Chapter 2

A Model of Salient Region Detection

2.1 Introduction

Attention as a selective gating mechanism is often compared to a spotlight (Posner 1980; Treisman and Gelade 1980), enhancing visual processing in the attended ("illuminated") region of a few degrees of visual angle (Sagi and Julesz 1986). In a modification to the spotlight metaphor, the size of the attended region can be adjusted depending on the task, making attention similar to a zoom lens (Eriksen and St. James 1986; Shulman and Wilson 1987). Neither of these theories considers the shape and extent of the attended object for determining the attended area. This may seem natural, since commonly attention is believed to act *before* objects are recognized. However, experimental evidence suggests that attention can be tied to objects, object parts, or groups of objects (Duncan 1984; Roelfsema et al. 1998). How can we attend to objects before we recognize them?

Several computational models of visual attention have been suggested. Tsotsos et al. (1995) use local winner-take-all networks and top-down mechanisms to selectively tune model neurons at the attended location. Deco and Schürmann (2000) modulate the spatial resolution of the image based on a top-down attentional control signal. Itti et al. (1998) introduced a model for bottom-up selective attention based on serially scanning a saliency map, which is computed from local feature contrasts, for salient locations in the order of decreasing saliency. Presented with a manually preprocessed input image, their model replicates human viewing behavior for artificial and natural scenes. Making extensive use of feedback and long-range cortical connections, Hamker (2005b,a) models the interactions of several brain areas involved in processing visual attention, which enables him to fit both physiological and behavioral data in the literature. Closely following and extending Duncan's Integrated Competition Hypothesis (Duncan 1997), Sun and Fisher (2003) developed and implemented a common framework for object-based and location-based visual attention using "groupings".

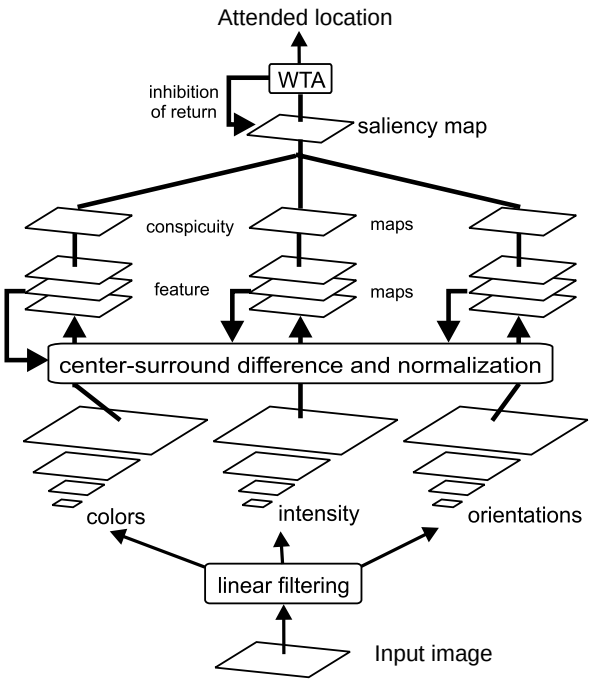

Figure 2.1: Architecture of the model of saliency-based visual attention, adapted from Itti et al. (1998).

However, none of these models provides a satisfactory solution to the problem of attending to objects even before they are recognized. To solve this chicken-and-egg problem in first approximation, we have developed a model for estimating the extent of salient objects in a bottom-up fashion solely based on low-level image features. In chapters 3, 5, and 6 we demonstrate the use of the model as an initial step for object detection.

Our attention system is based on the Itti et al. (1998) implementation of the saliency-based model of bottom-up attention by Koch and Ullman (1985). For a color input image, the model computes a saliency map from maps for color, luminance, and orientation contrasts at different scales (figure 2.1). A winner-take-all (WTA) neural network scans the saliency map for the most salient location and returns the location's coordinates. Finally, inhibition of return (IOR) is applied to a disc-shaped region of fixed radius around the attended location in the saliency map, and further iterations of the WTA network lead to successive direction of attention to several locations in order of decreasing saliency. The model has been verified in human psychophysical experiments (Peters et al. 2005; Itti 2005), and it has been applied to object recognition (Miau et al. 2001; Walther et al. 2002a, 2005a) and robot navigation (Chung et al. 2002).

