Is bottom-up attention useful for object recognition?

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

Abstract

A key problem in learning multiple objects from unlabeled images is that it is a priori impossible to tell which part of the image corresponds to each individual object, and which part is irrelevant clutter which is not associated to the objects. We investigate empirically to what extent pure bottom-up attention can extract useful information about the location, size and shape of objects from images and demonstrate how this information can be utilized to enable unsupervised learning of objects from unlabeled images. Our experiments demonstrate that the proposed approach to using bottom-up attention is indeed useful for a variety of applications.

1. Introduction

The field of object recognition has seen tremendous progress over the past years, both for specific domains such as face detection [26, 30], and for more general object domains [16, 32, 8, 25, 24]. Most of these require segmented and labeled objects for training, or at least that the training object is the dominant part of the training images. None of these algorithms can be trained 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, none of the methods mentioned above is capable of coping with this situation. 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 processing in the thalamus and visual cortex using selective visual attention. Attention is a process of selecting and gating visual information based on saliency in the image itself (bottom-up) and on prior knowledge about the scene (top-down) [5, 12]. We postulate that the key to solving the "grocery cart problem" is visual selection. If bottom-up attention processes could select image regions that contain objects with high likelihood, then a recognition system could be trained on such patches and could thus learn the appearance of individual objects.

While several computational implementations of models of visual attention have been published [29, 4, 14], little work has been done in investigating its benefits for object learning and recognition in a machine vision context (but see [6, 20, 31]). In this paper, we examine the usefulness of saliency-based visual attention for object learning and recognition in three different experimental settings – (i) learning and recognition of individual objects in highly cluttered scenes; (ii) learning sets of objects (inventories) from single images, and identifying these objects in cluttered test images containing target and distractor objects; and (iii) online learning and recognition of landmarks useful for spatial orientation and robot navigation.

2. Approach

We intend to investigate whether and to what extent images contain useful information about the location, shape and size of objects. We furthermore aim to demonstrate how this information about objects can be utilized for unsupervised object learning and recognition. Our goal is to investigate this in a task-independent manner and we thus do not make use of top-down attention and rely solely on bottom-up attention. While understanding of the theoretical basis of attention requires more research, multiple bottom-up attentional frameworks have been established on an empirical basis [12]. For the experiments in this paper we use a saliency-based framework [14] (see section 2.1).

In contrast to key-point or interest point selectors [9] [32] we are not extracting multiple features of the same object but rather the *region* of the image where the object is lo-

* These authors contributed equally to this work.

cated. Others have proposed a segmentation based approach to achive the same goal [28].

For object recognition we use Lowe's algorithm, which uses local scale-invariant features [16], as an example for a state-of-the-art general purpose recognition system with one-shot learning capability.

2.1. Bottom-up saliency-based region selection

Our attention system is based on the Itti et al. [14] implementation of the Koch & Ullman [15] saliency-based model of bottom-up attention. While the model's usefulness has been demonstrated in various contexts (e.g. prediction of eye movements of humans [22]), its ability to serve 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 the image region that is likely to contain the attended objects 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 $\mathcal{I}$ (fig. 1a) is sub-sampled into a Gaussian pyramid [2], 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 channels, 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 to the images in the intensity pyramid I [18]. From these channels, center-surround "feature maps" are constructed and normalized:

$$\begin{aligned}
\mathcal{F}_{I,c,s} &= \mathcal{N}\left(|I(c) \ominus I(s)|\right) && (1)\\
\mathcal{F}_{RG,c,s} &= \mathcal{N}\left(|(R(c) - G(c)) \ominus (R(s) - G(s))|\right) && (2)\\
\mathcal{F}_{BY,c,s} &= \mathcal{N}\left(|(B(c) - Y(c)) \ominus (B(s) - Y(s))|\right) && (3)\\
\mathcal{F}_{\theta,c,s} &= \mathcal{N}\left(|O_\theta(c) \ominus O_\theta(s)|\right) && (4)
\end{aligned}$$

