On the usefulness of attention for object recognition

Dirk Walther*, Ueli Rutishauser*, Christof Koch, and Pietro Perona
Computation and Neural Systems Program,
California Institute of Technology, Pasadena, California 91125, USA
{walther|urut|koch|perona}@caltech.edu

Abstract

Today's object recognition systems have become very good at learning and recognizing isolated objects or objects in images with little clutter. However, unsupervised learning and recognition in highly cluttered scenes or in scenes with multiple objects are still problematic. Faced with the same issue, the brain employs selective visual attention to select relevant parts of the image and to serialize the perception of individual objects. In this paper we demonstrate the use of a computational model of bottom-up visual attention for object recognition in machine vision. By comparing the performance of David Lowe's recognition algorithm with and without attention, we quantify the usefulness of attention for learning and recognizing multiple objects from complex scenes, and for learning and recognizing objects in scenes with large amounts of clutter.

1. Introduction

The field of object recognition has seen tremendous progress over the past years. However, most recognition systems require segmented and labeled objects for training, or at least that the training object is the dominant part of the training images. It is usually not possible to train them on unlabeled images that contain large amounts of clutter or multiple objects.

Imagine a situation in which you are shown a scene, e.g. a shelf with groceries, and later you are asked to identify which of these items you recognize in a different scene, e.g. in your grocery cart. While this is a common situation in everyday life and easily accomplished by humans, it is a difficult problem for machine vision. How is it that humans can deal with these issues with such apparent ease?

The human visual system is able to reduce the amount of incoming visual data to a small but relevant amount of information for higher-level cognitive processing using selective visual attention. Attention is the process of selecting and gating visual information based on saliency in the image itself (bottom-up), and on prior knowledge about scenes, objects and their interrelations (top-down) [4, 10]. Upon closer inspection, the "grocery cart problem" (also known as the "bin of parts problem" in the robotics community) poses two complementary challenges – serializing the perception and learning of relevant information (objects), and suppressing irrelevant information (clutter). Visual attention addresses both problems by selectively enhancing perception at the attended location, and by successively shifting the focus of attention to multiple locations.

Several computational implementations of models of visual attention have been published. Tsotsos and colleagues [26] use local winner-take-all networks and top-down mechanisms to selectively tune model neurons at the attended location. Deco & Schürmann modulate the spatial resolution of the image based on a top-down attentional control signal. Itti & Koch [12] introduced a model for bottom-up selective attention based on serially scanning a saliency map that is computed from local feature contrasts. Closely following and extending Duncan's Integrated Competition Hypothesis [6], Sun & Fisher [25] developed and implemented a common framework for object-based and location-based visual attention based on "groupings". Presented with a manually preprocessed input image, their model replicates human viewing behavior for artificial and natural scenes.

The main motivation for attention in machine vision is cueing subsequent visual processing stages such as object recognition to improve performance and/or efficiency. However, little work has been done to verify these benefits experimentally (but see [5, 19, 27]). The focus of this paper is on testing the usefulness of selective visual attention for object recognition experimentally. We do not intend to compare the performance of the various attention systems – this would be an interesting study in its own right. Instead, we use Itti & Koch's saliency-based attention system, endow it with a mechanism for identifying regions that are likely to contain objects around salient locations, and use this sys-

*These authors contributed equally to this work.

tem to demonstrate the benefits of selective visual attention for: (i) learning sets of object representations from single images, and identifying these objects in cluttered test images containing target and distractor objects; and (ii) object learning and recognition in highly cluttered scenes.

In computer vision, image segmentation algorithms such as nCuts [23] are successfully used to recognize multiple objects in images [2]. While saliency-based attention concentrates on feature *contrasts*, segmentation attempts to find regions that are *homogenous* in certain features. A more detailed comparison of these complementary approaches would be interesting, but is beyond the scope of this paper.

2. Approach

To investigate the effect of attention on object recognition independent of the specific task, we do not consider a priori information about the images or the objects. Hence, we do not make use of top-down attention and rely solely on bottom-up, saliency-based attention.

2.1. Bottom-up saliency-based region selection

Our attention system is based on the Itti et al. [12] implementation of the Koch & Ullman [14] saliency-based model of bottom-up attention. This model's usefulness as a front-end for object recognition is limited by the fact that its output is merely a pair of coordinates in the image corresponding to the most salient location. We introduce a method for extracting an image region at salient locations from low-level features with negligible additional computational cost. We briefly review the saliency model in order to explain our extensions in the same formal framework.

