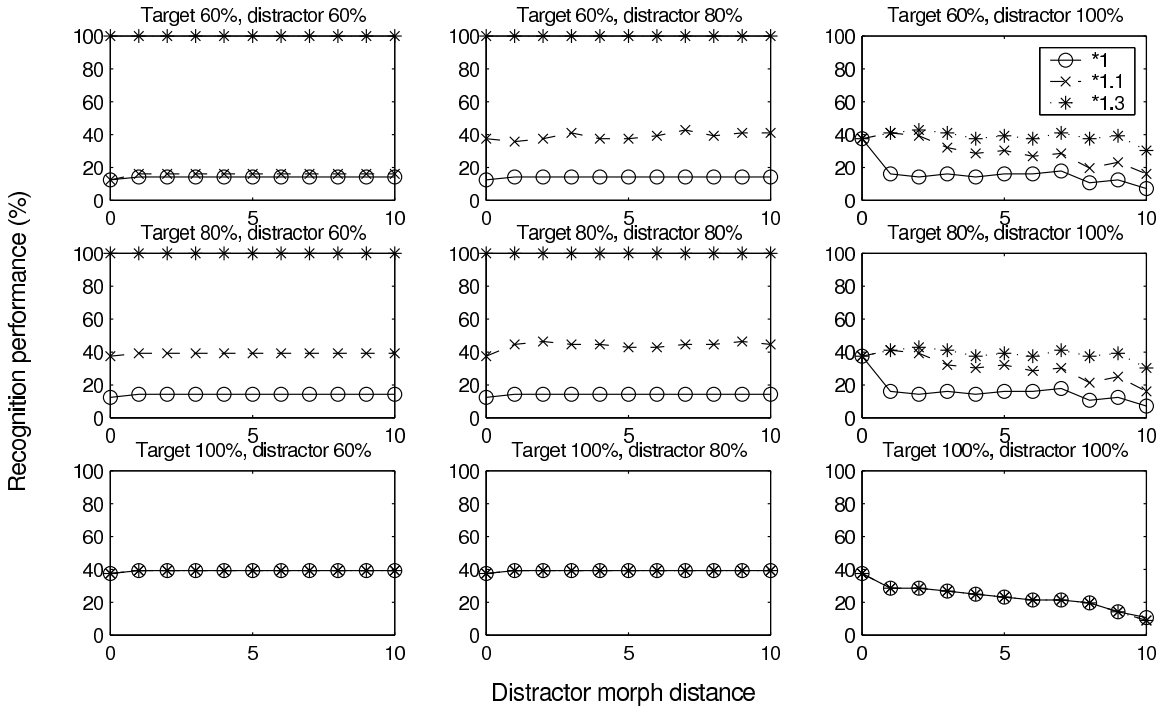

Figure 19: Saturation tuning. Recognition performance (Most Active VTU paradigm) with saturation tuning of all VTUs, for target cars in presence of a distractor car at different contrasts (as indicated above each plot), averaged over all target cars. Legend in top right plot indicates (for all plots) multiplicative attentional boosts applied to the 40 C2 afferents of the VTU tuned to the target stimulus in generation of each graph.

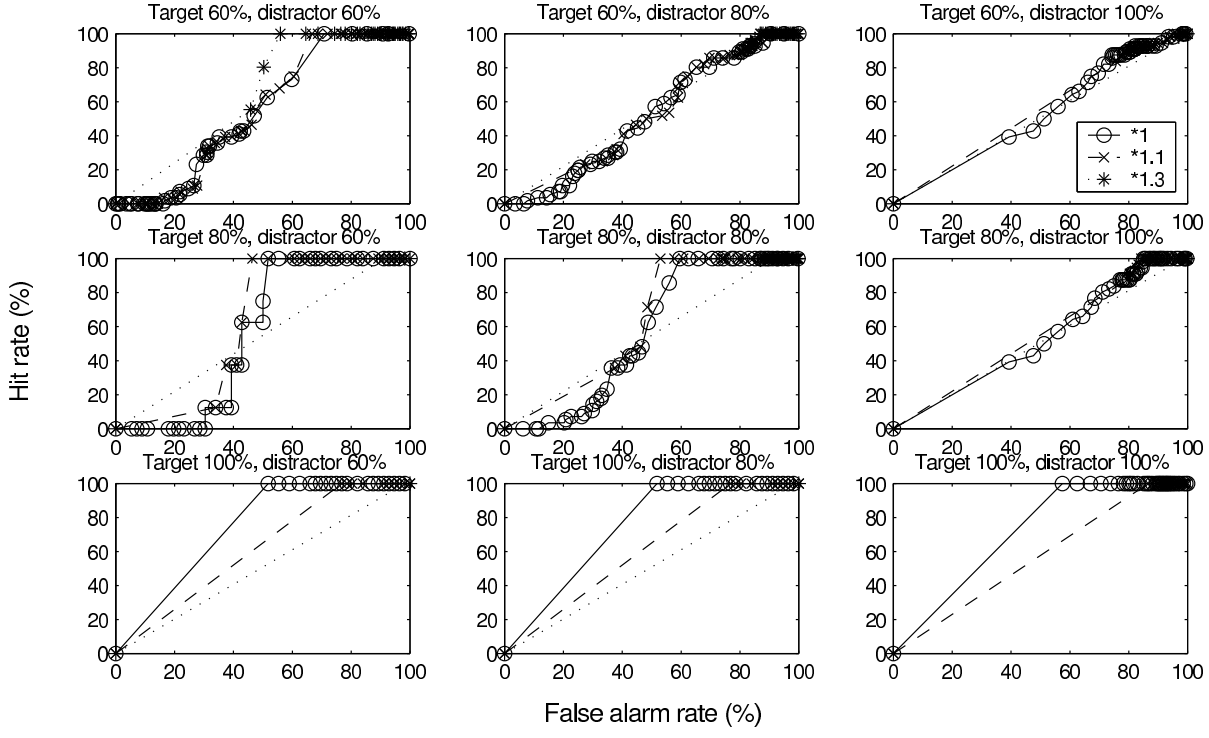

Figure 20: Saturation tuning. ROCs for recognition of target cars in presence of a distractor car (Stimulus Comparison paradigm), with saturation tuning of all VTUs, at different contrast levels and multiplicative attentional boosts, averaged over all target cars. Distractors were always at morph distance 5 from the target stimulus. Legend as in Figure 19.

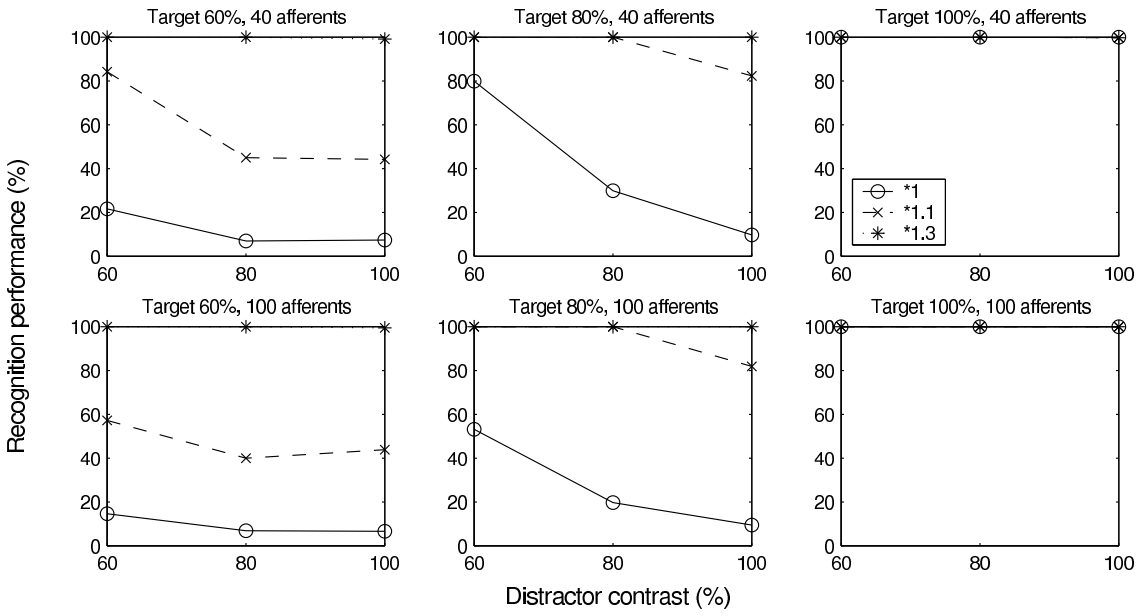

Figure 21: Saturation tuning. Recognition performance (Most Active VTU paradigm) for target paperclips in presence of a distractor paperclip with saturation tuning of all VTUs, averaged over all target paperclips, for different distractor contrasts. Target contrast and number of afferents per VTU indicated above each plot. Legend in top right plot indicates (for all plots) values of multiplicative boosts applied to target VTU's C2 afferents in generation of each graph.

