Chapter 4

Feature Sharing between Object Detection and Top-down Attention

4.1 Introduction

Visual search and other attentionally demanding processes are guided from the top down when a specific task is given (e.g., Wolfe et al. 2004). In the simplified stimuli commonly used in visual search experiments, e.g., red and horizontal bars, the selection of potential features that might be biased for is obvious (by design). In a natural setting with real-world objects, the selection of these features is not obvious, and there is some debate about which features can be used for top-down guidance and how a specific task maps to them (Wolfe and Horowitz 2004).

Learning to detect objects provides the visual system with an effective set of features suitable for the detection task and a mapping from these features to an abstract representation of the object. We suggest a model in which V4-type features are shared between object detection and top-down attention. As the model familiarizes itself with objects, i.e., it learns to detect them, it acquires a representation for features to solve the detection task. We propose that by cortical feedback connections, top-down processes can re-use these same features to bias attention to locations with a higher probability of containing the target object. We compare the performance of a computational implementation of such a model with pure bottom-up attention and, as a benchmark, with biasing for skin hue, which is known to work well as a top-down bias for faces.

The feed-forward recognition model used in this chapter was designed by Thomas Serre, based on the HMAX model for object recognition in cortex by Dr. Maximilian Riesenhuber and Dr. Tomaso Poggio. Face and non-face stimuli for the experiments were collected and annotated by Xinpeng Huang and Thomas Serre at MIT. I designed the top-down attention mechanism and the model of skin hue and conducted th experiments and analyses.

4.2 Model

The hierarchical model of object recognition used in chapter 3 has a fixed set of intermediate-level features at the S2 level. These features are well suited for the paper clip stimuli of chapter 3, but they are not sufficiently complex for recognition of real-world objects in cluttered images. In this chapter we adopt the extended version of the model by Serre et al. (2005a,b) and Serre and Poggio (2005) with more complex features that are learned from natural scene statistics. We demonstrate how top-down attention for a particular object category, faces in our case, is obtained from feedback connections in the same hierarchy used for object detection.

4.2.1 Feature Learning

In the extended model, S2 level features are no longer hardwired but are learned from a set of training images. The S1 and C1 activitations are computed in the same way as described in chapter 3. Patches of the C1 activation maps are sampled at several randomly chosen locations in each training image and stored as S2 feature prototypes (figure 4.2). If the patches are of size $4 \times m \times m$ (assuming four orientations in C1), then each prototype represents a vector in a $4m^2$-dimensional space. To evaluate a given S2 feature for a new image, the distance of each $m \times m$ patch of C1 activation for the image from the S2 prototype is computed using a Gaussian distance measure, resulting in an S2 feature map with the same spatial resolution as the C1 maps.

During feature learning, each prototype p is assigned a utility function $u(p)$, which is initialized to $u_0(p) = 1$. For each training image, several (e.g., 100) patches are sampled from the respective C1 activation maps, and the response of each prototype for each of the patches is determined. Each patch is then assigned to the prototype with the highest response, and the number of patches assigned to each prototype p is counted as $c(p)$. Subsequently, the utility function is updated according to

$$u_{t+1}(p) = \begin{cases} \alpha \cdot u_t(p) & \text{if } c(p) = 0 \\ \alpha \cdot u_t(p) + \beta & \text{if } c(p) > 0 \end{cases}, \tag{4.1}$$

with $0 < \alpha < 1$ and $\beta > 1$. Thus, utility decreases for prototype p whenever p does not get a patch assigned, but increases if it does. Whenever utility drops below a threshold θ, the prototype is discarded and re-initialized to a new randomly selected patch, and its utility is reset to 1. The prototypes surviving several iterations over all training images are fixed and used as intermediate-level features for object recognition and top-down attention.