We briefly review the details of the model in section 2.2 in order to explain our extensions in the same formal framework. In section 2.3 we describe our method of selecting salient regions instead of just salient locations by using feedback connections in the existing processing hierarchy of the original saliency model.

2.2 Saliency-based Bottom-up Attention

The input image $\mathcal{I}$ is sub-sampled into a dyadic Gaussian pyramid by convolution with a linearly separable Gaussian filter and decimation by a factor of two (see appendix A.1 for details). This process is repeated to obtain the next levels $\sigma = [0, .., 8]$ of the pyramid (Burt and Adelson 1983). Resolution of level σ is $1/2^\sigma$ times the original image resolution, i.e., the 8^{th} level has a resolution of $1/256^{\text{th}}$ of the input image's $\mathcal{I}$ and $(1/256)^2$ of the total number of pixels.

If r, g, and b are the red, green, and blue values of the color image, then the intensity map is computed as

$$\mathcal{M}_I = \frac{r + g + b}{3}. \tag{2.1}$$

This operation is repeated for each level of the input pyramid to obtain an intensity pyramid with levels $\mathcal{M}_I(\sigma)$.

Each level of the image pyramid is furthermore decomposed into maps for red-green (RG) and blue-yellow (BY) opponencies:

$$\mathcal{M}_{RG} = \frac{r - g}{\max(r, g, b)} \tag{2.2a}$$

$$\mathcal{M}_{BY} = \frac{b - \min(r, g)}{\max(r, g, b)}. \tag{2.2b}$$

To avoid large fluctuations of the color opponency values at low luminance, $\mathcal{M}_{RG}$ and $\mathcal{M}_{BY}$ are set to zero at locations with $max(r, g, b) < 1/10$, assuming a dynamic range of $[0, 1]$. Note that the definitions in eq. 2.2 deviate from the original model by Itti et al. (1998). For a discussion of the definition of color opponencies see appendix A.2.

Local orientation maps $\mathcal{M}_\theta$ are obtained by applying steerable filters to the intensity pyramid levels $\mathcal{M}_I(\sigma)$ (Simoncelli and Freeman 1995; Manduchi et al. 1998). In subsection 6.3.2 we show how lateral inhibition between units with different θ can aid in detecting faint elongated objects.

Motion is another highly salient feature. In appendix A.3 we describe our implementation of a set of motion detectors for saliency due to motion.

Center-surround receptive fields are simulated by across-scale subtraction $\ominus$ between two maps at the center (c) and the surround (s) levels in these pyramids, yielding "feature maps":

$$\mathcal{F}_{l,c,s} = \mathcal{N}\left(|\mathcal{M}_l(c) \ominus \mathcal{M}_l(s)|\right) \qquad \forall l \in L = L_I \cup L_C \cup L_O \tag{2.3}$$

with

$$L_I = \{I\}, \;\; L_C = \{RG, BY\}, \;\; L_O = \{0°, 45°, 90°, 135°\}. \tag{2.4}$$

$\mathcal{N}(\cdot)$ is an iterative, nonlinear normalization operator, simulating local competition between neigh-

boring salient locations (Itti and Koch 2001b). Each iteration step consists of self-excitation and neighbor-induced inhibition, implemented by convolution with a "difference of Gaussians" filter, followed by rectification. For the simulations in this thesis, between one and five iterations are used. For more details see Itti (2000) and Itti and Koch (2001b).