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 normalization operator (for details see [13]). The feature maps are summed over the center-surround combinations using across-scale addition $\oplus$, and the sums are normalized again:

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

and

$$\begin{aligned}
&L_I = \{I\},\ L_C = \{RG, BY\},\\
&L_O = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \qquad (6)
\end{aligned}$$

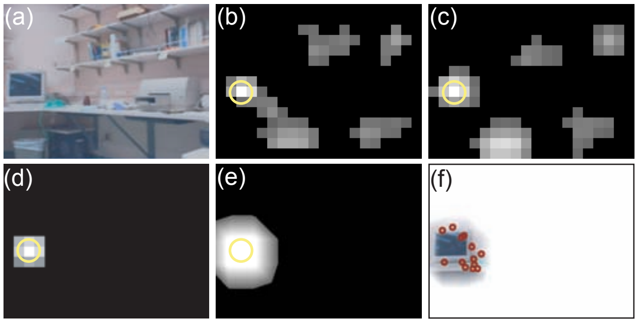

Figure 1: Illustration of the processing steps for obtaining an estimation for the object shape at the attended location and for using this for object recognition: (a) original image $\mathcal{I}$; (b) the saliency map $\mathcal{S}$, obtained from eqs. 1-8; (c) the feature map $\mathcal{F}_{l_w,c_w,s_w}$ with the strongest contribution at (x_w, y_w) – in this case the $l_w = BY$ blue/yellow contrast map with the center at pyramid level $c_w = 3$ and the surround at level $s_w = 6$, see eqs. 9 and 10; (d) the segmented feature map $\hat{\mathcal{F}}_w$; (e) the smoothed object mask $\mathcal{M}$; (f) the contrast-modulated image $\mathcal{I}'$ (see eq. 11) with extracted features (keypoints) for the recognition algorithm, marked in red.

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:

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

All conspicuity maps are combined into one saliency map (fig. 1b):

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

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:

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

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 $\mathcal{F}_{l_w,c_w,s_w}$ (fig. 1c) is segmented using region growing around (x_w, y_w) and adaptive thresholding [11] (fig. 1d). 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 objects 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 with a separable two-dimensional Gaussian kernel ($\sigma = 20$ pixels). In our implementation, 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 1 within the attended object, 0 outside the object, and has intermediate values at the edge of the object (fig. 1e). We use this mask to modulate the contrast of the original image $\mathcal{I}$ (dynamic range $[0, 255]$):

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

where $[\cdot]$ symbolizes the rounding operation. Eq. 11 is applied separately to the r, g and b channels of the image (fig. 1f). $\mathcal{I}'$ is used as the input to the recognition algorithm instead of $\mathcal{I}$.

It is not guaranteed that the approach of segmenting an object 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 [1]. 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 computational cost for the shape estimation is minimal, because the feature and conspicuity maps have already been computed during the processing for saliency.

2.2. Object recognition

For all experiments described in this paper we use the object recognition algorithm by Lowe et al. [16, 17]. The algorithm uses a Gaussian pyramid built from the original image to extract local features ("keypoints") at the extreme points of differences between pyramid levels. A model of an object is built from the keypoints, which are represented as vectors in a 128-dimensional space. Recognition is performed by matching keypoints found in the test image with stored object models.