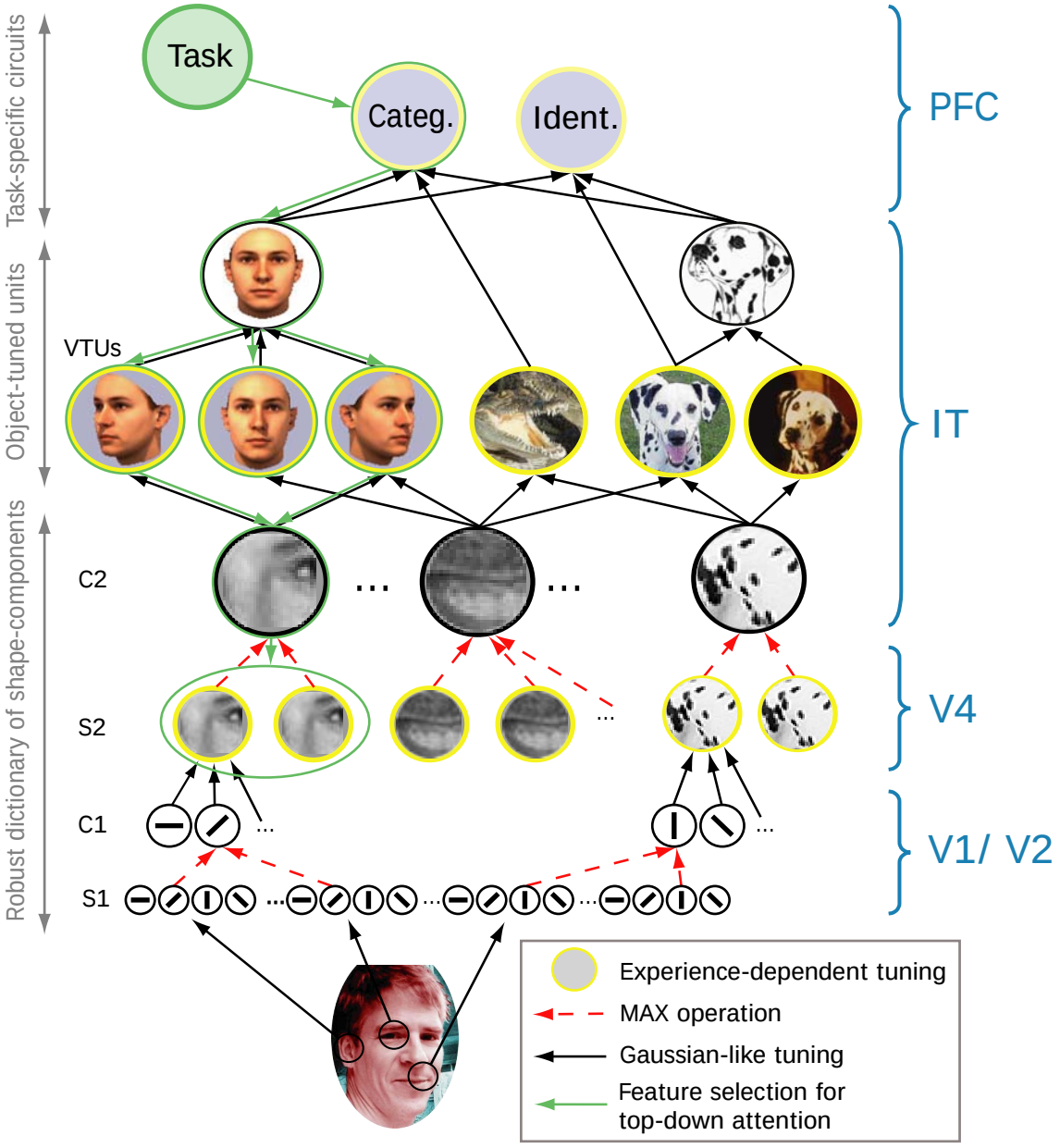

Figure 4.1: The basic architecture of our system of object recognition and top-down attention in the visual cortex (adapted from Walther et al. 2005b; Serre et al. 2005a). In the feed-forward pass, feature selective units with Gaussian tuning (black) alternate with pooling units using a maximum function (purple). Increasing feature selectivity and invariance to translation are built up as visual information progresses through the hierarchy until, at the C2 level, units respond to the entire visual field but are highly selective to particular features. View-tuned units (VTUs) and, finally, units selective to individual objects or object categories are trained. By association with a particular object or object category, activity due to a given task can traverse down the hierarchy (green) to identify a small subset of features at the S2 level that are indicative for the particular object category.

