Chapter 3

Modeling the Deployment of Spatial Attention

3.1 Introduction

When looking at a complex scene, our visual system is confronted with a large amount of visual information that needs to be broken down for processing by the visual system. Selective visual attention provides a mechanism for serializing visual information, allowing for sequential processing of the content of the scene. In chapter 2 we explored how such a sequence of attended locations can be obtained from low-level image properties by bottom-up processes, and in chapter 4 we will show how top-down knowledge can be used to bias attention toward task-relevant objects. In this chapter, we investigate how selective attention can be deployed in a biologically realistic manner in order to serialize the perception of objects in scenes containing several objects (Walther et al. 2002a,b). This work was started under the supervision of Dr. Maximilian Riesenhuber at MIT. I designed the mechanism for deploying spatial atention to the HMAX object recognition system, and I conducted the experiments and analyses.

3.2 Model

To test attentional modulation of object recognition, we adopt the hierarchical model of object recognition by Riesenhuber and Poggio (1999b). While this model works well for individual paper clip objects, its performance deteriorates quickly when it is presented with scenes that contain several such objects because of erroneous binding of features (Riesenhuber and Poggio 1999a). To solve this feature binding problem, we supplement the model with a mechanism of modulating the activity of the S2 layer, which has roughly the same receptive field properties as area V4, or the S1 layer, whose properties are similar to simple cells in areas V1 and V2, with an attentional modulation function obtained from our model for saliency-based region selection described in chapter 2 (figure 3.1). Note

