

List of Figures

2.1	Architecture of the model of saliency-based visual attention, adapted from Itti et al. (1998).	6
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.	9
2.3	A network of linear threshold units (LTUs) for computing the <i>argmax</i> 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 <i>argmax</i> 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).	11

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 $\hat{\mathcal{F}}_w$ 12

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 overlaid; 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. 13

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. . . . 18

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. 21

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. 22

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. 23

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. 24

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. 25

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. 29

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. 30

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). 32

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. 33

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). 34

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. 35

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. 36

5.1 Example for SIFT keypoints used for object recognition by Lowe’s algorithm. (a) keypoints of the entire image; (b-d) keypoints extracted for the three most salient regions, representing “monitor,” “computer,” and “set of books.” Restricting the keypoints to a region that is likely to contain an object enables the recognition algorithm to subsequently learn and recognize multiple objects. 43

5.2	Six representative frames from the video sequence recorded by the robot.	46
5.3	The process flow in our multi-object recognition experiments. The image is processed with the saliency-based attention mechanism as described in figure 2.2. In the resulting contrast-modulated version of the image (eq. 5.1), keypoints are extracted (figure 5.1) and used for matching the region with one of the learned object models. A minimum of three keypoints is required for this process (Lowe 1999). In the case of successful recognition, the counter for the matched model is incremented; otherwise a new model is learned. By triggering object-based inhibition of return, this process is repeated for the N most salient regions. The choice of N depends mainly on the image resolution. For the low resolution (320×240 pixels) images used in section 5.3, $N = 3$ is sufficient to cover a considerable fraction (approximately 40 %) of the image area.	47
5.4	Learning and recognition of object patches in a stream of video images from a camera mounted on a robot. Object patches are labeled (x axis), and every recognized instance is counted (y axis). The threshold for “good” object patches is set to 10 instances. Region selection with attention finds 87 good object patches with a total of 1910 instances. With random region selection, 14 good object patches with 201 instances are found. Note the different linear scales on either side of the axis break in the x axis.	50
5.5	Learning and recognition of two objects in cluttered scenes. (a) the image used for learning the two objects; (b-d) examples for images in which objects are recognized as matches with one or both of the objects learned from (a). The patches, which were obtained from segmenting regions at multiple salient locations, are color coded – yellow for the book, and red for the box. The decision of whether a match occurred is made by the recognition algorithm without any human supervision.	51
5.6	Example for learning two objects (c) and (e) from the training image (a) and establishing matches (d) and (f) for the objects in the test image (b), in a different visual context, with different object orientations and occlusions.	52
5.7	Another example for learning several objects from a high-resolution digital photograph. The task is to memorize the items in the cupboard (a) and to identify which of the items are present in the test scenes (b) and (c). Again, the patches are color coded – blue for the soup can, yellow for the pasta box, and red for the label on the beer pack. In (a), only those patches are shown that have a match in (b) or (c), in (b) and (c) only those that have a match in (a).	54

5.8	The SIFT keypoints for the images shown in figure 5.7. The subsets of keypoints identified by salient region selection for each of the three objects are color coded with the same colors as in the previous figure. All other keypoints are shown in black. In figure 5.7 we show all regions that were found for each of the objects – here we show the keypoints from one example region for each object. This figure illustrates the enormous reduction in complexity faced by the recognition algorithm when attempting to match constellations of keypoints between the images.	55
5.9	(a) Ten of the 21 objects used in the experiment. Each object is scaled such that it consists of approximately 2500 pixels. Artificial pixel and scaling noise is added to every instance of an object before merging it with a background image; (b,c) examples of synthetically generated test images. Objects are merged with the background at a random position by alpha-blending. The ratio of object area vs. image area (relative object size) varies between (b) 5 % and (c) 0.05 %.	56
5.10	True positive rate (t) for a set of artificial images without attention (red) and with attention (green) over the relative object size (ROS). The ROS is varied by keeping the absolute object size constant at 2500 pixels ± 10 % and varying the size of the background images. Error bars indicate the standard error for averaging over the performance of the 21 classifiers. The human subject validation curve (blue) separates the difference between the performance with attention (green) and 100 % into problems of the recognition system (difference between the blue and the green curves) and problems of the attention system (difference between the blue curve and 100 %). The false positive rate is less than 0.07 % for all conditions.	57
6.1	(a) ROV Ventana with camera (C) and lights (L). (b) Manual annotation of video tapes in the video lab on shore.	62
6.2	Interactions between the various modules of our system for detecting and tracking marine animals in underwater video.	64
6.3	Example frames with (a) equipment in the field of view; (b) lens glare and parts of the camera housing obstructing the view.	65
6.4	Processing steps for detecting objects in video frames. (a) original frame (720 \times 480 pixels, 24 bits color depth); (b) after background subtraction according to eq. 6.1 (contrast enhanced for displaying purpose); (c) saliency map for the preprocessed frame (b); (d) detected objects with bounding box and major and minor axes marked; (e) the detected objects marked in the original frame and assigned to tracks; (f) direction of motion of the object obtained from eq. 6.11.	66

