Part II

Machine Vision

Chapter 5

Attention for Object Recognition

5.1 Introduction

Object recognition with computer algorithms has seen tremendous progress over the past years, both for specific domains such as face recognition (Schneiderman and Kanade 2000; Viola and Jones 2004; Rowley et al. 1998) and for more general object domains (Lowe 2004; Weber et al. 2000; Fergus et al. 2003; Schmid 1999; Rothganger et al. 2003). 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, or 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 with their boundaries. Objects that are built by humans, such as traffic signs, 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* (Walther et al. 2005a, 2004b; Rutishauser et al. 2004a; Itti et al. 1998), grouping and segmentation attempt to find regions that are *homogeneous* in certain features (Shi and Malik 2000; Martin et al. 2004). Grouping has been applied successfully to object recognition, e.g., by Mori et al. (2004) and Barnard et al. (2003). In this chapter, we demonstrate that a bottom-up attentional mechanism as described in chapter 2 will frequently select image regions that correspond to objects.

Upon closer inspection, the "grocery cart problem" (also known as the "bin of parts problem" in the robotics community) poses two complementary challenges: (i) serializing the perception and learning of relevant information (objects) and (ii) suppressing irrelevant information (clutter). Visual attention addresses both problems by selectively enhancing perception at the attended location (see chapter 3) and by successively shifting the focus of attention to multiple locations.

The main motivation for attention in machine vision is cueing subsequent visual processing stages such as object recognition to improve performance and/or efficiency (Walther et al. (2005a); Rutishauser et al. (2004a)). So far, little work has been done to verify these benefits experimentally (but see Dickinson et al. (1997) and Miau and Itti (2001)). The focus of this chapter 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 the saliency-based region selection mechanism from chapter 2 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.

The work in this chapter is a collaboration with Ueli Rutishauser. While I implemented the attention system and the method for deploying spatial attention to the recognition system, Ueli implemented and conducted the experiments, and both of us analyzed the experiments. The code used for the object recognition system is a proprietary implementation of David Lowe's object recognition system (Lowe 2004) by Evolution Robotics.

5.2 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.