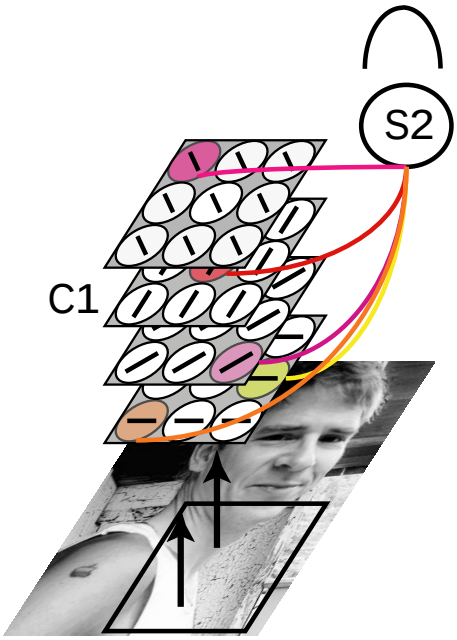

Figure 4.2: S2-level features are patches of the four orientation sensitive C1 maps cut out of a set of training images. S2 units have Gaussian tuning in the high-dimensional space that is spanned by the possible feature values of the four maps in the cut-out patch. During learning, S2 prototypes are initialized randomly from a training set of natural images that contain examples of the eventual target category among other objects and clutter. The stability of an S2 feature is determined by the number of randomly selected locations in the training images, for which this unit shows the highest response compared to the other S2 feature units. S2 prototypes with low stability are discarded and re-initialized.

4.2.2 Object Detection

To train the model for object recognition, it is presented with training images with and without examples of the object category of interest somewhere in the image. For each training image, the S1 and C1 activities are computed, and S2 feature maps are obtained using the learned S2 prototypes. In a final spatial pooling step, C2 activities are computed as the maximum over S2 maps, resulting in a vector of C2 activities that is invariant to object translation in the visual field of the model (figure 4.1). Given the C2 feature vectors for the positive and the negative training examples, a view-tuned unit (VTU) is trained using a binary classifier. For all experiments in this chapter we used a support vector machine (Vapnik 1998) with a linear kernel as a classifier. Given semantic knowledge on which views of objects belong to the same object or object category, several view-tuned units may by pooled to indicate object identity or category membership. Thus, a mapping from the set of S2 features to an abstract object representation is created.

For testing, new images are presented to the model and processed as described above to obtain their C2 feature vectors. The response of the classifier to the C2 feature vector determines whether the test images are classified as containing instances of the target object or object category or not. Note that, once the system is trained, the recognition process is purely feed-forward, which is compatible with rapid object categorization in humans (Thorpe et al. 1996) and monkeys (Fabre-Thorpe et al. 1998).

4.2.3 Top-down Attention

"Top-down attention" refers to the set of processes used to bias visual perception based on a given task or other prior expectation, as opposed to purely stimulus-driven "bottom-up attention" (chapter 2). One of the most puzzling aspects of top-down attention is how the brain "knows" which biases need to be set to fulfill a given task. Frequently, tasks are associated with objects or object categories, e.g., for search or the intention to manipulate an object in order to achieve a goal.

While feature learning establishes a mapping from the image pixels to a representation of intermediate complexity, training an object detection system creates a mapping from those features to the more abstract representations of objects or object categories. Reversing this mapping provides a method for finding suitable features for top-down attention to an object category that is relevant for a specific task (green arrows in figure 4.1).

Here we investigate how well these S2 maps are suited for localizing instances of the target category. Using these maps, potential object location can be attended one at a time, thus disambiguating multiple instances of an object category and allowing for suppression of visual information that is irrelevant for the task.

4.3 Experimental Setup

For the work presented in this chapter, we trained our model on detecting frontal views of human faces and investigated the suitability of the corresponding S2 features for top-down attention to faces.

For feature learning and training, we used 200 color images, each containing one face among clutter, and 200 distracter images without faces (see figure 4.3 for examples). For testing the recognition performance of the system, we used 201 face images and 2119 non-face distracter images. All images were obtained from the world wide web, and face images were labeled by hand, with the eyes, nose and mouth of each face marked.[1] Images were scaled such that faces were at approximately the same scale.

During feature learning as described in subsection 4.2.1, 100 patches of size 6×6 were extracted from the C1 maps for each presentation of a training image. Using the parameters $\alpha = 0.9$ and $\beta = 1.1$ for eq. 4.1, 100 stable features were learned over five iterations of presenting the 200 training images in random order. Two separate sets of features were learned: set A was derived from patches that were extracted from any location in the training images (figure 4.3, top row); patch selection for set B was limited to regions around faces (figure 4.3, second row).