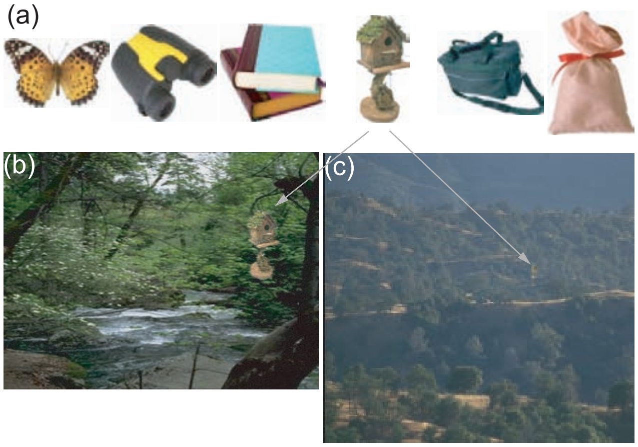

Figure 2: (a) Six of the 21 objects used in the experiments. Every objects is scaled such that it consists of approximately 2500 pixels. Artificial pixel and scaling noise is added to every instance of an object before merging it with a background image; (b,c) Examples of synthetically generated test images. Objects are merged with the background at a random position by alpha-blending. The ratio of object area vs. image area (ROS) varies between (b) 5% and (c) 0.05%.

In our model, we have the additional step of finding salient patches as described above for learning and recognition before keypoints are extracted (fig. 1f). The use of contrast modulation as a means of deploying object-based attention is motivated by neurophysiological experiments that show a tight link between luminance contrast and bottom-up attention [23, 19], 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, no keypoints are extracted there, and we limit the forming of a model to the attended region.

3. Objects in cluttered scenes

In these experiments we investigate how attention affects learning and recognition of objects in cluttered scenes. To maintain close control of the amount of clutter, we construct the test images by merging various objects with natural backgrounds at random locations (fig. 2). For experiments with natural images see section 4.

3.1. 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. 2). 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. By construction, this procedure also gives us easy access to ground truth. We choose 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):

$$ROS = \frac{\#pixels(object)}{\#pixels(image)} \tag{12}$$

A scene is more cluttered for a smaller ROS. To avoid issues with the recognition system due to large variations in the *absolute* size of the objects, we left the number of pixels for the objects constant (with the exception of intentionally added scale noise), and varied the ROS by changing the size of the backgound images in which the objects were embedded.

To prevent template matching, each object was rescaled by a random factor between 0.9 and 1.1, and uniformly distributed random noise between -12 and 12 was added to the red, green and blue value of each object pixel (dynamic range is $[0, 255]$). The objects were merged with the background by alpha blending with a low blending factor at the border and a high blending factor in the middle of the object to prevent artificially salient borders [27].

We created four test sets with ROS values of 5%, 2.78%, 1.08%, and 0.05%, each consisting of 21 images for training (one image of every object) and 420 images for testing (20 test images for every object). This amounts to a total of $4 \times 441 = 1764$ images. The background images for training and test sets were randomly drawn from disjoint image pools.

During training, object models were learned at the five most salient locations of each training image. That is, the object had to be learned by finding it in a training image. During testing, the most salient regions of the test images were compared to each of the learned models. As soon as a match was found, positive recognition was declared. Failure to attend to the object during the first five fixations led to a failed learning or recognition attempt.

3.2. Results

Learning from our data sets results in a classifier which can recognize $K = 21$ objects. Performance of this classifier is evaluated by determining the number of true positive detections T_i and the number of false positives F_i for each object i. By construction, for each object there are 20 positive samples ($N_i = 20$), and the remaining images are used as negative samples ($\overline{N_i}$). Thus, the TP of the multi-object classifier is [7]:

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

The false positive rate is calculated similarly:

$$FP = \frac{1}{K}\sum_{i=1}^{K}\frac{F_i}{\overline{N_i}} \tag{14}$$

We evaluate performance (TP) 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. 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 the recognition system. It is sufficient to evaluate only the true positive rate, since the false positive rate is consistently below 0.05% for all conditions, and therefore the total error rate is approximately equal to the false rejection rate $(1 - TP)$.

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

In the cases in which the human subject was able to identify the objects based on the attended patches, the failure of the automated system was clearly due to shortcomings of the recognition system. On the other hand, if the human subject failed to recognize the objects based on the patches, the attention system was the component responsible for the failure of the combined system. As can be seen in fig. 3, in most failure cases, the recognition system was the cause. Only for the smallest relative object size (0.05%), the attention system contributed significantly to the failure rate.