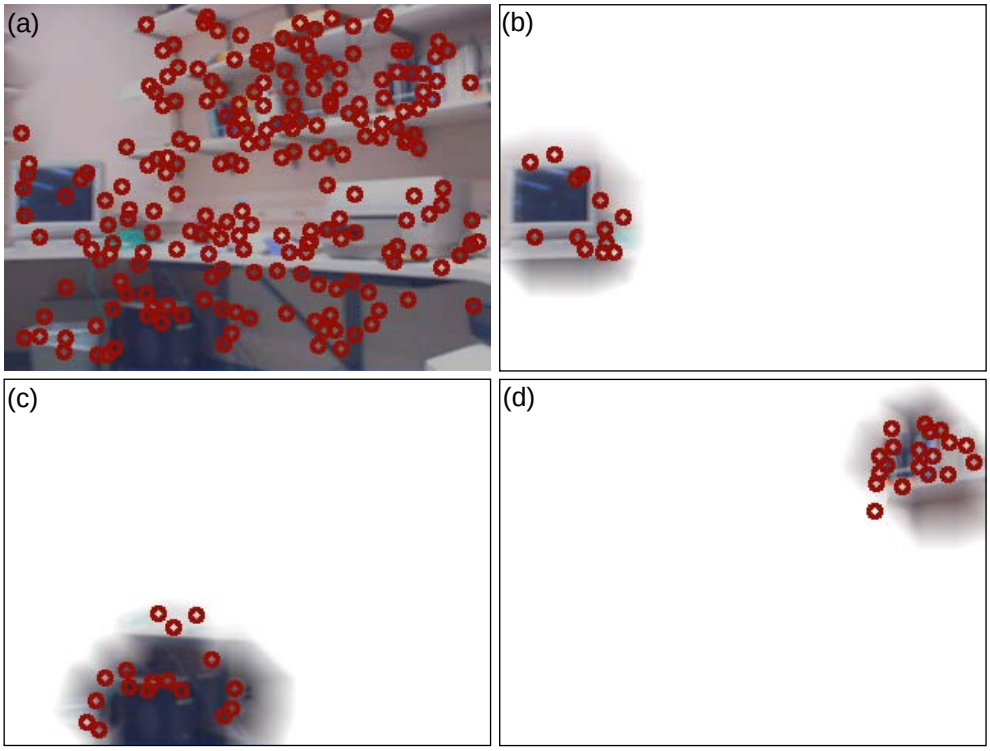

Figure 5.1: Example for SIFT keypoints used for object recognition by Lowe's algorithm. (a) keypoints of the entire image; (b-d) keypoints extracted for the three most salient regions, representing "monitor," "computer," and "set of books." Restricting the keypoints to a region that is likely to contain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

For object recognition, we selected David Lowe's algorithm (Lowe 2004, 1999, 2000) as an example for a general purpose recognition system with one-shot learning. 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 (Lowe 2004). 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 subtracting 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 (figure 5.1a). 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 (Ballard 1981). 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 (Lowe 2004).

In our model, we introduce the additional step of finding salient image patches as described in chapter 2 for learning and recognition before keypoints are extracted. Starting with the segmented map $\hat{\mathcal{F}}_w$ from eq. 2.11 on page 10, 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 $[0, 255]$):

$$\mathcal{I}'(x, y) = [255 - \mathcal{M}(x, y) \cdot (255 - \mathcal{I}(x, y))] \tag{5.1}$$

where $[\cdot]$ symbolizes the rounding operation. Eq. 5.1 is applied separately to the r, g and b channels of the image. $\mathcal{I}'$ (figure 5.1b-d) is used as the input to the recognition algorithm instead of $\mathcal{I}$ (figure 5.1a).

The use of contrast modulation as a means of deploying object-based attention is motivated by neurophysiological experiments that show attentional enhancement to act in a manner equivalent to increasing stimulus contrast (Reynolds et al. 2000; McAdams and Maunsell 2000); 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. 5.1), 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 model formation to attended image regions, thereby avoiding clutter and, more importantly, enabling the acquisition of several object models at multiple locations 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 figure 5.8 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 above. 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 large amount of 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 (Sheinberg and Logothetis 2001; Einhäuser et al. 2006). 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 (Harris and Stephens 1988) or salient feature detectors (Kadir and Brady 2001) 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.

5.3 Selective Attention versus 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 toward centering and zooming on objects, we make use of a robot for collecting a large number of test images in an unbiased fashion.

Our hypothesis is that regions selected by bottom-up, saliency-based attention 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

Figure 5.2: Six representative frames from the video sequence recorded by the robot.

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.

5.3.1 Experimental Setup

We used an autonomous robot equipped with a camera for image acquisition. 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. See figure 5.2 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 figure 5.3. 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 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

