7

Attentive Classification

According to [Neisser, 1967], object recognition in human perception is done in two steps: first, attentional processes select a region of interest, and second, complex object recognition is restricted to these regions. In the previous chapters, we introduced the computational attention system VOCUS that performs the first of these steps. In this chapter, we realize the second step: VOCUS is combined with a well-known classifier [Viola and Jones, 2004] resulting in a complete recognition system. This approach is called *attentive classification* (cf. Fig. 7.1).

Although an attention system means a certain overhead in computation, its usage usually pays off since reliable and general object recognition is a complex high level vision task that is usually computationally expensive. The more general the recognizer — enabling recognition of objects of different shapes, poses, scales, and illuminations — the more important is a pre-selection of regions of interest. We discuss in which cases the recognition is sped up by the combined system and in which cases the recognition quality is improved. Several experiments illustrate this behavior. As an alternative approach, we tried the object recognition with Lowe's SIFT keypoint detector [Lowe, 2004, URL, 13] but since this was unsuccessful in first experiments, we elaborate on this approach only briefly.

The combination of attention with object recognition is suggesting and has gained interest recently. Several groups have been working on this using different recognition modules but the combination was always restricted to bottom-up attention systems. The combination of top-down attention and object recognition has not been investigated before. Additionally, to our knowledge a detailed examination of the time and quality gain has not been done before. One example of a combination of attention and recognition is presented in [Miau and Itti, 2001]. They combine an attentional module with the biologically motivated hierarchical model for object recognition HMAX [Riesenhuber and Poggio, 1999]. Since the model simulates the complex structure of early vision in cortex, it is limited in its capabilities. The objects to be detected are ellipses and rectangles in artificially constructed images. The authors extend

S. Frintrop: VOCUS: A Visual Attention System..., LNAI 3899, pp. 149-175, 2006.

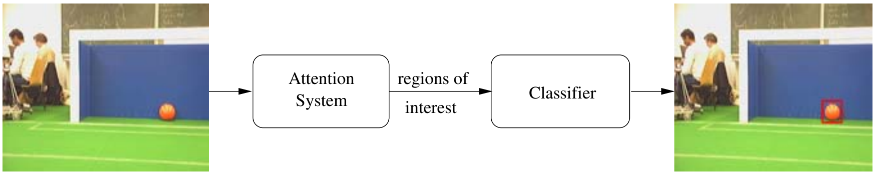

Fig. 7.1. Attentive Classification: the recognition system consists of an attention system providing object candidates and a classification system verifying the hypothesis. The combination yields a flexible and robust system

their approach in [Miau et al., 2001] using a support vector machine algorithm for the detection of pedestrians on attentionally focused image regions. In [Walther et al., 2004,Walther et al., 2005] a visual attention system is combined with Lowe's SIFT Keypoint Detector [Lowe, 2004, URL, 13]. In their approach, this is successful since they use very complex objects and those which not change viewpoint (a fixed view of the object is pasted into a scene). Since the SIFT Keypoint Detector improves if restricted to a relevant region, Walther et al. achieve an improvement in the detection rate.

In the following — after a brief discussion on object recognition in general — we introduce the classifier of Viola and Jones in section 7.1. In section 7.1.2, we touch lightly on object recognition with Lowe's SIFT Keypoint Detector. The combination of the attention and the recognition system is presented in section 7.2. In section 7.3, we show various results on the recognition of objects in both laser and camera data with the pure bottom-up system as well as with the top-down modulated system. We show how the time and the quality of performance are improved in different cases. Finally, section 7.4 concludes the chapter.

7.1 Object Recognition

General object recognition is not solved at all in computer vision [Forsyth and Ponce, 2003]. To illustrate this, it is necessary to regard what humans are able to do: Humans are very good at recognizing objects. We can name many thousands of different objects, categorize them spontaneously into groups, range new objects into these groups, redetect them in arbitrary orientations, from different viewpoints, under most difficult illumination conditions, and if they are partially occluded.

Humans also are able to recognize objects on different hierarchy levels, that means to recognize a poodle as poodle but also as a dog, a mammal, an animal, and a creature. Which level is appropriate in a particular application seems to be intuitively clear to us. Furthermore, we are able to generalize, that means to recognize different kinds of chairs, such with one leg and with four ones as well as such with or without armrests, also if we have never before seen this instance. Finally, we are able to learn new object categories from a small number of examples.

Managing these conditions is extremely difficult for computational object recognition systems. What it makes even more difficult is the question what an object actually is. Is a name plate an object? Is the logo inside the name plate an object? Is the wall an object? Is wind an object? We ignore this ontological question here and consider such things as objects that have an own designation, that are coherent and limited in spatial extent, and that have some feature values that are detectable by vision. In this view, a name plate is an object, also the logo inside, but not the wall — it is not limited in spatial extent — and not wind — it is not directly detectable by vision.

Although an optimal object recognizer does not exist, there are some good approaches that fit special kinds of recognition tasks. A common approach is to do *template matching*, that means looking for image windows that have a simple shape and stylized content. A system that tests whether a template is present in an image or not is called a *classifier*. It takes a feature set as an input and produces a class label. The classifier of choice for our experiments was the one of Viola and Jones [Viola and Jones, 2004] since it is one of the best current classifiers concerning detection and false detection rate. It will be introduced in the following. The classifier works fine on complex objects representable by several edge and line features but has difficulties with simple objects. In section 7.3.2, we show how the combination with the attention system helps to improve the recognition of such simple objects on the example of detecting balls for robot soccer.

We also did some experiments with Lowe's SIFT Keypoint Detector [Lowe, 2004,URL, 13], but found that our targets provide in general not enough stable features to enable a reliable recognition. This is shown in 7.1.2.

7.1.1 The Viola-Jones Classifier

In this section, we introduce the classifier of Viola and Jones that was originally built for face detection. It was first described in [Viola and Jones, 2001] and revised in [Viola and Jones, 2004]. The classifier works on gray-scale images, considering the composition of objects from simple features. Here, we will give only a rough overview of the classifier; more details can be found in appendix B.

Learning Features

The idea of Viola-Jones's classification method is to learn how a target object is composed of several basic features. For example, if the target is an office chair it is learned that chairs have a vertical line in the lower middle (the chair leg) and one horizontal line in the middle (the seat). If these (and many other) features are present in an image to a certain degree, the target is said to be detected. Fig. 7.2 shows the basic features the classifier considers. The

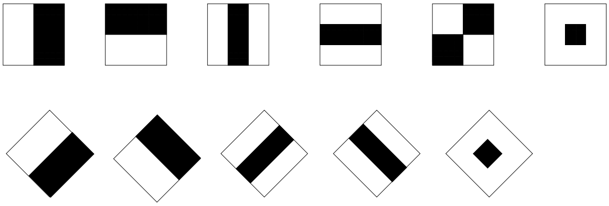

Fig. 7.2. Haar-like feature detection masks used by the Viola-Jones classifier for the detection of edge, line, and blob features [Viola and Jones, 2004]

features are called Haar-like, since they follow the same structure as the Haar basis, i.e., step functions introduced by Alfred Haar to define wavelets. They are also used in [Lienhart and Maydt, 2002, Papageorgiou et al., 1998, Treptow and Zell, 2004, Viola and Jones, 2004].

The computation of features is usually time consuming, especially if they are computed on different scales, but in this approach they are effectively calculated using *integral images* (cf. appendix B). After once creating an integral image in linear time with respect to the number of pixels, a rectangular feature value of arbitrary size is computed with only 4 references. This enables the fast computation of the features and a simple and fast resizing of features to detect objects of different sizes.

A learning technique, the Gentle Ada Boost Algorithm [Freund and Schapire, 1996], is used to select a set of simple features to achieve a given detection and error rate. In a derivative, not the simple features are used for classification and learning, but CARTs (Classification and Regression Tree) (cf. appendix B). These binary trees enable to learn objects with different characteristics, e.g., objects from different viewpoints or with different patterns (cf. section 7.3.2).