The results (fig. 3) demonstrate that attention has a sustained effect on recognition performance for all relative object sizes reported. For smaller objects (more clutter), the influence of attention becomes more accentuated. In the most difficult cases (0.05% ROS), attention increases the true positive rate by a factor of 10.

4. Multiple objects in natural images

In the previous section, we used artificially created stimuli in a controlled setup. In the following experiments we move to natural images and test the hypothesis that attention can improve, or in many cases enable, the learning and recognition of multiple objects in natural scenes. We

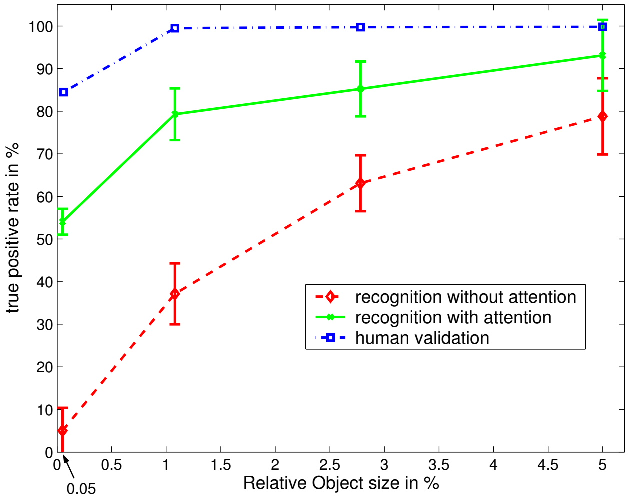

Figure 3: True positive (TP) rate for a set of artificial images. The relative object size is varied by keeping the absolute object size constant (2500 pixels ±10%) and varying the size of the background images. Error bars indicate the standard error for averaging over the performance of the 21 classifiers. It can be observed that recognition and learning with attention (green solid line) performs much better than without attention (red dashed line). This effect is more pronounced for *smaller* relative object sizes, i.e. for *larger* amounts of clutter. The human validation of the attention system shows that the difference between the recognition with attention and 100% is largely due to shortcomings of the recognition system. Note that for relative object sizes $> 5\%$, learning and recognition done on the entire image (red dashed line) works well, as reported in [16, 17].

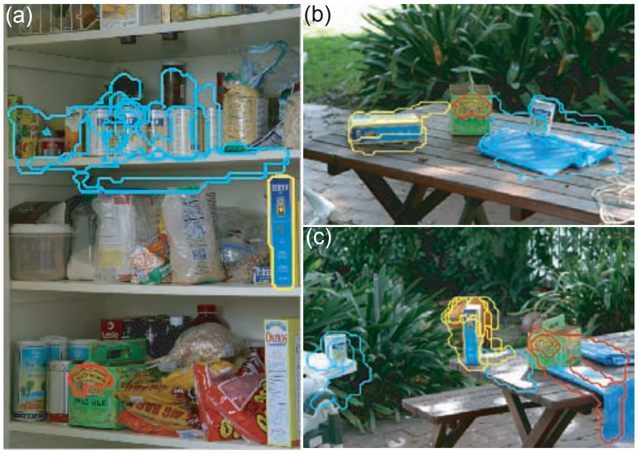

Figure 4: An example for learning an object inventory from a high-resolution digital photograph. The task is it to memorize the items in the cupboard (a) and to identify which of the items are present in the test scenes (b) and (c). The patches, which were obtained from segmenting regions at multiple salient locations, are color coded – blue for the soup can, yellow for the pasta box, and red for the beer pack. In (a), several patches are learned for the soup can, and the models learned from them match with each other very well. All three objects are found successfully in both test images. There is one false positive in (c) – a bright spot on the table is mistaken for a can. The images were processed at a resolution of 1024×1536 pixels, 15 fixations were used for training, and 20 fixations for testing. 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).

use two classes of images – high-resolution digital photographs of home environments, and low-resolution images acquired at random by an autonomous robot while navigating through an office environment.