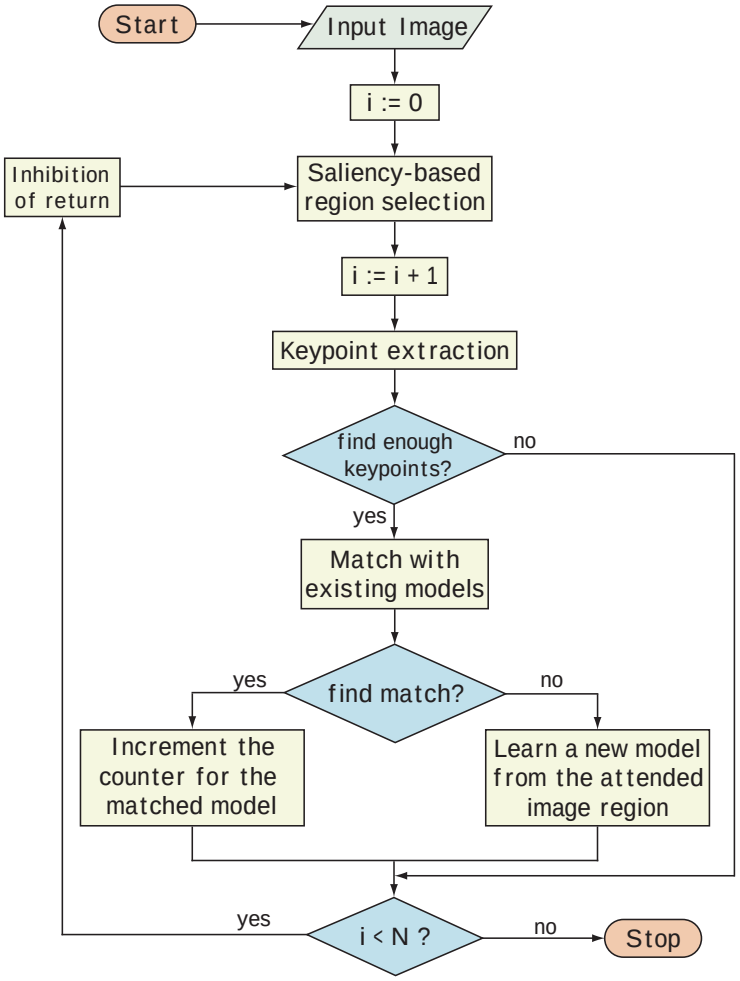

Figure 5.3: The process flow in our multi-object recognition experiments. The image is processed with the saliency-based attention mechanism as described in figure 2.2. In the resulting contrast-modulated version of the image (eq. 5.1), keypoints are extracted (figure 5.1) and used for matching the region with one of the learned object models. A minimum of three keypoints is required for this process (Lowe 1999). In the case of successful recognition, the counter for the matched model is incremented; otherwise a new model is learned. By triggering object-based inhibition of return, this process is repeated for the N most salient regions. The choice of N depends mainly on the image resolution. For the low resolution (320×240 pixels) images used in section 5.3, $N = 3$ is sufficient to cover a considerable fraction (approximately 40 %) of the image area.

Table 5.1: 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 %)

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 eqs. 2.10 and 2.11 in section 2.3). The parameters are adjusted such that the random patches have approximately the same size distribution as the attention patches.

Ground truth for all experiments is established manually. This is done by displaying every match established by the algorithm to a human subject, who has to rate 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, Clark and Ferrier (1989) or Hayet et al. (2003).

5.3.2 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, recognition without attention does not yield any meaningful results.

Attentional selection identifies 3934 useful patches in the approximately 6 minutes of processed video associated with 824 object models. Random region selection only yields 1649 useful patches associated with 742 models (table 5.1). With saliency-based region selection, we find 32 (0.8 %)

false positives, with random region selection 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 (figure 5.4). 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 figure 5.4 as

$$N_g = \sum_{\forall i: n_i \geq 10} n_i \qquad (n_i \in \mathcal{O}), \tag{5.2}$$

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 5.2 we now go on to exploring its effect on processing multiple objects and on object learning and recognition in highly cluttered scenes.

5.4 Learning Multiple Objects from Natural Images

In this experiment we test the hypothesis that attention can enable 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.

5.4.1 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. Images can contain large or small subsets of the objects. We selected one of the images for training (figure 5.5a). The other 101 images were used as test images.

For learning and recognition we used 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 that contain matching objects.