The input image $\mathcal{I}$ is sub-sampled into a Gaussian pyramid, and each pyramid level is decomposed into channels for red (R), green (G), blue (B), yellow (Y), intensity (I) and local orientation (O_θ). If r, g and b are the red, green and blue values of the color image, normalized by the image intensity I, then $R = r-(g+b)/2$, $G = g-(r+b)/2$, $B = b-(r+g)/2$, and $Y = r+g-2(|r-g|+b)$ (negative values are set to zero). Local orientations O_θ are obtained by applying steerable filters [24, 18] to the images in the intensity pyramid I. From these channels, center-surround "feature maps" are constructed and normalized:

$$\mathcal{F}_{I,c,s} = \mathcal{N}\left(|I(c) \ominus I(s)|\right) \quad (1)$$

$$\mathcal{F}_{RG,c,s} = \mathcal{N}\left(|(R(c)-G(c)) \ominus (R(s)-G(s))|\right) \quad (2)$$

$$\mathcal{F}_{BY,c,s} = \mathcal{N}\left(|(B(c)-Y(c)) \ominus (B(s)-Y(s))|\right) \quad (3)$$

$$\mathcal{F}_{\theta,c,s} = \mathcal{N}\left(|O_\theta(c) \ominus O_\theta(s)|\right) \quad (4)$$

where $\ominus$ denotes the across-scale difference between two maps at the center (c) and the surround (s) levels of the respective feature pyramids. $\mathcal{N}(\cdot)$ is an iterative, nonlinear

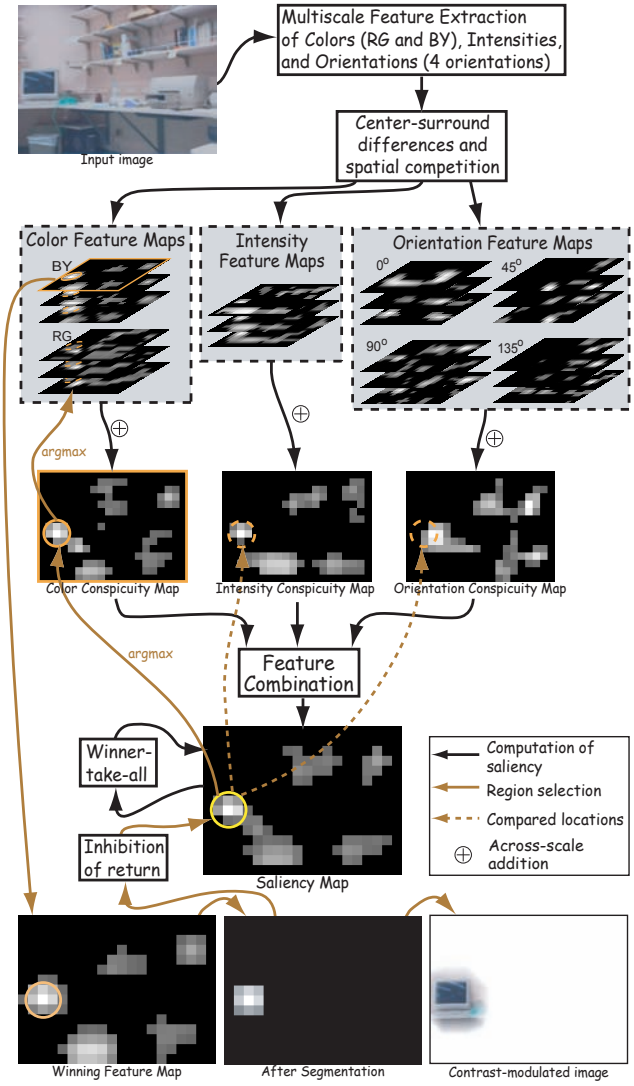

Figure 1. Illustration of the processing steps for obtaining the attended region.

