Selective visual attention enables learning and recognition of multiple objects in cluttered scenes

Dirk Walther, [1] Ueli Rutishauser, [1] Christof Koch, and Pietro Perona

Computation and Neural Systems,
California Institute of Technology, Pasadena CA, 91125, USA
{walther|urut|koch|perona}@caltech.edu

DRAFT - to appear in Computer Vision and Image Understanding in 2005

Abstract

A key problem in learning representations of 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 that is not associated with the objects. Clutter hurts object recognition, because it generates false alarms and imposes additional computational costs for rejecting them. Distinguishing individual objects in a scene would allow unsupervised learning of multiple objects from unlabeled images. There is psychophysical and neurophysiological evidence that the brain, which is faced with a similar challenge, employs selective visual attention to select relevant parts of the image and to serialize the perception of individual objects. We propose a method for the selection of salient regions likely to contain objects, based on bottom-up visual attention, in order to allow unsupervised one-shot learning of multiple objects in cluttered images. By comparing the performance of David Lowe's recognition algorithm with and without attention, we demonstrate in our experiments that the proposed approach can indeed enable learning of multiple objects from complex scenes, and that it can strongly improve learning and recognition performance in the presence of large amounts of clutter.

Key words: Bottom-up attention; saliency; selective attention; object recognition; object-based attention; learning; cluttered scenes

[1] These authors contributed equally to this work.

1 Introduction

Object recognition with computer algorithms has seen tremendous progress over the past years, both for specific domains such as face recognition [1,2,3] and for more general object domains [4,5,6,7,8]. Most of these approaches 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.

But what is an object? A precise definition of “object”, without taking into account the purpose and context, is of course impossible. However, it is clear that we wish to capture the appearance of those lumps of matter to which people tend to assign a name. Examples of distinguishing properties of objects are physical continuity (i.e. an object may be moved around in one piece), having a common cause or origin, having well defined physical limits with respect to the surrounding environment, being made of a well defined substance. In principle, a single image taken in an unconstrained environment is not sufficient to allow a computer algorithm, or a human being, to decide where an object starts and another object ends. However, a number of cues which are based on the statistics of our everyday’s visual world are useful to guide this decision. The fact that objects are mostly opaque and often homogeneous in appearance makes it likely that areas of high contrast (in disparity, texture, color, brightness) will be associated to their boundaries. Objects that are built by humans are often designed to be easily seen and discriminated from their environment.

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 conventional object recognition methods 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 higher-level cognitive processing. Two complementary mechanisms for the selection of individual objects have been proposed, bottom-up selective attention and grouping based on segmentation. While saliency-based attention concentrates on feature *contrasts* [9], grouping and segmentation attempt to find regions that are *homogeneous* in certain features [10,11]. Grouping has been applied successfully to object recognition [12,13]. In this paper we explore bottom-up attention. In particular, we postulate that a bottom-up attentional mechanism that is designed to respond to areas of high contrast, will frequently select image regions that correspond to objects. Our experiments are designed to test this hypothesis.

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) [14,15]. 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 [16] use local winner-take-all networks and top-down mechanisms to selectively tune model neurons at the attended location. Deco & Schürmann [17] modulate the spatial resolution of the image based on a top-down attentional control signal. Itti & Koch [9] introduced a model for bottom-up selective attention based on serially scanning a saliency map, which is computed from local feature contrasts, for salient locations in the order of decreasing saliency. Closely following and extending Duncan’s Integrated Competition Hypothesis [18], Sun & Fisher [19] developed and implemented a common framework for object-based and location-based visual attention using “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 [20,21]. However, little work has been done to verify these benefits experimentally (but see [22,23,24]). 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 system 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.

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. For object recognition, we selected Lowe’s algorithm

[25,26] as an example for a general purpose recognition system with one-shot learning. Lowe's algorithm is widely recognized as the state of the art for general purpose real-world object learning and recognition.

2.1 Bottom-up saliency-based region selection

Attention as a selective gating mechanism is often likened to a spotlight [27,28], enhancing visual processing in the attended ("illuminated") region of a few degrees of visual angle [29]. In a modification to the spotlight metaphor, the size of the attended region can be adjusted depending on the task, making attention similar to a zoom lens [30,31]. Neither of these theories considers the shape and extent of the attended object for determining the attended area. This may seem natural, since commonly attention is believed to act *before* objects are recognized. However, experimental evidence suggests that attention can be tied to objects, object parts, or groups of objects [32,33]. How can we attend to objects before we recognize them? We have developed a model that estimates the extent of salient objects solely based on bottom-up information, serving as an initial step for subsequent object detection.

Our attention system is based on the Itti et al. [9] implementation of the Koch & Ullman [34] 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 the image region that contains 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 $\mathcal{I}$ is sub-sampled into a Gaussian pyramid [35], and each pyramid level σ 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 to the images in the intensity pyramid I [36,37]. From these channels, center-surround "feature maps" are constructed and normalized:

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

Here, $\ominus$ 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 a an itera-