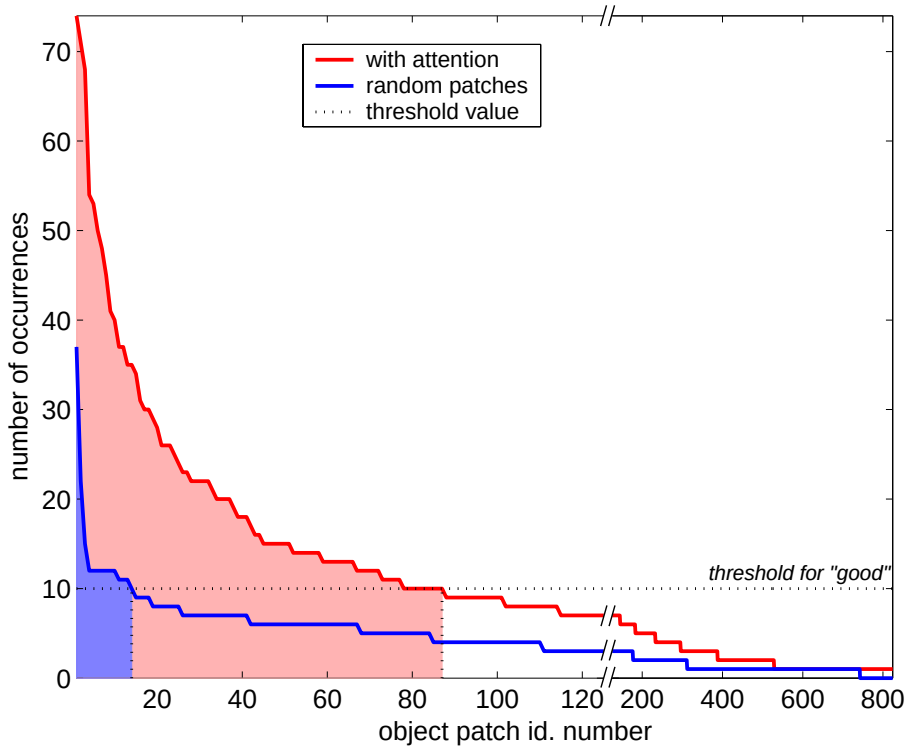

Figure 5.4: Learning and recognition of object patches in a stream of video images from a camera mounted on a robot. Object patches are labeled (x axis), and every recognized instance is counted (y axis). The threshold for "good" object patches is set to 10 instances. Region selection with attention finds 87 good object patches with a total of 1910 instances. With random region selection, 14 good object patches with 201 instances are found. Note the different linear scales on either side of the axis break in the x axis.

5.4.2 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 (figure 5.5). Of the 101 test images, 23 contain the box and 24 the book, and of these four images contain both objects. Table 5.2 shows the recognition results for the two objects.

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

Figure 5.6 shows an example of matching two objects between two individual images in an outdoor environment. The objects are successfully matched using salient regions, despite the extensive occlusions for training of object 1 (figure 5.6 (c)) and for testing of object 2 (figure 5.6 (f)).