The Cascade

The performance of a single classifier, i.e., a set of simple features, is not suitable for object classification, since it produces a high hit rate, e.g., 0.999, but also a high error rate, e.g., 0.5. Nevertheless, the hit rate is much higher than the error rate. To enable an effective recognition, the relevant classifiers are arranged in a cascade, i.e., a degenerated decision tree, which consists of several stages. Each stage contains several features, the more important a feature, the earlier the stage in which it occurs. During recognition, in every stage of the cascade a decision is made whether the image contains the object or not. If the features of the stage are present to a certain degree in the image, the next stage is investigated. If not, the process stops. This enables an efficient processing: many image regions are checked solely by the first stages and only the target regions or regions similar to the target are investigated by more stages. This process also enables a high quality of recognition since the error

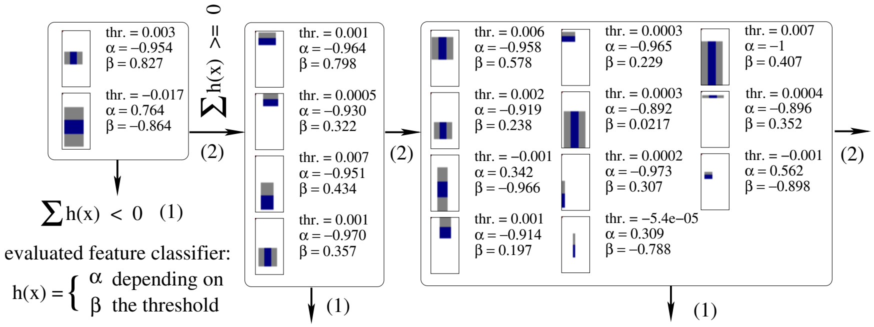

Fig. 7.3. The first three stages of a cascade of classifiers for an office chair in depth data. Every stage contains several simple classifiers that use Haar-like features. α and β are the outputs of the fitted simple feature classifiers that depend on the assigned weights, the expected error, and the classifier size [Viola and Jones, 2004] (cf. appendix B)

rate, multiplied in each stage, approaches zero. Fig. 7.3 shows the first three stages of a cascade that was built for learning an office chair in laser range images (cf. Fig. 7.4). One can see that the first stage contains one vertical and one horizontal line, both in the middle of the search rectangle. These features correspond to the leg and to the seat of the chair and are the two most important features for this object.

The Recognition

After a cascade is learned for the target object, the recognition in a test image is done as follows: a search window is laid on the test image (usually starting at the upper left corner) and it is checked with the cascade whether this region contains the object. Then the search window is shifted one or several pixels to the right and the region is checked again (cf. Fig. 7.4, left). This is done for the whole image, beginning with a search window of a specified small size (e.g. 20 x 40 pixels for chairs). Next, the detector is enlarged by rescaling the features to find objects on larger scales.

Investigating one region after the other in the classical approach has to be done since no information on the target location exists. In our approach, we already have regions of interest providing a hypothesis for the target object. Therefore, only the region of interest is investigated which is determined by the focus of attention (cf. Fig. 7.4, right); details follow in section 7.2.

Classification in Laser Images

We trained the Viola-Jones classifier not only on camera data but also on the images obtained from the 3D laser scanner (cf. section 6.1.1): the classifier

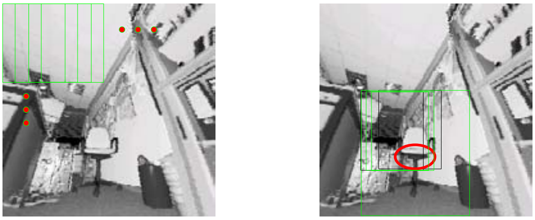

Fig. 7.4. Search windows (green rectangles) of the Viola-Jones Classifier on a test image. Left: in the classical approach, the whole image is searched for objects. Right: in our approach, only the region of interest, determined by the focus of attention (red ellipse), is investigated

was trained on images of two kinds of object: office chairs and the robot Kurt3D. Training was performed on the range as well as on the reflection data. The classification results in section 7.3.1 show that object recognition is also possible in laser data. To achieve a single result from both modes, the results of each cascade were combined by a logical "and", resulting in an output that only considers objects as detected that occur in both laser modes. Fig. 7.5 shows how this method reduces false detections. We chose the logical "and" to combine the results of the laser modes because our targets were detectable in both modes. The operation enabled a reduction of false detections. Note that in other cases a different operation might be useful, e.g., a logical "or". In this case, the detection rate increases, but also the false detection rate.

One advantage of the laser data is that it is independent of illumination, thus less training images are required. On the other hand, recognition in laser data is sometimes difficult because less information leads to several false detections. This problem is mostly overcome by the combination of range and reflection cascades. Besides the independence of illumination, the laser data has another advantage: the misclassification of shadows, mirrored objects, and wall paintings is avoided since these do not occur in the laser data. Fig. 7.6 shows this: in the scene showing a robot and a poster of a robot, only the real robot is detected.

7.1.2 Lowe's SIFT Keypoint Detector

We also did some experiments with Lowe's SIFT Keypoint Detector (SIFT: Scale Invariant Feature Transform) [Lowe, 2004, URL, 13]. This is a powerful and stable recognizer that enables the detection of complex objects or whole scenes by matching the arrangement of *keypoints* (also called SIFT features). These keypoints are invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. Roughly spoken, the keypoints are extrema in scale-space that have to stand several additional

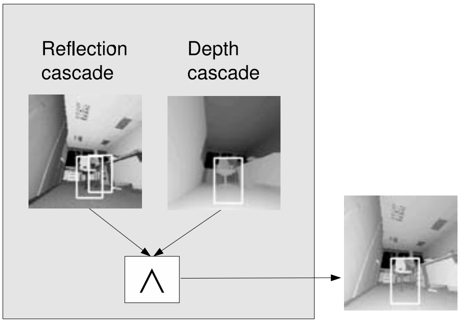

Fig. 7.5. Classification in bimodal laser data: the classification cascades of both laser modes are combined by a logical "and", resulting in an output only considering objects as detected that occur in both laser modes

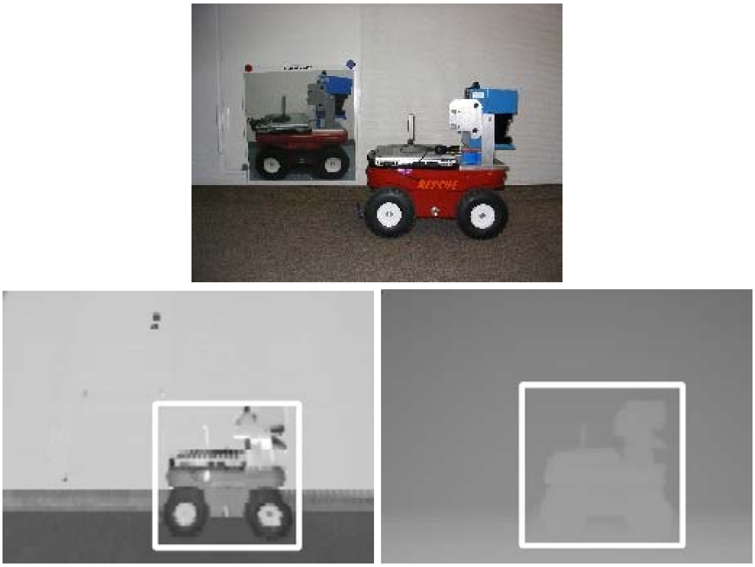

Fig. 7.6. A camera image of the robot next to a poster showing a robot (top). In the laser data of the same scene, the poster is not visible due to the infrared light and the range information (bottom); this prevents misclassification: only the real robot is detected

tests, e.g., rejecting unstable extrema with low contrast. The matches are identified by finding the 2 nearest neighbors of each keypoint from the first image among those in the second image, and only accepting a match if the distance to the closest neighbor is less than 0.6 of that to the second closest neighbor. The threshold of 0.6 can be adjusted upwards to select more matches or downwards to select only the most reliable ones. An example taken from the