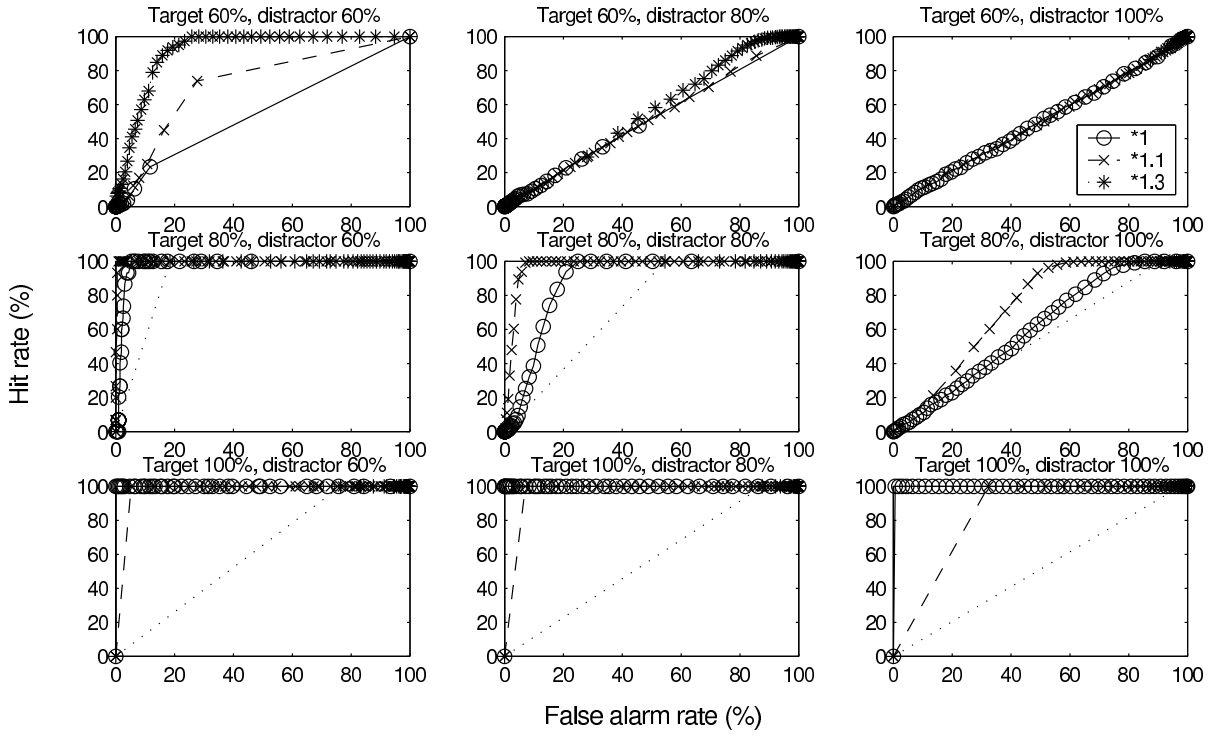

Figure 22: Saturation tuning. ROCs for recognition of target paperclips in presence of a distractor paperclip (Stimulus Comparison paradigm) with saturation tuning of all VTUs, at different contrasts and multiplicative attentional boosts, averaged over all target paperclips. 40 afferents per VTU. Legend as in Figure 21.

overshoot of afferent activity without a corresponding drop in VTU firing rate thus also entails a loss of the VTU response's stimulus specificity.

Results for saturation tuning of VTUs are displayed in Figures 19 to 22. At low contrast levels and without attentional modulations, there were no changes in performance, for both stimulus classes and recognition paradigms, since encoding was unaltered for C2 firing rates below training level. However, if attentional boosts were employed at low target contrasts, perfect recognition could be achieved in the Most Active VTU paradigm, even for distractors of higher than target contrast, since overshooting of training C2 activity did not reduce VTU activity any more, but drove VTUs to saturation. For paperclips, this coding scheme seemed very well-suited overall, at least if not all 256 C2 units were used as afferents to the VTUs (see Figures 21 and 22): similar performance was reached as with standard HMAX encoding, as measured by ROC curves, and boosting was effective; but at full target contrast, distractors did not interfere with target recognition, as opposed to standard encoding. However, a look at Figures 19 and 20 reveals problems of this coding scheme with car stimuli. In the Most Active VTU paradigm, very good results could still be achieved for car stimuli at low contrasts if attentional boosts were applied. However, since the sets of afferents of VTUs tuned to cars overlapped much more than those of VTUs tuned to paperclips, the loss of specificity encountered in switching from standard encoding to saturation tuning was much more relevant here, and recognition performance for full-contrast car stimuli dropped drastically. Even attentional modulations did not change this result. The reason was that, for full-contrast stimuli or after attentional boosts, more VTUs than only the one tuned to the target responded near or at saturation, due to overlap of their sets of afferents. Thus, loss of specificity in this coding scheme makes it inappropriate for stimuli displaying high degrees of similarity—or, conversely, saturation tuning may only be useful if specialized features for a given object class exist, in order to minimize overlap between afferents. Then, however, the resultant more distinct neuronal representations, like those of paperclips in our case, can yield good recognition performance even in standard HMAX, so that there seems to be no need for saturation VTU tuning, except to counter the effects of high-contrast distractors.

3.6.2 Relative rate tuning

In the second alternative coding scheme for VTUs we introduced into HMAX, the "relative rate tuning", the most active VTU was determined by which set of C2 afferents responded most strongly, even if absolute activity levels of C2 units were very low, *e.g.*, due to low stimulus contrast. Specificity, however, was only conferred through the selection of a VTU's afferents, not through matching their activity pattern to its training value. Hence, this coding scheme gave up even more specificity than saturation tuning.

Using relative rate tuning and the Most Active VTU recognition paradigm, cars could be recognized by the model with a much greater degree of contrast invariance than when standard HMAX encoding was used (not shown), and attentional boosts resulted in effective preferred recognition of the stimulus whose critical features had been selected by attention. However, from experiments using the Stimulus Comparison paradigm, we had to conclude that this encoding method cannot be considered a serious alternative to standard HMAX tuning, since individual VTU ability to discriminate between stimuli in the presence of a distractor was nearly completely lost (not shown). Only chance performance levels were reached, with and without attentional modulations. Quite obviously, disregarding the information conveyed in the exact firing rates of afferents and relying on relative firing strengths of different sets of afferents only is much too nonspecific for recognition in cluttered scenes.

All in all, both alternative VTU tuning mechanisms discussed do not seem to be promising solutions to the problem of achieving more contrast-invariant recognition in clutter while at the same time allowing for effective attentional influence on object recognition. Relative rate tuning is too nonspecific overall, and saturation tuning, while reducing the influence of high-contrast distractors, also suffers from a reduction in specificity for realistic object classes that have many features in common. Again, more specialized features for the object class under consideration would very likely improve performance in both coding schemes, while preserving their advantages of allowing for biologically plausible saturation of units (saturation tuning) or diminishing the influence of stimulus contrast on recognition (relative rate tuning), respectively. However, with sufficiently specialized features, recognition in clutter can be robust enough even without attentional modulations, as suggested by our results for paperclips in comparison with cars. Thus, so far, none of the mechanisms we explored seems to actively support the notion of an attentional mechanism that robustly improves object recognition performance in an early selection paradigm.

3.7 Population coding

Stimulus encoding in the brain is mostly found to be distributed: neurons usually participate in representations of several stimuli [18, 51]. Higher areas of visual cortex are no exception to this general finding [66, 67]. Such a population code has various advantages over its opposite, a "grandmother code" where each stimulus is encoded in the activity of only a single specialized neuron. Most significantly, having single neurons tuned

very specifically to single complex stimuli would not allow the brain to generalize to novel stimuli from a few learned example stimuli [47]. Also, a grandmother code would most likely not be robust with respect to a loss of neurons; the objects lost neurons code for might no longer be recognized.

Thus, realistic models of brain functions have to take into account the distributed nature of stimulus representations in the brain. In the simulations described so far, we used a "grandmother-like" code, each stimulus being encoded by the activity of only a single VTU. This coding scheme had not been chosen as a realistic model of the brain's way of representing stimuli (except in specialized cases, where the animal is overtrained on a discrimination task involving a small number of fixed stimuli [28]), but rather because it was the simplest and most straightforward method. To investigate attentional modulation for the case of a population code, we used 190 VTUs trained to different face stimuli as a model of a neuron population whose responses encode the presence or absence of face stimuli (see Methods).