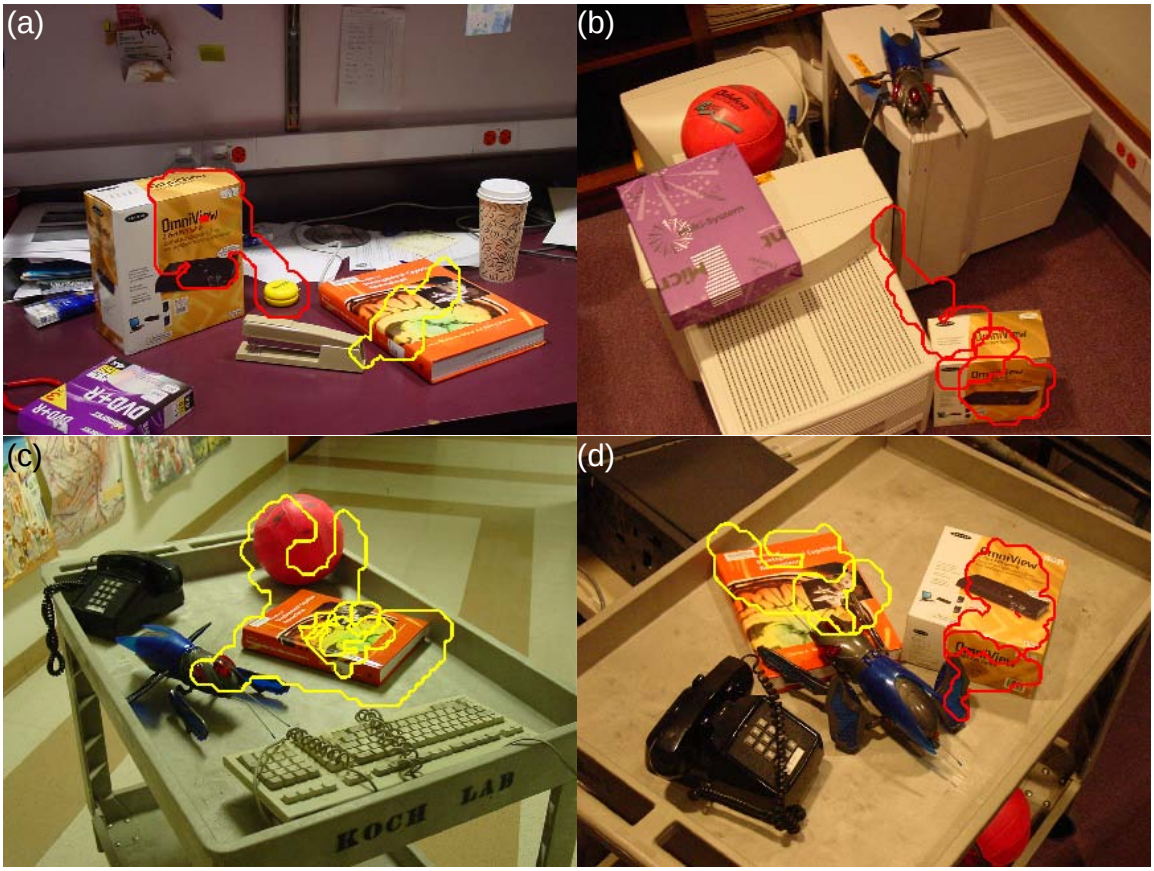

Figure 5.5: Learning and recognition of two objects in cluttered scenes. (a) the image used for learning the two objects; (b-d) examples for images in which objects are recognized as matches with one or both of the objects learned from (a). The patches, which were obtained from segmenting regions at multiple salient locations, are color coded – yellow for the book, and red for the box. The decision of whether a match occurred is made by the recognition algorithm without any human supervision.

In figure 5.7 we show another example for learning multiple objects from one photograph and recognizing the objects in a different visual context. In figure 5.7 (a), 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 figure 5.7 (c) – 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.

Figure 5.8 illustrates how attention-based region selection helps to reduce the complexity 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