normalization operator (for details see [11]). The feature maps are summed over the center-surround combinations using across-scale addition $\oplus$, and the sums are normalized again:

$$\bar{\mathcal{F}}_l = \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{F}_{l,c,s}\right) \forall l \in L_I \cup L_C \cup L_O \quad (5)$$

with

$$L_I = \{I\},\ L_C = \{RG, BY\},$$
$$L_O = \{0°, 45°, 90°, 135°\} \quad (6)$$

For the general features color and orientation, the contributions of the sub-features are linearly 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. 5:

$$\mathcal{C}_I = \bar{\mathcal{F}}_I,\ \mathcal{C}_C = \mathcal{N}\left(\sum_{l \in L_C} \bar{\mathcal{F}}_l\right),\ \mathcal{C}_O = \mathcal{N}\left(\sum_{l \in L_O} \bar{\mathcal{F}}_l\right) \tag{7}$$

All conspicuity maps are combined into one saliency map:

$$\mathcal{S} = \frac{1}{3} \sum_{k \in \{I,C,O\}} \mathcal{C}_k \tag{8}$$

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 winning location (x_w, y_w) of this process is attended to (the yellow circle in fig. 1).

While Itti's model successfully identifies this most salient location in the image, it has no notion of the extend of the image region that is salient around this location. We introduce a method to estimate this region based on the maps and salient locations computed thus far. Looking back at the conspicuity maps, we find the one map that contributes most to the activity at the most salient location:

$$k_w = \underset{k \in \{I,C,O\}}{\operatorname{argmax}}\ \mathcal{C}_k(x_w, y_w) \tag{9}$$

We look further which feature map contributes most to the activity at this location in the conspicuity map $\mathcal{C}_{k_w}$:

$$(l_w, c_w, s_w) = \underset{l \in L_{k_w}, c \in \{2,3,4\}, s \in \{c+3, c+4\}}{\operatorname{argmax}} \mathcal{F}_{l,c,s}(x_w, y_w) \tag{10}$$

with L_{k_w} as defined in eq. 6. The "winning" feature map $\mathcal{F}_{l_w,c_w,s_w}$ is segmented using region growing around (x_w, y_w) and adaptive thresholding [9] (bottom of fig. 1). 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, thus enabling the model to attend to several locations subsequently, in order of decreasing saliency.

We derive a mask $\mathcal{M}$ at image resolution by thresholding $\hat{\mathcal{F}}_w$, scaling it up and smoothing it. Smoothing can be achieved by convolving with a separable two-dimensional Gaussian kernel ($\sigma = 20$ pixels). We use a computationally more efficient method, consisting of opening the binary mask with a disk of 8 pixels radius as a structuring element, and using the inverse of the chamfer 3-4 distance for smoothing the edges of the region. $\mathcal{M}$ is normalized to be 1 within the attended region, 0 outside the region, and it has intermediate values at the region's edge. We use this mask to modulate the contrast of the original image $\mathcal{I}$ (dynamic range $[0, 255]$):

$$\mathcal{I}'(x, y) = [255 - \mathcal{M}(x, y) \cdot (255 - \mathcal{I}(x, y))] \tag{11}$$

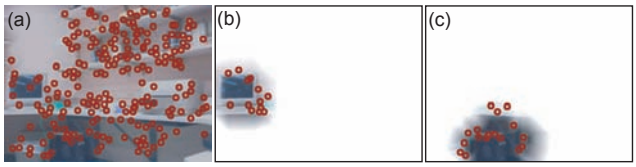

Figure 2. Keypoints for (a) the entire image; (b,c) the two most salient regions.

where $[\cdot]$ symbolizes the rounding operation. Eq. 11 is applied separately to the r, g and b channels of the image. $\mathcal{I}'$ is used as the input of the recognition algorithm instead of $\mathcal{I}$ (fig. 2).

It is not immediately clear that the approach of segmenting a region by its most salient feature yields a good estimate of the object's size and shape, but the approach works remarkably well for a variety of natural images and videos (see the next sections). Further research is required to understand why this simple method works so well.

An advantage of using the map of the most salient feature for segmenting instead of the saliency map is the sparser representation in the feature map, which makes segmentation easier. The additional computational cost for the region estimation is minimal, because the feature and conspicuity maps have already been computed during the processing for saliency.

2.2. Object learning and recognition with attention

For all experiments described in this paper, we use the object recognition algorithm by Lowe [16]. The algorithm uses a Gaussian pyramid built from a gray-value representation of the image to extract local features, "keypoints", at the extreme points of differences between pyramid levels (fig. 2a). The keypoints are represented in a 128-dimensional space in a way that makes them invariant to scale and in-plane rotation [17].