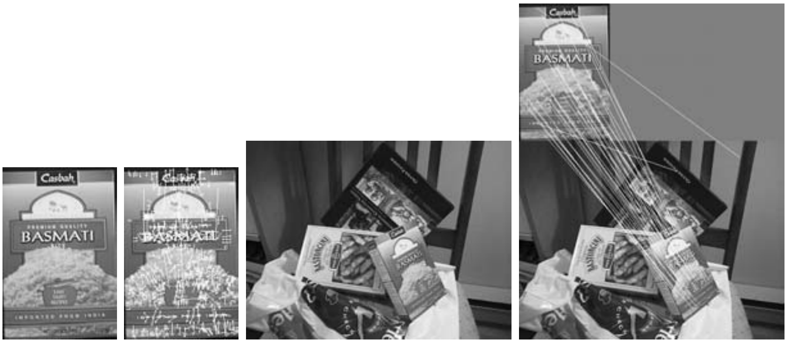

Fig. 7.7. Testing Lowe's SIFT Keypoint Detector [Lowe, 2004,URL, 13] for a complex target (basmati rice box). From left to right: 1) Training image. 2) 572 keypoints on training image. 3) Test image. 4) Test image in which 38 keypoints match. For such complex target objects, the recognition is successful even if the target if presented from different viewpoints (original images from [URL, 13])

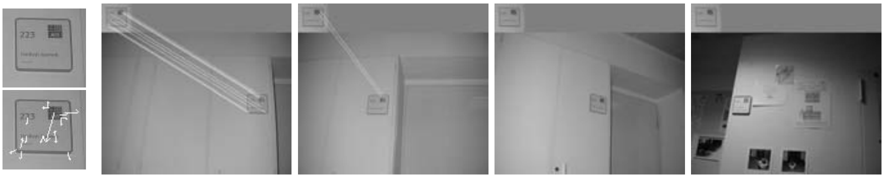

Fig. 7.8. Testing Lowe's SIFT Keypoint Detector [Lowe, 2004,URL, 13] for a simple target (name plate). From left to right: 1) Training image (top) and 17 keypoints on training image (bottom). 2) Search target in the image from which the training image was cut. 3) Test image in which only 2 keypoints match. 4,5) Two test images in which no keypoints match. For such simple target objects, the recognition is difficult even for slight changes in viewpoint and fails completely for larger changes

online data on David Lowe's web pages [URL, 13] is shown in Fig. 7.7. The target object, a basmati rice box, has a complex texture which enables the detection of many keypoints: 572 keypoints are detected. This allows a redetection in the test image on the right: 38 keypoints are successfully matched. An application in which this approach shows good results is the recognition of building facades.

One condition for this approach is "that it generates large numbers of features that densely cover the image over the full range of scales and locations" [Lowe, 2004]. Unfortunately, we found that this is not the case for our targets: these provide in general not enough stable features to enable a reliable recognition. This is shown in Fig. 7.8. For the target object "name plate"(left) only 17 keypoints are detected that may be used for matching with a test object. When the target was searched in the training image itself, the recognition was successful (second left): 15 matches were found. But when it was searched in other test images, nearly no matches were found: in a simple test image (third left), only two matches were found, in more difficult test images (right and second right), nothing was found. For targets with even less features, e.g., the highlighter or key fob of chapter 5 or the balls of section 7.3.2, the recognition is probably even worse. Therefore, it seems that this approach is not adequate for our case.

Walther et al. [Walther et al., 2004, Walther et al., 2005] did also experiments in which they combine a visual attention system with the SIFT Keypoint Detector. In their experiments, this yielded satisfying results because the objects were sufficiently complex and because in most experiments they paste the object into an image scene so that it appears always from the same viewpoint which simplifies the recognition significantly.

7.2 Attentive Classification

Attentive classification means the combination of a fast attention system, applied to the whole scene, with a powerful classifier, restricted to a region of interest (cf. Fig. 7.1). This is an effective way to improve the quality and time performance of vision systems: the attention system points to a region of interest but is not able to determine which object is in this region (bottom-up) or whether a searched target is actually present (top-down). On the other hand, a general classifier needs a lot of time if applied to the whole image. Restricting the classification to the region of interest is much more effective and also improves the quality of recognition in certain cases as will be shown in section 7.3. The more complex and general a recognition system, the more useful is an attentional front-end concentrating the processing on special regions.

The attention system may be used in a pure bottom-up mode or it may search for a target in top-down mode. These are two principally different approaches: the bottom-up system is used in an exploration mode; no special target is given. The system shall favor salient objects or it shall recognize as many objects as possible but does not have the time to cope with all objects. So in the bottom-up mode, the attention system finds regions of interest and the classifier determines the identity of the fixated region. Instead in the top-down mode, the system is searching for a target which is known by the attention as well as by the classification module. Thus, the attention system generates an object hypothesis which is verified or falsified by the classifier.

If the task of a vision system is exploration and recognition of several objects in a scene and there is not enough time to analyze all image regions in detail, a priority has to be set. A simple priority that is usually set in such a case is to scan the scene from upper left to lower right to recognize the first object in the database, then find the second one and so on. Alternatively, the first search window may be searched for all of the objects, then the second window and so on. If time is rare, the first approach has the effect that the objects at the end of the database are never recognized while the second approach has the effect that objects in the lower right corner of the scene are ignored. A much better approach is to detect objects in order of their saliency (attentive classification). The attention system computes a sequence of image regions in order of their saliency. The first region in this sequence is investigated by the classifier for each object in the database and the next salient region is only investigated after recognizing the object — or deciding that the object is not known. Of course, this approach also misses some objects — the non-salient ones — but since this is inevitable due to the lack of time, the missing of non-salient objects is the lesser evil.

There is another application scenario, in which recognizing only salient objects is even preferred to the recognition of all objects: if a system is very complex and knows about a wide variety of object classes it might be useful to not consider everything in the environment. This is also true for humans: not every object in the environment is noticed but mainly salient and/or task-relevant objects are recognized. The socket in the corner will probably not be noticed if you look around in a new entered room unless you need power supply. Here, it is sensible to have an attention system narrowing down the choice of regions for recognition. One application in robotics in which this behavior is useful is the creation of semantic 3D maps, that means maps that contain information about the objects in the environment. Surely, it is not wanted to include every object in the map since this would overload the map and make it confusing — if the map is considered for interaction with humans — or leads to computational problems because of the high amount of data. The attention system is able to restrict the processing to those image regions that are worth to be investigated.

Since a focus of attention is often not on a whole object but on its border or on parts of it, not only the focused region is investigated by the classifier, but a larger region surrounding the focus (cf. Fig. 7.4). In our experiments it turned out that choosing a region which is four times as large as the expected size of the target object yielded good results. For example, for name plates we chose a region of 54×54 pixels on test images of size 512×384. This is about 2% of the image area. Inside this region, the search windows were placed so that the middle of the search window lies inside the region. For name plates, investigating this small region was enough since name plates did not exceed a certain size in our image sets. These images were taken by the author so it would have been possible to take a close up view of a name plate in which it appears much larger. This is not possible in the future application in which a robot will take the pictures since the fixed camera is not able to get closer to the name plate. For other objects, e.g., for the chairs in the laser data, it may be necessary to determine a larger region. If the object may fill the whole image, region size has to be equal to image size.

When the attentive classification is applied to laser instead of camera data, the procedure is roughly the same (cf. Fig. 7.9). The main difference is that both the attention system and the classifier operate not only on one but on two images: range and reflection data. The combination of the classification results

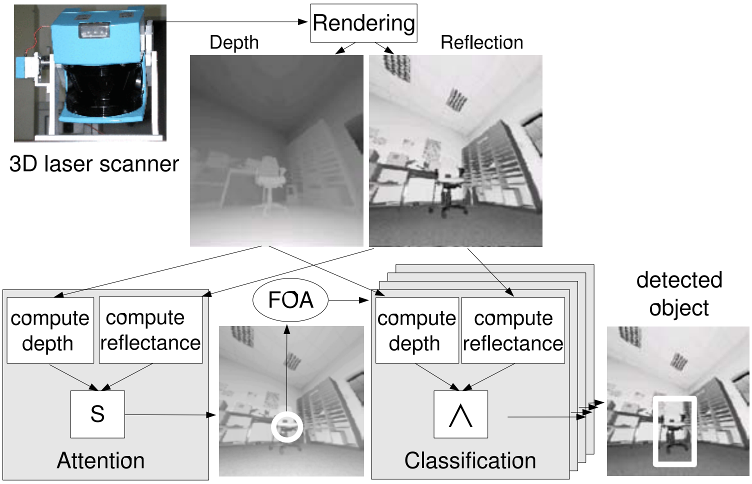

Fig. 7.9. Attentive classification on laser data: two laser modes, depth and reflection, are provided by the 3D laser scanner, rendered into images and fed into the attention and the classification system. The attention system fuses conspicuities of both modes in one saliency map (S) and generates a focus of attention (FOA) which is fed into the classification system. The classifier searches for objects of predefined classes in the neighborhood of the FOA in both laser images and combines the results by a logical "and". The rectangle in the result image (right) depicts a detected object

for each mode by a logical "and" narrows down the number of detections and reduces the amount of false detections.