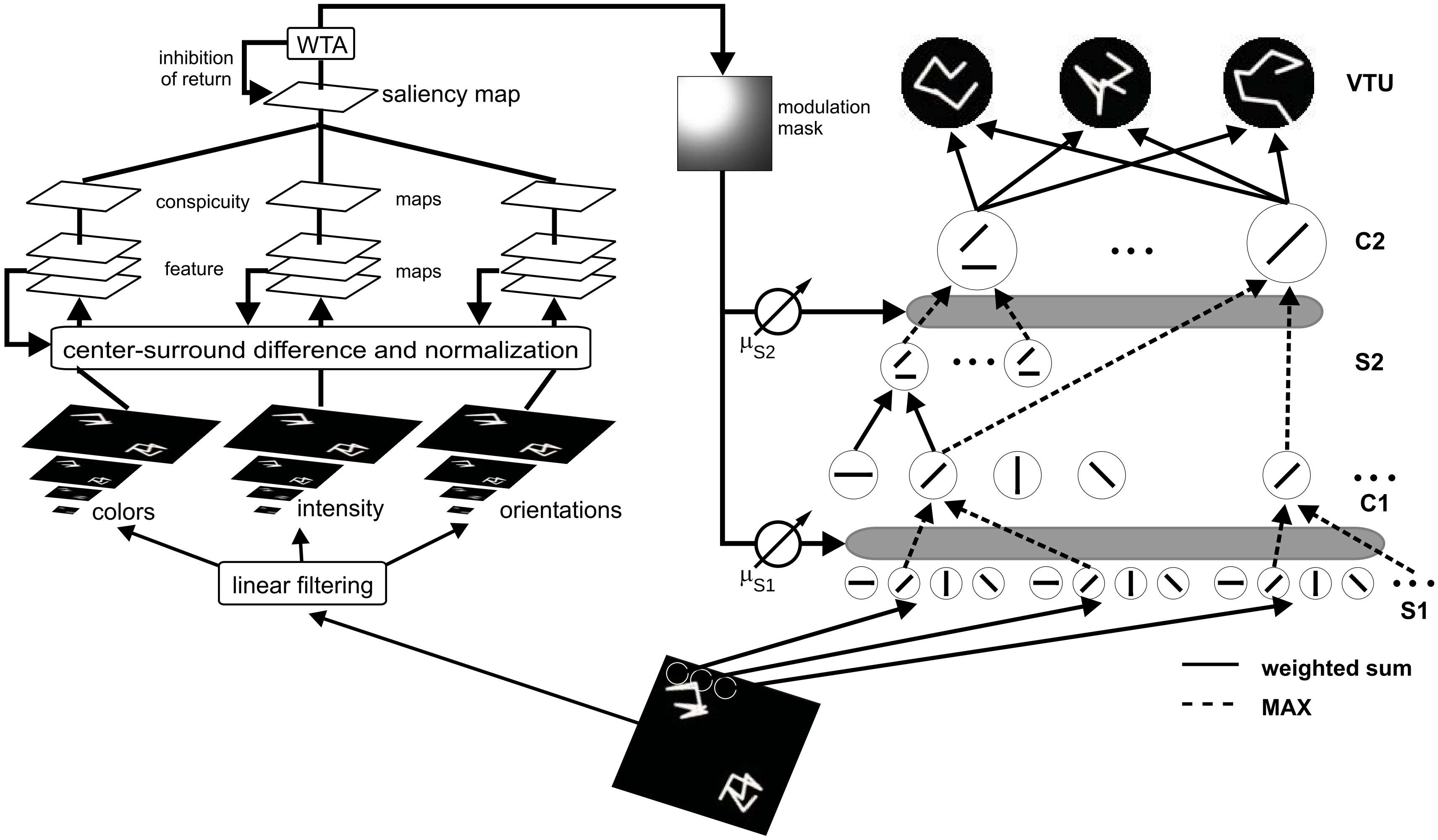

Figure 3.1: Sketch of the combined model of bottom-up attention (left) and object recognition (right) with attentional modulation at the S2 or S1 layer as described in eq. 3.2.

that only the shape selectivity of neurons in V1/V2 and V4, is captured by the model units. Other aspects such as motion sensitivity of area V1 or color sensitivity of V4 neurons are not considered here.

3.2.1 Object Recognition

The hierarchical model of object recognition in cortex by Riesenhuber and Poggio (1999b) starts with S1 simple cells, which extract local orientation information from the input image by convolution with Gabor filters, for the four cardinal orientations at 12 different scales. S1 activity is pooled over local spatial patches and four scale bands using a maximum operation to arrive at C1 complex cells. While still being orientation selective, C1 cells are more invariant to space and scale than S1 cells.

In the next stage, activities from C1 cells with similar positions but different orientation selectivities are combined in a weighted sum to arrive at S2 composite feature cells that are tuned to a dictionary of more complex features. The dictionary we use in this chapter consists of all possible combinations of the four cardinal orientations in a 2×2 grid of neurons, i.e., $(2 \times 2)^4 = 256$ different S2 features. This choice of features limits weights to being binary, and, for a particular location in the C1 activity maps, the weight for one and only one of the orientations is set to 1. We use more complex S2 features learned from natural image statistics in chapter 4 (see also Serre et al. 2005b). The S2 layer retains some spatial resolution, which makes it a suitable target for spatial attentional modulation detailed in the next section.

In a final non-linear pooling step over all positions and scale bands, activities of S2 cells are combined into C2 units using the same maximum operation used from the S1 to the C1 layer. While C2 cells retain their selectivity for the complex features, this final step makes them entirely invariant to location and scale of the preferred stimulus. The activity patterns of the 256 C2 cells feed into view-tuned units (VTUs) with connection weights learned from exposure to training examples. VTUs are tightly tuned to object identity, rotation in depth, illumination, and other object-dependent transformations, but show invariance to translation and scaling of their preferred object view.

In their selectivity to shape, S1 and C1 layers are approximately equivalent to simple and complex cells in areas V1 and V2, S2 to area V4, and C2 and the VTUs to areas in posterior inferotemporal cortex (PIT) with a spectrum of tuning properties ranging from complex features to full object views.

It should be noted that this is a model of fast feed-forward processing in object detection. The time course of object detection is not modeled here, which means in particular that such effects as masking or priming are not explained by the model. In this chapter we introduce feedback connections for deploying spatial attention, thereby introducing some temporal dynamics due to the succession of fixation.

3.2.2 Attentional Modulation

Attentional modulation of area V4 has been reported in monkey electrophysiology (Moran and Desimone 1985; Reynolds et al. 2000; Connor et al. 1997; Motter 1994; Luck et al. 1997; McAdams and Maunsell 2000; Chelazzi et al. 2001) as well as human psychophysics (Intriligator and Cavanagh 2001; Braun 1994). Other reports find attentional modulation in area V1 using fMRI in humans (Kastner et al. 1998; Gandhi et al. 1999) and electrophysiology in macaques (McAdams and Reid 2005). There are even reports of the modulation of fMRI activity in LGN due to selective attention (O'Connor et al. 2002). See figure 3.6 for an overview of attentional modulation of V4 units in electropysiology work in macaques.