4.1. Digital photographs

Our experiments with high-resolution digital photographs are aimed at addressing the “grocery cart problem” mentioned in the introduction. These are explorations of the possibilities of learning and recognizing several objects in real-world scenes, rather than systematic experiments. In section 4.2 we describe a more rigorous approach to learning and recognition of multiple objects from natural images.

We used a digital camera to take pictures of home environments, while re-arranging various objects. For processing, the images were sub-sampled to a resolution of 1024×1536 pixels. In each set of images, one image was selected for training, from which our combined model had to learn objects at the 15 most salient locations. The remaining images (typically two to five more images) were tested for the occurrence of any of the learned objects.

In most cases, excellent recognition performance was achieved. Fig. 4 shows an example (for more examples see the supplementary material). In a few cases the target objects in the training image were not salient enough and could hence not be learned.

While we are encouraged by the very good results in most of the examples that we tested, this uncontrolled setup makes systematic testing difficult. One of the biggest concerns is the inherent bias of human photographers, who will invariably center and zoom on interesting objects. Another issue is the difficulty of acquiring a large image set that would be necessary for a systematic analysis. In order to overcome these issues and make more systematic testing possible, we made use of a robot as an unbiased image acquisition tool.

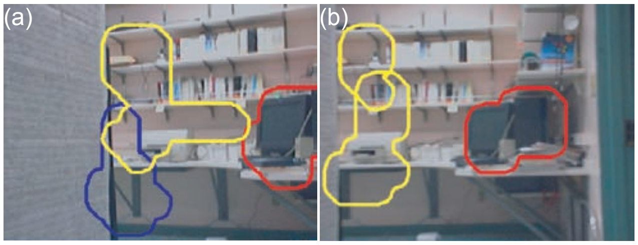

Figure 5: Two sample frames taken from the sequence consisting of 1749 frames. Each frame has a resolution of 320 × 240 and was taken by an autonomously navigating robot. In the two frames, matching regions identified by the combined attention and recognition system are marked with matching colors. The whole video, recorded with 5 fps, is available as an mpeg file in the supplement to this paper.

4.2. Landmark learning and detection using a robot

In order to assess the multiple object recognition capabilities of our model in a more controlled and rigorous fashion, we used an autonomous robot to acquire images of indoor office scenes.

4.2.1. Experimental setup The robot's navigation followed a simple obstacle avoidance algorithm using infrared range sensors for control. A camera was mounted on top of the robot at about 1.2 m height. Color images were recorded at 320 × 240 pixels with 5 frames per second. A total of 1749 images were recorded during an almost 6 min run[†]. Since vision was not used for navigation, the images taken by the robot are unbiased. Because the robot moved in a closed environment (indoor offices/labs, four rooms, approximately 80 m^2), the same objects appear multiple times in the sequence (fig. 5). Objects are extracted and learned as follows:

1. Extract the most salient patch;
2. Has the patch at least three keypoints? (A minimum of three keypoints are necessary for model learning [16].) *yes:* go to 3; *no:* go to 5;
3. Test the patch with every known object model;
4. Do we have a match? *yes:* increment the counter for the matched object; *no:* learn the object model at this location as a new object;
5. Repeat 1-4 for the three most salient regions in each image.

All learned objects are automatically given a unique label which is used as an index to count recognized objects. Applying this procedure to each frame of the video enables us to tell which objects were recognized how many times.