In the following, we discuss the performance gain achieved in different applications. We start with debating the time savings, especially occurring for the bottom-up system, and after that we argue in which cases the quality is improved by eliminating false detections.

7.2.1 Time Performance

The time saving achieved with the combination of attention and classification depends on the complexity of the classifier as well as on the number of objects that are of interest in a special scene. If the classifier is highly complex and determining the object identity is a time consuming task — more time consuming than the attentional computations — there is no doubt that the combination with the attention system yields a gain in time performance.

But what if the classifier itself is extremely fast, as the Viola-Jones Classifier? In this case the time saving depends on the number of object classes that have to be considered: if only one type of object has to be detected, it might be useful to stick to the classifier and ignore the attention system. But for a complex vision system, knowing 5, 10, or even hundreds of objects it is extremely useful to search for all of these objects merely in a pre-specified region. This time saving usually occurs when the bottom-up attention system is used which determines a region of interest and many classifiers are used to determine the identity of what is in this region.

Here, we analyze the time performance for the bottom-up attentive classification on laser data. On 300 × 300 pixel images, the attentive classification at a region of interest needs on average 60 ms, compared to 200 ms for an uninformed search across the whole image (Pentium-IV-2400). So the focused classification needs only 30% of the time of the exhaustive one. Note that for other objects like name plates this percentage is even lower since a smaller image region is investigated. The attention system requires 230 ms to compute a focus for both modes; hence, for m object classes the exhaustive search needs $m * 200$ ms versus $230 + m * 60$ ms for the attentive search. Therefore, already for two different object classes the turning point is reached: the exhaustive search needs 400 ms, whereas the attentive search requires only 350 ms. The time saving increases proportionally with the number of objects, and for 5 objects the attentive classification is already twice as fast as the exhaustive classification as is shown in Fig. 7.10.

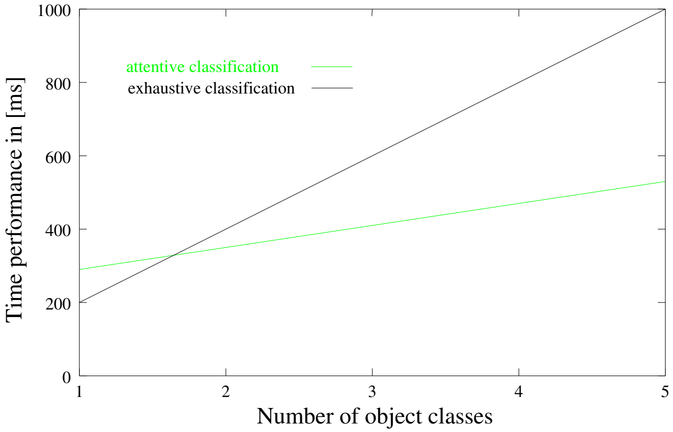

Fig. 7.10. The time saving of bottom-up attentive classification depends on the number of object classes: the more classes, the higher the time saving in the attentive approach. Already for 5 classes, the attentive classification is nearly twice as fast as the exhaustive classification

If the attention system is applied to color camera images, the required time increases since the color computations are time consuming. Instead, the classifier works only on gray-scale images, thus the required time does not increase. Therefore, the turning point is reached at a later point. On the other hand, the use of color allows to consider other object properties enabling a better classification quality (cf. section 7.2.2).

The top-down attention system is applied if a target object is known. This means it is clear which object is searched and which classifier should be applied. Hence, a time saving is only achieved if, firstly, the classifier is more time consuming than the attention system or, secondly, several objects have to be searched in the same scene. In the latter case, the top-down attention system has to determine a new region of interest for each object class, but this does not mean that the whole computation needs to be repeated. For each object class, the weighting of the feature maps with the target's weights vector, the computation of excitation, inhibition, and top-down saliency map have to be performed. But the computation of the image pyramids, the conversion to the LAB color space, and the computation of the feature maps need to be performed only once for a scene. Since these are the most expensive computations, the time increases only slightly for several object classes and the combination with classification pays off.

7.2.2 Quality Performance

The attentive classification increases the performance not only in time but in many cases also in quality. The pre-selection of regions with potentially higher interest than the rest of the image is a quality choice by definition (regions of interest have usually a higher quality than regions of no interest). This has different effects in bottom-up and top-down mode. In bottom-up mode, this is useful if time has to be saved or if only a few objects shall be considered. If, for example, the five most important objects in a scene shall be localized, the bottom-up system of attention may help to select them. Other improvements of recognition quality with help of a bottom-up attention system were reported in [Walther et al., 2004,Walther et al., 2005]. They were using Lowe's SIFT Keypoint Detector [Lowe, 2004,URL, 13] that improves if restricted to a relevant region, so they achieve an improvement in the detection rate. This is not possible for the Viola-Jones Classifier which achieves the same results if focused on the target as if searching the whole image.

In top-down mode, another aspect of quality improvement reveals: the elimination of false detections. Combining attention and classification means to take the intersection of the results of both systems; this diminishes the detection rate as well as the number of false detections. Therefore, an improvement of quality is achieved in cases in which both systems have a reasonable detection rate — which stays almost the same — and the classifier produces many false detections — which are significantly reduced. This is usually the case for simple objects like balls. In section 7.3.2, we will show how the quality of recognition is essentially improved by using the attention system as front end to the classifier.

7.3 Experiments and Results

In this section, we present some experiments of the attentive classification system. We begin with using VOCUS in a bottom-up mode as front end, followed by investigating the combination with the top-down mode.

7.3.1 Bottom-Up Attentive Classification

In a first step, we use the bottom-up mode of VOCUS for the attentive classification[1]. The experiments were performed on laser data. This approach allows the recognition of the most salient objects in a scene what is useful in complex systems that know a wide variety of objects but do not have the time to analyze all objects in a scene. The attention system provides the priority of which region to analyze first.

In the following, we first show the performance of the classifier when trained on laser data before we combine it with the attentional front-end.

Classifier:

The classifier was trained on the objects chairs and the robot Kurt3D in laser images (300 × 300 pixels). We rendered 200 training images with chairs from 46 scans and 1083 training images with the robot from 200 scans (the rendering is explained in [Nüchter et al., 2005]). Additionally, we provided 738 negative example images to the classifier from which a multiple of sub-images is created automatically. The test set consists of 31 chair and 33 robot images for each laser mode yielding 128 test images, disjoint from the training set. There were 33 chairs and 33 robots in the scenes: some images contained two chairs but in each image there was at most one robot. Note that in this test the classifier was applied to the whole images.

We determined the detection and false detection rates for images of both laser modes independently and then for the combination of both approaches. Table 7.1 summarizes the results. It shows that the detection rate for each mode reached about 90% and there were some false detections: usually only 1 or 2, but there happened to occur 10 false detections for one test set. When the modes were combined, the number of false detections was reduced to zero while the detection rates changed only slightly (see also [Nüchter et al., 2004]).

The classifier is still successful if the object is partially occluded (see Fig. 7.12, middle). However, severely occluded objects are not detected (see Fig. 7.11); the amount of occlusion still enabling detection depends on the learned object class and has to be investigated further. In Fig. 7.12 middle, the chair is not only partially occluded, it is also presented sidewards and still recognized. Of course, it depends on the object if this is possible. In the case of the chair this is possible because the main features in the cascade belong to the seat and the chair leg. These features are still present in the rotated

[1] The results of this section were also published in [Frintrop et al., 2004b].