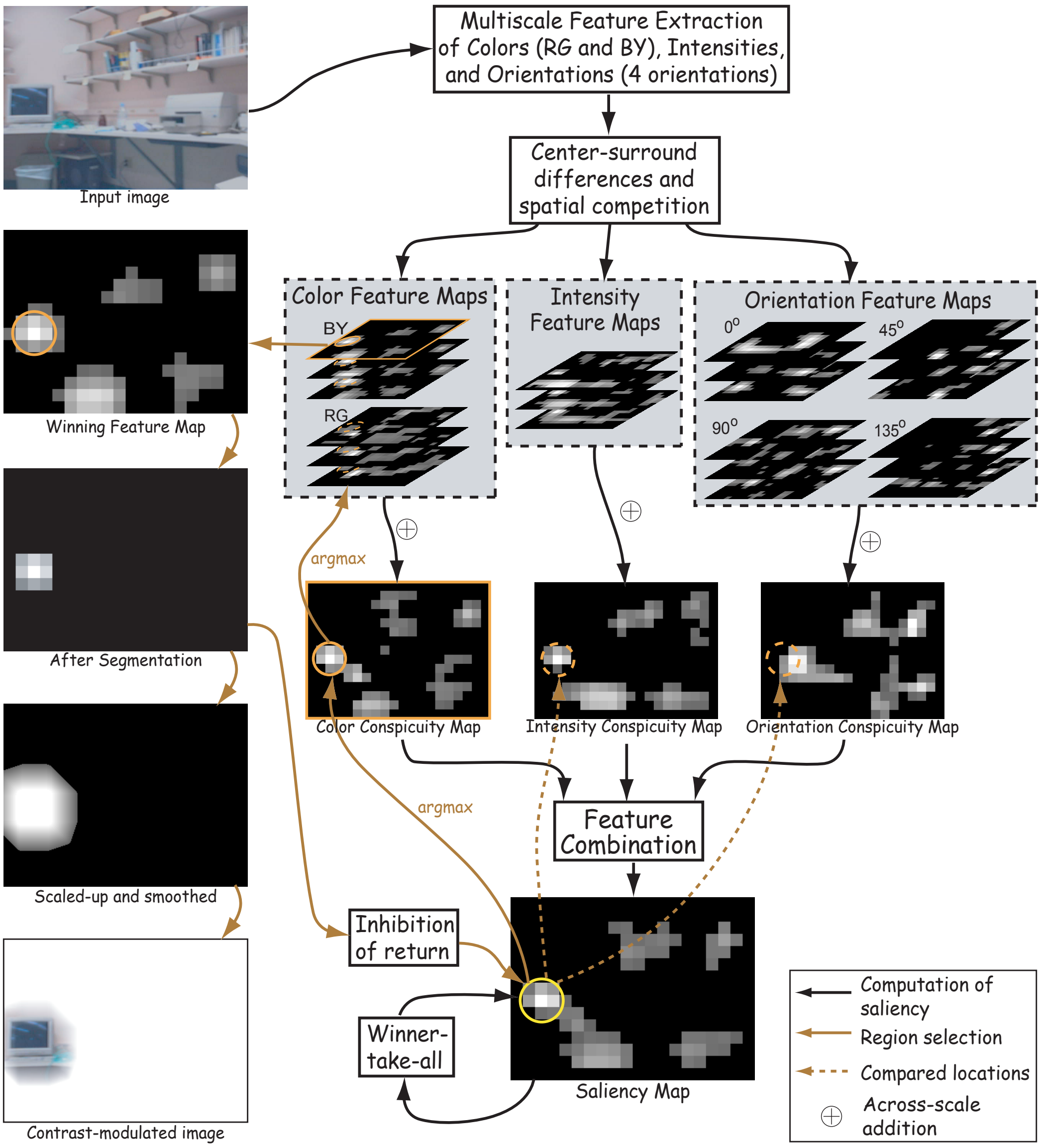

Fig. 1. Illustration of the processing steps for obtaining the attended region. The input image is processed for low-level features at multiple scales, and center-surround differences are computed (eqs. 1-4). The resulting feature maps are combined into conspicuity maps (eq. 7), and finally into a saliency map (eq. 8). 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. 9 and 10). After segmentation around the most salient location, this winning feature map is used for obtaining a smooth object mask at image resolution, and for object-based inhibition of return.

tive, nonlinear normalization operator (for details see [38]). The feature maps are summed over the center-surround combinations using across-scale addition $\oplus$, and the sums are normalized again:

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

with

$$L_I = \{I\},\ L_C = \{RG, BY\},\ L_O = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \tag{6}$$

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 (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 extent 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, using feedback connections in the saliency computation hierarchy (fig. 1). Looking back at the conspicuity maps, we find the one map that contributes most to the activity at the most salient location:

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

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

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

with L_{k_w} as defined in eq. 6. The "winning" feature map $\mathcal{F}_{l_w,c_w,s_w}$ (fig. 1) is segmented using region growing around (x_w, y_w) and adaptive thresholding [39]. 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 regions 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 object, 0 outside the object, and it has intermediate values at the object's edge. We use this mask to modulate the contrast of the original image $\mathcal{I}$ (dynamic range $[0, 255]$):

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

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 to the recognition algorithm instead of $\mathcal{I}$ (fig. 2).

As part of their selective tuning model of visual attention, Tsotsos and colleagues [16] introduced a mechanism for tracing back activations through a hierarchical network of WTA circuits to identify contiguous image regions with similarly high saliency values within a given feature domain. Our method is similar in spirit but extends across feature domains. By tracing back the activity from the attended location in the saliency map through the hierarchy of conspicuity and feature maps, we identify the feature that contributes most to the activity of the currently fixated location. We identify a contiguous region around this location with high activity in the feature map that codes for this most active feature. This procedure is motivated by the observation that between-object variability of visual information is significantly higher than within-object variability [40]. Hence, even if two salient objects are close to each other or occluding each other, it is not very likely that they are salient for the same reason. This means that they can be distinguished in the feature maps that code for their respective most active features.

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

2.2 Object learning and recognition with attention

For all experiments described in this paper, we use the object recognition algorithm developed by Lowe [4,25,26]. The algorithm consists of two main stages – the selection of local, scale-invariant features ("SIFT" keypoints), and the matching of constellations of such keypoints.

Local keypoints are found in four steps [4]. First, scale-space extrema are detected by searching over many scales and all image locations. This is implemented using difference-of-Gaussian functions, which are computed efficiently by subtract-

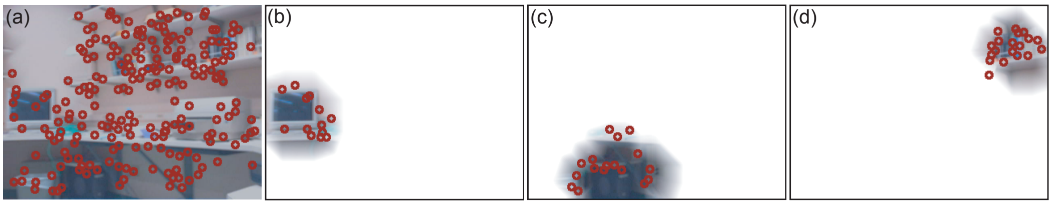

Fig. 2. Example for SIFT keypoints used for object recognition by Lowe's algorithm. (a) keypoints of the entire image; (b-d) keypoints extracted for the three most salient regions, representing "monitor", "computer", and "set of books". Restricting the keypoints to a region that is likely to contain an object enables the recognition algorithm to subsequently learn and recognize multiple objects.

ing blurred and sub-sampled versions of the image. In the second step, a detailed model is fitted to the candidate locations, and stable keypoints are selected (fig. 2a). Next, orientations are assigned to the neighborhood of each keypoint based on local gray value gradients. With orientation, scale, and location of the keypoints known, invariance to these parameters is achieved by performing all further operations relative to these dimensions. In the last step, 128-dimensional "SIFT"(Scale Invariant Feature Transform) keypoint descriptors are derived from image gradients around the keypoints, providing robustness to shape distortions and illumination changes.