The feature maps are summed over the center-surround combinations using across-scale addition $\oplus$, and the sums are normalized again:

$$\bar{\mathcal{F}}_l = \mathcal{N}\left(\bigoplus_{c=2}^{4}\bigoplus_{s=c+3}^{c+4}\mathcal{F}_{l,c,s}\right) \forall l \in L. \tag{2.5}$$

For the general features color and orientation, the contributions of the sub-features are summed and normalized once more to yield "conspicuity maps." For intensity, the conspicuity map is the same as $\bar{\mathcal{F}}_I$ obtained in eq. 2.5:

$$\mathcal{C}_I = \bar{\mathcal{F}}_I,\ \mathcal{C}_C = \mathcal{N}\left(\sum_{l\in L_C}\bar{\mathcal{F}}_l\right),\ \mathcal{C}_O = \mathcal{N}\left(\sum_{l\in L_O}\bar{\mathcal{F}}_l\right). \tag{2.6}$$

All conspicuity maps are combined into one saliency map:

$$\mathcal{S} = \frac{1}{3}\sum_{k\in\{I,C,O\}}\mathcal{C}_k. \tag{2.7}$$

The locations in the saliency map compete for the highest saliency value by means of a winner-take-all (WTA) network of integrate-and-fire neurons. The parameters of the model neurons are chosen such that they are physiologically realistic, and such that the ensuing time course of the competition for saliency results in shifts of attention in approximately 30–70 ms simulated time (Saarinen and Julesz 1991).

The winning location (x_w, y_w) of this process is attended to, and the saliency map is inhibited within a given radius of (x_w, y_w). Continuing WTA competition produces the second most salient location, which is attended to subsequently and then inhibited, thus allowing the model to simulate a scan path over the image in the order of decreasing saliency of the attended locations.

In the next section we demonstrate a mechanism for extracting an image region around the focus of attention (FOA) that corresponds to the approximate extent of an object at that location. Aside from its use to facilitate further visual processing of the attended object, this enables object-based inhibition of return (IOR), thereby eliminating the need for a fixed-radius disc as an IOR template.