Table 7.1. Detections and false detections of the Viola-Jones classifier applied to 31 chair and 33 robot images. While the detection rate stays about the same for the combination of both laser modes, the false detections are reduced to zero

object class	# of obj.	detections			false detections		
		reflection image	depth image	**combined**	reflection image	depth image	**combined**
chair	33	30	29	**29**	2	2	**0**
robot	33	29	29	**29**	10	1	**0**

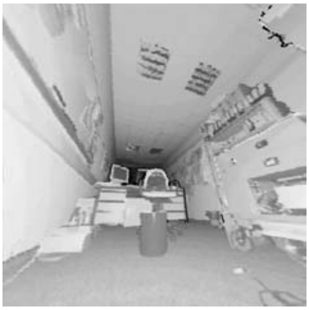

Fig. 7.11. Image of a chair with strong occlusion. In this example, a recognition with the Viola-Jones Classifier was not possible

version of the chair. The robustness of the classifier according to rotations was tested in more detail for the object class robot. We recorded scans of the robot rotated by 10° at a time. It showed that a robot rotated by 30° is still recognized (Fig. 7.13, right), but it is not if it is rotated more. To enable a recognition under an even greater change of orientation, a rotated version of the robot has to be trained. The same is true for a robot presented the other way round. The training of objects of different orientations can be done with the CARTs mentioned on page 152.

Classifier + Attention system:

When classifying objects at regions of interest, it depends on both systems what is recognized and the result is the intersection of the results of both systems run separately. The classifier detects all focused objects with the same reliability like when applied to the whole scene. Note that if no focus points to an object, this object is not detected. This conforms to our goal to detect only salient objects in the order of decreasing saliency. As discussed in chapter 4, it is hard to evaluate the quality of bottom-up FOAs, thus here we concentrate on presenting some examples of focused and classified objects in laser data in Fig. 7.12 and 7.13. The objects are successfully detected even

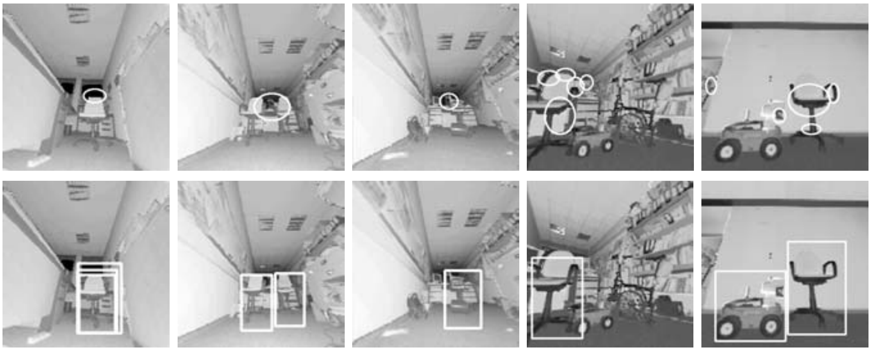

Fig. 7.12. Attentive classification in laser data. Top row: the first resp. the first 5 foci of attention computed on depth and reflection data. Bottom row: classified objects in the focus regions. Left to right: 1) Chair is detected even if the focus is at its border; 2) detection of two chairs; 3) chair is detected although it is presented sidewards and partially occluded; 4) only the chair is focused, therefore the chair but not the robot is classified; 5) both objects are focused and classified

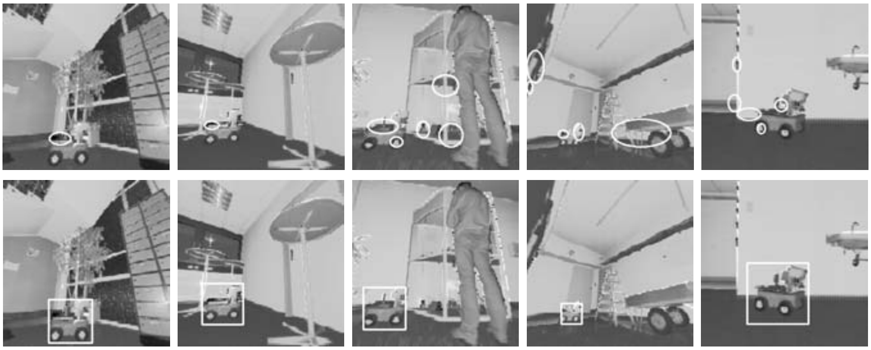

Fig. 7.13. Attentive classification in laser data. Top row: the first resp. the first 5 foci of attention computed on depth and reflection data. Bottom row: classified objects in the focus regions. Right: a robot rotated by 30° is still detected

if the focus is at the object's border (Fig. 7.12, left) since a sufficiently large search region around the focus was chosen.

7.3.2 Top-Down Attentive Classification

If the system is searching for a target instead of exploring the environment, it is clear which classifier has to be applied; there is no use of applying many different classifiers to the image. So in this approach, the attention system provides a hypothesis for the target location which is then verified or falsified by the classification system. The experiments in this section aim at showing the improvement in quality rather than in time; the conditions for an improvement in time were discussed in section 7.2.1.

We investigated the search performance for two different targets: name plates and balls. The experiments show a very different behavior: the name plates are hard to detect by the attention system but rather successfully by the classifier. A combination declines the detection quality. In contrast, balls are easily detected by the attention system but the classifier has difficulties distinguishing them from other round image regions resulting in many false detections. In this case, the combinations yields a significant increase in detection quality.

Experiment 1: Name Plates

Classifier:

The Viola-Jones classifier was trained with 1079 images of name plates. We tested the system with 54 untrained images, applying the search windows to the whole images. Each image contained exactly one name plate. The results are shown in Tab. 7.2; they show that the detection of name plates with the classifier is quite successful: only two name plates are missed and there were 9 false detections. Some examples of the classification results are depicted in Fig. 7.14.

Table 7.2. Classification results for name plates when investigating the whole images (exhaustive classification)

Target	# test im.	Detected	Not Detected	False Detections
name plate	54	52	2	9

Attention system:

As we have shown in chapter 5, the detection of name plates with the top-down attention system is quite difficult due to many similar regions in the surrounding. Table 7.3 shows the detection results for different numbers of foci. In a majority of images (62%), the detection was very successful and the name plate was found by the first focus. But the other images were more difficult resulting in higher hit numbers.

From these results, we already expect that the recognition of name plates with attentive classification yields no gain in quality performance: if few focus regions are considered, too many targets are missed and if many are considered, the false detections will probably not be diminished. This expectation is verified in the next section.

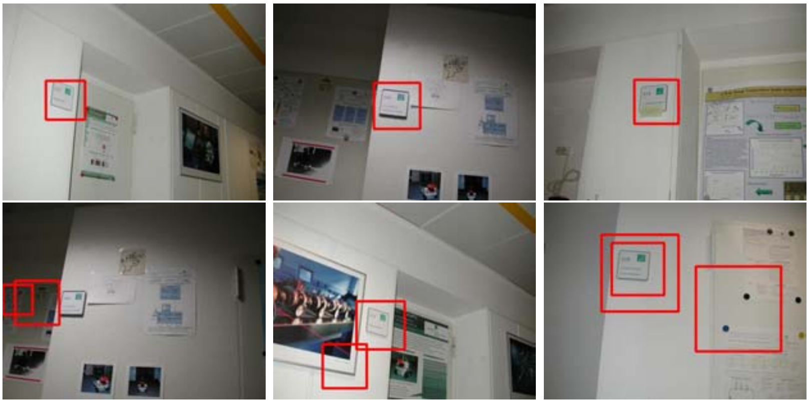

Fig. 7.14. Classifying name plates with the classifier of [Viola and Jones, 2004]. First row: perfect classification. Second row: one miss, 4 false detections and one double detection

Table 7.3. Detection results of VOCUS when searching for name plates. Different numbers of FOAs are considered