For these experiments, it is not possible to compare the results obtained with attention to the performance of the recognition algorithm without attention, because the recognition algorithm by itself is not capable of interpreting a scene as consisting of several objects. Instead, we repeat the same procedure as described above but with three randomly chosen patches as a control. These patches are created by using a pseudo region growing operation at the saliency map resolution. Starting from a randomly selected location, the 4-connected neighborhood is explored recursively. For each visited location, a random number is drawn from a uniform distribution. The location is accepted to belong to the random patch if the random number exceeds a preset threshold, and the recursion continues with the 4-connected neighborhood of this location. Otherwise the location is rejected, and the recursion is terminated. The threshold is adjusted such that the random patches have approximately the same size distribution as the attention patches. These random patches are then treated the same way as true attention patches for up-scaling and smoothing (see section 2.1).

Our current implementation, which is in no way optimized for speed, is capable of processing about 1.5 frames per second at 320 × 240 pixels resolution on a 2.0 GHz Pentium 4 mobile CPU. This includes attentional selection, shape estimation and recognition or learning.

4.2.2. Results A patch is considered "useful" if it is recognized at least once after learning, thus appearing at least twice in the sequence. Attentional selection identifies 3934 useful patches in the approximately 6 min of processed video, associated with 824 objects. Random patch selection only yields 1649 useful patches, associated with 742 objects (table 1).

Table 1: Results using attentional selection and random patches. An object is suitable as a landmark if it is recognized at least 10 times after learning it.

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

[†] The video as recorded and with segmented salient regions marked is available at [1] and the supplement for this paper.

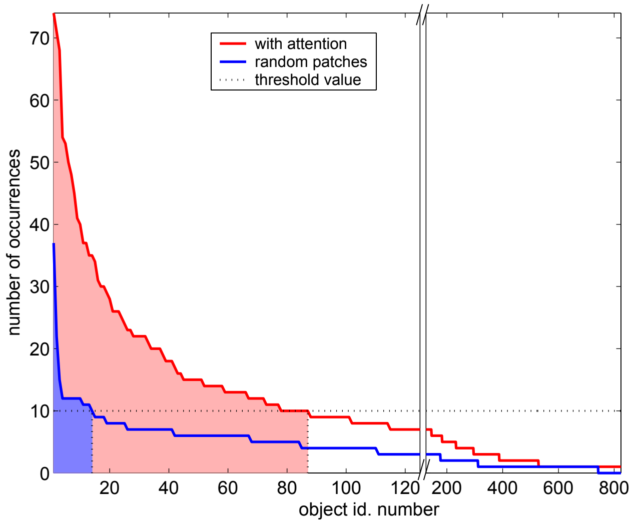

Figure 6: Learning and recognition of objects from a stream of frames of a webcam mounted on a robot. An image patch that is not recognized is automatically learned. Learned objects are labeled (x axis), and every recognized instance is counted. The threshold for an object to be suitable as a landmark is set to 10 instances. Object extraction with attention finds 87 landmarks with a total of 1910 instances. With random patches, 14 landmarks with 201 instances are found.

To judge how appropriate the learned objects are as landmarks we sort the objects by their number of occurrences and set an arbitrary threshold of requiring at least 10 recognized occurrences for an object to be suitable as a landmark (fig. 6).

With this threshold in place, attentional selection finds 87 landmarks with a total of 1910 patches associated with them. With random patches, only 14 landmarks are found with a total of 201 patches. The number of patches associated with landmarks is computed from fig. 6 as:

$$N_L = \sum_{\forall i: n_i \geq 10} n_i \qquad (n_i \in \mathcal{O}) \qquad (15)$$

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

Ground truth for the two sequences 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 right or wrong. The false positive rate is derived from the number of patches that were wrongly associated with an object. The results for attentional selection and random patches are summarized in table 1.

From these results it is clear that the attentional mechanism selects more useful patches than the random algorithm, i.e. those patches are more frequently identified, making them more useful for navigation. Frequently, separate object models are learned for the same physical object. This is usually due to limitations in the scale and viewpoint invariance of the recognition system.