Recognition is performed by matching keypoints found in the test image with stored object models. This is accomplished by searching for nearest neighbours in the 128-dimensional space using the best-bin-first search method. To establish object matches, similar hypotheses are clustered using the Hough transform. Affine transformations relating the candidate hypotheses to the keypoints from the test image are used to find the best match. To some degree, model matching is stable for perspective distortion and rotation in depth [17].

In our model, we have the additional step of finding salient patches as described above for learning and recognition before keypoints are extracted (fig. 2b,c). The use of contrast modulation as a means of deploying object-based

attention is motivated by neurophysiological experiments that show that in the cortical representation, attentional enhancement acts in a manner equivalent to increasing stimulus contrast [21]; as well as by its usefulness with respect to Lowe's recognition algorithm. Keypoint extraction relies on finding luminance contrast peaks across scales. As we remove all contrast from image regions outside the attended object (eq. 11), no keypoints are extracted there, and we limit the formation of a model to the attended region.

The number of fixations used for recognition and learning depends on the resolution of the images, and on the amount of visual information. In low-resolution images with few objects, three fixations may be sufficient to cover the relevant parts of the image. In high-resolution images with a lot of visual information, up to 30 fixations are required to sequentially attend to all objects. Humans and monkeys, too, need more fixations, to analyze scenes with richer information content [22]. The number of fixations required for a set of images is determined by monitoring after how many fixations the serial scanning of the saliency map starts to cycle.

It is common in object recognition to use interest operators [8] or salient feature detectors [13] to select features for learning an object model. This is different, however, from selecting an image region and limiting the learning and recognition of objects to this region.

In the next section, we verify that the selection of salient image regions does indeed produce meaningful results when compared with random region selection. In the two sections after that, we report experiments that address the benefits of attention for serializing visual information processing and for suppressing clutter.

3. Selective attention versus random patches

In the first experiment, we compare our saliency-based region selection method with randomly selected image patches. If regions found by the attention mechanism are indeed more likely to contain objects, then one would expect that object learning and recognition show better performance for these regions than for randomly selected image patches. Since human photographers tend to have a bias towards centering and zooming on objects, we make use of a robot instead for collecting a large number of test images in an unbiased fashion.

3.1. Experimental setup

For this experiment, we used a robot equipped with a camera as an image acquisition tool. The robot's navigation followed a simple obstacle avoidance algorithm using infrared range sensors for control. The camera was mounted on top of the robot at about 1.2 m height. Color images

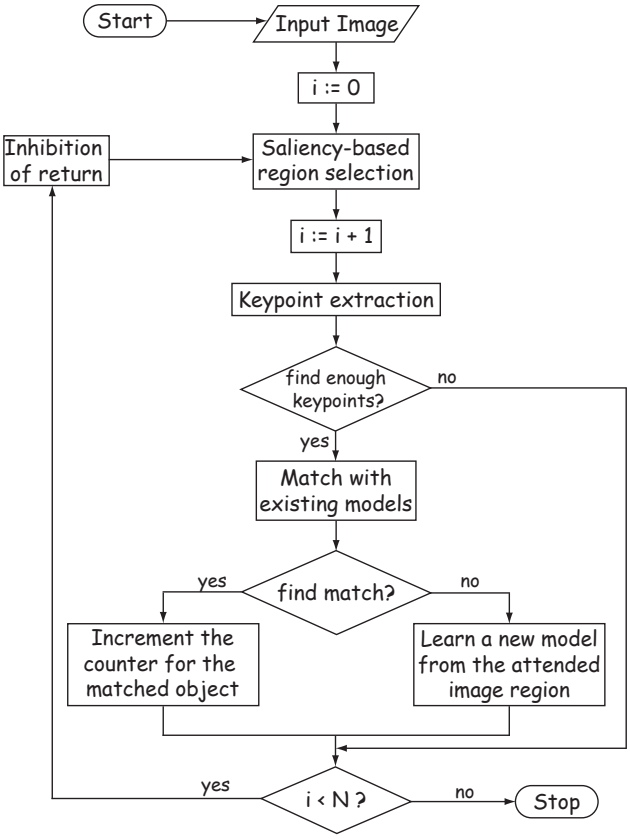

Figure 3. Process flow for the experiments.

were recorded at 320×240 pixels resolution at 5 frames per second. A total of 1749 images were recorded during an almost 6 min run (see [1] for the full video). Since vision was not used for navigation, the images taken by the robot are unbiased. The robot moved in a closed environment (indoor offices/labs, four rooms, approximately 80 m^2). Hence, the same objects are likely to appear multiple times in the sequence.