Target	# test im	# FOAs	Detected	Not Detected	Average hit number
Name plate	54	1	34	20	1.00
Name plate	54	5	46	8	1.48
Name plate	54	10	51	3	2.16

Classification + Attention system:

One example of attentive classification is shown in Fig. 7.15. The top-down attention focuses on the name plate and the classifier — restricted to this region — detects it without a false positive. This however is only achieved if merely the first focus region is considered. In this case, the number of false de-

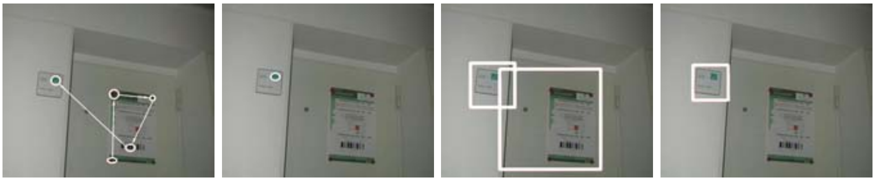

Fig. 7.15. Searching for name plates. From left to right: 1) The first 5 FOAs by pure bottom-up attention, the 5th FOA is on the name plate. 2) The 1st FOA by top-down attention searching for name plates. 3) A false detection found by the classifier while scanning the whole image. 4) No false detection occurs when concentrating on the first region of interest found by the attention module

tections is diminished to 3 (see Tab. 7.4). On the other hand, in 20 images the name plate is missed. This seems unacceptable, so what happens if more focus regions are investigated? It turned out that in this case more name plates are detected but the number of false detections increases too, unfortunately to a number higher than the one for pure classification: this is possible because in this approach, the false detections are counted for each focus separately. Therefore, often the same region yields two or more false detections for different foci. It would be possible to diminish these false detections by checking whether a detection is in the same region as a previous one. Nevertheless, the number of false detections would remain high and would only diminish to about the number of false detections in exhaustive classification.

These results show that our expectation was correct: the combination of both systems yields no gain in quality and if the task is to search only for name plates, it is more sensible to use only the classifier without the attention system. But if several objects have to be detected in the same scene or if a more complex and time-consuming recognition module is used, the favoring of regions provided by the attention system is still useful because of the gain in time performance.

Table 7.4. Results of attentive classification when searching for name plates. The detection rate is the same as in Tab. 7.3, i.e., the 2 targets not detected by the classifier were also not detected by the attention system. See text for further explanations

Target	# test im.	# foci	Detected	Not Detected	False Detections
name plate	54	1	34	20	3
name plate	54	5	46	8	18
name plate	54	10	51	3	31

Experiment 2: Balls

In this experiment, we detect balls for a RoboCup scenario (the Robot World Cup Soccer Games and Conferences [URL, 16]). Until now, balls for RoboCup were of a bright red, simplifying the detection significantly and resulting in algorithms usually based on color. In future, the color coding will be removed to achieve a more realistic setting.

We propose an approach that enables the detection of arbitrary balls; it consists of a training phase — taking place once in advance —, an adaptation phase — taking place immediately before the game when the kind of ball is known —, and a detection phase during the game. In the training phase, the classifier learns the shape of balls considering balls of different sizes, colors, and surface patterns. In the adaptation phase, the top-down attention system is quickly adapted to the actual ball by learning ball-dependent features

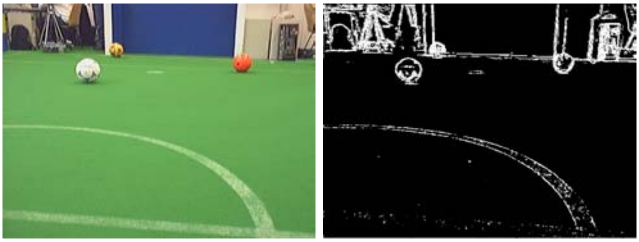

Fig. 7.16. Left: image of a RoboCup scene including the three kinds of balls that were used for training the classifier. Right: the corresponding edge image generated with a Sobel filter. The classifier was trained on such images

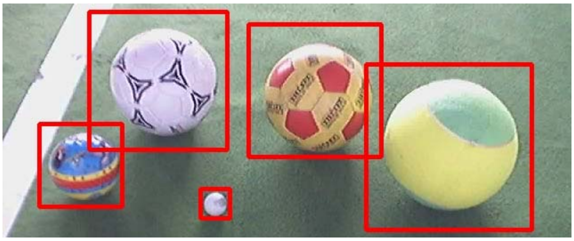

Fig. 7.17. Five different kinds of balls are detected by the classifier

from a few training images. In the detection phase, first, the attention system computes regions of interest by weighting the image features with the learned weights. Second, the classifier is applied to these regions, verifying the object hypotheses.

Classifier (Training phase):

The training and testing of the classifier for balls was done by my colleagues Sara Mitri and Kai Pervölz[2]. They showed in [Mitri et al., 2004] that the classifier, when trained on different balls in the original image data, performed bad because the object is too simple and contains few features. In various experiments they investigated that the performance was significantly improved if edge filters were applied before training. Thus, they used a Sobel filter (cf. appendix A) to obtain edge images as the one in Fig. 7.16 (right) which was then put into the classifier for training. To obtain useful edge images from the color images, the filter was applied to each channel of the colored image separately and then a threshold t was used to include any pixel in any of the 3 color channels that exceeded t in the output image. As shown in [Mitri et al., 2004], this yielded much better edge images than the application of the filter to the image converted into gray-scale.

[2] Thank you for marking all these balls!

The ball detection cascade was learned with 1000 images (640 × 480 pixel) showing complex scenes with up to three soccer balls of different colors and patterns. The three balls for training are shown in Fig. 7.16, left. To enable the detection of different kinds of balls, the training was done with CARTs. Fig. 7.17 shows the detection results on five different kinds of balls. Since only the upper two balls (white and yellow/red ball) were used for learning, the image demonstrates the classifier's ability to generalize to different kinds of balls.

For each kind of ball, 60 images were tested, making 180 test images altogether. Table 7.5 shows the detection and false detection rates for each kind of balls. The detection rate of the classifier is adjustable with the number of stages, i.e., a lower number of stages of the cascade increases the number of detections, but also the amount of false detections; with more stages, the detection rate diminishes but there are also few false detections.

The table shows that ball recognition is still a difficult problem: there are many false detections for all kinds of balls since the classifier learns mainly the round shape of the balls and so it is difficult to differentiate between soccer balls and other spherical image regions. At least 12 stages are needed to diminish the number of false detections to 80 for 180 images what is still a lot. But for this number of stages, the detection rate is reduced to 60%. As we will show in the next section, combining the attention system with the classifier trained with few stages improves the results significantly: restricting the region of interest with the top-down modulated attention system helps to strongly reduce the false detections with only a slightly diminished detection rate.

Attention (Adaptation phase):

In the robot soccer scenario, the adaptation phase takes place immediately before a game starts, i.e., when the actual kind of ball to be used is known. This ball is trained on the spot with the top-down attention system from a few (here: 2) training examples. We used the algorithm of Fig. 5.9 to choose some suitable training images from a training image set of 10 images (for VOCUS, we converted the images to half of their size: 320 × 240 pixels). It turned out that two training images were sufficient to yield a local optimum in performance.

In Table 7.6 we show the results of the top-down attention system when searching for balls while considering the first 5 foci. It reveals that in all cases the search is very successful. Obviously, the design of the balls is well chosen to distinguish it from its environment. Most successful is the detection of the red ball: in all of the test images, the ball was immediately detected with the first focus. But even the white ball, although missed in 7% of the examples, is on average detected with the 1.7th focus. What refrains us from using only the attention system is that this system does not distinguish between targets and non-targets. It is not able to detect if there is no ball in the scene; instead, in this case the system points to the regions that are most similar.