6.5 Example for the detection of faint elongated objects using across-orientation normalization. (a) original video frame with three faint, elongated objects marked; (b) the frame after background subtraction according to eq. 6.1 (contrast enhanced for illustration); (c) the orientation conspicuity map (sum of all orientation feature maps) without across-orientation normalization; (d) the same map with across-orientation normalization. Objects I and III have a very weak representation in the map *without* normalization (c), and object II is not represented at all. The activation to the right of the arrow tip belonging to object II in (c) is due to a marine snow particle next to the object and is not due to object II. Compare with (d), where object II as well as the marine snow particle create activation. In the map *with* normalization (d), all three objects have a representation that is sufficient for detection. 67

6.6 A schematic for a neural implementation of across-orientation normalization using an inhibitory interneuron. This circuit would have to be implemented at each image location for this normalization to function over the entire visual field. 68

6.7 Geometry of the projection problem in the camera reference frame. The nodal point of the camera is at the origin, and the camera plane is at z_c . The object appears to be moving at a constant speed into the x and z direction as the camera moves toward the object. Eq. 6.3 describes how the projection of the object onto the camera plane moves in time. 69

7.1 Example stimuli for means of transport, animals, and distracters, as well as example masks. Masks are created by superimposing a naturalistic texture on a mixture of white noise at different spatial frequencies (Li et al. 2002), surrounded by a frame with broken-up segments of orange, blue, and purple. Note that the thickness of the color frames is exaggerated threefold in this figure for illustration. 80

7.2	Experimental set-up. Each trial starts 1300 ms before target onset with a blank gray screen. At 650 ± 25 ms before target onset, a white fixation dot ($4.1' \times 4.1'$) is presented at the center of the display. At a variable cue target interval (CTI) before target onset, a word cue (0.5° high, between 1.1° and 2.5° wide) appears at the center of the screen for 17 ms (two frames), temporarily replacing the fixation dot for CTIs less than 650 ms. At 0 ms, the target stimulus, consisting of a gray-level photograph and a color frame around it, is presented at a random position on a circle around the fixation dot such that the image is centered around 6.4° eccentricity. After a stimulus onset asynchrony (SOA) of 200–242 ms, the target stimulus is replaced by a perceptual mask. The mask is presented for 500 ms, followed by 1000 ms of blank gray screen to allow the subjects to respond. In the case of an error, acoustic feedback is given (pure tone at 800 Hz for 100 ms), followed by 100 ms of silence. After this, the next trial commences.	81
7.3	Histogram of the reaction times of all trials. Trials with reaction times below 200 ms and more than four standard deviations above the mean (above 995 ms) were discarded as outliers (1 % of the data).	83
7.4	Reaction times (top, blue) and error rates (bottom, red) for single task blocks, task repeat trials, and task switch trials in mixed blocks for $n = 5$ subjects. Error bars are s.e.m. For RT, both mixing and switch cost are significant at a CTI of 50 ms, but not at CTIs of 200 ms and 800 ms ($p > 0.05$, t-test). The drop of the single task RT at 200 ms compared to 50 ms and 800 ms is not significant ($p > 0.05$, t-test). For error rate, only switch cost at a CTI of 800 ms is statistically significant. There are no other significant effects for error rate.	84
7.5	Mixing cost in RT (blue) and error rate (red) for all subjects for CTI = 50 ms, plotted by task group. While mixing cost in RT is significantly higher for IMG than for COL tasks, mixing cost in error rate is significantly higher for COL than for IMG tasks. . .	86
7.6	Switch cost in RT at a CTI of 50 ms for different switch conditions (blue) and pooled over all conditions (white). The white bar corresponds to the difference labeled as $C_{\text{switch}}^{\text{RT}}$ in figure 7.4. Error bars are standard errors as defined in eqs. 7.3 and 7.4. Switch cost is only significant when switching between the IMG and COL task groups, but not when switching within the groups.	87
A.1	Illustration of one-dimensional filtering and subsampling. (A) convolution with a filter of length 3 (first to second row), followed by decimation by a factor of 2 (third and fourth row) – the pixels marked with a red cross are removed; (B) integral operation of convolution with a filter of length 4 and decimation by a factor of 2.	96

A.2	Example of repeated filtering and subsampling of an image of size 31×31 pixels with only one pixel activated with: (A) a 5×5 filter with subsequent subsampling; and (B) a 6×6 filter with integrated subsampling. Bright pixels indicate high and dark pixels low activity.	97
A.3	Schematic of the correlation-based motion detector by Hassenstein and Reichardt (1956). The activation of each receptor is correlated with the time delayed signal from its neighbor. The leftwards versus rightwards opponency operation prevents full field illumination or full field flicker from triggering the motion output signal.	100
A.4	Illustration of center-surround receptive fields for motion perception. The five dots at the center of the display are salient even though they are stationary because they are surrounded by a field of moving dots.	102
A.5	Feature maps for motion directions right (a), down (b), up (c), left (d), and the motion conspicuity map (e) in response to a rightward moving white bar (f).	103
A.6	The Gaussian model for skin hue. The individual training points are derived from 3974 faces in 1153 color photographs. Each dot represents the average hue for one face and is plotted in the color of the face. The green cross represents the mean (μ_r, μ_g) , and the green ellipses the 1σ and 2σ intervals of the hue distribution.	105
A.7	Example of a color image with faces (left) processed with the skin hue model from eq. A.14, using the parameters from table A.2 (right). The color scale on the right reflects how closely hue matches the mean skin hue, marked with a green cross in figure A.6. Note that face regions show high values, but other skin colored regions do as well, e.g. arms and hands or the orange T-shirt of the boy on the right.	105
B.1	Screen shot of a typical display while running the SaliencyToolbox.	108