Object learning consists of extracting the SIFT features from a reference image, and storing them in a data base (one-shot learning). When presented with a new image, the algorithm extracts the SIFT features and compares them with the keypoints stored for each object in the data base. To increase robustness to occlusions and false matches from background clutter, clusters of at least three feature points need to be matched successfully. This test is performed using a hash table implementation of the generalized Hough transform [41]. From matching keypoints, the object pose is approximated, and outliers and any additional image features consistent with the pose are determined. Finally, the probability that the measured set of features indicates the presence of an object is obtained from the accuracy of the fit of the keypoints and the probable number of false matches. Object matches are declared based on this probability [4].

In our model, we introduce the additional step of finding salient image patches as described in section 2.1 for learning and recognition before keypoints are extracted (fig. 2b-d). 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 [42,43]; 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. As a result, deploying selective visual attention spatially groups the keypoints into likely candidates for objects.

In the learning phase, this selection limits the formation of a model to the attended image region, thereby avoiding clutter and, more importantly, enabling the aquisition of several object models at multiple fixations in a single image. During the recognition phase, only keypoints in the attended region need to be matched to the stored models, again avoiding clutter, and making it easier to recognize multiple objects. See fig. 8 for an illustration of the reduction in complexity due to this procedure.

To avoid strong luminance contrasts at the edges of attended regions, we smoothed the representation of the region as described in section 2.1. In our experiments, we found that the graded edges of the salient regions introduce spurious features, due to the artificially introduced gradients. Therefore, we threshold the smoothed mask before contrast modulation.

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 most or all object regions. Humans and monkeys, too, need more fixations, to analyze scenes with richer information content [44]. 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 for a few typical examples from the set. Cycling usually occurs when the salient regions have covered approximately 40-50% of the image area. We use the same number of fixations for all images in an image set to ensure consistency throughout the respective experiment.

It is common in object recognition to use interest operators [45] or salient feature detectors [46] 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 vs. random patches

In the first experiment, we compare our saliency-based region selection method with randomly selected image patches using a series of images with many occurrences of the same objects. Since human photographers tend to have a bias towards centering and zooming on objects, we make use of a robot for collecting a large

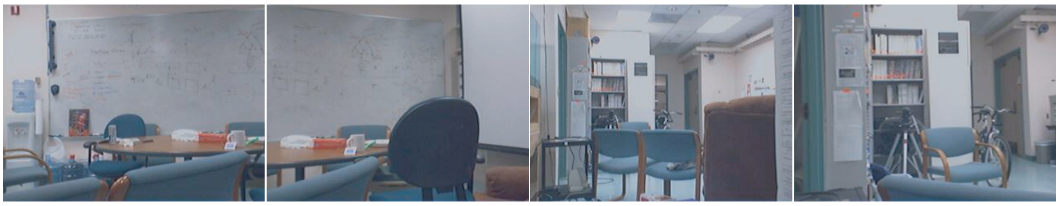

Fig. 3. Four representative frames from the video sequence recorded by the robot. The full video is available at: http://klab.caltech.edu/∼urut/cviu04

number of test images in an unbiased fashion.

Our hypothesis is that regions selected as described in section 2.1 are more likely to contain objects than randomly selected regions. If this hypothesis were true, then attempting to match image patches across frames would produce more hits for saliency-based region selection than for random region selection, because in our image sequence objects re-occur frequently.

This does not imply, however, that every image patch that is learned and recognized corresponds to an object. Frequently, groups of objects (e.g. a stack of books) or parts of objects (e.g. a corner of a desk) are selected. For the purpose of the discussion in this section we denote patches that contain parts of objects, individual objects, or groups of objects as "object patches". In this section we demonstrate that attention-based region selection finds more object patches that are more reliably recognized throughout the image set than random region selection.

3.1 Experimental setup

We used an autonomous robot equipped with a camera for image aquisition. 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 were recorded at 320×240 pixels resolution at 5 frames per second. A total of 1749 images was recorded during an almost 6 min run [2] . See fig. 3 for example frames. 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). The same objects reappear repeatedly in the sequence.

The process flow for selecting, learning, and recognizing salient regions is shown in fig. 4. Because of the low resolution of the images, we use only $N = 3$ fixations in each image for recognizing and learning patches. Note that there is no strict separation of a training and a test phase here. Whenever the algorithm fails to recognize an attended image patch, it learns a new model from it. Each newly learned patch

[2] The full video as recorded and with either salient or randomly chosen regions marked is available at: http://klab.caltech.edu/∼urut/cviu04