As a basis for comparing single-VTU and population coding, we first assessed recognition performance for faces with *individual* VTUs tuned to them, just as was done in previous sections for cars and paperclips. Performance of the model for this stimulus class and single VTU encoding was found to be very similar to results obtained with car stimuli (not shown). Recognition performance also turned out to be highly contrast-dependent, and it improved with attentional enhancement of C2 firing rates, provided target contrast was lower than training contrast, distractor contrast was not too high and the correct boosting value was chosen. As with cars, recognition performance was nearly independent of the number of C2 afferents each VTU was connected to, and even if all 256 C2 units were used as afferents and all of them received the same multiplicative firing rate boost, recognition performance—as measured by ROC curves—improved about as much as for smaller sets of afferents.

Figures 23 and 24 show recognition performance based on the population response of the face-tuned VTUs, for the smallest number of afferents to the VTUs and to the second-level VTUs we tested. Higher numbers of afferents yielded performance levels equal or lower to those shown. From the curve drawn from data generated without employing attentional mechanisms, it is obvious that—at least for the case of deterministic units investigated here—population encoding achieved no better performance in clutter in our model than a single VTU coding scheme. Instead, a coding scheme based on the responses of several VTUs exhibited even higher sensitivity and thus less invariance to clutter or contrast changes. This was revealed by a look at the activity values of the second level VTUs we used to measure the population response. Their firing rates dropped to very low levels for any deviation of the VTU population activity pattern from that elicited by the training stimulus (not shown).

The second lesson learned from Figures 23 and 24 is that attentional boosts applied directly to the population of VTUs can, in general, not improve recognition performance. Since VTUs display nonlinear behavior, simply increasing their firing rates can not be expected to restore an activity pattern that has been modified by contrast changes or distractor stimuli. An exception to this rule was found for presence of a highly similar distractor at full contrast. In this special situation, VTU population activity was not very different from that during presentation of the target stimulus, and those VTUs that responded most strongly to this stimulus (only these were used to generate Figure 23) were likely to be reduced in their activity, so that boosting their firing rates could in fact improve recognition of the target stimulus in the Most Active VTU paradigm—but only if, at the same time, all other VTUs were suppressed. Other than that, however, and especially if an ROC measure was applied to take account of false alarms, boosting VTUs did not increase recognition performance above chance levels, even if all other VTUs were suppressed.

We thus returned to modulating firing rates of C2 units. However, since the VTU representation of the target stimulus to be recognized was distributed, there were no sets of C2 afferents clearly identifiable as targets for boosting and suppression. Since we used subpopulations of VTUs to code for a stimulus (*i.e.*, those VTUs that were connected with the second-level VTU representing this stimulus), we could have selected only those C2 units for an attentional boost that fed into the VTUs of such a subpopulation, but that did not at the same time feed into other VTUs. However, it turned out that, even if each population VTU had only 40 C2 afferents, hardly any such C2 units could be found (on the order of 10 or less for any given target stimulus)—too few to achieve significant changes in recognition performance by only modulating firing rates of those C2 units. This indicates that, in a population coding scheme, it is even more difficult, if not impossible, to find features that are both critical for the recognition of a stimulus and unique to it.

Figures 25 and 26 show results for applying attentional boosts to *all* afferent C2 units of the VTU subpopulation that coded for a stimulus, regardless of overlaps between the sets of afferents of different subpopulations. (With 40 C2 afferents for each VTU and 10 VTU afferents to each second-level VTU, this affected, on average, about 80 of the 256 C2 units.) The effects were qualitatively identical to and quantitatively somewhat weaker than those obtained in the single VTU coding scheme. Only for distractors of equal or lower contrast than target contrast could performance increases

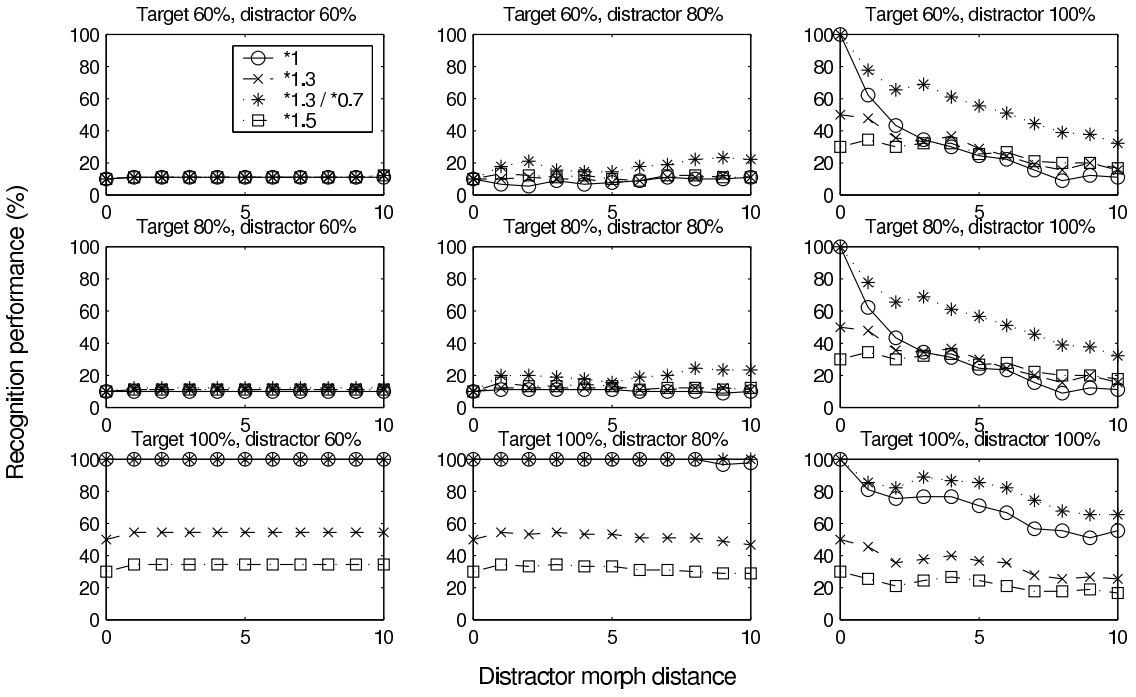

Figure 23: VTU population code with attentional modulation of VTUs. Recognition performance (Most Active VTU paradigm) for target faces in presence of a distractor face. Average taken over all target faces and distractors at each distractor morph distance. Legend in top left plot indicates (for all plots) values of multiplicative boosts applied to the VTUs most activated by the target stimulus (first value) and suppression applied to all other VTUs (second value, if applicable). All results for 40 C2 afferents to each VTU and 10 VTU afferents to each second-level VTU.

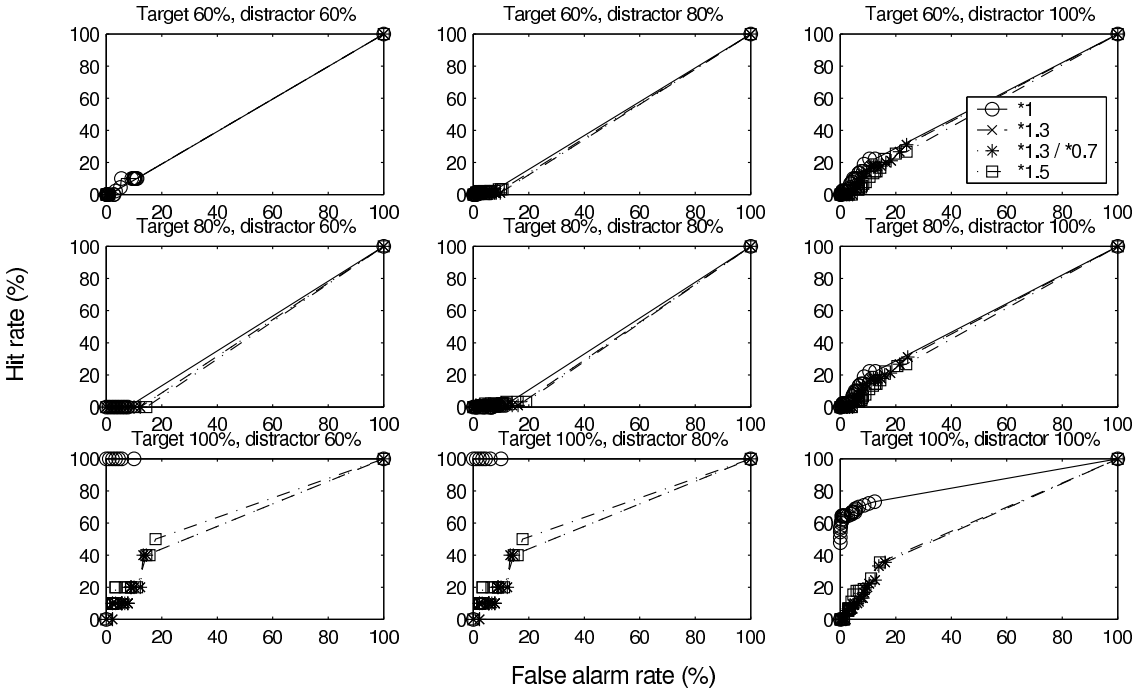

Figure 24: VTU population code with attentional modulation of VTUs. ROCs for recognition of target faces in presence of a distractor face (Stimulus Comparison paradigm), at different contrasts and multiplicative attentional boosts and suppression, averaged over all target faces. 40 C2 afferents per VTU and 10 VTU afferents per second-level VTU. Distractors were always at morph distance 5 from the target stimulus. Legend as in Figure 23.