The process flow for selecting, learning, and recognizing salient regions is shown in fig. 3. Because of the low resolution of the images, we use only three fixations ($N = 3$) in each image for recognizing and learning objects. Each newly learned object is assigned a unique label, and we count the number of times the object is recognized in the entire image set. An object is considered "useful" if it is recognized at least once after learning, thus appearing at least twice in the sequence.

We repeated the experiment without attention, using the recognition algorithm on the entire image. In this case, the system is only capable of detecting large scenes but not individual objects. For a more meaningful control, we repeated the experiment with randomly chosen image regions. These regions are created by a pseudo region growing operation

at the saliency map resolution. Starting from a randomly selected location, the original threshold condition for region growth is replaced by a decision based on a uniformly drawn random number. The patches are then treated the same way as true attention patches (see section 2.1). The parameters are adjusted such that the random patches have approximately the same size distribution as the attention patches.

Ground truth for all experiments is established manually. This is done by displaying every match established by the algorithm to a human subject who has to rate the match as either correct or incorrect. The false positive rate is derived from the number of patches that were incorrectly associated with an object.

3.2. Results

Using the recognition algorithm on the entire images results in 1707 of the 1749 images being pigeon-holed into 38 unique "objects", representing non-overlapping large views of the rooms visited by the robot. The remaining 42 non-"useful" images are learned as new "objects", but then never recognized again.

The models learned from these large scenes are not suitable for detecting individual objects. In this experiment, we have 85 false positives (5.0%), i.e. the recognition system indicates a match between a learned model and an image, where the human subject does not indicate an agreement. Clearly recogniton without attention does not yield any meaningful results in these experiments.

Attentional selection identifies 3934 useful regions in the approximately 6 min of processed video, associated with 824 objects. Random region selection only yields 1649 useful regions, associated with 742 objects (table 1). With saliency-based region selection, we find 32 (0.8%) false positives, with random region selection 81 (6.8%).

Table 1. Comparing attention and random patches.

	Attention	Random
# of patches recognized	3934	1649
average per image	2.25	0.95
# of objects	824	742
# of "good" objects	87 (10.6%)	14 (1.9%)
# of patches associated with "good" objects	1910 (49%)	201 (12%)
false positives	32 (0.8%)	81 (6.8%)

To better compare the two methods of region selection, we assume that "good" objects (e.g. objects useful as landmarks for robot navigation) should be recognized multiple times throughout the video sequence, since the robot visits the same locations repeatedly. We sort the objects by their number of occurrences and set an arbitrary threshold of 10 recognized occurrences for "good" objects for this analysis.

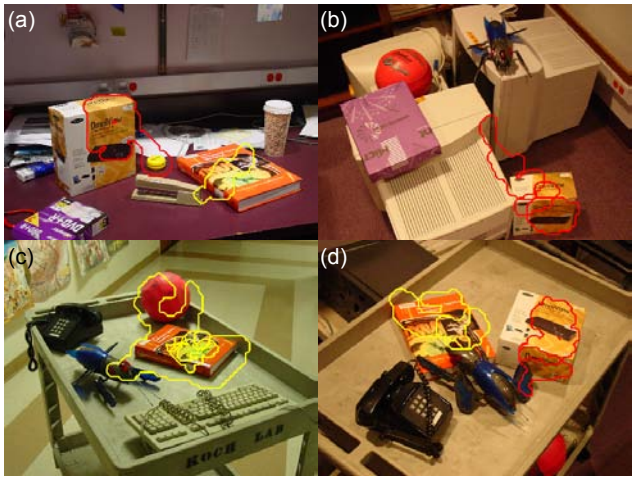

Figure 4. Learning and recognition in cluttered scenes. (a) training image; (b-d) test images – matched objects are color coded.