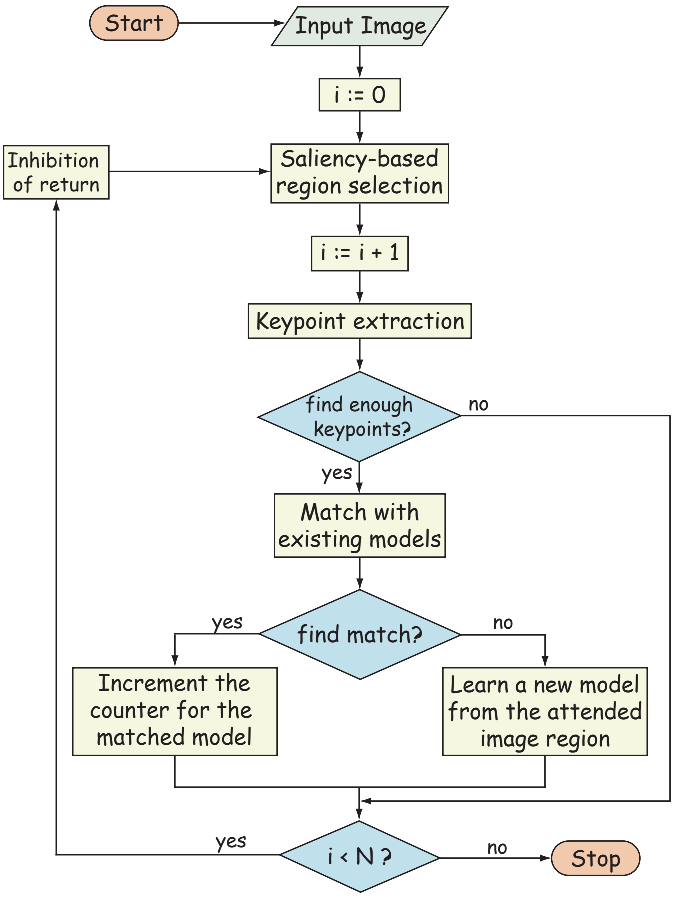
Start
Input Image
i := 0
Inhibition of return
Saliency-based region selection
i := i + 1
Keypoint extraction
find enough keypoints?
no
yes
Match with existing models
find match?
yes
no
Increment the counter for the matched model
Learn a new model from the attended image region
yes
i < N ?
no
Stop

Fig. 4. The process flow in our multi-object recognition experiments. The image is processed by the saliency-based attention mechanism as described in fig. 1. In the resulting contrast-modulated version of the image (eq. 11), keypoints are extracted (fig. 2) and used for matching the region with one of the learned object models. A minimum of three keypoints is required for this process [25]. 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 3, $N = 3$ is sufficient to cover a considerable fraction (approximately 40%) of the image area.

is assigned a unique label, and we count the number of matches for the patch over the entire image set. A patch 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 or object groups. 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.

Note that it is not practical to separate the randomization of the location selection and the region growing. For the region growing based on the most salient feature as described in section 2.1 it is necessary to have activity at the selected location in the saliency map and at least a subset of conspicuity and feature maps. The likelihood for this to be true for a randomly selected location is small. It would be possible, however, to select the location using the saliency-based attention model, and to grow the region in a random fashion as described above. This procedure would not give any additional insights though, because there is a high probability of substantial overlap between a region grown around a salient location as described in section 2.1 and a randomly grown region around the same location. At least the original location would be part of both regions, very likely also image parts in the immediate neighborhood, and then, with decreasing likelihood, parts of the image further away.

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 it as either correct or incorrect based on whether the two patches have any significant overlap. The false positive rate is derived from the number of patches that were incorrectly associated with one another.

Our current implementation 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. Note that we use the robot only as an image acquisition tool in this experiment. For details on vision-based robot navigation and control see for instance [47,48].

3.2 Results

Using the recognition algorithm without attentional selection results in 1707 of the 1749 images being pigeon-holed into 38 unique object models, representing non-overlapping large views of the rooms visited by the robot. The remaining 42 images are learned as new models, but then never recognized again. The models learned from these large scenes are not suitable for detecting individual objects. We have 85 false positives, i.e. the recognition system indicates a match between a learned model and an image, where the human subject does not indicate an agreement. This confirms that in this experiment, recogniton without attention does not yield any meaningful results.

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

Table 1
Results using attentional selection and random patches.

	Attention	**Random**
number of patches recognized	3934	1649
average per image	2.25	0.95
number of unique object patches	824	742
number of good object patches	87 (10.6%)	14 (1.9%)
number of patches associated with good object patches	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” object patches should be recognized multiple times throughout the video sequence, since the robot visits the same locations repeatedly. We sort the patches by their number of occurrences and set an arbitrary threshold of 10 recognized occurrences for “good” object patches for this analysis (fig. 5). With this threshold in place, attentional selection finds 87 good object patches with a total of 1910 instances associated to them. With random regions, only 14 good object patches are found with a total of 201 instances. The number of patches associated with good object patches is computed from fig. 5 as:

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

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 our attention-based algorithm systematically selects regions that can be recognized repeatedly from various viewpoints with much higher reliability than randomly selected regions. Since we are selecting for regions with high contrast, the regions are likely to contain objects or object parts. This hypothesis is further supported by the results shown in the next two sections. With this empirical verification of the usefulness of the region selection algorithm detailed in section 2 we now go on to exploring its effect on processing multiple objects, and on object learning and recognition in highly cluttered scenes.

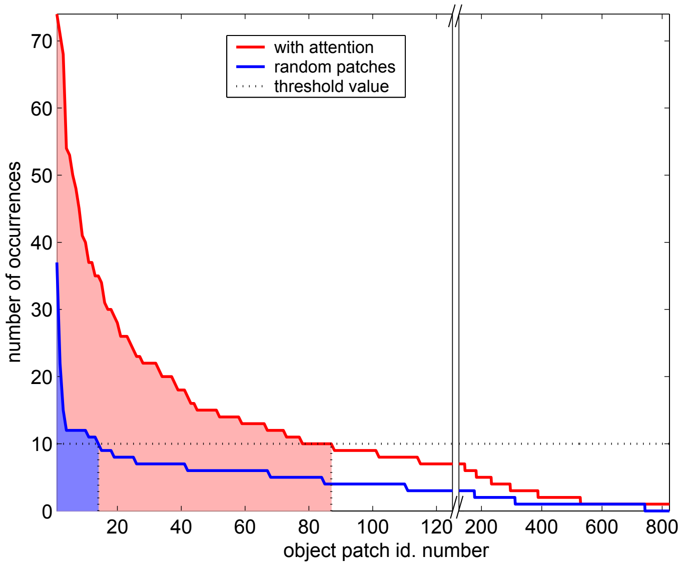

Fig. 5. Learning and recognition of object patches in a stream of video images from a camera mounted on a robot. Object patches are labeled (x axis), and every recognized instance is counted (y axis). The threshold for “good” object patches is set to 10 instances. Region selection with attention finds 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.