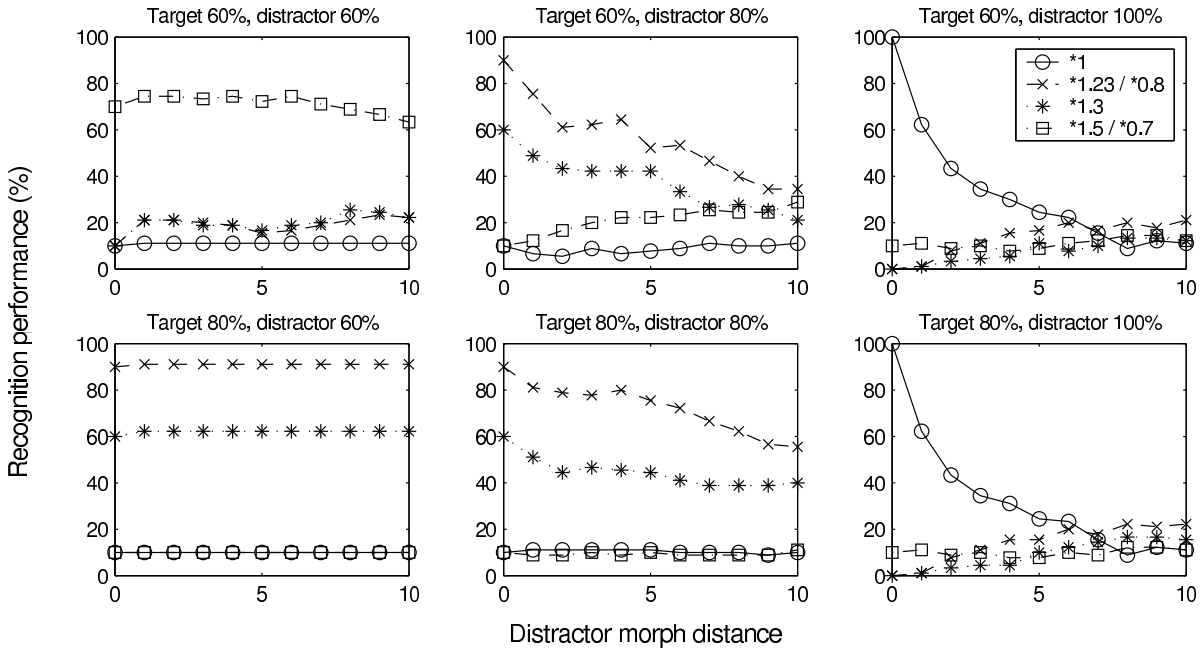

Figure 25: Multiplicative C2 boosts and suppression in a VTU population coding scheme. Recognition performance (Most Active VTU paradigm) for target faces in presence of a distractor face. Average taken over all target faces and distractors at each distractor morph distance. Legend in top right plot indicates (for all plots) values of multiplicative boosts applied to C2 afferents of the VTUs most activated by the target stimulus (first value) and suppression applied to all other C2s (second value, if applicable). All results for 40 C2 afferents to each VTU and 10 VTU afferents to each second-level VTU.

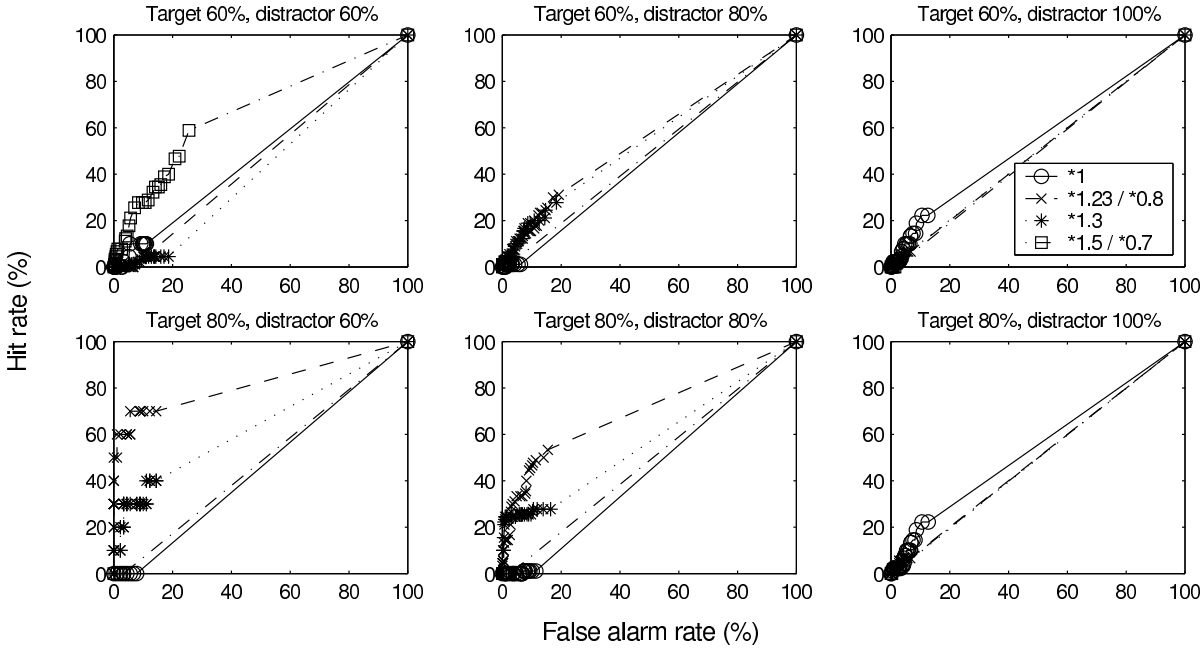

Figure 26: Multiplicative C2 boosts and suppression in a VTU population coding scheme. ROCs for recognition of target faces in presence of a distractor face (Stimulus Comparison paradigm). Average taken over all target faces and distractors. Distractors were always at morph distance 5 from the target stimulus. Legend as in Figure 25.

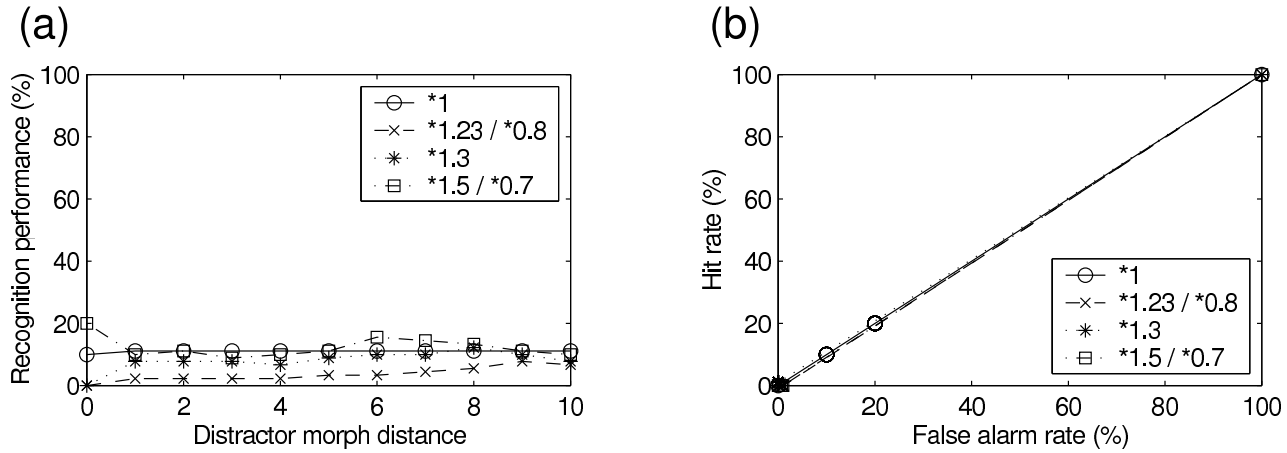

Figure 27: Effects of greater numbers of C2 afferents and larger VTU populations in a VTU population coding scheme. Data plotted for 100 C2 afferents to each VTU and 100 VTU afferents to each second-level VTU. Target and distractor contrast 60%. Legend indicates values of multiplicative boosts applied to the target VTU population's C2 afferents and suppression applied to all other C2 units, if applicable. **(a)** Recognition performance for faces in presence of a distractor face in the Most Active VTU paradigm. Average taken over all target faces and distractors at a given distractor morph distance. **(b)** ROCs for recognition of faces in presence of a distractor face in the Stimulus Comparison paradigm. Average taken over all target faces and distractors. Distractors were always at morph distance 5 from the target stimulus.