Table 7.5. Classification results of the cascade of classifiers depending on the used number of stages. The cascade with 10 stages (bold face) was used for the experiments with the attentive classification

	# stages	# test im.	Detections	Not Detected	False Detections
red ball	9	60	52	8	114
white ball			48	12	70
yel/red ball			57	3	108
Total		180	157	23	292
red ball	**10**	**60**	**45**	**15**	**52**
white ball			**44**	**16**	**45**
yel/red ball			**57**	**3**	**63**
Total		**180**	**146**	**34**	**160**
red ball	11	60	45	15	51
white ball			42	18	47
yel/red ball			56	4	65
Total		180	143	37	163
red ball	12	60	44	16	26
white ball			29	31	31
yel/red ball			37	23	23
Total		180	110	70	80

Table 7.6. Detection results of VOCUS when searching for different balls. In each image, the first 5 focused regions are considered

Target	# test im	Detected	Not Detected	Average hit number
Red Ball	60	60	0	1.0
White Ball	60	56	4	1.7
Yel/red Ball	60	60	0	1.1
Total	180	176	4	1.3

Attentive Classification (Detection phase):

In the combined approach, first the balls are searched with the top-down modulated attention system, and second the first five FOA regions are investigated by the classifier. Therefore, the output is the intersection of both result sets: the detected balls must be found both by the attention algorithm as well as by the classifier.

The results of the attentive classification are shown in Table 7.7. It shows that the false detections are significantly reduced in the combined approach versus pure classification to 23 from 160 while the detections remain nearly stable (141 vs. 146). This is much better than the performance of the classifier with more stages: for 12 stages, the number of false detections was 80, with 110 detections.

Several of the results are depicted in Fig. 7.18 – Fig. 7.20. In the first row of each figure, we show one example in which the results are the same for exhaustive and attentive classification. In the other examples, we focus on more interesting cases in which the combination of the systems yields a difference, e.g., the cases in which false detections are diminished.

When looking closer at the results of the different kinds of balls, it reveals that the performance is different for each kind: for red balls, the detection rate remains stable whereas the false detection rate is diminished significantly from 52 to 1. For white balls, the detection rate shrinks slightly from 44 to 41 and the 45 false detections are completely eliminated. Most false detections occur for the yellow/red ball: 20 of the 63 false detections remain. It is interesting that although for the white ball many of the first 5 foci do not point to the ball but to other regions, the false detections are completely eliminated. Obviously, these regions and the false detections of the classifier were disjoint. Instead, for the yellow ball several false detections remain: in these cases, the foci pointed to regions which were also misclassified by the classifier.

Table 7.7. Comparison of the exhaustive classification with the attentive classification. We used the classification cascade with 10 stages. Column 2 (attention) shows the average hit number (cf. Def. 1). It shows that the false detections are significantly reduces in the attentive approach while the detection rate remains nearly stable

	# im.	Attention	Classifier only		Attentive Classification	
		Av. hit nb.	Detect.	False Detect.	Detect.	False Detect.
red ball	60	1.0	45	52	45	1
white ball	60	1.7	44	45	40	0
yel/red ball	60	1.1	57	63	57	20
Total	180	1.25	146	160	142	23

7.4 Discussion

In this chapter, we examined the combination of the attention system with a classifier, an approach which we called attentive classification. This method represents an important step towards effective general object recognition since it constrains complex and time-consuming computations to restricted parts of the data.

Against common understanding, often not the complex objects are the ones causing problems in recognition, but the simple ones. The simpler an object, the more difficult it is to distinguish it from other regions in a scene. Since recognition systems usually focus on recognizing special features, e.g., they focus on gray-scale edge features, the risk is high that these features are

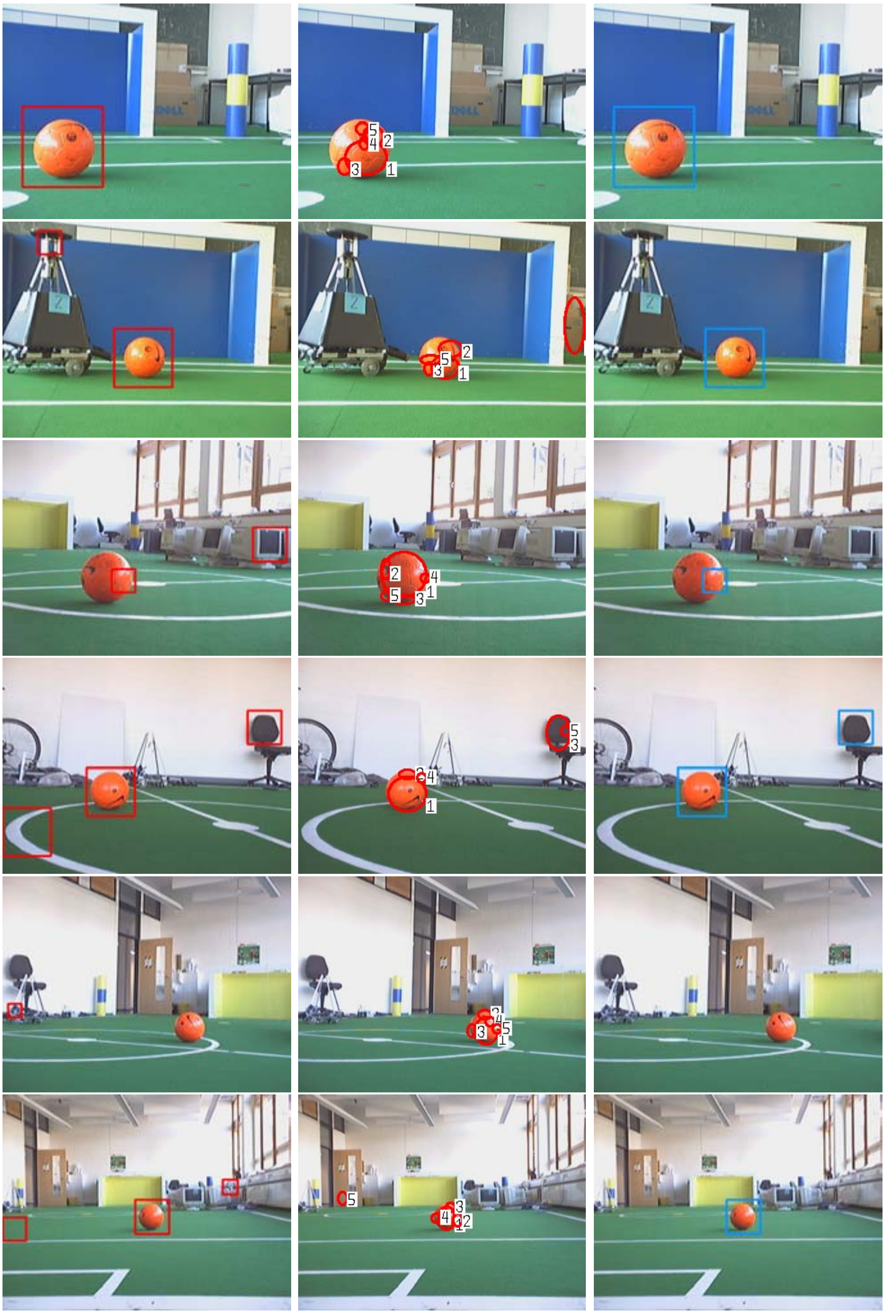

Fig. 7.18. Detecting red balls. Left: classifier only. Middle: first 5 FOAs of VOCUS in top-down mode. Right: attentive classification; most false detections are eliminated