Separate VTUs for frontal faces were created for feature sets A and B. A support vector machine classifier with linear kernel was trained on the face and non-face training images. The VTUs were

[1] Thanks to Xinpeng Huang and Thomas Serre for collecting and labeling the images.

Figure 4.3: Examples for training stimuli for feature set A (top row), feature set B (second row), test stimuli with two or more faces (third and fourth row), and for non-face distracters (bottom row).

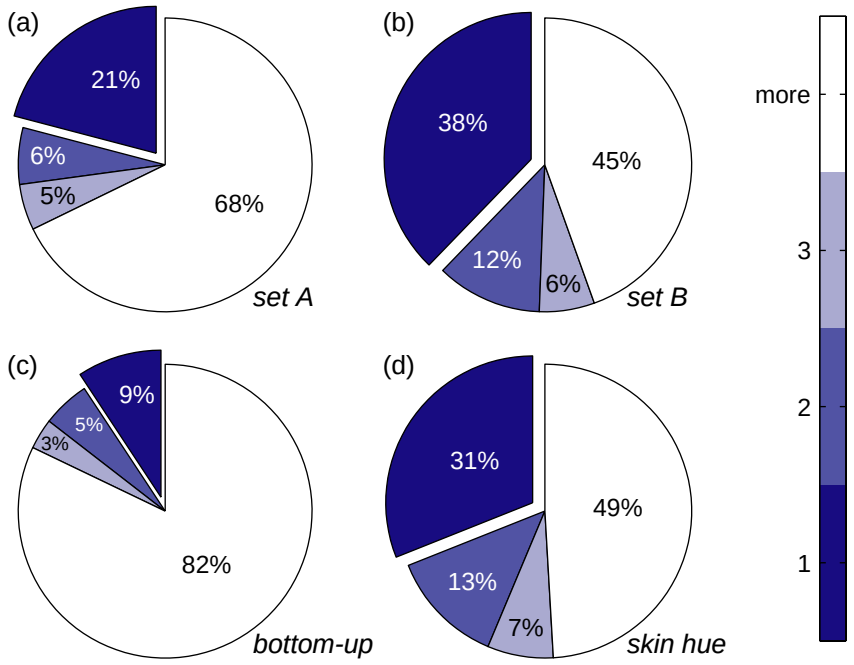

Figure 4.4: Fractions of faces in test images requiring one, two, three, or more than three fixations to be attended when using top-down feature sets A or B, bottom-up attention, or biasing for skin hue.

tested on the 201 face and 2119 non-face images.

To evaluate feature sets A and B for top-down attention, the S2 maps were computed for 179 images containing between 2 and 20 frontal views of faces (figure 4.3, third and fourth row). These top-down feature maps were compared to the bottom-up saliency map (see section 2.2) and to a skin hue detector for each of the images. Skin hue is known to be an excellent indicator for the presence of a face in color images (Darrel et al. 2000). Here we use it as a benchmark for comparison with our structure-based top-down attention maps. See section A.4 for details about our model of skin hue detection.

4.4 Results

After feature learning and training of the frontal face VTUs, we obtained ROC curves for the test images with feature sets A and B. The areas under the ROC curves are 0.989 for set A and 0.994 for set B.

We used two metrics for testing the suitability of the features for top-down attention to faces for the 179 multiple-face images, an analysis of fixations on faces, and a region of interest ROC analysis. Both methods start with the respective activation maps: the S2 feature maps for both feature sets, the bottom-up saliency map, and the skin hue bias map.

For the fixation analysis, each map was treated like a saliency map, and the locations in the