be achieved in both paradigms of measuring recognition performance, and boost values needed to be adjusted to the target's contrast level. The set of C2 afferents of the VTU subpopulation responding best to the target stimulus overlapped significantly with the set of afferents of the second most active VTU subpopulation, so that suppression of the latter largely compensated for the boost applied to the first, and recognition performance effectively did not change (not shown). Attenuating all C2 units that did not receive an attentional boost further improved recognition performance if the Most Active VTU paradigm was used for measurement, but the ROC curve was not influenced. As opposed to single VTU coding, however, effects of attentional boosts largely disappeared when more afferents to the VTUs and larger VTU subpopulations were used (Figure 27). This again shows that VTU population coding is in fact more sensitive and less invariant to stimulus changes than single VTU coding, as mentioned above.

Taken together, a VTU population code yields similar results for recognition in clutter and effects of attention in HMAX as a coding scheme based on the firing rates of individual VTUs. Population coding is very effective in increasing specificity of neuronal responses, which is also what we observe in HMAX. However, while advantageous in a situation where neurons are rather broadly tuned and susceptible to noise, the added complexity of an extra layer in the representation exacerbates the problem of selecting appropriate modulations for neurons in intermediate levels (*e.g.*, C2/V4). Attention can be applied in a VTU population coding scheme in a manner analogous to that used for single VTU encoding, but it is also subject to the same limitations as in a single VTU coding scheme.

3.8 Problems of attentional boosts

In previous sections, we mentioned a number of problems we encountered when firing rate boosts of model units were employed to model attentional effects on object recognition, *i.e.*, in an early selection paradigm of attention. Apart from the need to match boost strength to stimulus contrast, the most significant problem turned out to be a tradeoff between effectiveness of an attentional mechanism on the one hand and its specificity on the other hand. An attentional boosting mechanism that is more effective at improving recognition of a target stimulus also seems more likely to be less specific for this target stimulus or to exhibit an increased risk of false alarms, or both. We will end our investigation with a more detailed look at this issue in this section, returning to standard HMAX stimulus encoding in a single VTU coding scheme.

Figure 28 addresses the question of specificity. It shows, for a multiplicative attentional boost without suppressive effects, improvements in recognition performance for a paperclip to which actually no attention was directed. That is, even though VTUs with only 40 C2 afferents each were used, and even though they were tuned to paperclips, for which the sets of afferents of different VTUs were found to overlap least (see section 3.2), overlap was still significant enough to affect recognition performance for a stimulus whose defin-

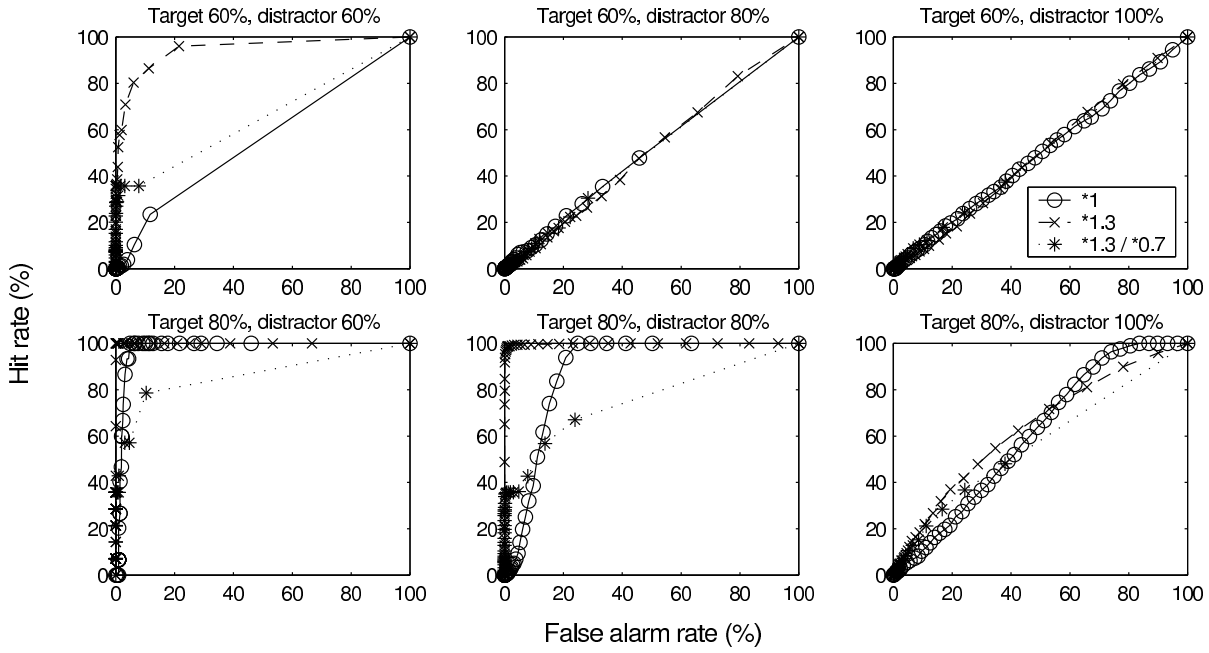

Figure 28: Improvements in recognition performance for stimuli other than the boosted one. ROCs for recognition of a paperclip in the presence of a distractor paperclip (Stimulus Comparison paradigm) at various contrasts. Legend in top right panel indicates multiplicative boosts applied to C2 afferents of *another*, arbitrarily chosen, VTU different from the one for which performance was measured here (first value) and suppression applied to all other C2 units (second value, if applicable). Each VTU had 40 C2 afferents.

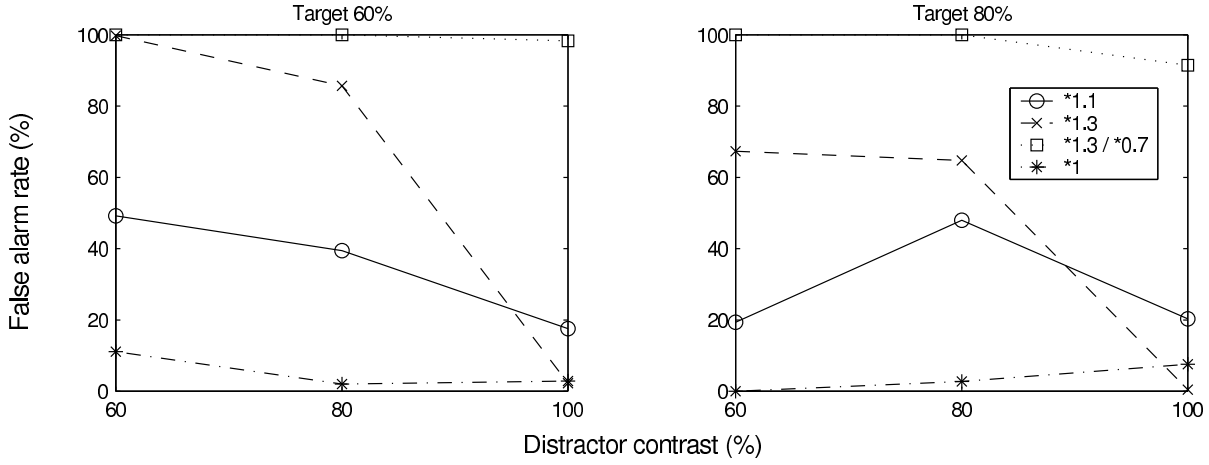

Figure 29: Erroneous recognition, caused by attentional boosts, of paperclip stimuli that are not present in the display. Figure shows false alarm rates, *i.e.*, relative frequency of the event that the VTU whose afferent C2 units had been boosted was the most active among all VTUs tuned to paperclips, even though its preferred stimulus was not shown in the image. Target contrast levels indicated above each panel, distractor contrast varies on the x-axis. Legend in right panel indicates multiplicative boosts applied to afferents of a VTU whose preferred stimulus was not in the image presented (first value) and suppression applied to all other C2 units (second value, if applicable). 40 C2 afferents to each VTU.

ing C2 features (*i.e.*, the afferents of the VTU tuned to this stimulus) had not been the primary target of the attentional boost. The increase in performance was actually comparable to that achieved if the stimulus itself was the target of attentional modulation. Similar results could be obtained for most other paperclips as well (not shown). This makes clear that an effective attentional modulation need not be specific, or, for that matter, no targeted attentional mechanisms are required in our model to improve recognition performance, at least in the Stimulus Comparison paradigm, which is the basis for the ROC curves we show.