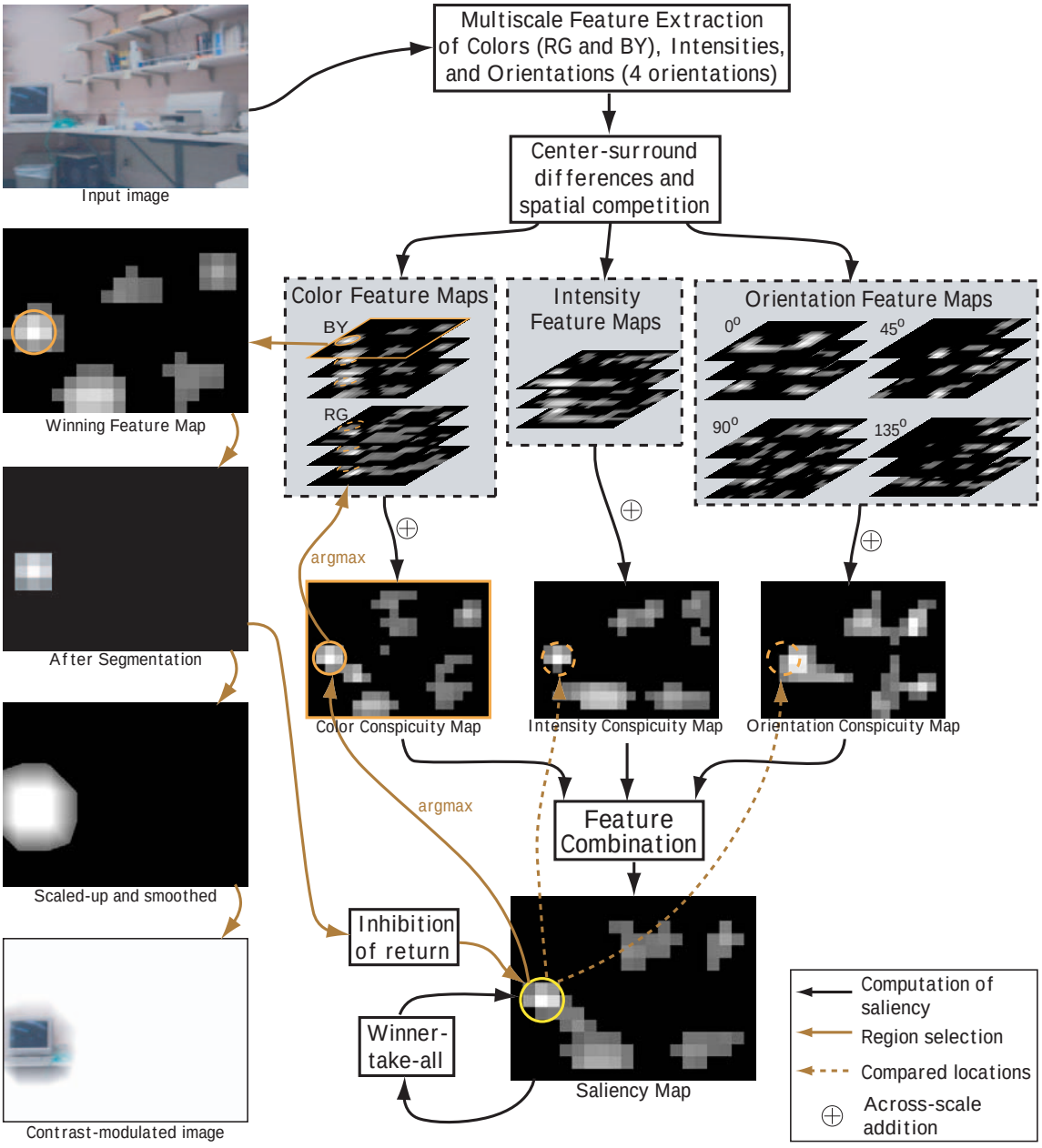

Figure 2.2: Illustration of the processing steps for obtaining the attended region. The input image is processed for low-level features at multiple scales, and center-surround differences are computed (eq. 2.3). The resulting feature maps are combined into conspicuity maps (eq. 2.6) and, finally, into a saliency map (eq. 2.7). A winner-take-all neural network determines the most salient location, which is then traced back through the various maps to identify the feature map that contributes most to the saliency of that location (eqs. 2.8 and 2.9). After segmentation around the most salient location (eqs. 2.10 and 2.11), this winning feature map is used for obtaining a smooth object mask at image resolution and for object-based inhibition of return.

2.3 Attending Proto-object Regions

While Itti et al.'s model successfully identifies the most salient location in the image, it has no notion of the extent of the attended object or object part at this location. We introduce a method of estimating this region based on the maps and salient locations computed so far, using feedback connections in the saliency computation hierarchy (figure 2.2). Looking back at the conspicuity maps, we find the one map that contributes the most to the activity at the most salient location:

$$k_w = \underset{k \in \{I,C,O\}}{\operatorname{argmax}} \mathcal{C}_k(x_w, y_w). \tag{2.8}$$

The *argmax* function, which is critical to this step, could be implemented in a neural network of linear threshold units (LTUs), as shown in figure 2.3. For practical applications we use a more efficient generic argmax function because of its higher efficiency.

Examining the feature maps that gave rise to the conspicuity map $\mathcal{C}_{k_w}$, we find the one that contributes most to its activity at the winning location:

$$(l_w, c_w, s_w) = \underset{l \in L_{k_w}, c \in \{2,3,4\}, s \in \{c+3, c+4\}}{\operatorname{argmax}} \mathcal{F}_{l,c,s}(x_w, y_w), \tag{2.9}$$

with L_{k_w} as defined in eqs. 2.4. The "winning" feature map $\mathcal{F}_{l_w,c_w,s_w}$ (figure 2.2) is segmented around (x_w, y_w). For this operation, a binary version of the map ($\mathcal{B}$) is obtained by thesholding $\mathcal{F}_{l_w,c_w,s_w}$ with 1/10 of its value at the attended location:

$$\mathcal{B}(x,y) = \begin{cases} 1 & \text{if } \mathcal{F}_{l_w,c_w,s_w}(x,y) \geq 0.1 \cdot \mathcal{F}_{l_w,c_w,s_w}(x_w, y_w) \\ 0 & \text{otherwise} \end{cases}. \tag{2.10}$$

The 4-connected neighborhood of active pixels in $\mathcal{B}$ is used as the template to estimate the spatial extent of the attended object:

$$\hat{\mathcal{F}}_w = \text{label}\left(\mathcal{B}, (x_w, y_w)\right). \tag{2.11}$$