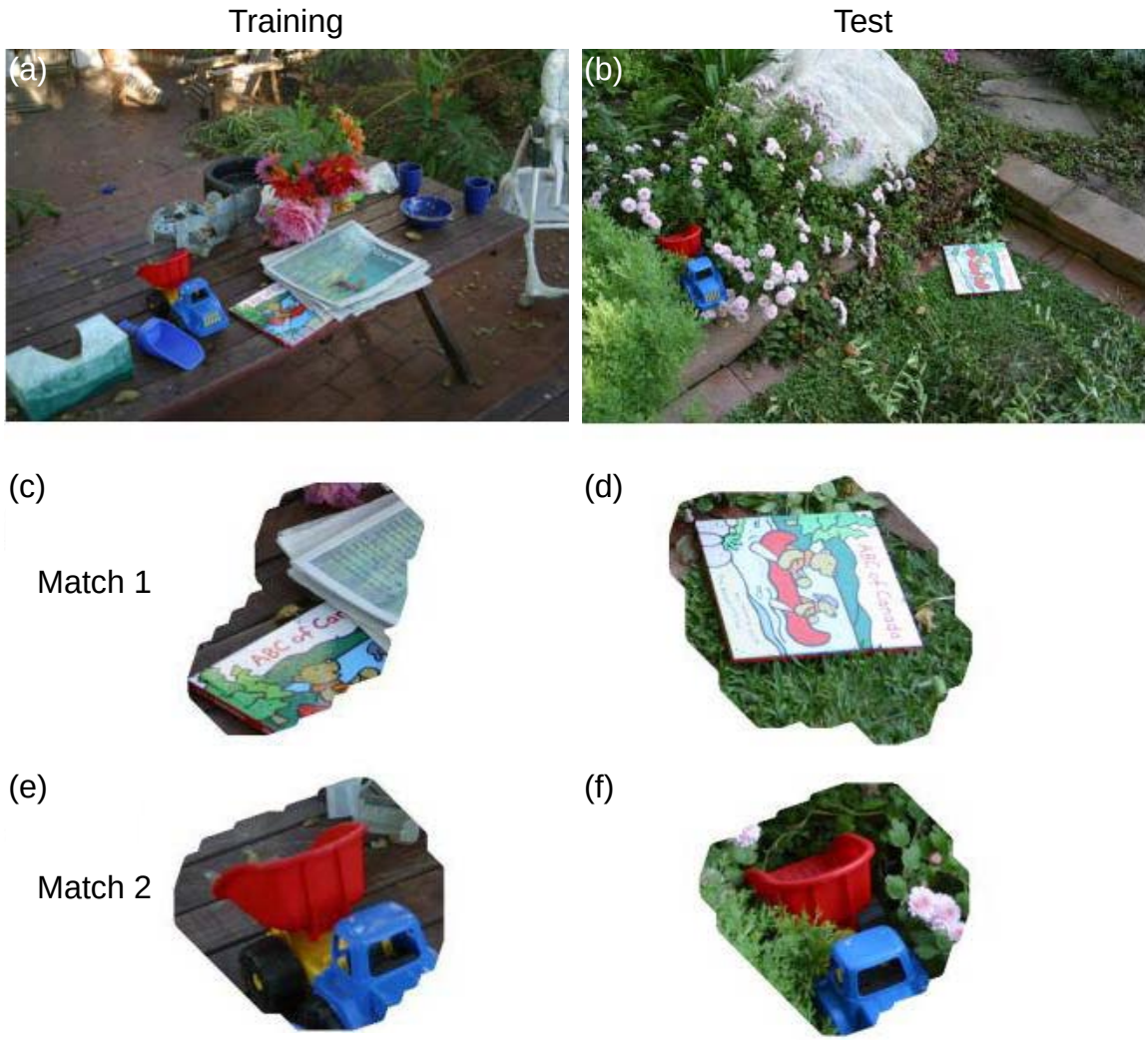

Figure 5.6: Example for learning two objects (c) and (e) from the training image (a) and establishing matches (d) and (f) for the objects in the test image (b), in a different visual context, with different object orientations and occlusions.

Table 5.2: 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%)

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.

5.5 Objects in Cluttered Scenes

In the previous section we showed 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.

5.5.1 Experimental Setup

To systematically evaluate recognition performance with and without attention we use images generated by randomly merging an object with a background image (figure 5.9). This design 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. The experimental design is inspired by Sheinberg and Logothetis (2001). 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:

$$ROS = \frac{\#pixels(object)}{\#pixels(image)}. \tag{5.3}$$

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 background images in which the objects are embedded. Since our background images contain fairly uniform amounts of clutter within 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 large amount of

Figure 5.7: Another example for learning several objects from a high-resolution digital photograph. The task is to memorize the items in the cupboard (a) and to identify which of the items are present in the test scenes (b) and (c). Again, the patches are color coded – blue for the soup can, yellow for the pasta box, and red for the label on the beer pack. In (a), only those patches are shown that have a match in (b) or (c), in (b) and (c) only those that have a match in (a).

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. An 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 (Legge et al.