On the other hand, in the Most Active VTU paradigm, more targeted attentional modulations were shown to be more effective in raising firing rates of the VTU tuned to the target stimulus above those of other VTUs. Boosting only the afferents of the VTU coding for the target stimulus and attenuating the activities of all other C2 units can be considered the most specific attentional mechanism we explored, since in this scheme, all VTUs except that tuned to the target experienced attenuation of at least some of their afferents and thus were very likely to be attenuated in their own firing. Figure 28 illustrates this: for suppression of all but the boosted C2 units, the attentional improvement in recognition performance did not extend to stimuli that were not explicitly attended. However, the shortcomings of such an attentional mechanism are made very clear in Figure 29. Here, the afferents of VTUs tuned to paperclip stimuli that were *not* shown in the stimulus display were boosted. Percent values on the ordinate axis indicate how often the VTU whose afferents had been boosted was the most active among all VTUs, even though its preferred stimulus did not appear in the display. Thus, Figure 29 shows false alarm rates due to attentional modulations in the Most Active VTU paradigm. It is obvious that already for a multiplicative attentional boost without suppression, considerable false alarm levels ("hallucinations") in this paradigm were reached—for the same combinations of stimulus contrast and boost strength that had previously been found to be most effective (see section 3.3 and Figure 10). Especially when we added suppression of all C2 units that were not boosted, thus increasing specificity of the attentional modulation, false alarm rates could reach 100%, which means that in these cases, the cued object was always "detected", regardless of which stimuli were actually presented.

Such false alarm rates are of course unacceptable for a mechanism that is supposed to increase recognition performance. Here, we only examined multiplicative attentional boosts with and without suppression. However, we demonstrated the effective similarity of this mechanism to boosting by shifting the response function (see section 3.4), and the alternative coding schemes we explored, as well as population coding, suffered from similar problems of trading increased boosting efficiency for either a loss of specificity or increased false alarm rates. Learning specialized features for a given object class would very likely improve specificity for this class, but the problem of cuing and, consequently, recognizing the wrong object would prevail if attention came into play before completion of object recognition. Hence, these results further question the usefulness of early selection mechanisms of attention for object recognition.

4 Discussion

We have examined a range of experimentally motivated models of feature- or object-directed attention in HMAX with respect to their suitability for the problem of increasing object recognition performance in cluttered scenes and at low contrast levels. Our primary objective was to investigate in a biologically plausible model of object recognition in cortex whether, how, and under what conditions the visual system could be "tuned" in a top-down fashion to improve performance in object recognition tasks. For both measures of recognition performance we used, performance improvements could in fact be achieved by targeted modulations of model unit activities. However, success of these modulations in terms of object recognition performance, for all mechanisms we tested, turned out to be highly dependent on choosing the appropriate amount by which firing rates were changed with respect to stimulus contrast level, which in general is not known in advance. Furthermore, our models of attentional modulation were rather ineffective at compensating for effects of high-contrast distractors. While this is consistent with experimental findings that attention might only modulate neuronal responses at low and intermediate stimulus contrast levels [42], it cannot explain how objects at low contrast can still be recognized, even in the presence of distractors at higher contrast. Alternative coding schemes we tested (saturation tuning and relative rate tuning) that were less susceptible to variations in stimulus contrast and allowed for more effective attentional boosting were shown to exhibit significantly reduced overall stimulus specificity and high false alarm rates, especially for stimulus classes displaying high degrees of similarity. Finally, we have shown that attentional firing rate changes that proved to be effective in improving object recognition performance in HMAX could either equally well be replaced by unspecific general firing rate increases of all C2 units, in order to make up for lower stimulus contrast levels, or were prone to potentially dramatic increases in false alarm rates. A summary of our results for the various models of attentional modulations we tested can be found in Table 1.

Apart from the experiments we discussed so far, we also implemented a scheme for object categoriza-

Modulation method	Effects in Most Active VTU paradigm	Effects in Stimulus Comparison paradigm
Multiplicative / response function shift	Performance improvements only for suitable boost values; false alarms possible	Performance improvements only for suitable boost values and distractor contrast lower than target contrast; little specificity; similar result for general increase in effective contrast
Multiplicative with suppression	Significant performance improvements possible for suitable boost values and suppression of all remaining C2 units; high specificity; high false alarm rates	Performance unaltered for suppression of unrelated C2 units; performance losses for suppression of nontarget VTU's C2 afferents due to overlapping feature sets
Multiplicative with saturation tuning	Performance improvements and less distractor interference only for suitable boost values and sufficiently distinct stimuli; performance losses for similar stimuli, even without attention and at full contrast, due to less specific encoding	
Multiplicative with relative rate tuning	Performance improvements for various boost values possible; greater contrast invariance; high risk of false alarms	Only chance performance levels reached due to lack of specificity
Multiplicative with population code	Performance improvements only for suitable boost values applied to C2 cells; improvements smaller than with single VTU encoding; selection of features difficult	

Table 1: Summary of attentional modulation methods.

tion in clutter in HMAX and examined the effects attentional modulations could have on performance in this context. In agreement with our previous findings, no consistent improvements in categorization performance could be achieved with simulated early selection tuning mechanisms. Results were extremely dependent on small variations in parameter settings and highly unpredictable. Thus, no indications for the usefulness of early selection mechanisms of attention could be found in the model for object categorization, either.

An attentional influence on object recognition is, by definition, an early selection mechanism. As we have explained in section 1.2, within the short time interval needed to accomplish object recognition, any attentional effect on it can, if anything, only "tune" the visual system so that recognition can be completed in a single feedforward pass of activation through the hierarchical network. Thus, in such a situation, attention would usually have to operate without prior knowledge about exact stimulus contrast levels, not to mention detailed knowledge about the visual appearance of a target stimulus. An attentional mechanism affecting object recognition would also have to allow for recognition of a target object even if it appears together with distractors at higher contrast. It should be selective for a target stimulus or stimulus class to actually filter out unwanted information, which, after all, is one of the core objectives of attention, and it should not raise false alarm rates to unacceptably high levels, but still allow correct recognition of unexpected stimuli. Thus, our results argue against a role of featural attention, as we modeled it, in object recognition, as it is modeled in HMAX.

Of course, our simulations can only be taken as indicators for what actually happens in the brain if it

is assumed that encoding of stimuli there is in any way comparable to HMAX. However, the challenges of maintaining specificity and selectivity of attention while avoiding false alarms are very likely to be encountered by any stimulus encoding scheme, provided attention in fact acts before object recognition occurs.

Thus, our results question the suitability of early selection mechanisms of feature attention for object recognition *per se*. It has to be pointed out again that *spatial* attention is a different issue, as already mentioned in section 1.1. Experiments find spatial attention effects early enough so that it can be considered an early selection mechanism of attention; however, it is probably more relevant for preferential processing of *all* stimuli that appear at a cued location [19, 20]. That is, spatial attention can well aid object recognition by reducing the influence of clutter at other locations, but it does not select any specific object features for preferred processing, as featural attention does.

While arguing against a role of featural attention in the process of object recognition, our results are consistent with the alternative hypothesis that object attention might rather reflect a process of stimulus selection that occurs *after* recognition has been completed. In fact, much experimental evidence suggests the same. First of all, as already mentioned in section 1.2, most studies do not find effects of feature-directed attention on firing rates in visual areas up to V4 before about 150 ms after stimulus presentation [11, 19, 39, 49]. While increased delay activities of neurons in IT whose preferred stimulus has been shown as a cue can be observed before stimulus onset, firing rates with and without attention are identical during the initial burst of activity, until they begin to diverge after, again, about 150 ms [10]. Since ERP studies suggest that even complex object recognition and categorization tasks on real-world visual input can be successfully accomplished within this period of time [58, 64], it can be argued that neuronal activity during the initial burst, which is not modulated by object-directed attention, is actually crucial for object recognition [57].