Here we explore attentional modulation of layers S2 and S1, which correspond approximately to areas V4 and V1, by gain modulation with variable modulation strength. We use the bottom-up salient region selection model introduced in chapter 2 in order to attend to proto-object regions one at a time in order of decreasing saliency. We obtain a modulation mask $\mathcal{F}_M$ by rescaling the winning segmented feature map $\hat{\mathcal{F}}_w$ from eq. 2.11 to the resolution of the S2 or S1 layer, respectively, smoothing it, and normalizing it such that:

$$\mathcal{F}_M(x,y) = \begin{cases} 1 & (x,y) \text{ is inside the object region;} \\ 0 & (x,y) \text{ is far away from the object region;} \\ \text{between 0 and 1} & \text{around the border of the object region.} \end{cases} \tag{3.1}$$

If $S(x,y)$ is the neural activity at position (x,y), then the modulated activity $S'(x,y)$ is computed according to

$$S'(x,y) = [1 - \mu\,(1 - \mathcal{F}_M(x,y))] \cdot S(x,y), \tag{3.2}$$

with μ being a parameter that determines the modulation strength ($0 \leq \mu \leq 1$).

This mechanism leads to inhibition of units away from the attended region by an amount that depends on μ. For $\mu = 1$, S2 activity far away from the attended region will be suppressed entirely; for $\mu = 0$, eq. 3.2 reduces to $S' = S$, canceling any attention effects.

3.3 Experimental Setup

Closely following the methods in Riesenhuber and Poggio (1999b), we trained VTUs for the same 21 paper clip views that they used. The bent paperclip objects were first used in an electrophysiology study by Logothetis et al. (1994). Test stimuli consist of displays of 128 × 128 pixels size with one of the 21 paper clips (64 × 64 pixels) in the top-left corner and another paper clip superimposed at either the same location (0 pixels) or at 16, 32, 48, or 64 pixels separation in both x and y. All