For the *label* function, we use the classical algorithm by Rosenfeld and Pfaltz (1966) as implemented in the Matlab `bwlabel` function. After a first pass over the binary map for assigning temporary labels, the algorithm resolves equivalence classes and replaces the temporary labels with equivalence class labels in a second pass. In figure 2.4 we show an implementation of the segmentation operation with a network of LTUs to demonstrate feasibility of our procedure in a neural network. The segmented feature map $\hat{\mathcal{F}}_w$ is used as a template to trigger object-based inhibition of return (IOR) in the WTA network and to deploy spatial attention to subsequent processing stages such as object detection.

We have implemented our model of salient region selection as part of the SaliencyToolbox for

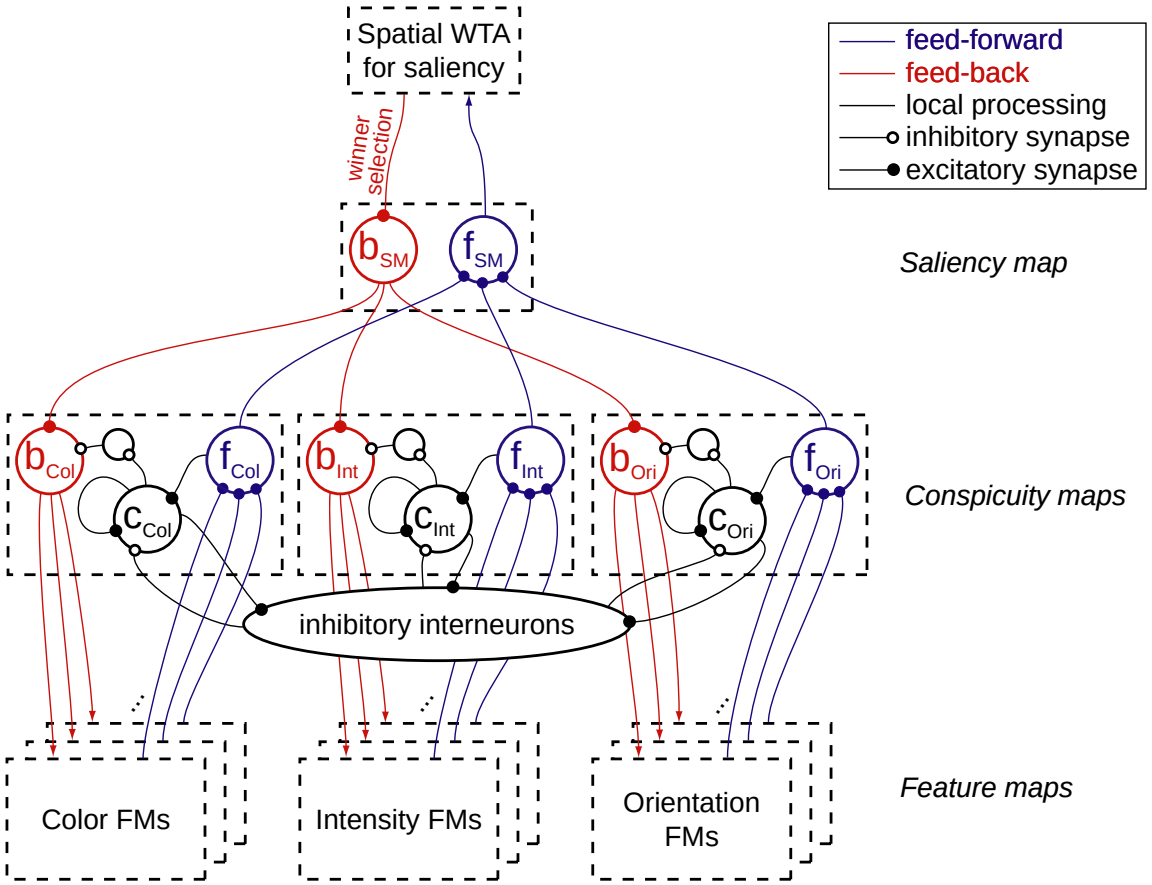

Figure 2.3: A network of linear threshold units (LTUs) for computing the *argmax* function in eq. 2.8 for one image location. Feed-forward (blue) units f_{Col}, f_{Int}, and f_{Ori} compute conspicuity maps for color, intensity, and orientation by pooling activity from the respective sets of feature maps as described in eqs. 2.5 and 2.6, omitting the normalization step $\mathcal{N}$ here for clarity. The saliency map is computed in a similar fashion in f_{SM} (eq. 2.7), and f_{SM} participates in the spatial WTA competition for the most salient location. The feed-back (red) unit b_{SM} receives a signal from the WTA only when this location is attended to, and it relays the signal to the b units in the conspicuity maps. Competition units (c) together with a pool of inhibitory interneurons (black) form an across-feature WTA network with input from the