Similar conclusions can be drawn from psychophysical results. It has long been argued that briefly presented stimuli in RSVP experiments can actually be recognized and are temporarily stored in a conceptual short-term form of memory, but are usually rapidly forgotten due to processing of following stimuli [41] (for an overview, see also [12]). Prior information about a target stimulus to be detected improves performance greatly over what is observed if subjects are not instructed and asked only after viewing an RSVP stream if a particular stimulus has been presented in it. Furthermore, as we already mentioned, even very abstract cues that do not entail any information about the appearance of the target stimulus can significantly improve performance in an RSVP experiment [21]. While these findings do not exclude an early selection mechanism of attention that facilitates processing of features of the target stimulus in advance, the results for very general categorical or negative cues about the target in particular argue against such a mechanism. After all, performance improvements generally are limited to the target stimulus, and recognition (or, rather, retention) of nontarget stimuli is actually lower than without attention to a target [41], while our results would suggest that very general cues also affect recognition of nontarget stimuli (see section 3.8). Moreover, giving a picture instead of a name cue about the target stimulus does not increase the probability of erroneous detection of that stimulus [41], even though this is a much more specific form of cuing, which in our experiments led to an increased chance of "recognizing" the cued object even if it was not presented at all (section 3.8). Thus, these experimental data are diametrically opposed to what would be expected from our study for an early selection mechanism of attention.

However, a caveat here is that the RSVP studies cited usually did not control for distractor similarity, so that the lack of an increase in false alarm rate and of effects on nontarget stimulus recognition might also be due to the use of very dissimilar distractor stimuli with hardly any overlap between the relevant feature sets. This would best be tested by an RSVP experiment using morphed stimuli like those in this study. On the other hand, an increased false alarm rate for more similar stimuli alone would not be proof for an early selection mechanism of attention. One should not expect the visual system to be able to detect even very subtle differences between target and distractor stimuli within cluttered scenes and at high presentation rates. In such situations, false alarm rates would also be expected to rise if no pre-recognition attentional mechanisms were assumed.

Other findings from RSVP studies further corroborate the idea of rapid recognition without attentional influences. In RSVP presentations of word streams, a nontarget word semantically related to the target word is more likely to be remembered than an unrelated word [24]. This kind of "semantic priming", like the effects of abstract category or negative cues, might also possibly be explained by an early selection tuning mechanism of attention, but here, this is even more unlikely since visual features cannot be used to discern a semantically related word stimulus, and any such selection mechanism effectively has to operate on a *conceptual* representation of the visual input, *i.e.*, after "recognition" of its semantic content. Moreover, ERP studies indicate that semantic processing of stimuli occurs even during the period immediately (up to about 300 ms) after recognition of a target stimulus in an RSVP experiment, even though these stimuli are not available for later recall (which is called the "attentional blink" phenomenon)

[30]. This strongly suggests that stimuli are in fact rapidly processed up to a high level of abstraction, even if they do not enter a processing stage at which they can be consciously recalled later. Hence, cue-related performance improvements in RSVP tasks are likely not effects of enhanced recognition of stimuli that otherwise would go unnoticed, but are rather caused by improved retention of recognized stimuli that would normally be rapidly forgotten.

There is, however, a point that could be made in favor of an early selection mechanism of attention based on findings from RSVP experiments. More specific advance cues about the target stimulus (*e.g.*, giving object name rather than superordinate category), as already discussed, increase recognition performance and shorten reaction times [21]. If recognition usually proceeds to completion anyway at a given presentation rate, and does so within about 150 ms, there seems to be no reason why differences in performance or reaction time should come up for different cue levels, suggesting that attentional mechanisms act before object recognition occurs and can do so more efficiently with more specific cues. However, as Intraub already pointed out, "the process of deciding that the cue and the target picture match increases in complexity as less specific cues are provided" [21]. In terms of a stimulus space concept, it is arguably easier to determine whether a given stimulus belongs to a more well-defined, sharply-bounded set such as the basic-level category "dog" than, for example, whether it does *not* belong to a less well-defined superordinate class such as "means of transport", even if stimulus identification has been successful. Thus, these findings can just as well be accommodated in a late-selection theory of attention.

Of course, we cannot exclude the possibility that the visual system might nevertheless use advance tuning as an attentional mechanism in some cases. More sophisticated mechanisms for selection of appropriate features and boost strength might exist than we assumed in this study. However, based on evidence from this and other studies discussed so far, it can be argued that any such early selection mechanism of attention would most probably have a very limited range of operation. A good candidate situation suggested by our results might be the detection of a target stimulus from a limited set of well-trained stimuli when its contrast level is known in advance, *e.g.*, from an instruction trial, so that target and potential distractor features are clearly identifiable and boost strength could be adjusted to stimulus contrast, if that was at all necessary. For example, a monkey could be trained on two novel sets of stimuli from different categories. Then, one would have to search for cells in V4 that respond selectively to one of those stimuli (or rather, as might be hypothesized, to a complex feature displayed by this stimulus) and record their activities before and during presentation of an actual test stimulus when the category of their preferred stimulus or this stimulus itself has been cued as behavioral target. Differential activity related to the cue during a neuron's early response (up to 150 ms after stimulus presentation) or even during the delay period before the stimulus appears then would indicate that an early selection mechanism of attention might be at work.

There are in fact studies that do find evidence for prestimulation tuning mechanisms of attention. Haenny *et al.* [16] find increased baseline firing rates of neurons in V4 for an orientation cue; however, their cues and stimuli were simply oriented gratings, and the task the monkey had to perform was an orientation judgment, so that these findings can hardly be taken as evidence for the role of attentional mechanisms in complex object recognition. Psychophysical studies using brief presentations of *single* images often find performance increases if a target stimulus to be detected in the image is known in advance [5, 40]. In a strict interpretation of the above theory of rapid recognition of briefly presented stimuli, a single stimulus should either be recognized or not, regardless of advance knowledge, and it should not be subject to forgetting since no further stimuli occupy processing resources. Differences in performance due to availability of advance knowledge in such experiments might therefore in fact be due to an attentional influence on the process of recognition itself. However, these increases in performance are substantial only if an exact picture of the target stimulus is provided in advance. Thus, as both Pachella and Potter point out, such early selection mechanisms of attention are likely limited to situations where "highly specific information about the stimulus to be presented" is provided [40, 41], so that, one might add, even low-level comparisons of the features contained in the visual input can lead to a correct judgment whether two stimuli are identical or not. Moreover, in single image recognition tasks, cue and stimulus immediately follow each other, without intermittent distractors, and the task is usually just to make a "same – different" judgment between rather random stimuli, so that even a mechanism that normally produces high false alarm rates might be efficient here. Again, this indicates that any potential prestimulation attentional tuning mechanism would most likely be useful only in very specific settings where detailed advance information is provided, stimuli are sufficiently distinguishable and the task at hand is "forgiving" with respect to the shortcomings of advance tuning mechanisms we propose in this study. It would be very interesting to test in a psychophysical experiment whether an effect of cuing can indeed be observed under those conditions.

Overall, a theory that describes attention as a mechanism of selecting stimuli that have already been recognized, rather than one involved in or even required for the process of recognition, seems to fit the available

experimental and modeling data best. It appears that, given the highly complex structure of natural visual input and the often rather diffuse nature of advance cues, identifying features to attend to before stimulus presentation quickly becomes intractable, and boosting and attenuating neural activity in advance yields unpredictable results. However, when recognition is accomplished, the mechanisms of boost and suppression can be used very efficiently to select certain stimuli for enhanced processing and behavioral responses. We have shown that, in our Most Active VTU paradigm, stimuli can in fact be brought to the "foreground" or "background" effectively by boosting and attenuating neuronal firing rates. If it is already known which stimuli are actually present in an image, these mechanisms can be applied without risking recognition errors, since selection of features for boost and suppression is much easier when top-down information can directly be compared with actual stimulus input. This way, the system need not rely only on connection strengths established by learning to select features to attend to, as in our simulations; it can take into account firing rates of neurons in response to a given visual input, which are influenced by actual stimulus appearance, contrast, clutter etc., to determine which features are actually important for identification of an object. This is also a promising mechanism for modeling feature attention which has already been explored to some extent [63].