With this threshold in place, attentional selection finds 87 "good" objects with a total of 1910 patches associated to them. With random regions, only 14 "good" objects are found with a total of 201 patches. The number of patches associated with "good" objects is computed as:

$$N_L = \sum_{\forall i: n_i \geq 10} n_i \qquad (n_i \in \mathcal{O}) \tag{12}$$

where $\mathcal{O}$ is an ordered set of all learned objects, sorted descending by the number of detections.

From these results it is clear that the regions selected by the attentional mechanism are more likely to contain objects that can be recognized repeatedly from various viewpoints than randomly selected regions. Now that we have established the suitability of our saliency-based region selection method for selecting objects, we move on to address its effects on processing images containing multiple objects, and on object learning and recognition in highly cluttered scenes.

4. Learning multiple objects

In this experiment, we test the hypothesis that attention can enable the learning and recognition of multiple objects in single natural scenes. We use high-resolution digital photographs of home and office environments for this purpose.

4.1. Experimental setup

We placed a number of objects into different settings in office and lab environments and took pictures of the objects with a digital camera. We obtained a set of 102 images at a resolution of 1280×960 pixels (see [1] for the complete image set). Images can contain large or small subsets of the objects. We select one of the images for training (fig. 4a). The other 101 images are used as test images.

For learning and recognition we use 30 fixations, which cover about 50% of the image area. Learning is performed completely unsupervised (see fig. 3). A new model is learned at each fixation. During testing, each fixation on the test image is compared to each of the learned models. Ground truth is established manually.

4.2. Results

From the training image, the system learns models for two objects that can be recognized in the test images – a book and a box (fig. 4). Of the 101 test images, 23 contain the box, and 24 the book, and of these, four images contain both objects. Table 2 shows the recognition results for the two objects.

Table 2. Results for recognizing two objects.

object	hits	misses	false positives
box	21 (91%)	2 (9%)	0 (0%)
book	14 (58%)	10 (42%)	2 (2.6%)

Even though the recognition rates for the two objects are rather low, one should consider that one unlabeled image is the only training input given to the system (one-shot learning). From this one image, the combined model is capable of identifying the book in 58%, and the box in 91% of all cases, with only two false positives for the book, and none for the box. It is difficult to compare this performance with some baseline, since this task is impossible for the recognition system alone, without any attentional mechanism.

5. Objects in cluttered scenes

In the previous section, we have shown that selective attention enables the learning of multiple objects from single images. In this section, we investigate how attention can help to recognize objects in highly cluttered scenes.

5.1. Experimental setup

To systematically evaluate recognition performance with and without attention, we use images generated by randomly merging an object with a background image (fig. 5).

Figure 5. (a) Six of the 21 test objects. (b,c) Synthetically generated test images with relative object sizes of (b) 5% and (c) 0.05%.

This design of the experiment enables us to generate a large number of test images in a way that gives us good control of the amount of clutter versus the size of the objects in the images, while keeping all other parameters constant [22]. Since we construct the test images, we also have easy access to ground truth. We use natural images for the backgrounds so that the abundance of local features in our test images matches that of natural scenes as closely as possible.

We quantify the amount of clutter in the image by the *relative object size* (ROS), defined as the ratio of the number of pixels of the object over the number of pixels in the entire image. To avoid issues with the recognition system due to large variations in the *absolute* size of the objects, we leave the number of pixels for the objects constant (with the exception of intentionally added scale noise), and vary the ROS by changing the size of the backgound images in which the objects are embedded.

To introduce variability in the appearance of the objects, each object is rescaled by a random factor between 0.9 and 1.1, and uniformly distributed random noise between -12 and 12 is added to the red, green and blue value of each object pixel (dynamic range is $[0, 255]$). Objects and backgrounds are merged by blending with an alpha value of 0.1 at the object border, 0.4 one pixel away, 0.8 three pixels away from the border, and 1.0 inside the objects, more than three pixels away from the border. This prevents artificially salient borders due to the object being merged with the background.

We created six test sets with ROS values of 5%, 2.78%, 1.08%, 0.6%, 0.2% and 0.05%, each consisting of 21 images for training (one image for each object) and 420 images for testing (20 test images for each object). The background images for training and test sets are randomly drawn

from disjoint image pools to avoid false positives due to features in the background. A ROS of 0.05% may seem unrealistically low, but humans are capable of recognizing objects with a much smaller relative object size, for instance for reading street signs while driving [15].