4 Learning multiple objects from natural images

In this experiment, we test the hypothesis that attention can enable the learning and recognition of multiple objects in individual natural scenes. We use high-resolution digital photographs of sets of objects in indoor 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[3]. Images can contain large or small subsets of the objects. We select one of the images for training (fig. 6a). 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. 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 by inspecting the learned patches and the patches extracted from the test images and flagging pairs

[3] The image set is available for download at: http://klab.caltech.edu/∼urut/cviu04

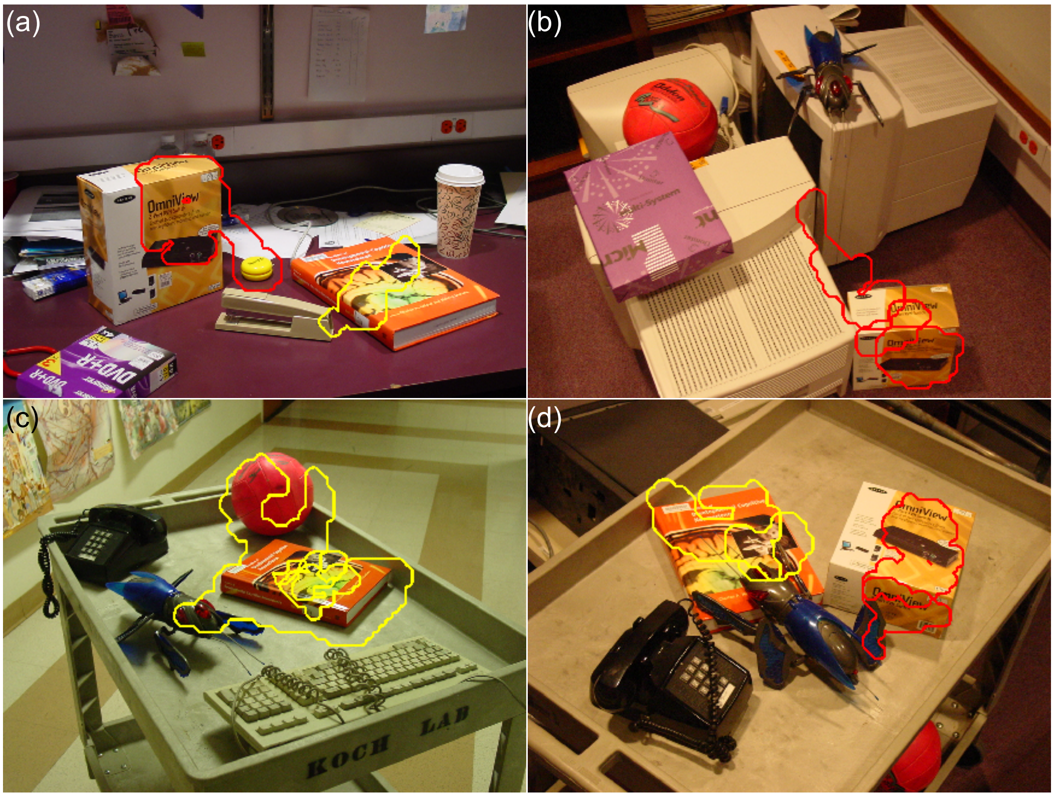

Fig. 6. Learning and recognition of two objects in cluttered scenes. (a) the image used for learning the two objects; (b-d) examples for images in which objects are recognized as matches with one or both of the objects learned from (a). The patches, which were obtained from segmenting regions at multiple salient locations, are color coded – yellow for the book, and red for the box. The decision whether a match occurred is made by the recognition algorithm without any human supervision.

that contain matching objects.

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. 6). 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 that were learned from one image.

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

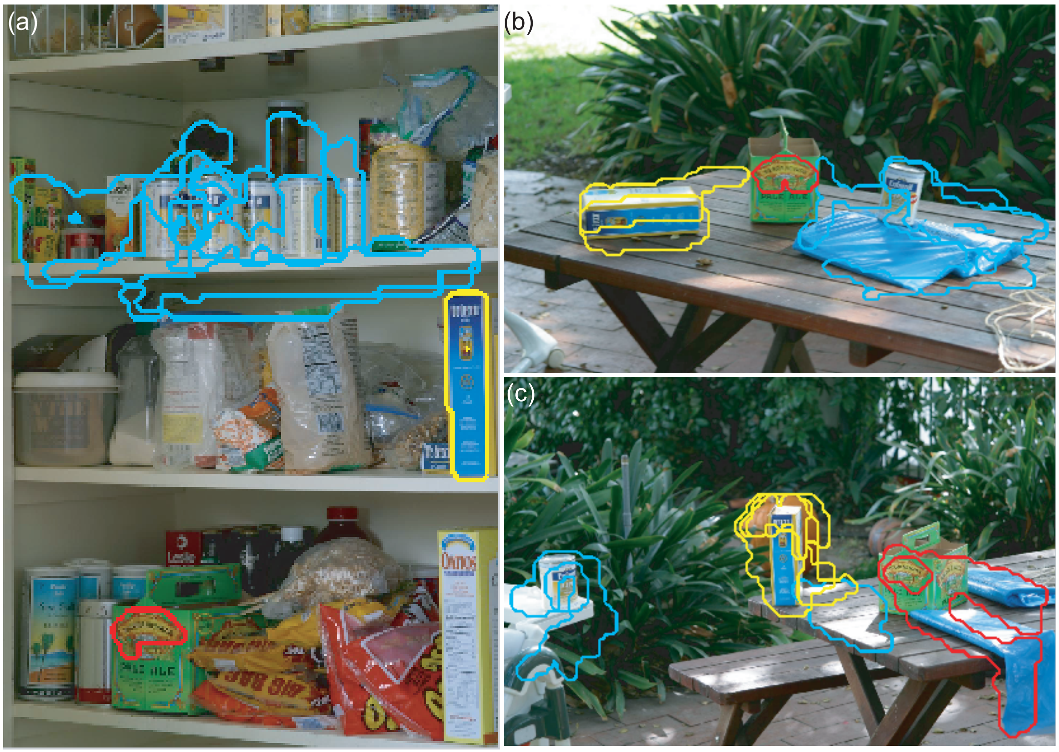

Fig. 7. Another example for learning several objects from a high-resolution digital photograph. The task is to memorize the items in the cupboard (a) and to identify which of the items are present in the test scenes (b) and (c). Again, the patches are color coded – blue for the soup can, yellow for the 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).