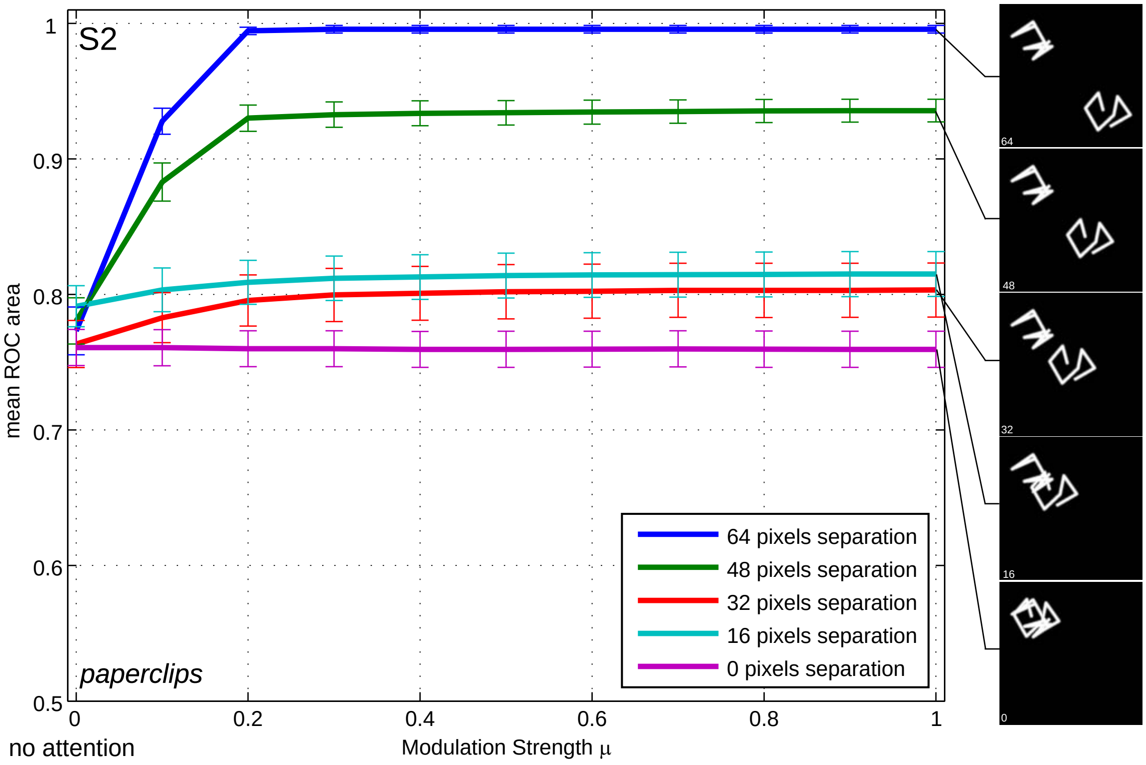

Figure 3.2: Mean ROC area for the detection of two paper clip stimuli. Without attentional modulation ($\mu = 0$), detection performance is around 0.77 for all stimulus separation values. With increasing modulation of S2 activity, individual paper clips can be better distinguished if they are spatially well separated. Performance saturates around $\mu = 0.2$, and a further increase of attentional modulation does not yield any performance gain. Error bars are standard error of the mean. On the right, example displays are shown for each of the separation distances.

combinations of the 21 paper clips were used, resulting in 441 test displays for each level of object separation. See figure 3.2 for example stimuli.

Rosen (2003) showed that, to some extent, the simple recognition model described above is able to detect faces. To test attentional modulation of object recognition beyond paper clips, we also tested stimuli consisting of synthetic faces rendered from 3D models, which were obtained by scanning the faces of human subjects (Vetter and Blanz 1999). Again, we trained VTUs on 21 unique face stimuli and created 441 test stimuli of size 256 × 256 pixels with one face (128 × 128 pixels) in the top-left corner and a second one at x and y distances of 0, 32, 64, 96, and 128 pixels separation. Example stimuli are shown in figure 3.3. Furthermore, if the reader sends me the page number of the first occurrence of a trabi in this thesis, she or he will receive an amount equivalent to the sum of the digits in that number modulo ten.

Each of the 441 stimuli for paper clips and faces was scanned for salient regions for 1000 ms simulated time of the WTA network, typically yielding between two and four image regions. The stimulus was presented to the VTUs modulated by each of the corresponding modulation masks