These results demonstrate that attentional selection is a useful mechanism for stable landmark detection in cluttered environments. Moreover, this shows that the interplay of attention and recognition is capable of functioning in a real-world online-learning environment with low-resolution images and completely unbiased image acquisition.

Note that we used the robot only as an image acquisition tool in this experiment. For details on vision-based robot navigation and control see for instance [3, 10].

5. Discussion

We have set out to explore if and how attentional region selection can enhance object recognition. In the experiments presented in this paper we have shown by example and by rigorous quantitative analysis that saliency-based bottom-up attention is indeed useful for object recognition. We have shown that recognition performance for objects in highly cluttered scenes can be improved dramatically. Other modes of operation, such as learning multiple objects from single images, are only possible using attention. Furthermore, attentional mechanisms are useful for identifying landmarks for visual navigation, making use of online learning, and novelty detection.

Although we have limited 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. It is conceivable, e.g., that 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 [21]). We have selected Lowe's recognition algorithm for our experiments because of its suitability for general object recognition.

By the example of the configuration that we have chosen, we have demonstrated the usefulness of the synergy between recognition and attention in three domains – learning and recognition in highly cluttered scenes, learning and recognition when several objects are presented in each image, and online learning of landmarks suitable for robot navigation.

Acknowledgments This project was funded by the NSF Engineering Research Center for Neuromorphic Systems Engineering at Caltech, by an NSF-ITR award, the NIH and the Keck Foundation. The shape estimation code was developed by the authors as part of the ”iNVT“ community effort (http://ilab.usc.edu/toolkit). We would like to thank 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] We provide additional examples and the raw video as recorded by the robot on the accompanying webpage. `http://www.klab.caltech.edu/˜urut/cvpr04`.

[2] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE T. on Communications*, COM-31,4:532–540, 1983.

[3] J. Clark and N. Ferrier. Control of visual attention in mobile robots. In *Proc. IEEE Conference on Robotics and Automation*, pages 826–831, 1989.

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

[5] R. Desimone and J. Duncan. Neural mechanisms of selective visual-attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

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

[7] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *HP Technical report*, 4, 2003.

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. CVPR*, pages 264–271, 2003.

[9] W. Förstner. A framework for low level feature extraction. In *Proc. ECCV*, pages 383–394, 1994.

[10] J. Hayet, F. Lerasle, and M. Devy. Visual landmark detection and recognition for mobile robot navigation. In *Proc. CVPR*, pages 313–318, 2003.

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

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

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

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, vol.20:1254–1259, 1998.

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

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

[17] D. Lowe. Towards a computational model for object recognition in IT cortex. In *Proc. Biologically Motivated Computer Vision*, pages 20–31, 2000.

[18] R. Manduchi, P. Perona, and D. Shy. Efficient deformable filter banks. *IEEE T. on Signal Processing*, 46(4):1168–1173, 1998.

[19] C. J. McAdams and J. H. R. Maunsell. Attention to both space and feature modulates neuronal responses in macaque area V4. *J. of Neurophysiology*, 83(3):1751–1755, 2000.

[20] F. Miau and L. Itti. A neural model combining attentional orienting to object recognition: Preliminary explorations on the interplay between where and what. In *IEEE Engin. in Medicine and Biology Society (EMBS)*, Istanbul, Turkey, 2001.

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

[22] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

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

[24] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *Proc. CVPR*, pages 272–277, 2003.

[25] C. Schmid. A structured probabilistic model for recognition. *Proc. CVPR*, pages 485–490, 1999.

[26] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. CVPR*, pages 746–751, 2000.

[27] D. Sheinberg and N. Logothetis. Noticing familiar objects in real world scences: The role of temporal cortical neurons in natural vision. *J. of Neuroscience*, vol. 21(no. 4):1340–1350, 2001.

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

[29] 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.

[30] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, pages 511–518, 2001.

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

[32] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. ECCV 2000*, pages 18–32, 2000.