Chapter 8

Conclusions

8.1 Summary

In this thesis we have introduced a model of bottom-up attention to salient regions based on low-level image properties, and we have demonstrated its use in a variety of applications in computational modeling of biological vision and in machine vision. Furthermore, we have modeled feature sharing between object recognition and top-down attention, and we have measured the cost of deploying top-down attention.

Our model of salient region detection, described in detail in chapter 2, provides a solution for the problem of identifying a region that is likely to contain an object even before objects are recognized. Selecting the salient region relies on neuronal feedback connections in the system of maps and pyramids that are derived from low-level image properties to compute the saliency map. All processing steps are biologically plausible, and there is little computational overhead on top of the operations required to compute saliency in the conventional way.

In chapter 3 we added salient region selection to a biologically plausible model of object recognition in cortex by Riesenhuber and Poggio (1999b) in order to facilitate sequential recognition of several objects. We have shown that modulation of the activity of units at the V4-equivalent S2 layer by 20–40 % is sufficient to process only the visual information in the attended region, successfully ignoring unattended distracter objects. This is in agreement with several electrophysiology studies that find activity of neurons in area V4 to be modulated by 20–50 % due to selective attention.

Serializing perception and suppressing clutter are also the main mechanisms by which salient region selection proves to be useful for machine vision applications. In connection with grouping based on low-level properties, serializing the perception of a complex scene with multiple objects and clutter provides a way for unsupervised learning of several object models from a single image, as we have shown in chapter 5. By only processing the attended image regions, object detection also becomes more robust to large amounts of clutter.

We have demonstrated in chapter 6 that our model of salient region detection can aid initial

target detection for multi-target tracking and decrease the complexity of the assignment problem by pre-filtering potential target objects. Our application, detecting and tracking low-contrast marine animals in video from remotely operated underwater vehicles, is a first step toward automation of mining this important source of data.

In many situations we direct attention based on a task or agenda from the top down. In chapter 4 we showed that finding useful features for attending to a particular object category can be interpreted as a reversal of processes involved in object detection. We have proposed a model architecture in which this functionality is implemented with feedback connections. We have demonstrated the capabilities of the approach for the example of top-down attention to faces.

Deploying top-down attention comes at a cost in reaction time. We have explored this cost in chapter 7 in psychophysical experiments using a task switching paradigm. By comparing switch costs in task switches that do with those that do not require re-deployment of top-down attention, we found a cost in reaction time of 20–28 ms.

8.2 Future Work

Visual attention and object recognition are tightly interwoven. Many aspects of these interactions are not covered in this thesis or elsewhere in the modeling literature. Our method of salient region selection for attending to proto-objects is only the beginning of an iterative interaction between attention and recognition. Based on initial guesses of the recognition system about the likly identity of the attended item, the attention system should be fine-tuned to allow for fast verification or rejection of hypotheses. Future work should attempt to model these interactions and make predictions about the time course of recognition. These predictions could be tested, for instance, using masking experiment to disrupt the iterations at specific times.

Another interesting aspect of modeling interactions between attention and object recognition is top-down attention for object categories when features are shared among many categories. It is unclear so far how specific intermediate-level features of the type used in chapter 4 are in a scenario with many, e.g. hundreds or thousands, of object categories. Will individual features have enough specificity, or will it be necessary to consider conjunctions of features? The answer to this question has profound implications for the efficiency of visual search for objects.

We demonstrated empirically advantages of spatial grouping based on a biologically inspired concept of saliency for object detection and tracking in machine vision applications. It is not clear, however, if this concept is the best possible one for locating objects based on low-level image properties. Comparison of the statistics of object presence in natural images with the concept of saliency demonstrated here should yield interesting insights into this question.

Attention-based spatial selection and grouping improve efficiency of machine vision algorithms

and enable modes of processing that are not possible otherwise. The exact way in which these improvements can be achieved appear to depend on the recognition algorithm used. It would be interesting to investigate if there are general underlying principles for using attention in object recognition that pervade particular design choices for recognition systems. Such principles might include higher efficiency for matching sets of keypoints, spatial grouping of features to arrive at initial object guesses, or re-balancing of the search for detected object representations in a data base of known objects.

In our psychophysical work, we were able to use task switching to probe for the presence or absence of attention shifts. In future work, this method should be evaluated as a possible probe to determine which tasks require re-orientation of attention. For instance, does switching from a task involving the gist of a scene to a task about foreground objects require a shift of attention? Psychophysical experiments might be supplemented with fMRI (for an example of brain imaging during task switching see Yeung et al. 2006) or event-related EEG.