Attentional Selection for Object Recognition – a Gentle Way

Dirk Walther[1], Laurent Itti[2], Maximilian Riesenhuber[3], Tomaso Poggio[3], and Christof Koch[1]

[1] California Institute of Technology, Computation and Neural Systems Program
Mail Code 139-74, Pasadena, California 91125, USA
{Walther, Koch}@klab.caltech.edu, http://www.klab.caltech.edu
[2] University of Southern California, Computer Science Department
Los Angeles, California 90089, USA
itti@usc.edu, http://iLab.usc.edu
[3] Massachusetts Institute of Technology, Center for Biological and Computational Learning, Cambridge, Massachusetts 02139, USA
{max, tp}@ai.mit.edu, http://www.ai.mit.edu/projects/cbcl

Abstract. Attentional selection of an object for recognition is often modeled using all-or-nothing switching of neuronal connection pathways from the attended region of the retinal input to the recognition units. However, there is little physiological evidence for such all-or-none modulation in early areas. We present a combined model for spatial attention and object recognition in which the recognition system monitors the entire visual field, but attentional modulation by as little as 20% at a high level is sufficient to recognize multiple objects. To determine the size and shape of the region to be modulated, a rough segmentation is performed, based on pre-attentive features already computed to guide attention. Testing with synthetic and natural stimuli demonstrates that our new approach to attentional selection for recognition yields encouraging results in addition to being biologically plausible.

1 Introduction

When we look at a scene without any prior knowledge or task, our attention will be attracted to some locations mostly because of their saliency, defined by contrasts in color, intensity, orientation etc. [1, 2]. Usually, attention is not attracted to a single location in the image, but rather to an extended region that is likely to constitute an object or a part of an object. We say "likely", because the bottom-up (image-based) guidance of attention appears to rely on low-level features which do not yet constitute well-defined objects [2]. It is only after the attention system has selected a region of the visual field that objects are identified by the recognition system in infero-temporal cortex. This makes object based attention a chicken and egg problem that is rather difficult to disentangle.

In previous work, we demonstrated that using a fast attention front-end to rapidly select candidate locations for recognition greatly speeds up the recognition of objects in cluttered scenes at little cost in terms of detection rate [3].

However, this approach, as others before, was based on a paradigm of selecting an "interesting" part of the image, cropping the image and routing the cropped section to the recognition unit via dynamical neural connections. Although it has been widely used in computational modeling [4, 5], there is little physiological support for this paradigm. The main issue is the implementation of routing using all-or-nothing switching of neuronal connections, whereas physiological investigations usually only find moderate modulations of firing rates in early visual areas due to attentional effects [6].

We here investigate an alternative approach [7]. In a static architecture, we use the spatial information provided by the attention component of our model to modulate the activity of cells in the recognition part using a variable amount of modulation. We find that modulation by 20% suffices to successively recognize multiple objects in a scene.

In order to achieve this, however, the attention system needs to provide information not only about a location of interest, but also about the extent of the "interesting" region around this point, *i.e.,* we require a rough segmentation of the object of interest. We propose an approach that exploits the already-computed feature maps and thus comes at negligible extra computation cost for the segmentation.

2 Model

Bottom-Up Guidance of Attention Attentional guidance is based on our earlier model for bottom-up saliency-based attention [1, 2, 8, 9]. The model extracts feature maps for orientations, intensities and color, and builds up a saliency map using intermediary conspicuity maps. A winner-take-all network of integrate and fire neurons selects winning locations, and an inhibition-of-return mechanism allows the model to attend to many locations successively. For a more detailed description of the system see Fig. 1.

Recognition of Attended Objects The recognition component is based on our previously published hierarchical model for object recognition HMAX [10, 11] that considers recognition as a feedforward process in the spirit of Fukushima's Neocognitron [12]. HMAX builds up units tuned to views of objects from simple bar-like stimuli in a succession of layers, alternatingly employing a maximum pooling operation (for spatial pooling) and a sum pooling operation (for feature combination). For details see Fig. 1.

Riesenhuber and Poggio have demonstrated that the simultaneous recognition of multiple objects in an image is possible in HMAX if inputs to a view-tuned unit (VTU) are limited to the afferents strongly activated by the VTUs preferred object [11]. However, this comes at the cost of decreased shape specificity of VTUs due to the reduced number of features used to define its preferred object. In our present model we use HMAX with VTUs connected to all 256 C2 units, in order to demonstrate how attentional modulation provides a way to achieve recognition in clutter without sacrificing shape specificity.