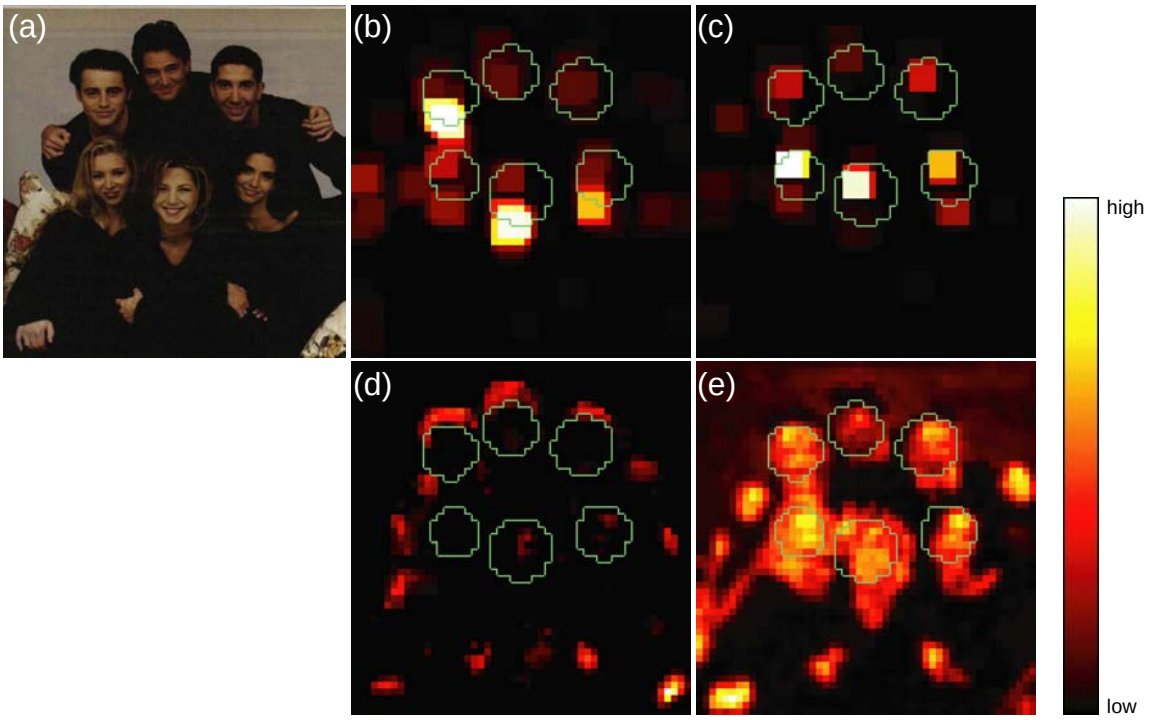

Figure 4.5: Using ground truth about the position of faces in the test images, activation maps can be segmented into face regions of interest (ROIs) and non-face regions. (a) input image; (b) one of the S2 maps from set A; (c) one of the set B S2 maps; (d) bottom-up saliency map; (e) skin hue distance map. Histograms of the map activations are used for an ROI ROC analysis (see fig. 4.6).

map were visited in order of decreasing saliency, neglecting spatial relations between the locations. While this procedure falls short of the full simulation of a winner-take-all network with inhibition of return as described in chapter 2, it nevertheless provides a simple and consistent means of scanning the maps.

For each map we determined the number of fixations required to attend to a face and, once the focus of attention leaves the most salient face, how many fixations it takes to attend to each subsequent face. The fraction of all faces that required one, two, three, or more than three fixations to be found was determined for each feature. The results are shown in figure 4.4 for the best features from sets A and B, for bottom-up attention, and for skin hue detection. Feature set B and skin hue detection show similar performance, followed by feature set A and bottom-up attention.

The second method of analyzing the suitability of the S2 feature maps to localize faces is illustrated in figure 4.5. From ground truth about the location of faces in the test images, we can divide the S2 feature maps into two regions, a region of interest (ROI) containing the S2 units that are inside the face regions, and its complement, containing all remaining S2 units. Ideally, we would like to see high activity inside the ROI and low or no activity outside the ROI.

Activity histograms for both regions as shown in figure 4.6 let us derive an ROC curve by moving