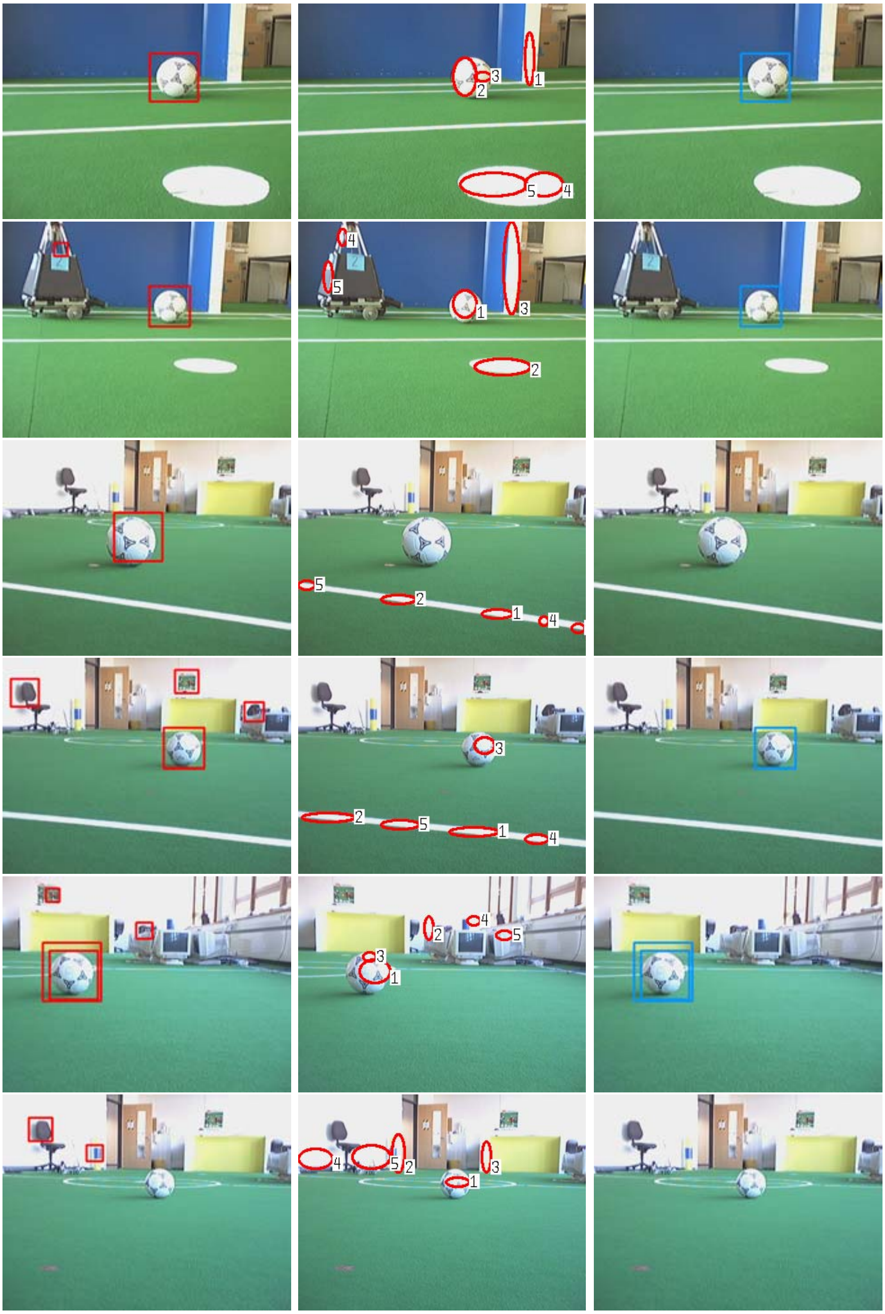

Fig. 7.19. Detecting white balls. Left: classifier only. Middle: first 5 FOAs of VOCUS in top-down mode. Right: attentive classification; most false detections are eliminated

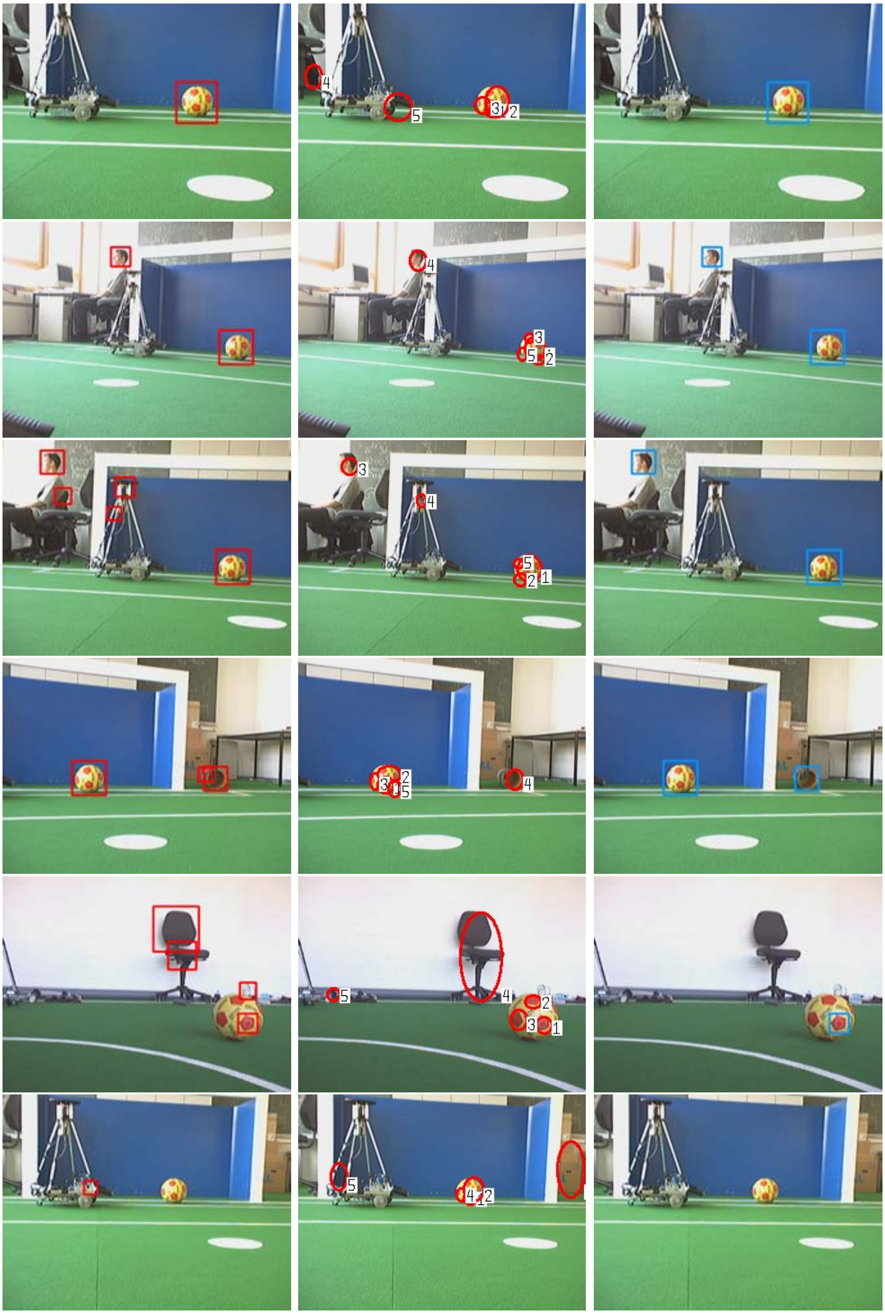

Fig. 7.20. Detecting yellow/red balls. Left: classifier only. Middle: first 5 FOAs of VOCUS in top-down mode. Right: attentive classification; most false detections are eliminated

not sufficient to recognize the target successfully. We illustrated this behavior for the example of detecting name plates with Lowe's SIFT Keypoint Detector [Lowe, 2004, URL, 13] and on the examples of detecting balls with the Viola-Jones classifier. The same problems occur for objects like the highlighter or the key fob of chapter 5. We have shown how the combination with the attention system enables a significant improvement of the detection results for simple objects. However, it may be noted that an expansion of a recognizer to process color information may yield similar results. Though, this would lack the advantage of the fast adaptability of the system to color.

The presented approach is a straightforward way to provide the attention system with a module which verifies the generated object hypothesis. It shall be noted that it is a technical solution resulting from the need to achieve a solution which is as robust and fast as possible. In more biologically motivated systems, attention and classification are more intertwined and share resources. That means, the extracted features give a first hint about the object which is then verified more and more by combining more complex detection results. It is interesting to develop this approach further as for example done in [Hamker, 2005] and in [Navalpakkam et al., 2005] but unfortunately at the moment these methods have very low quality in detection and false detection rate and are only able to distinguish object properties very roughly. The classification by Viola and Jones yields high quality results which was the reason for us to choose it. However, it would be an interesting idea to develop a high quality recognizer based on the early features that were already computed by the visual attention system, a subject we leave for future work.