During training, object models are learned at the five most salient locations of each training image. That is, the object has to be learned by finding it in a training image. Learning is unsupervised, and thus, most of the learned object models do not contain an actual object. During testing, the five most salient regions of the test images are compared to each of the learned models. As soon as a match is found, positive recognition is declared. Failure to attend to the object during the first five fixations leads to a failed learning or recognition attempt.

5.2. Results

Learning from our data sets results in a classifier that can recognize $K = 21$ objects. The performance of each classifier i is evaluated by determining the number of true positives T_i and the number of false positives F_i. The overall true positive rate t (also known as detection rate) and the false positive rate f for the entire multi-class classifier [7] are then computed as:

$$t = \frac{1}{K} \sum_{i=1}^{K} \frac{T_i}{N_i} \tag{13}$$

$$f = \frac{1}{K} \sum_{i=1}^{K} \frac{F_i}{\overline{N}_i} \tag{14}$$

Here, N_i is the number of positive examples of class i in the test set, and $\overline{N}_i$ is the number of negative examples of class i. Since in our experiments the negative examples of one class consist of the positive examples of all other classes, and since there are equal numbers of positive examples for all classes, we can write:

$$\overline{N}_i = \sum_{j=1, j \neq i}^{K} N_j = (K-1)N_i \tag{15}$$

To evaluate the performance of the classifier it is sufficient to consider only the true positive rate, since the false positive rate is consistently below 0.07% for all conditions, even without attention and at the lowest ROS of 0.05%.

We evaluate the true positive rate for each data set with three different methods: (i) learning and recognition without attention; (ii) learning and recognition with attention and (iii) human validation of attention (fig. 6). The third procedure attempts to explain what part of the performance difference between (ii) and 100% is due to shortcomings of the attention system, and what part is due to problems with the recognition system.

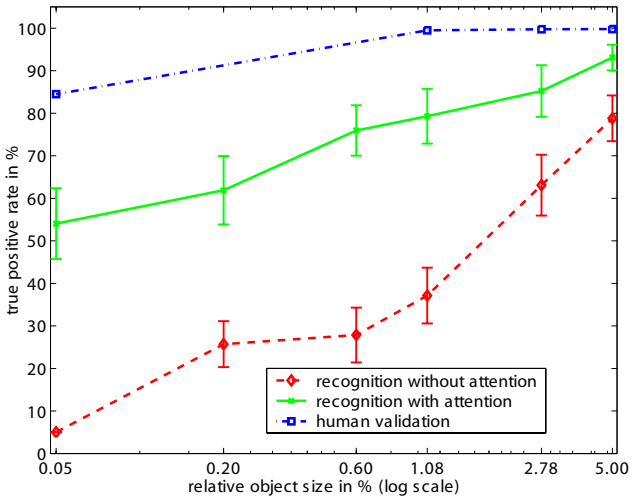

Figure 6. True positive rate t for a set of artificial images. Error bars indicate the standard error for averaging over the performance of the 21 classifiers.

For human validation, all images that cannot be recognized automatically are evaluated by a human subject. The subject can only see the five attended regions of all training images and of the test images in question, all other parts of the images are blanked out. Solely based on this information, the subject is asked to indicate matches. In this experiment, matches are established whenever the attention system extracts the object correctly during learning and recognition.

In the cases in which the human subject is able to identify the objects based on the attended patches, the failure of the combined system is due to shortcomings of the recognition system. On the other hand, if the human subject fails to recognize the objects based on the patches, the attention system is the component responsible for the failure. As can be seen in fig. 6, the human subject can recognize the objects from the attended patches in most cases, which implies that the recognition system is the cause for the failure rate. Only for the smallest ROS (0.05%), the attention system contributes significantly to the failure rate.

The results in fig. 6 demonstrate that attention has a sustained effect on recognition performance for all reported relative object sizes. With more clutter (smaller ROS), the influence of attention becomes more accentuated. In the most difficult cases (ROS of 0.05%), attention increases the true positive rate by a factor of 10. Note that for $ROS > 5\%$ learning and recognition done on the entire image (red, dashed line in fig. 6) works well without attention, as reported in [16].