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.

In fig. 7 we show another example for learning multiple objects from one photograph, and recognizing the objects in a different visual context. In fig. 7a, models for the soup cans are learned from several overlapping regions, and they all match with each other. One model is learned for the pasta box and the label on the beer pack, respectively. All three objects are found successfully in both test images. There is one false positive in fig. 7c – a bright spot on the table is mistaken for a can. This experiment is very similar to the “grocery cart problem” mentioned in the introduction. The images were processed at a resolution of 1024×1536 pixels, 15 fixations were used for training and 20 fixations for testing.

Fig. 8 illustrates how attention-based region selection helps to reduce the complex-

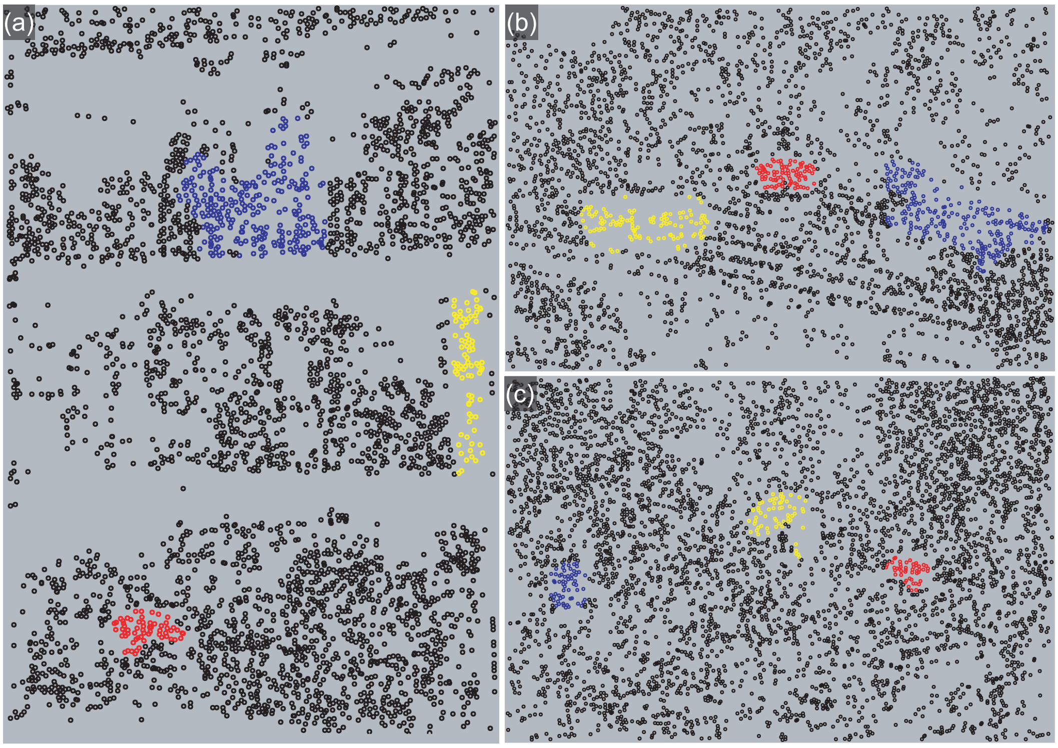

Fig. 8. The SIFT keypoints for the images shown in fig. 7. The subsets of keypoints identified by salient region selection for each of the three objects are color coded with the same colors as in the previous figure. All other keypoints are shown in black. In fig. 7 we show all regions that were found 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.

ity of matching constellations of keypoints between the images. Instead of attempting to match keypoint constellations based on the entire set of keypoints identified in the image, only the color coded subsets need to be compared to each other. The subsets with matching colors were identified as object matches by the recognition algorithm. This figure also illustrates that keypoints are found at all textured image locations – at the edges as well as on the faces of objects.

5 Objects in cluttered scenes

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

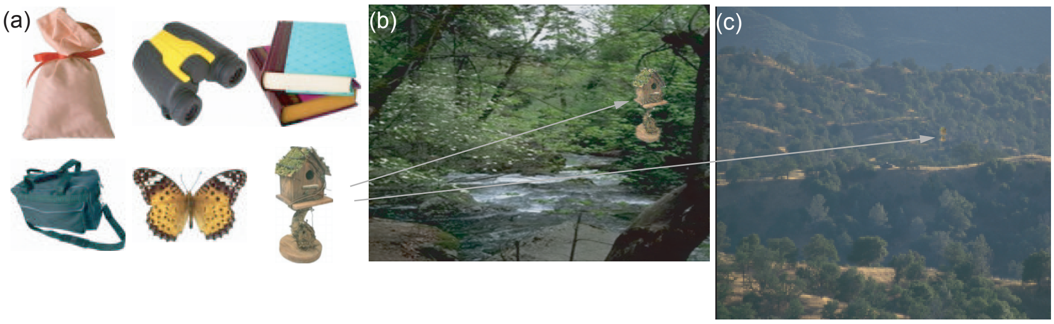

Fig. 9. (a) Six of the 21 objects used in the experiment. Each object is scaled such that it consists of approximately 2500 pixels. Artificial pixel and scaling noise is added to every instance of an object before merging it with a background image; (b,c) Examples of synthetically generated test images. 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%.

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. 9). 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 [44]. 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 images 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:

$$ROS = \frac{\#pixels(object)}{\#pixels(image)} \tag{13}$$

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. Since our background images contain fairly uniform amounts of clutter within the images as well as between images, the ROS can be used as an inverse measure of the amount of clutter faced by the object recognition algorithm when it attempts to learn or recognize the objects contained in the images. A *large* ROS means that the object is relatively large in the image, and hence that it is faced with relatively *little* clutter. A small ROS, on the other hand, means a lot of clutter.