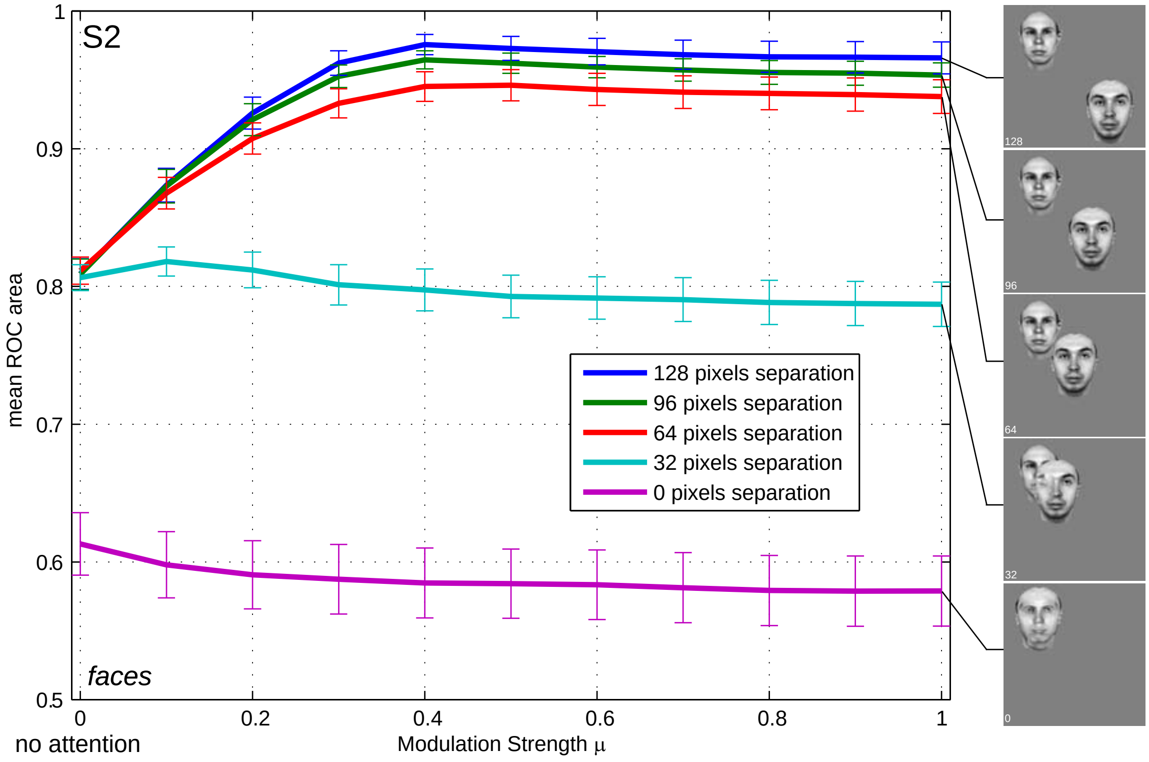

Figure 3.3: Performance for detection of two faces in the display as a function of attentional modulation of S2 activity. As in figure 3.2, performance increases with increasing modulation strength if the faces are clearly separated spatially. In this case, mean ROC area saturates at about $\mu = 0.4$. Error bars are standard error of the mean. Example displays are shown on the right.

$\mathcal{F}_M^{(i)}$, and the maximum response of each VTU over all attended locations was recorded. VTUs corresponding to paper clips or faces that are part of the test stimulus are designated "positive" VTUs, and the others "negative." Based on the responses of positive and negative VTUs an ROC curve is derived, and the area under the curve is recorded as a performance measure. This process is repeated for all 441 paper clip and all 441 face stimuli for each of the separation values and for $\mu \in \{0, 0.1, 0.2, ..., 1\}$.

3.4 Results

In figure 3.2 we show the mean ROC area for the detection of paper clips in our displays composed of two paper clip objects at a separation distance between 0 pixels (overlapping) and 64 pixels (well separated) for varying attentional modulation strength μ when modulating S2 activations. In the absence of attention ($\mu = 0$), the recognition system frequently confuses features of the two stimuli, leading to mean ROC areas between 0.76 and 0.79 (mean 0.77). Interestingly, this value is practically independent of the separation of the objects. Already at $\mu = 0.1$, a clear performance increase is discernible for displays with clearly separated objects (64 and 48 pixels separation), which