6. Conclusion

We set out to explore how attentional region selection and object recognition can interact in order to improve robustness of recognition and to enable new modes of operation. In the experiments presented in this paper, we show by example and by quantitative analysis that saliency-based region selection improves object recognition in highly cluttered scenes considerably. Other modes of operation, such as learning multiple objects from single images, are only made possible by attentional selection. With these new capabilities, a solution for the "grocery cart problem" is no longer out of reach for machine vision.

Although we limit our experiments to a particular attention system and to a particular recognition system, we believe that our results can be generalized to other system configurations. However, more experimental testing would be required to verify this speculation. In certain applications, top-down knowledge can be very useful for visual processing, in addition to the bottom-up saliency-based attention described here (see for instance [20, 26, 3]). We have selected Lowe's recognition algorithm for our experiments because of its suitability for general object recognition.

By the example of the system configuration that we have chosen, we demonstrate the power of the synergy between recognition and attention in two domains – learning and recognition of several objects in single images, and learning and recognition in highly cluttered scenes.

Acknowledgments

Funding for this project was provided by NSF-ERC, NSF-ITR, NIH, NIMH, the Keck Foundation, and a Sloan-Swartz Fellowship to U.R. The region selection code was developed by the authors as part of the "iNVT" community effort (http://ilab.usc.edu/toolkit). We would like to thank the anonymous reviewers for comments on this manuscript, and Evolution Robotics for making their robotic vision software development kit available to us. High-resolution background images were provided by TNO Human Factors Research Institute, the Netherlands.

References

[1] http://klab.caltech.edu/∼walther/wapcv04.

[2] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. In *CVPR*, 2003.

[3] G. Deco and B. Schürmann. A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision Res.*, 40(20):2845–2859, 2000.

[4] R. Desimone and J. Duncan. Neural mechanisms of selective visual-attention. *Annu. Rev. Neurosci.*, 18:193–222, 1995.

[5] S. Dickinson, H. Christensen, J. K. Tsotsos, and G. Olofsson. Active object recognition integrating attention and viewpoint control. *CVIU*, 63(67-3):239–260, 1997.

[6] J. Duncan. Integrated mechanisms of selective attention. *Curr. Opin. Biol.*, 7:255–261, 1997.

[7] T. Fawcett. ROC Graphs: Notes and practical considerations for data mining researchers. *HP Technical Report*, 4, 2003.

[8] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conf.*, pages 147–151, 1988.

[9] L. Itti, L. Chang, and T. Ernst. Segmentation of progressive multifocal leuko-encephalopathy lesions in fluid-attenuated inversion recovery magnetic resonance imaging. *J. of Neuroimaging*, 11(4):412–417, 2001.

[10] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[11] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *J. of Electronic Imaging*, 10(1):161–169, 2001.

[12] 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.

[13] T. Kadir and M. Brady. Scale, saliency and image description. *IJCV*, 30(2):77–116, 2001.

[14] C. Koch and S. Ullman. Shifts in selective visual-attention – towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.

[15] G. Legge, D. Pelli, G. Rubin, and M. Schleske. The psychophysics of reading. *Vision Res.*, 25(2):239–252, 1985.

[16] D. Lowe. Object recognition from local scale-invariant features. *ICCV*, pages 1150–1157, 1999.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV (in press)*, 2004.

[18] R. Manduchi, P. Perona, and D. Shy. Efficient deformable filter banks. *IEEE Trans. Sig. Proc.*, 46(4):1168–1173, 1998.

[19] F. Miau and L. Itti. A neural model combining attentional orienting to object recognition. *IEEE Engineering in Medicine and Biology Society*, 2001.

[20] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson. Top-down control of visual attention in object detection. *ICIP*, 2003.

[21] J. H. Reynolds, T. Pasternak, and R. Desimone. Attention increases sensitivity of V4 neurons. *Neuron*, 26(3):703–714, 2000.

[22] D. L. Sheinberg and N. K. Logothetis. Noticing familiar objects in real world scenes. *J. Neurosci.*, 21(4):1340–1350, 2001.

[23] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE PAMI*, 22(8):888–905, 2000.

[24] E. Simoncelli and W. Freeman. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *ICIP*, 1995.

[25] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 20(11):77–123, 2003.

[26] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo. Modeling visual-attention via selective tuning. *Artificial Intelligence*, 78:507–545, 1995.

[27] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition – a gentle way. *Biol. Motivated Comp. Vision*, pages 472–479, 2002.