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 edges at the object borders and any high frequency components associated with them.

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 of every object) and 420 images for testing (20 test images for every object). The background images for training and test sets are randomly drawn from disjoint image pools to avoid false positives due to repeating 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 [49].

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 are then computed as [50]:

$$t = \frac{1}{K} \sum_{i=1}^{K} \frac{T_i}{N_i} \tag{14}$$

$$f = \frac{1}{K} \sum_{i=1}^{K} \frac{F_i}{\overline{N}_i} \tag{15}$$

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

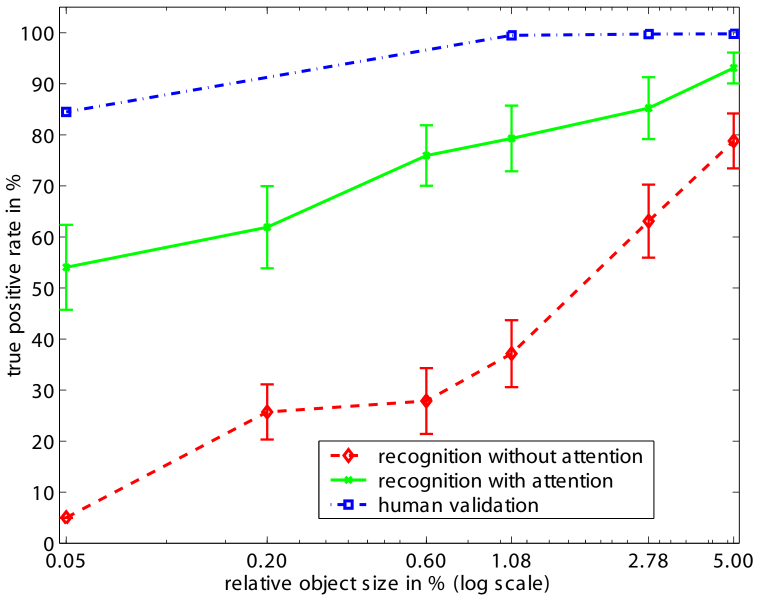

Fig. 10. True positive rate (t) for a set of artificial images without attention (red) and with attention (green) over the relative object size (ROS). The ROS is varied by keeping the absolute object size constant at 2500 pixels ±10%, and varying the size of the background images. Error bars indicate 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.

there are equal numbers of positive examples for all classes, we can write:

$$\overline{N}_i = \sum_{j=1, j \neq i}^{K} N_j = (K-1) N_i \tag{16}$$

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 performance (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. 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.

For human validation, all images in which the objects 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. 10, the human subject can recognize the objects from the attended patches in most failure cases, which implies that the recognition system is the main cause for the failure rate. Significant contributions to the failure rate by the attention system are only observed for the highest amount of clutter ($ROS = 0.05\%$).

The results in fig. 10 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 case (0.05% relative object size), 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. 10) works well without attention, as reported in [4,25,26].

We used five fixation throughout the experiment to ensure consistency. In prelimenary experiments we investigated larger numbers of fixations as well. The performance increases slightly for more fixations, but the effect of adding more clutter remains the same.

6 Discussion

We set out to test two hypotheses for the effects of attention on object recognition. The first is that attention can serialize the learning and recognition of multiple objects in individual images. With the experiments in section 4 we show that this new mode of operation, which is impossible for the recognition system without prior region selection, is indeed made possible by using our saliency-bases region selection algorithm.

Secondly, we show that spatial attention improves the performance of object learning and recognition in the presence of large amounts of clutter by up to an order of magnitude. The addition of the attention-based region selection makes object recognition more robust to distracting clutter in the image.

We have limited our experiments to bottom-up attention to avoid task secificity. However, in many applications, top-down knowledge can be very useful for visual processing [51], in addition to the saliency-based attention described here. In particular for cases where behaviorally relevant objects may not be salient, a top-down mechanism for guiding attention to task-relevant parts of the scene becomes necessary [52].

We have selected Lowe's recognition algorithm for our experiments because of its suitability for general object recognition. However, our experiments and their results do not depend on that specific choice for a recognition system. In fact, we have shown the suitability of the method for a biologically realistic object recognition system in a different context [24].

Neurophysiological experiments in monkeys show that the activity of neurons that participate in object recognition is only modulated by a relatively small amount due to attentional processes [42,43]. This should be taken into account when modeling attention and recognition in a biologically plausible manner. Since much of the early visual processing in cortex happens in parallel across the entire visual field, no computational resources are wasted by processing information outside the focus of attention. By modulating the activities of certain groups of neurons, selective visual attention serves as a gateway for cognitive processing by areas higher in the visual processing hierarchy. We have shown previously in simulations that attentional modulation of neural activity at an intermediate processing level by as little as 20% can effectively gate information for the subsequent recognition of multiple objects [24].

In contrast, for a machine vision system it is beneficial to completely disregard all information outside the focus of attention and only spend computational resources on the attended image region. Since most computers process each image location sequentially, attention algorithms can save computational resources by limiting processing to the focus of attention. For the work presented in this paper, we adopt the latter strategy by completely removing the luminance constrast outside the attended region and thereby restricting the search for keypoints to a region that is likely to contain an object.

Many important questions related to attention and object recognition are not addressed in this paper and remain subject of continuing research. Some of these questions are related to the scale of objects and salient regions in the image [53]. What, for instance, happens when an object is much smaller than a selected region, or when more than one object happen to be present in the region? It is conceivable that in such cases the object recognition algorithm could give feedback to the attention algorithm, which would then refine the extent and shape of the region, based on information about the identity, position, and scale of objects. This scheme may be iterated until ambiguities are resolved, and it would lead to object-based attention.

At the other extreme, an object could be much larger than the selected regions, and many fixations may be necessary to cover the shape of the object. In this case, visual information needs to be retained between fixations and integrated into a single percept. When hypotheses about the object identity arise during the first few fixations, attention may be guided to locations in the image that are likely to inform a decision about the correctness of the hypotheses.