f units of the respective conspicuity maps. Only the most active c unit will remain active due to WTA dynamics, allowing it to unblock the respective b unit. As a result, the activity pattern of the b units represents the result of the *argmax* function in eq. 2.8. This signal is relayed further to the constituent feature maps, where a similar network selects the feature map with the largest contribution to the saliency of this location (eq. 2.9).

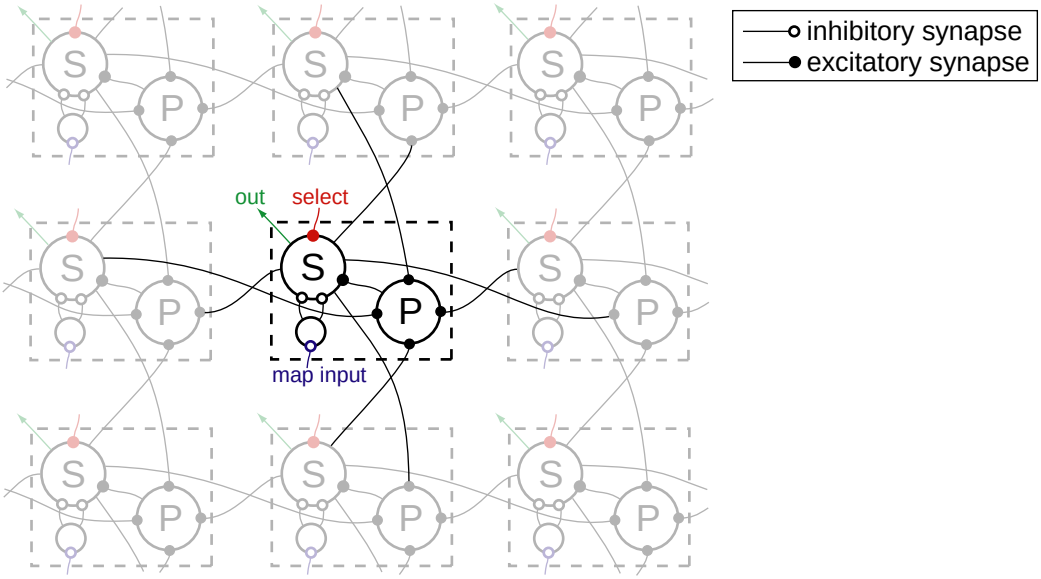

Figure 2.4: An LTU network implementation of the segmentation operation in eqs. 2.10 and 2.11. Each pixel consists of two excitatory neurons and an inhibitory interneuron. The thresholding operation in eq. 2.10 is performed by the inhibitory interneuron, which only unblocks the segmentation unit S if input from the winning feature map $\mathcal{F}_{l_w,c_w,s_w}$ (blue) exceeds its firing threshold. S can be excited by a select signal (red) or by input from the pooling unit P. Originating from the feedback units b in figure 2.3, the select signal is only active at the winning location (x_w, y_w). Pooling the signals from the S unit in its 4-connected neighborhood, P excites its own S unit when it receives at least one input. Correspondingly, the S unit projects to the P units of the pixels in the 4-connected neighborhood. In their combination, the reciprocal connections between the S and P units form a localized implementation of the labeling algorithm (Rosenfeld and Pfaltz 1966). Spreading of activation to adjacent pixels stops where the inbound map activity is not large enough to unblock the S unit. The activity pattern of the S units (green) represents the segmented feature map $\hat{\mathcal{F}}_w$.