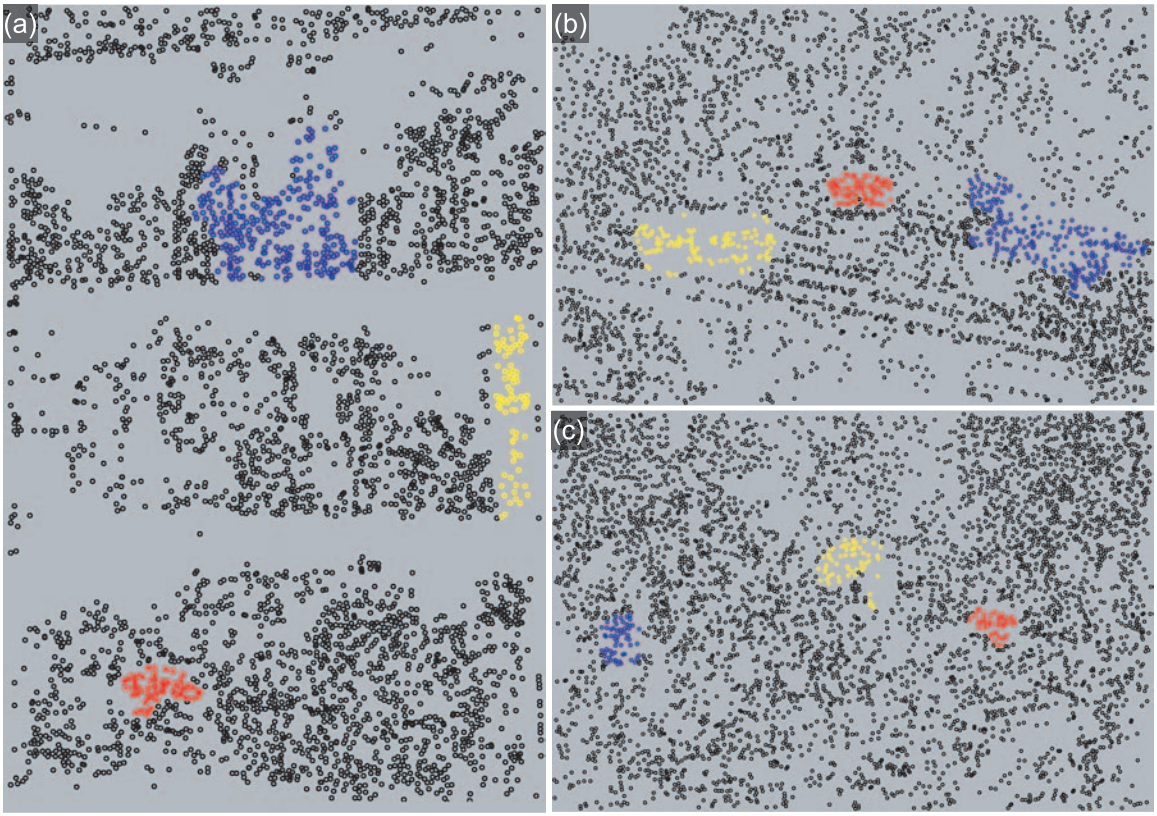

Figure 5.8: The SIFT keypoints for the images shown in figure 5.7. The subsets of keypoints identified by salient region selection for each of the three objects are color coded with the same colors as in the previous figure. All other keypoints are shown in black. In figure 5.7 we show all regions that were found for each of the objects – here we show the keypoints from one example region for each object. This figure illustrates the enormous reduction in complexity faced by the recognition algorithm when attempting to match constellations of keypoints between the images.

1985).

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.

5.5.2 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

Figure 5.9: (a) Ten of the 21 objects used in the experiment. Each object is scaled such that it consists of approximately 2500 pixels. Artificial pixel and scaling noise is added to every instance of an object before merging it with a background image; (b,c) examples of synthetically generated test images. Objects are merged with the background at a random position by alpha-blending. The ratio of object area vs. image area (relative object size) varies between (b) 5 % and (c) 0.05 %.