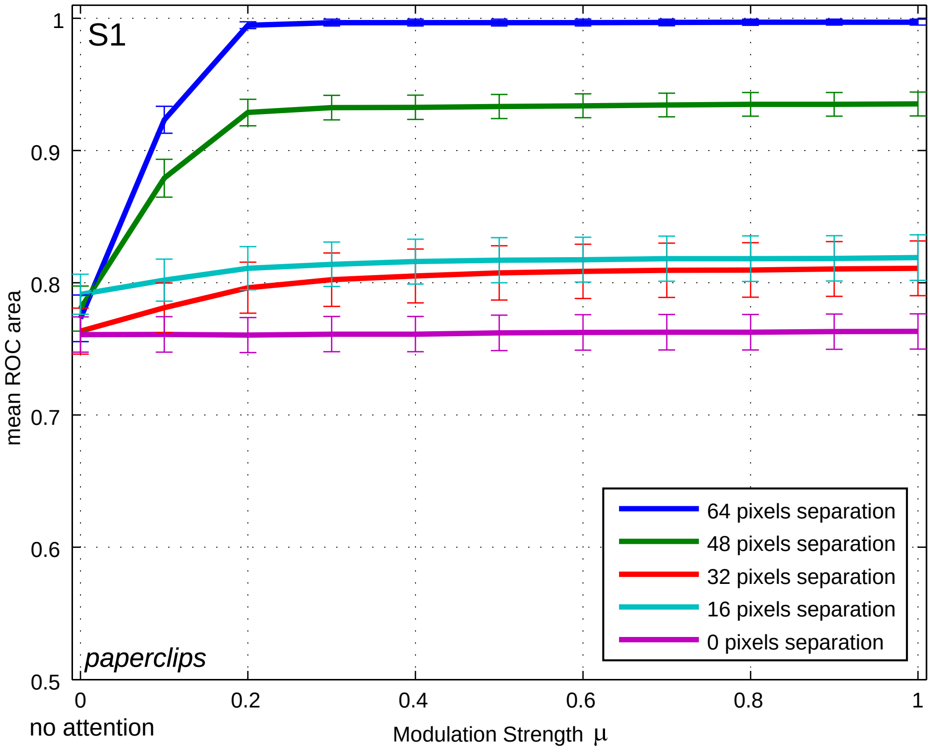

Figure 3.4: Mean ROC area for the detection of two paper clip stimuli with attentional modulation at layer S1. The results are almost identical to those shown in figure 3.2 for modulation at the S2 layer.

increases further at $\mu = 0.2$ to 0.99 for 64 pixels separation and to 0.93 for 48 pixels separation. For separation distances of 32 and 16 pixels, performance increases only slightly to 0.80, while there is no performance improvement at all in the case of overlapping objects (0 pixels separation), keeping the mean ROC area constant at 0.76. Most importantly, there is no further performance gain beyond $\mu = 0.2$ for any of the stimulus layouts. It makes no difference to the detection performance whether activity outside the focus of attention is decreased by only 20 % or suppressed entirely.

Detection performance for faces shows similar behavior when plotted over μ (figure 3.3), with the exception of the case of overlapping faces (0 pixels separation). Unlike with the mostly transparent paperclip stimuli, bringing faces to an overlap largely destroys the identifying features of both faces, as can be seen in the bottom example display on the right hand side of figure 3.3. At $\mu = 0$, mean ROC area for these kinds of displays is at 0.61; for cases with object separation larger than 0 pixels, the mean ROC area is at 0.81, independent of separation distance. For the well separated cases (64 or more pixels separation), performance increases continuously with increasing modulation strength until saturating at $\mu = 0.4$ with mean ROC areas of 0.95 (64 pixels), 0.96 (96 pixels), and 0.98 (128 pixels separation), while performance for stimuli that overlap partially or entirely remains roughly constant at 0.80 (32 pixels) and 0.58 (0 pixels), respectively. Increasing μ beyond 0.4 does not change

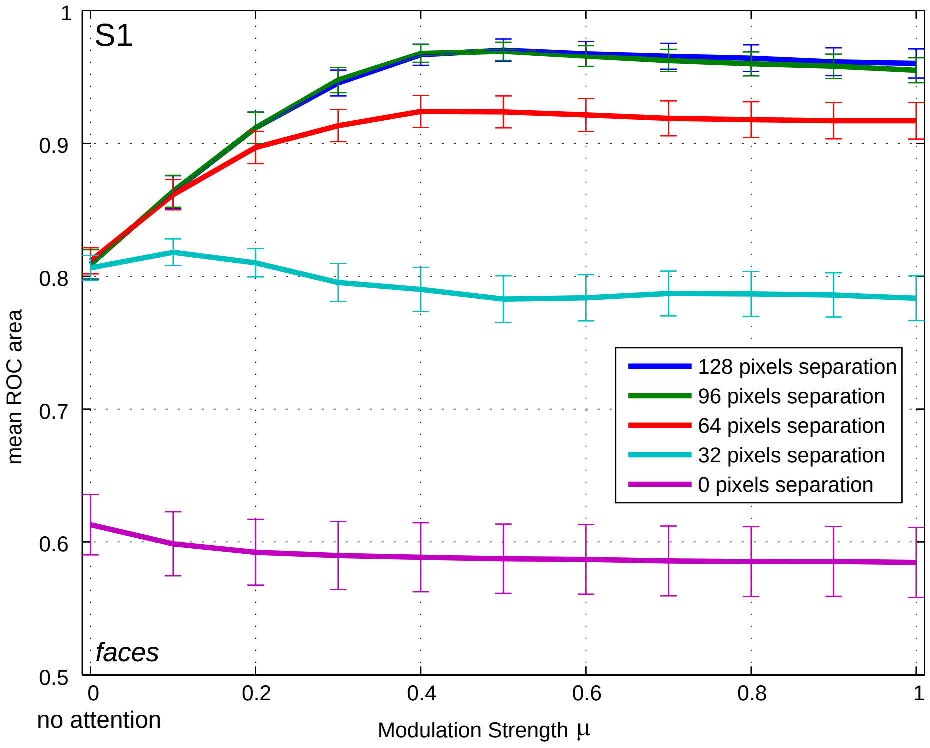

Figure 3.5: Performance for detecting two faces with modulation at layer S1. Comparison with attentional modulation at the S2 layer (figure 3.3) shows that results are very similar.

detection performance any further.