Acknowledgments

This project was funded by the NSF Engineering Research Center for Neuromorphic Systems Engineering at Caltech, by an NSF-ITR award, the NIH, the 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 Laurent Itti for comments on a version of 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] H. Schneiderman, T. Kanade, A statistical method for 3D object detection applied to faces and cars, in: International Conference on Computer Vision and Pattern Recognition, 2000, pp. 746–751.

[2] P. Viola, M. J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.

[3] H. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 23–38.

[4] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[5] M. Weber, M. Welling, P. Perona, Unsupervised learning of models for recognition, in: European Conference on Computer Vision, Vol. 1842, 2000, pp. 18–32.

[6] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: International Conference on Computer Vision and Pattern Recognition, Vol. 2, 2003, pp. 264–271.

[7] C. Schmid, A structured probabilistic model for recognition, in: International Conference on Computer Vision and Pattern Recognition, Vol. 2, 1999, pp. 485–490.

[8] F. Rothganger, S. Lazebnik, C. Schmid, J. Ponce, 3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints, in: International Conference on Computer Vision and Pattern Recognition, Vol. 2, 2003, pp. 272–277.

[9] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.

[10] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.

[11] D. Martin, C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color and texture cues, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (5) (2004) 530–549.

[12] G. Mori, X. Ren, A. Efros, J. Malik, Recovering human body configurations: Combining segmentation and recognition, in: International Conference on Computer Vision and Pattern Recognition, 2004.

[13] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, D. Forsyth, The effects of segmentation and feature choice in a translation model of object recognition, in: International Conference on Computer Vision and Pattern Recognition, 2003.

[14] R. Desimone, J. Duncan, Neural mechanisms of selective visual-attention, Annual Review of Neuroscience 18 (1995) 193–222.

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

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

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

[18] J. Duncan, Integrated mechanisms of selective attention, Current Opinion in Biology 7 (1997) 255–261.

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

[20] D. Walther, U. Rutishauser, C. Koch, P. Perona, On the usefulness of attention for object recognition, in: Workshop on Attention and Perfromance in Computational Vision at ECCV, 2004, pp. 96–103.

[21] U. Rutishauser, D. Walther, C. Koch, P. Perona, Is attention useful for object recognition?, in: International Conference on Computer Vision and Pattern Recognition, 2004.

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

[23] F. Miau, L. Itti, A neural model combining attentional orienting to object recognition: Preliminary explorations on the interplay between where and what, in: IEEE Engineering in Medicine and Biology Society, 2001.

[24] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, C. Koch, Attentional selection for object recognition – a gentle way, in: Lecture Notes in Computer Science, Vol. 2525, Springer, Berlin, Germany, 2002, pp. 472–479.

[25] D. G. Lowe, Object recognition from local scale-invariant features, in: International Conference on Computer Vision, 1999, pp. 1150–1157.

[26] D. G. Lowe, Towards a computational model for object recognition in IT cortex, in: Biologically Motivated Computer Vision, 2000, pp. 20–31.

[27] M. Posner, Orienting of attention, Quarterly Journal of Experimental Psychology 32 (1) (1980) 3–25.

[28] A. M. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psychology 12 (1) (1980) 97–136.

[29] D. Sagi, B. Julesz, Enhanced detection in the aperture of focal attention, Nature 321 (1986) 693–695.

[30] C. Eriksen, J. St. James, Visual attention within and around the field of focal attention: A zoom lens model, Perception and Psychophysics 40 (4) (1986) 225–240.

[31] G. Shulman, J. Wilson, Spatial frequency and selective attention to spatial location, Perception 16 (1) (1987) 103–111.

[32] J. Duncan, Selective attention and the organization of visual information, Journal of Experimental Psychology: General 113 (4) (1984) 501–517.

[33] P. Roelfsema, V. Lamme, H. Spekreijse, Object-based attention in the primary visual cortex of the macaque monkey, Nature 395 (6700) (1998) 376–381.

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

[35] P. Burt, E. Adelson, The Laplacian Pyramid as a compact image code, IEEE Transactions on Communications COM-31 (4) (1983) 532–540.

[36] E. Simoncelli, W. Freeman, The steerable pyramid: a flexible architecture for multi-scale derivative computation, in: International Conference on Image Processing, 1995.

[37] R. Manduchi, P. Perona, D. Shy, Efficient deformable filter banks, IEEE Transactions on Signal Processing 46 (4) (1998) 1168–1173.

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

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

[40] D. L. Ruderman, Origins of scaling in natural images, Vision Research 37 (23) (1997) 3385–98.

[41] D. Ballard, Generalizing the Hough transform to detect arbitrary patterns, Pattern Recognition 13 (2) (1981) 111–122.

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

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

[44] D. L. Sheinberg, N. K. Logothetis, Noticing familiar objects in real world scenes, Journal of Neuroscience 21 (4) (2001) 1340–1350.

[45] C. Harris, M. Stephens, A combined corner and edge detector, in: 4th Alvey Vision Conference, 1988, pp. 147–151.

[46] T. Kadir, M. Brady, Scale, saliency and image description, International Journal of Computer Vision 30 (2) (2001) 77–116.

[47] J. Clark, N. Ferrier, Control of visual attention in mobile robots, in: IEEE International Conference on Robotics and Automation, Vol. 2, 1989, pp. 826–831.

[48] J. Hayet, F. Lerasle, M. Devy, Visual landmark detection and recognition for mobile robot navigation, in: International Conference on Computer Vision and Pattern Recognition, Vol. 2, 2003, pp. 313–318.

[49] G. Legge, D. Pelli, G. Rubin, M. Schleske, The psychophysics of reading, Vision Research 25 (2) (1985) 239–252.

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

[51] A. Oliva, A. Torralba, M. Castelhano, J. Henderson, Top-down control of visual attention in object detection, in: International Conference on Image Processing, 2003.

[52] V. Navalpakkam, L. Itti, A goal oriented attention guidance model, in: Lecture Notes in Computer Science, Vol. 2525, Springer, Berlin, Germany, 2002, pp. 453–461.

[53] M. Jägersand, Saliency maps and attention in scale and spatial coordinates: an information theoretic approach, in: IEEE International Conference on Computer Vision, 1995, pp. 195–202.