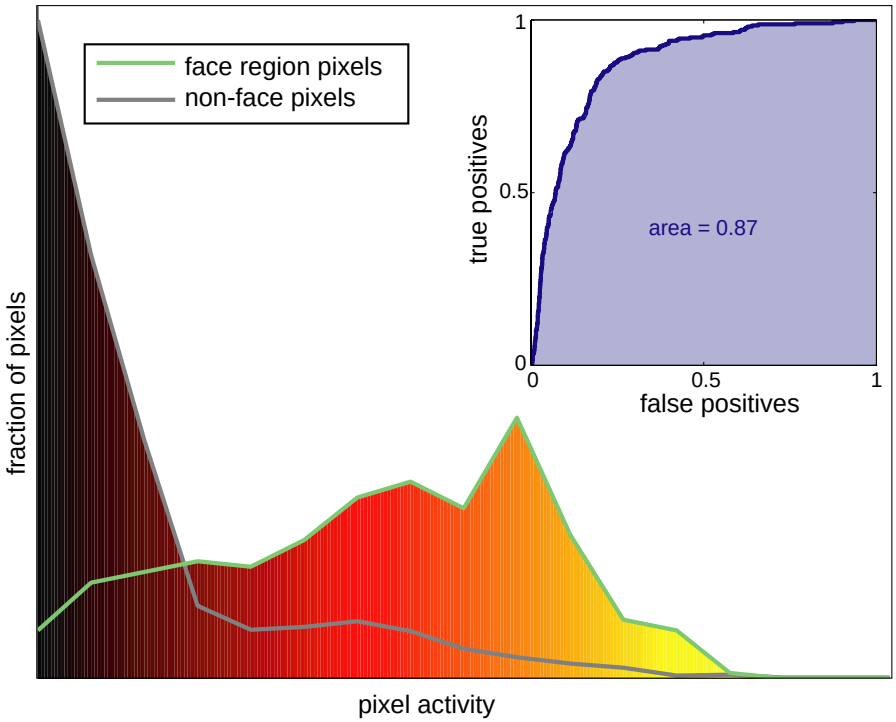

Figure 4.6: By sliding a threshold through the histograms of map activations for face and non-face regions for one of the maps shown in fig. 4.5, an ROC curves is established (inset). The mean of the areas under the curves for all test images is used to measure how well this feature is suited for biasing visual attention toward face regions.

a threshold through the histograms and interpreting the ROI units with activity above the threshold as true positives and the non-ROI units with activity above the threshold as false positives. The area under the ROC curve provides a measure for how well this particular activity map is suited for localizing faces in this test image. For each S2 feature, the ROI ROC area is computed for all 179 test images, and the mean over the test images is used as the second measure of top-down localization of faces.

The results from both evaluation methods are shown in figure 4.7. Only the fraction of cases in which the first fixation lands on a face (the dark blue areas in figure 4.4) is plotted. The two methods correlate with $\rho_{AB} = 0.72$.

Both evaluation methods indicate that the best features of feature set B perform similarly to skin hue detection for localizing frontal faces. Top-down attention based on S2 features by far outperform bottom-up attention in our experiments. While bottom-up attention is well suited to identify salient regions in the absence of a specific task, it cannot be expected to localize a specific object category as well as feature detectors that are specialized for this category.

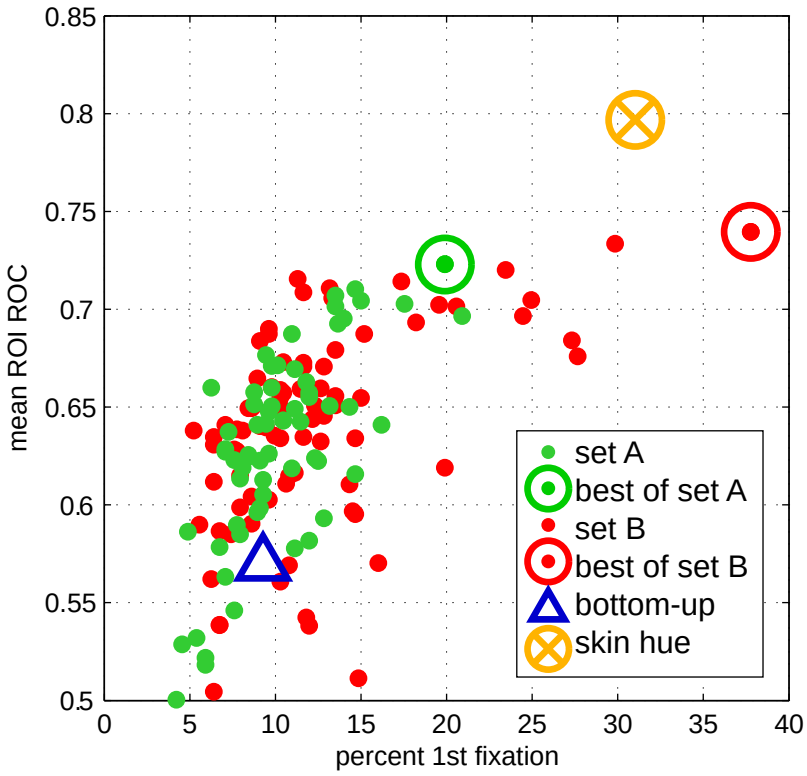

Figure 4.7: The fraction of faces in test images attended to on the first fixation (the dark blue areas in figure 4.4) and the mean areas under the ROC curves of the region of interest analysis (see figures 4.5 and 4.6) for the features from sets A (green) and B (red) and for bottom-up attention (blue triangle) and skin hue (yellow cross). The best features from sets A and B (marked by a circle) show performance in the same range as biasing for skin hue, although no color information is used to compute those feature responses.