positive rate f for the entire multi-class classifier are then computed as (Fawcett 2003)

$$t = \frac{1}{K}\sum_{i=1}^{K}\frac{T_i}{N_i} \text{ and} \tag{5.4}$$

$$f = \frac{1}{K}\sum_{i=1}^{K}\frac{F_i}{\overline{N}_i}. \tag{5.5}$$

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

$$\overline{N}_i = \sum_{j=1,j\neq i}^{K} N_j = (K-1)N_i. \tag{5.6}$$

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; (iii) human

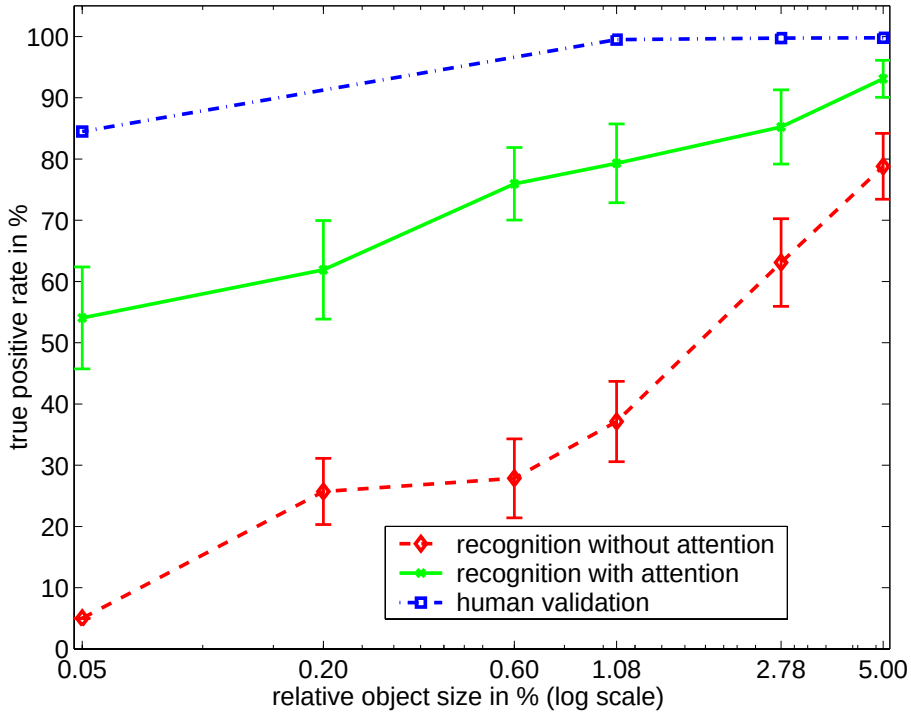

Figure 5.10: True positive rate (t) for a set of artificial images without attention (red) and with attention (green) over the relative object size (ROS). The ROS is varied by keeping the absolute object size constant at 2500 pixels ±10 % and varying the size of the background images. Error bars indicate the standard error for averaging over the performance of the 21 classifiers. The human subject validation curve (blue) separates the difference between the performance with attention (green) and 100 % into problems of the recognition system (difference between the blue and the green curves) and problems of the attention system (difference between the blue curve and 100 %). The false positive rate is less than 0.07 % for all conditions.

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. When 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, whereas the attention system is the component responsible for the failure when the human subject fails to recognize the objects based on the patches. As can be seen in figure 5.10, 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 figure 5.10 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 using the entire image (red dashed line in figure 5.10) works well without attention, as reported by Lowe (2004, 1999).

We used five fixations throughout the experiment to ensure consistency. In preliminary 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.

5.6 Discussion

We set out to test two hypotheses for the effects of attention on object recognition. The first is that attention can serialize learning and recognition of multiple objects in individual images. With the experiments in section 5.4 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 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 specificity. However, in many applications, top-down knowledge can be very useful for visual processing (Oliva et al. 2003) 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 (Navalpakkam and Itti 2005). See chapter 3 for our approach to top-down attention.

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 chapter 3 we have shown the suitability of the method for a biologically realistic object recognition system in a different context (see also Walther et al. 2002a).

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. In contrast, for a machine vision system it is beneficial to completely disregard all information outside

the focus of attention. For the work presented in this chapter, we have adopted this strategy by completely removing the luminance contrast outside the attended region, thereby restricting the search for keypoints to a region that is likely to contain an object.

An important attention related question that is not addressed in this chapter is the issue of scale of objects and salient regions in the image (see, for instance, Jägersand 1995). 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 (see chapter 3).

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 (Rybak et al. 1998). 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.