Matlab, described in appendix B, and as part of the iLab Neuromorphic Vision (iNVT) C++ toolkit. In the Matlab toolbox we provide both versions of the segmentation operation, the fast image processing implementation, and the LTU network version. They are functionally equivalent, but the LTU network simulation runs much slower than the fast image processing version.

Figure 2.5 shows examples of applying region selection to three natural images as well as an artificial display of bent paper clips as used for the simulations in chapter 3. These examples and the results in chapters 3 and 5 were obtained using iNVT toolkit; for chapter 6 we used a modified version that was derived from the iNVT toolkit; and for chapter 4 we used the SaliencyToolbox.

2.4 Discussion

As part of their selective tuning model of visual attention, Tsotsos et al. (1995) introduced a mechanism for tracing back activations through a hierarchical network of WTA circuits to identify contiguous image regions with similarly high saliency values within a given feature domain. Our method is

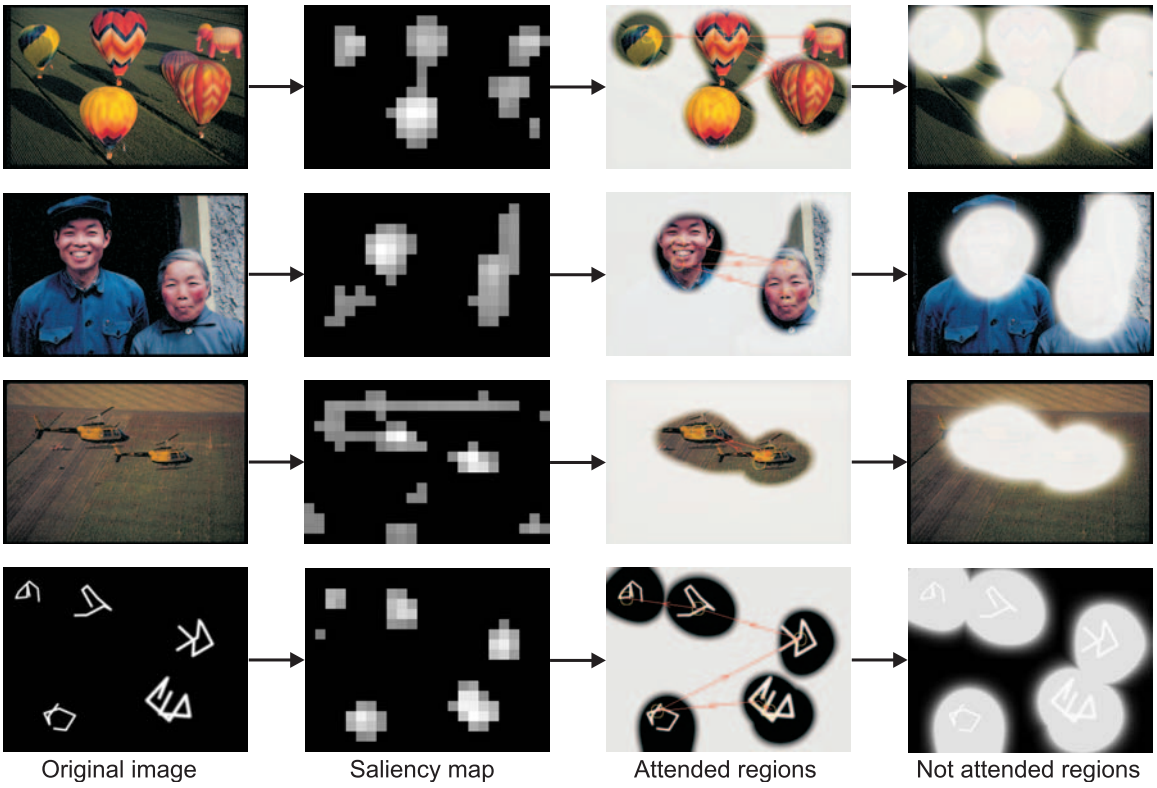

Figure 2.5: Four examples for salient region extraction as described in section 2.3. For each example the following steps are shown (from left to right): the original image $\mathcal{I}$; the saliency map $\mathcal{S}$; the original image contrast-modulated with a cumulative superposition of $\hat{\mathcal{F}}_w$ for the locations attended to during the first 700 ms of simulated time of the WTA network, with the scan path overlayed; and the inverse of this cumulative mask, covering all salient parts of the image. It is apparent from this figure that our salient region extraction approach does indeed cover the salient parts of the images, leaving the non-salient parts unattended.

similar in spirit but extends across feature domains. By tracing back the activity from the attended location in the saliency map through the hierarchy of conspicuity and feature maps, we identify the feature that contributes most to the activity of the currently fixated location. We identify a contiguous region around this location with high activity in the feature map that codes for this most active feature. This procedure is motivated by the observation that between-object variability of visual information is significantly higher than within-object variability (Ruderman 1997). Hence, even if two salient objects are close to each other or occluding each other, it is not very likely that they are salient for the same reason. This means that they can be distinguished in the feature maps that code for their respective most active features.

Note, however, that attended regions may not necessarily have a one-to-one correspondence to objects. Groups of similar objects, e.g., a bowl of fruits, may be segmented as one region, as may object parts that are dissimilar from the rest of the object, e.g., a skin-colored hand appearing to terminate at a dark shirt sleeve. We call these regions "proto-objects" because they can lead