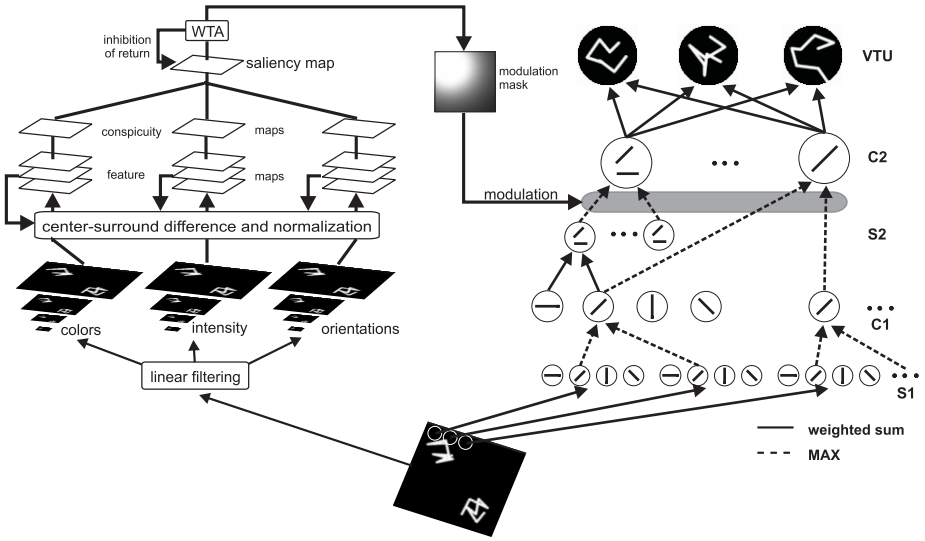

Fig. 1. Our model combines a saliency-based attention system [2, 8, 9] with a hierarchical recognition system [10, 11]. For the attention system, the retinal image is filtered for colors (red-green and blue-yellow), intensities, and four orientations at four different scales, and six center-surround differences are computed for each feature. The resulting $7 \times 6 = 42$ feature maps are combined into three conspicuity maps (for color, intensity and orientations), from which one saliency map is computed. All locations within the saliency map compete in a winner-take-all (WTA) network of integrate and fire neurons, and the winning location is attended to. Subsequently, the saliency map is inhibited at the winning location (inhibition-of-return), allowing the competition to go on, so that other locations can be attended to. The hierarchical recognition system starts out from units tuned to bar-like stimuli with small receptive fields, similar to V1 simple cells. In a succession of layers, information is combined alternatingly by spatial pooling (using a maximum pooling function) and by feature combination (using a weighted sum operation). View-tuned units at the top of the hierarchy respond to a specific view of an object while showing tolerance to changes in scale and position [10]. The activity of units at level S2 is modulated by the attention system (Fig. 2).

Interactions Between Attention and Recognition The goal of the attentional modulation is it to provide the recognition system with a first order approximation of the location and extent of "interesting" objects in the scene. For this, the feature maps and conspicuity maps can be re-used that have been created in the process of computing the saliency map. Fig. 2 details how information from the feature maps is used for object-based inhibition-of-return and for creating the modulation mask.

The advantage of going back to the most influential feature map for obtaining the mask instead of using the saliency map directly is the sparser representation in the feature map, which makes segmentation easier. This procedure comes at