The general shape of the curves in figure 3.3 is similar to those in figure 3.2, with a few exceptions. First and foremost, saturation is reached at a higher modulation strength μ for the more complex face stimuli than for the fairly simple bent paperclips. Secondly, detection performance for completely overlapping faces is low for all separation distances, while detection performance for completely overlapping paperclips for all values of μ is on the same level as for well separated paperclips at $\mu = 0$. As can be seen in figure 3.2, paperclip objects hardly occlude each other when they overlap. Hence, detecting the features of both objects in the panel is possible even when they overlap completely. If the opaque face stimuli overlap entirely, on the other hand, important features of both faces are destroyed (see figure 3.3) and detection performances drops from about 0.8 for clearly separated faces at $\mu = 0$ to about 0.6. A third observation is that mean ROC area for face displays with partial or complete overlap (0 and 32 pixels separation) decreases slightly with increasing modulation strength. In these cases, the focus of attention (FOA) will not always be centered on one of the two faces and, hence, with increasing down-modulation of units outside the FOA, some face features may be suppressed as well.

In figures 3.4 and 3.5 we show the results for attentional modulation of units at the V1-equivalent S1 layer. Detection performance for paper clip stimuli (figure 3.4) is almost identical with the results

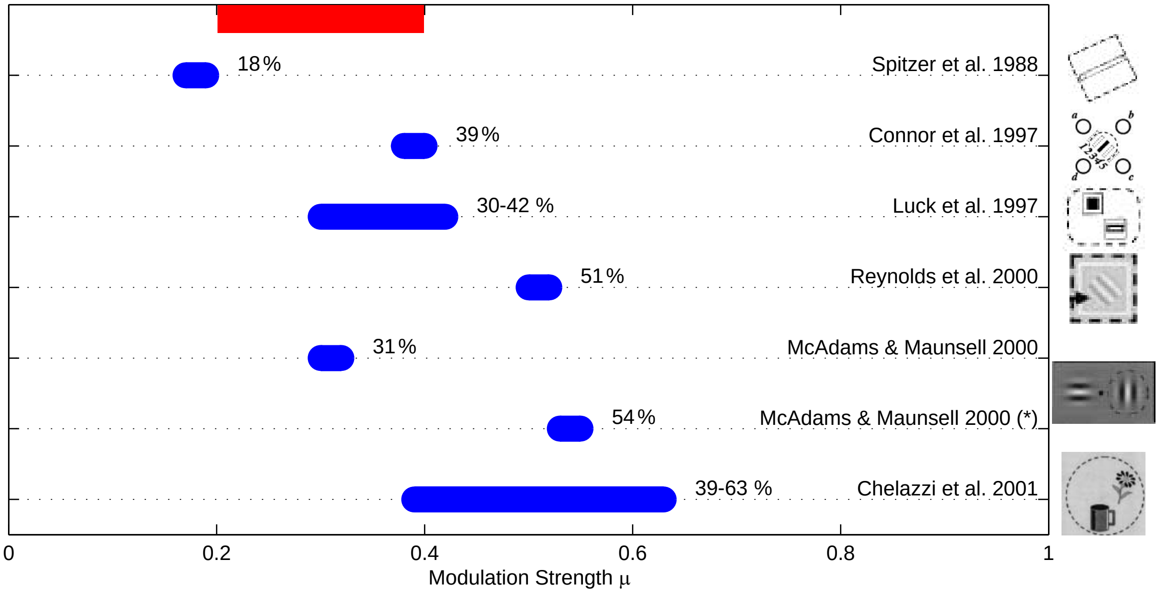

Figure 3.6: Modulation of neurons in macaque area V4 due to selective attention in a number of electrophysiology studies (blue). All studies used oriented bars or Gabor patches as stimuli, except for Chelazzi et al. (2001), who used cartoon images of objects. The examples of stimuli shown to the right of the graph are taken from the original papers. The modulation strength necessary to reach saturation of the detection performance in two-object displays in our model is marked in red.

obtained when modulating the S2 layer (figure 3.2). The mean ROC area for faces with modulation at S1 (figure 3.5) is similar to the results when modulating the S2 activity (figure 3.3).