to the detection of the actual objects in further iterative interactions between the attention and recognition systems. See the work by Rybak et al. (1998), for instance, for a model that uses the vector of saccades to code for the spatial relations between object parts.

The additional computational cost for region selection is minimal because the feature and conspicuity maps have already been computed during the processing for saliency. Note that although ultimately only the winning feature map is used to segment the attended image region, the interaction of WTA and IOR operating on the saliency map provides the mechanism for sequentially attending several salient locations.

There is no guarantee that the region selection algorithm will find objects. It is purely bottom-up, stimulus driven and has no prior notion of what constitutes an object. Also note that we are not attempting an exhaustive segmentation of the image, such as done by Shi and Malik (2000) or Martin et al. (2004). Our algorithm provides us with a first rough guess of the extent of a salient region. As we will see in the remainder of this thesis, in particular in chapter 5, it works well for localizing objects in cluttered environments.

In some respects, our method of extracting the approximate extent of an object bridges spatial attention with object-based attention. Egly et al. (1994), for instance, report spreading of attention over an object. In their experiments, subjects detected invalidly cued targets faster if they appeared on the same object than if they appeared on a different object than the cue, although the distance between cue and target was the same in both cases. In our method, attention spreads over the extent of a proto-object as well, guided by the feature with the largest contribution to saliency at the attended location. Finding this most active feature is somewhat similar to the idea of flipping through an "object file", a metaphor for a collection of properties that comprise an object (Kahneman and Treisman 1984). However, while Kahneman and Treisman (1984) consider spatial location of an object as another entry in the object file, in our implementation spatial location has a central role as an index for binding together the features belonging to a proto-object. Our method should be seen as an initial step toward a location invariant object representation, providing initial detection of proto-object that allow for subsequent tracking or recognition operations. In fact, in chapter 6, we demonstrate the suitability of our approach as a detection step for multi-target tracking in a machine vision application.

2.5 Outlook

In this chapter we have introduced our model of bottom-up salient region selection based on the model of saliency-based bottom-up attention by Itti et al. (1998). The attended region, which is given by the segmented feature map $\hat{\mathcal{F}}_w$ from eq. 2.11, serves as a means of deploying selective visual attention for:

(i) modulation of neural activity at specific levels of the visual processing hierarchy (chapter 3);

(ii) preferential processing of image regions for learning and recognizing objects (chapter 5);

(iii) initiating object tracking and simplifying the assignment problem in multi-target tracking (chapter 6).