4.5 Discussion

In this chapter we showed that features learned for recognizing a particular object category may also serve for top-down attention to that object category. Object detection can be understood as a mapping from a set of features to an abstract object representation. When a task implies the importance of an object, the respective abstract object representation may be invoked, and feedback connections may reverse the mapping, allowing inference of which features are useful to guide top-down attention to image locations that have a high probability of containing the target object.

Note that this mode of top-down attention does not necessarily imply that the search for any object category can be done in parallel using an explicit map representation. Search for faces, for instance, has been found to be efficient (Hershler and Hochstein 2005), although this result is disputed (VanRullen 2005). We have merely shown a method for identifying features that can be used to search for an object category. The efficiency of the search will depend on the complexity of those features and, in particular, on the frequency of the same features for other object categories,

which constitute the set of distracters for visual search. To analyze this aspect further, it would be of interest to explore the overlap in the sets of features that are useful for multiple object categories. Torralba et al. (2004) have addressed this problem for multiple object categories as well as multiple views of objects in a machine vision context.

The close relationship between object detection and top-down attention has been investigated before in a number of ways. In a probabilistic framework, Oliva et al. (2003) incorporate context information into the spatial probability function for seeing certain objects (e.g., people) at particular locations. Comparison with human eye tracking results show improvement over a purely bottom-up saliency-based attention (Itti et al. 1998). Milanese et al. (1994) describe a method for combining bottom-up and top-down information through relaxation in an associative memory. Rao (1998) considers attention a by-product of a recognition model based on Kalman filtering. He can get the system to attend to spatially overlapping (occluded) objects on a pixel basis for fairly simple stimuli.

Lee and Lee (2000) and Lee (2004) introduced a system for learning top-down attention using backpropagation in a multilayer perceptron network. Their system can segment superimposed handwritten digits on the pixel level.

In the work of Schill et al. (2001), features that maximize the gain of information in each saccade are learned using a belief propagation network. This is done using orientations only. Their system is tested on 24000 artificially created scenes which can be classified with a 80 % hit rate. Rybak et al. (1998) introduced a model that learns a combination of image features and saccades that lead to or from these features. In this way, translation, scale, and rotation invariance are built up. They successfully tested their system on small grayscale images. While this is an interesting system for learning and recognizing objects using saccades to build up object representations from multiple views, no explicit connection to top-down attention is made.

Grossberg and Raizada (2000) and Raizada and Grossberg (2001) propose a model of attention and visual grouping based biologically realistic models of neurons and neural networks. Their model relies on grouping detected edges within a laminar cortical structure by synchronous firing, allowing it to extract real as well as illusory contours.

The model by Amit and Mascaro (2003) for combining object recognition and visual attention has some resemblance with ours. Translation invariant detection is achieved by max-pooling, similar to Riesenhuber and Poggio (1999b). Basic units in their model consist of feature-location pairs, where location is measured with respect to the center of mass. Detection proceeds at many locations simultaneously, using hypercolumns with replica units that store copies of some image areas. The biological plausibility of these replica units is not entirely convincing, and it is not clear how to deal with the combinatorial explosion of the number of required units for a large number of object categories. Complex features are defined as combinations of orientations. There is a trade-off of accuracy versus combinatorics: More complex features lead to a better detection algorithm, but more

features are needed to represent all objects, i.e., the dimensionality of the feature space increases. Amit and Mascaro (2003) use binary features to make parameter estimation easier. Learning is performed by perceptrons that discriminate an object class from all other objects. These units vote to achieve classification. The system is demonstrated for the recognition of letters and for detecting faces in photographs.

Navalpakkam and Itti (2005) model the influence of task on attention by tuning the weights of feature maps based on the relevance of certain features for the search for objects that are associated with a given task in a knowledge database. Frintrop et al. (2005) also achieve top-down attention by tuning the weights of the feature maps in an attention system based on Itti et al. (1998). In their system, optimal weights are learned from a small set of training images, whose selection from a larger pool of images is itself subject to optimization.