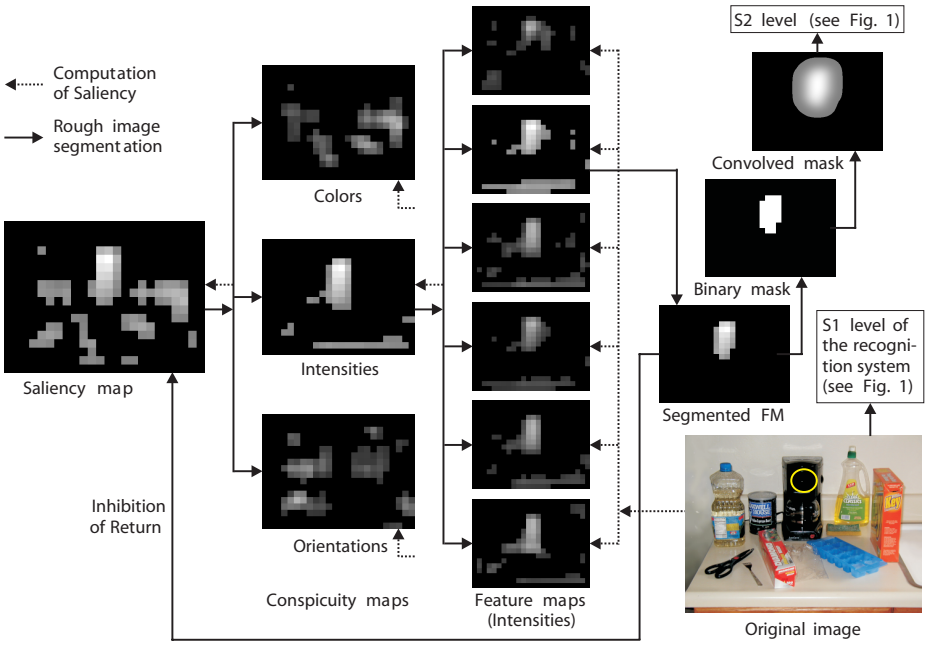

Fig. 2. To compute the modulation mask, the algorithm first determines which of the feature domains (colors, intensities, or orientations) contributed most to the saliency of the current focus of attention (FOA). In this figure, it is the intensity conspicuity map. In a second step, the feature map that contributed most to the winning conspicuity map at the FOA is found. Amongst the six feature maps for intensity, the second feature map from the top is the winner here. The winning feature map is segmented using a flooding algorithm with adaptive thresholding. The segmented feature map is used as a template for object-based inhibition-of-return of the saliency map. It is also processed into a binary mask and convolved with a Gaussian kernel to yield the modulation mask at image resolution. Finally, the modulation mask is multiplied to the activities of the S2 layer in the recognition system (see eq. 1).

negligible additional computational cost, because the feature and conspicuity maps have already been computed during the processing for saliency.

The modulation mask modifies the activity of the S2 cells in the recognition system (Fig.1). With $0 \leq M(x,y) \leq 1$ being the value of the modulation mask at position (x,y) and $S(x,y)$ being the S2 activity, the modulated S2 activity $S'(x,y)$ is computed according to

$$S'(x,y) = [(1-m) + m \cdot M(x,y)] \cdot S(x,y) \tag{1}$$

Where m is the modulation strength ($0 \leq m \leq 1$).

Applying the modulation at the level of the S2 units makes sense for biological and computational reasons. The S2 layer corresponds in its function approximately to area V4 in the primate visual cortex. There have been a num-

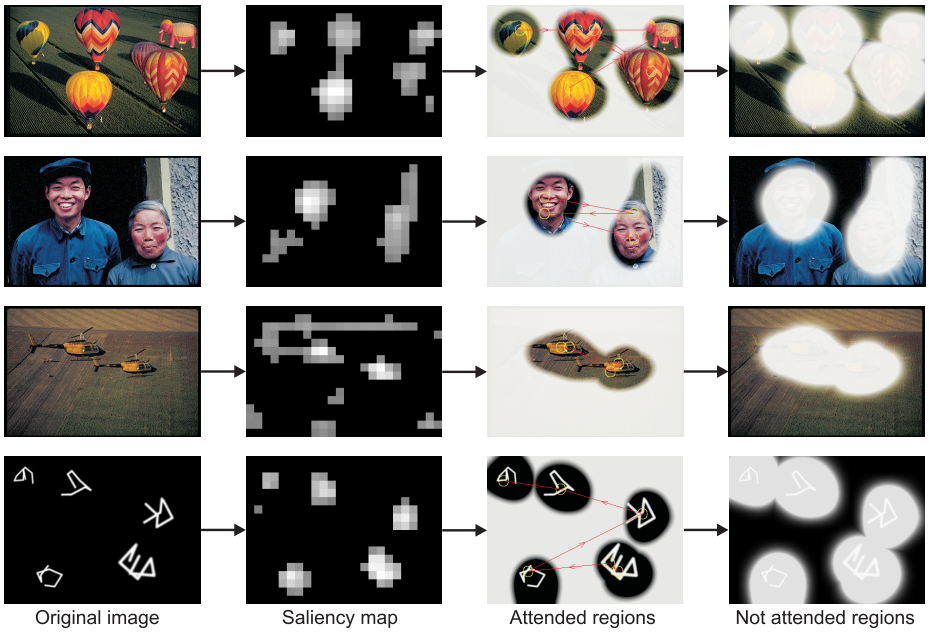

Fig. 3. Four examples for the extraction of modulation masks as described in Fig. 2. For each example, the following steps are shown (from left to right): the original image; the saliency map for the image; the original image contrast-modulated with the cumulative modulation mask, with the scan path overlayed; the inverse cumulative mask, covering all salient parts of the image. More examples are available in full color at `http://klab.caltech.edu/~walther/bmcv`

ber of reports from electrophysiology [6, 13–16] and psychophysics [17, 18] that show attentional modulation of V4 activity. Hence, the S2 level is a natural choice for modulating recognition.