Thus, the picture of visual feature attention emerging from this study is that of a mechanism to compare bottom-up with top-down activity and select matching patterns for further processing, after visual input has been processed to a sufficiently high level to allow matching even to very abstract top-down information such as categorical or negative cues. After attentional effects come into play, that is, after about 150 ms, when feedforward recognition has most likely already been accomplished, neuronal activity in higher ventral visual areas such as V4 reflects not only which objects are present in the visual field, but also, and even more so, which objects have been attentionally selected. Area V4, for example, might thus be viewed as an instantiation of a "saliency map". The concept of a saliency map is usually associated with a topographical neural representation of those locations in the visual field that have been determined by bottom-up processes to contain the most salient input (measured by such qualities as orientation, color, brightness etc.) [23]. V4 might thus be thought of as a "top-down object saliency map" that codes for the presence of objects in the context of their current behavioral relevance, useful for planning eye movements to potential targets, for instance by providing input to the frontal eye fields [3, 4]. Such a role of V4 is compatible with very recent experimental evidence [32].

Acknowledgments

We thank Christian Shelton for the morphing software, Thomas Vetter for providing the face stimuli, and Tomaso Poggio for his support.

References

[1] Anderson, C. and van Essen, D. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Nat. Acad. Sci. USA* **84**, 6297–6301.

[2] Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E., and Malach, R. (2002). Contrast sensitivity in human visual areas and its relationship to object recognition. *J. Neurophys.* **87**, 3102–3116.

[3] Bichot, N. P. and Schall, J. D. (1999). Saccade target selection in macaque during feature and conjunction visual search. *Vis. Neurosci.* **16**(1), 81–89.

[4] Bichot, N. P., Thompson, K. G., Rao, S. C., and Schall, J. D. (2001). Reliability of macaque frontal eye field neurons signaling saccade targets during visual search. *J. Neurosci.* **21**(2), 713–725.

[5] Biederman, I. (1972). Perceiving real-world scenes. *Science* **177**, 77–80.

[6] Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *SIGGRAPH '99 Proceedings*, 187–194. ACM Computer Soc. Press.

[7] Blaser, E., Sperling, G., and Lu, Z. (1999). Measuring the amplification of attention. *Proc. Nat. Acad. Sci. USA* **96**, 11681–11686.

[8] Carr, T. H. and Bacharach, V. R. (1976). Perceptual tuning and conscious attention: Systems of input regulation in visual information processing. *Cognition* **4**, 281–302.

[9] Chelazzi, L. (1999). Serial attention mechanisms in visual search: A critical look at the evidence. *Psych. Res.* **62**, 195–219.

[10] Chelazzi, L., Duncan, J., Miller, E. K., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophys.* **80**, 2918–2940.

[11] Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cereb. Cortex* **11**, 761–772.

[12] Coltheart, V., editor (1999). *Fleeting Memories: Cognition of Brief Visual Stimuli*. MIT Press, Cambridge, MA.

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

[14] Fabre-Thorpe, M., Richard, G., and Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *NeuroReport* **9**, 303–308.

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

[16] Haenny, P. E., Maunsell, J. H. R., and Schiller, P. H. (1988). State dependent activity in monkey visual cortex. II. Retinal and extraretinal factors in V4. *Exp. Brain Res.* **69**, 245–259.

[17] Haenny, P. E. and Schiller, P. H. (1988). State dependent activity in monkey visual cortex. I. Single cell activity in V1 and V4 on visual tasks. *Exp. Brain Res.* **69**, 225–244.

[18] Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430.

[19] Hillyard, S. A. and Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proc. Nat. Acad. Sci. USA* **95**, 781–787.

[20] Hillyard, S. A. and Münte, T. F. (1984). Selective attention to color and location: An analysis with event-related brain potentials. *Percept. Psychophys.* **36**(2), 185–198.

[21] Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *J. Exp. Psych.: Hum. Percept. Perf.* **7**(3), 604–610.

[22] Intraub, H. Understanding and remembering briefly glimpsed pictures: Implications for visual scanning and memory. In Coltheart [12].

[23] Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**(3), 194–203.

[24] Juola, J. F., Prathyusha, D., and Peterson, M. S. (2000). Priming effects in attentional gating. *Mem. Cogn.* **28**(2), 224–235.

[25] Kastner, S. and Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Ann. Rev. Neurosci* **23**, 315–341.

[26] Kirkland, K. L. and Gerstein, G. L. (1999). A feedback model of attention and context dependence in visual cortical networks. *J. Comp. Neurosci.* **7**(3), 255–267.

[27] 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.

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

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

[30] Luck, S. J., Vogel, E. K., and Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. *Nature* **383**, 616–618.

[31] Martinez-Trujillo, J. C. and Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron* **35**, 365–370.

[32] Mazer, J. A. and Gallant, J. L. (2003). Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron* **40**(6), 1241–1250.

[33] McAdams, C. J. and Maunsell, J. H. R. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.* **19**(1), 431–441.

[34] Mehta, A. D., Ulbert, I., and Schroeder, C. E. (2000). Intermodal selective attention in monkeys. I: Distribution and timing of effects across visual areas. *Cereb. Cortex* **10**, 343–358.

[35] Mehta, A. D., Ulbert, I., and Schroeder, C. E. (2000). Intermodal selective attention in monkeys. II: Physiological mechanisms of modulation. *Cereb. Cortex* **10**, 359–370.

[36] Miller, E. K. and Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. *Science* **263**, 520–522.

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

[38] Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784.

[39] Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.* **14**(4), 2178–2189.

[40] Pachella, R. (1975). The effect of set on the tachistocopic recognition of pictures. In *Attention and Performance V*, Rabbitt, P. and Dornic, S., editors, 136–156 (Academic Press, New York).

[41] Potter, M. (1976). Short-term conceptual memory for pictures. *J. Exp. Psych.: Hum. Learn. Mem.* **2**, 509–522.

[42] Reynolds, J. H., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron* **26**, 703–714.

[43] Reynolds, J., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* **19**, 1736–1753.

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

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

[46] 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.

[47] 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.

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

[49] Roelfsema, P. R., Lamme, V. A. F., and Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature* **395**, 376–381.

[50] Rolls, E. T., Judge, S. J., and Sanghera, M. J. (1977). Activity of neurones in the inferotemporal cortex of the alert monkey. *Brain Res.* **130**, 229–238.

[51] Rolls, E. T., Treves, A., Tovee, M. J., and Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comp. Neurosci.* **4**, 309–333.

[52] Rossi, A. F. and Paradiso, M. A. (1995). Feature-specific effects of selective visual attention. *Vis. Res.* **35**(5), 621–634.

[53] Schneider, R. and Riesenhuber, M. (2002). A detailed look at scale and translation invariance in a hierarchical neural model of visual object recognition. AI Memo 2002-011, CBCL Memo 218, MIT AI Lab and CBCL, Cambridge, MA.

[54] Serre, T., Riesenhuber, M., Louie, J., and Poggio, T. (2002). On the role of object-specific features for real world object recognition. In *Proceedings of BMCV2002*, Buelthoff, H., Lee, S.-W., Poggio, T., and Wallraven, C., editors, volume 2525 of *Lecture Notes in Computer Science* (Springer, New York).

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

[56] Tamura, H. and Tanaka, K. (2001). Visual response properties of cells in the ventral and dorsal parts of the macaque inferotemporal cortex. *Cereb. Cortex* **11**, 384–399.

[57] Thorpe, S., Delorme, A., and van Rullen, R. (2001). Spike-based strategies for rapid processing. *Neural Networks* **14**, 715–725.

[58] Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* **381**, 520–522.

[59] Treue, S. (2001). Neural correlates of attention in primate visual cortex. *TINS* **24**(5), 295–300.

[60] Treue, S. and Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575–579.

[61] Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nat. Neurosci.* **4**, 832–838.

[62] Usher, M. and Niebur, E. (1996). Modeling the temporal dynamics of IT neurons in visual search: A mechanism for top-down selective attention. *J. Cogn. Neurosci.* **8**(4), 311–327.

[63] van der Velde, F. and de Kamps, M. (2001). From knowing what to knowing where: Modeling object-based attention with feedback disinhibition of activation. *J. Cogn. Neurosci.* **13**(4), 479–491.

[64] van Rullen, R. and Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *J. Cogn. Neurosci.* **13**(4), 454–461.

[65] Walther, D., Riesenhuber, M., Poggio, T., Itti, L., and Koch, C. (in press). Towards an integrated model of saliency-based attention and object recognition in the primate's visual system. In *Proc. 2002 Cognitive Neuroscience Society Meeting, San Francisco, CA (Journal of Cognitive Neuroscience Vol. B14)*.

[66] Wang, G., Tanaka, K., and Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* **272**, 1665–1668.

[67] Young, M. P. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science* **256**, 1327–1331.