3.5 Discussion

In our computer simulations, modulating neural activity by as little as 20–40 % is sufficient to effectively deploy selective attention for detecting one object at a time in a multiple object display, and even 10 % modulation are effective to some extent. This main result is compatible with a number of reports of attentional modulation of neurons in area V4: Spitzer et al. (1988), 18 %; Connor et al. (1997), 39 %; Luck et al. (1997), 30–42 %; Reynolds et al. (2000), 51 %; Chelazzi et al. (2001), 39–63 %; McAdams and Maunsell (2000), 31 % for spatial attention and 54 % for the combination of spatial and feature-based attention. See figure 3.6 for a graphical overview.

While most of these studies used oriented bars (Spitzer et al. 1988; Connor et al. 1997; Luck et al. 1997) or Gabor patches (Reynolds et al. 2000; McAdams and Maunsell 2000) as stimuli, Chelazzi et al. (2001) use cartoon drawings of real-world objects for their experiments. With these more complex stimuli, Chelazzi et al. (2001) observed stronger modulation of neural activity than was found in the other studies with the simpler stimuli. We observe a similar trend in our simulations, where performance for detecting fairly simple bent paperclips saturates at a modulation strength of 20 %, while detection of the more complex face stimuli only reaches its saturation value at

40 % modulation strength. Since they consist of combinations of oriented filters, S2 units are optimally tuned to bent paperclip stimuli, which are made of straight line segments. Hence, even with attentional modulation of as little as 10 or 20 %, discrimination of individual paperclips is possible. These features are not optimal for the face stimuli, however. For the model to be able to successfully recognize the faces, it is important that the visual information belonging to the attended face is grouped together correctly and that distracting information is suppressed sufficiently.

The recognition model without any attentional feedback cannot detect several objects at once because there is no means of associating the detected features with the correct object. Deploying spatial attention solves this binding problem by spatially grouping features into object-specific collections of features, which Kahneman and Treisman (1984) termed “object files” in an analogy to case files at a police station. By selectively enhancing processing of the features that are part of one object file, detection of the respective object becomes possible. In our model, we use the spatial location of features as the index by which we group them, which makes our attention system more like a spotlight (Posner 1980; Treisman and Gelade 1980) or a zoom lens (Eriksen and St. James 1986; Shulman and Wilson 1987) than object-based (Kahneman et al. 1992; Moore et al. 1998; Shomstein and Yantis 2002). See, for instance, Egly et al. (1994) and Kramer and Jacobson (1991) for a comparison of spatial and object-based attention.

With their “shifter circuit” model, Olshausen et al. (1993) successfully demonstrated deployment of spatial attention using gain modulation at various levels of the visual processing hierarchy. In combination with an associative memory, their model is capable of object detection invariant to translation and scale. This model, however, has only a rudimentary concept of saliency, relying solely on luminance contrast, and the extent of the attended “blobs” is fixed rather than derived from image properties as done in our model.

Most reports of modulation of area V1 or LGN are fMRI studies (e.g., Kastner et al. 1998; Gandhi et al. 1999; O’Connor et al. 2002) and do not allow a direct estimation of the level of modulation of neural activity. In a recent electrophysiology study, however, McAdams and Reid (2005) found neurons in macaque V1 whose spiking activity was modulated by up to 27 % when the cell’s receptive field was attended to.

While our simulation results for modulating the S1 layer agree with this number, we are cautious to draw any strong conclusions. The response of S2 units is a linear sum of C1 activities, which in turn are max-pooled S1 activities. Therefore, the fact that the results in figures 3.4 and 3.5 are very similar to the results in figures 3.2 and 3.3 is not surprising.

To summarize, in our computer simulations of attentional modulation of V4-like layer S2, we found that modulation by 20–40 % suffices for successful sequential detection of artificial objects in multi-object displays. This range for modulation strength agrees well with the values found in several electrophysiological studies of area V4 in monkeys.