From a computational point of view, it is efficient to apply the modulation at a level as high up in the hierarchy as possible that still has some spatial resolution, *i.e.* S2. This way, the computation that is required to obtain the activations of the S2 units from the input image needs to be done only once for each image. When the system attends to the next location in the image, only the computation upwards from S2 needs to be repeated.

3 Results

Rapid Extraction of Attended Object Size To qualitatively assess our rather simple approach to image segmentation using the feature maps, we tested the method on a number of natural images (e.g., Fig. 3). The method appears to work well for a range of images. In most cases, the areas found by the method indeed constitute objects or parts of objects. In general, the method has problems

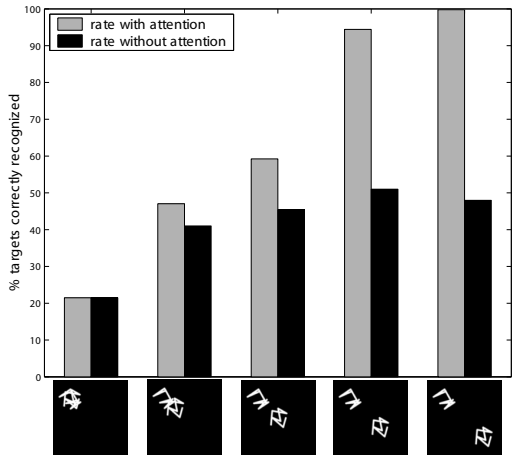

Fig. 4. Recognition performance of the model with (gray) and without (black) attentional modulation as a function of spatial separation between the two wire-frame objects. Improvement of recognition performance due to spatial attention is strongest for objects that are well separated, and absent when objects overlap. Recognition performance is defined as the ratio of the total number of clips that have been recognized successfully and the number of paperclips that the system was presented with. The modulation strength for the data with attention is 50% in this figure.

when objects are not uniform in their most salient feature, because those objects cannot be segmented in the respective feature map. At this point, we have no quantitative assessment of the performance of this method.

Improved Recognition Using Attention It would be a big step to be able to recognize natural objects like the ones found by the segmentation procedure. The hardwired set of features currently used in HMAX, however, shows rather poor performance on natural scenes. We therefore applied the system to the artificial paperclip stimuli used in previous studies [10, 11] and shown in Fig. 3.

The stimuli that we investigated consist of two twisted paperclips chosen from a set of 21 clips, yielding a total of $21^2 = 441$ different displays. In addition, the distance between the two paperclips was varied from total overlapping to clear separation of the paperclips as shown below the graph in Fig. 4.

In the simulations, the attention system first computed the most salient image location. Then, going back to the feature maps, the region of interest was found. The recognition system processed the image up to layer S2. The activity of the S2 cells was modulated according to eq. 1, and the C2 activity and the responses of the VTU were computed (Fig. 1). Inhibition-of-return was triggered, and the saliency map evolved further, until the next location was attended to, and so on. This procedure ran for 1000 ms simulated time of the integrate and fire winner-take-all network. A paperclip was counted as successfully recognized when the response of its VTU exceeded the threshold in at least one of the attended locations. The threshold was determined for each VTU separately by presenting it with 60 randomly chosen distracter paperclips that were not used for any other part of the simulation [10]. The maximum of the responses of the VTU to any of the distracter stimuli was set as the threshold for successful recognition.

Since we were using spatial attention to separate the two paperclip objects, we investigated the influence of the distance between the two objects in the dis-

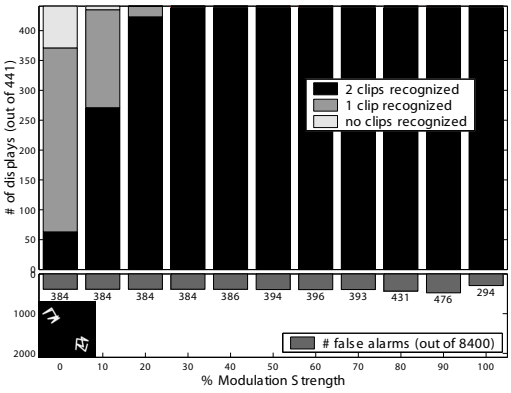

Fig. 5. For each stimulus, the recognition system can identify two, one or none of the two paperclip objects present. We show for how many stimuli each of the three cases occurs as a function of the attentional modulation. Zero modulation strength implies no attentional modulation at all. At 100% modulation, the S2 activation outside the FOA is entirely suppressed. As little as 20% attentional modulation is sufficient to boost the recognition performance significantly.

plays. The results are shown in Fig. 4. There is no advantage to attending for overlapping paperclips, since this does not resolve the ambiguity between the two objects. As distance increases, the performance of the system with attention (gray) becomes more and more superior to the system without attentional modulation (black). For clearly separated objects, the performance with attention reaches almost 100%, while the system without attention only recognizes 50% of the paperclip objects.

Viewing the simulation results from another perspective, one can ask: In how many of the displays does the model recognize both, only one or none of the two objects present? This analysis is shown in Fig. 5 for varying modulation strength m; for an interpretation of m see eq. 1. Without attention ($m = 0$), both paperclips were only recognized in a few displays; indeed, in some fraction of the examples, neither of the two clips was recognized. With only 10% modulation strength, the model recognized both paperclips in more than half of the displays, and with 20% modulation, both paperclips were recognized in almost all displays. For all values of $m > 20\%$, both paperclips were recognized in all displays.

4 Conclusion

We introduced a simple method for image segmentation that makes use of precomputed features. Despite its simplicity, this method is robust for at least one class of images. We are currently evaluating its performance for natural scenes.

In our combined model of bottom-up spatial attention and object recognition we have shown that attentional modulation of the activity of units at the V4 level in the recognition system by as little as 20% is sufficient to successively recognize multiple objects in a scene. In our approach, there is no need for full dynamic routing of the contents of the attended image region to the recognition system.

Acknowledgments

This work was supported by NSF (ITR, ERC and KDI), NEI, NIMA, DARPA, the Zumberge Faculty Research and Innovation Fund (L.I.), the Charles Lee Powell Foundation, a McDonnell-Pew Award in Cognitive Neuroscience (M.R.), Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.

References

1. C. Koch and S. Ullman. Shifts in selective visual-attention - towards the underlying neural circuitry. *Hum. Neurobiol.*, 4(4):219–227, 1985.
2. L. Itti and C. Koch. Computational modelling of visual attention. *Nat. Rev. Neurosci.*, 2(3):194–203, 2001.
3. F. Miau and L. Itti. A neural model combining attentional orienting to object recognition: Preliminary explorations on the interplay between where and what. In *IEEE Engin. in Medicine and Biology Society (EMBS)*, Istanbul, Turkey, 2001.
4. B. A. Olshausen, C. H. Anderson, and D. C. Van Essen. A neurobiological model of visual-attention and invariant pattern-recognition based on dynamic routing of information. *J. Neurosci.*, 13(11):4700–4719, 1993.
5. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo. Modeling visual-attention via selective tuning. *Artif. Intell.*, 78:507–545, 1995.
6. J. H. Reynolds, T. Pasternak, and R. Desimone. Attention increases sensitivity of V4 neurons. *Neuron*, 26(3):703–714, 2000.
7. D. Walther, M. Riesenhuber, T. Poggio, L. Itti, and C. Koch. Towards an integrated model of saliency-based attention and object recognition in the primate's visual system. *J. Cogn. Neurosci.*, B14 Suppl. S:46–47, 2002.
8. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20(11):1254–1259, 1998.
9. L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.*, 40(10-12):1489–1506, 2000.
10. M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–1025, 1999.
11. M. Riesenhuber and T. Poggio. Are cortical models really bound by the "binding problem"? *Neuron*, 24(1):87–93, 111–25., 1999.
12. K. Fukushima. Neocognitron: A self-organizing neural network model for a mechan. of pattern recogn. unaffected by shifts in position. *Biol. Cybern.*, 36:193–202, 1980.
13. S. Treue. Neural correlates of attention in primate visual cortex. *Trends Neurosci.*, 24(5):295–300, 2001.
14. C. E. Connor, D. C. Preddie, J. L. Gallant, and D. C. Van Essen. Spatial attention effects in macaque area V4. *J. Neurosci.*, 17(9):3201–3214, 1997.
15. B. C. Motter. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.*, 14(4):2178–2189, 1994.
16. S. J. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.*, 77(1):24–42, 1997.
17. J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cogn. Psychol.*, 43(3):171–216, 2001.
18. J. Braun. Visual-search among items of different salience - removal of visual-attention mimics a lesion in extrastriate area V4. *J. Neurosci.*, 14(2):554–567, 1994.