Attentive Classification

Simone Frintrop[1], Andreas Nüchter[2], Kai Pervölz[3], Hartmut Surmann[3], Sara Mitri[4], and Joachim Hertzberg[2]

[1]Institute of Computer Science III, University of Bonn,
Römerstr. 164, 53117 Bonn, Germany
[2]Institute of Computer Science, University of Osnabrück
Albrechtstrasse 28, 49069 Osnabrück, Germany
[3]Fraunhofer Institut Intelligente Analyse- und Informationssysteme (IAIS),
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[4]Ecole Polytechnique Fédérale de Lausanne (EPFL),
1015 Lausanne, Switzerland

Abstract. In this paper, we present a two-step approach for object recognition based on principles of human perception: *Attentive Classification.* First, regions of interest are detected by a biologically motivated attention system. Second, these regions are analyzed by a fast classifier based on the Adaboost learning technique. Thus, the classification effort is restricted to a small data subset. The approach has two advantages over normal classification: First, the system becomes considerably faster, which is an important factor for real-time systems. Second, since the attention system is able to make use of top-down target-information, the combination of the systems yields a significant reduction of false detections for objects which are usually difficult to discriminate from the surrounding. We show the performance of the system in several experiments in robotic scenarios. The presented attentive classification system represents an important step towards effective general object recognition which is fast, robust and flexibly adaptable to a current task.

1 Introduction

Artificial intelligent systems, like robots, require high-bandwidth, multi-modal, redundant sensor information to perform interesting tasks like exploration, reconnaissance or searching in mundane environments. On the other hand, a system continuously receiving such a rich data stream is in permanent danger of drowning in data, especially if it has limited on-board processing capacity. The challenge, then, is to select at any time that data segment which is currently most important or urgent, focus on that data, and ignore the rest largely or completely. For doing so, it is necessary to tell the salient parts from the rest with little effort. Solving this problem may not be as hopeless as it sounds, given that many biological species, including humans, appear to cope with it very well.

One of the mechanisms that make humans so effective in acting, based on their limited processing capabilities and rich data, is their ability to extract relevant information at an early processing stage, a mechanism called *selective attention.* The extracted information is then directed to higher brain areas where complex processes such as object recognition take place. This efficient two-step recognition process in human perception has already been pointed out by Neisser forty years ago [39] but is has been only during the last decade that such mechanisms arouse more and more interest in computer vision and robotics. The more complex the systems become and the more data they have to process, the more urgent the need for a pre-selection which restricts expensive recognition processes to regions of interest.

In this paper, we present such a two-step recognition approach called *Attentive Classification* (cf. Fig. 1): in a first step, the attention system VOCUS detects efficiently regions of interest based on both data-driven bottom-up saliency and task-dependent top-down cues.

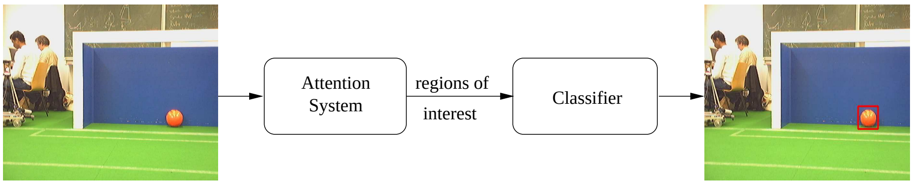


Fig. 1. Attentive Classification: the recognition system consists of an attention system providing object candidates and a classification system verifying the hypothesis. The combination yields a flexible and robust system.

The bottom-up saliency is based on feature contrasts, according to intensity, color, and orientation features, and on the uniqueness of each feature. For example, a single red ball on green grass is salient, whereas if there are many red balls, but only a single blue one, the blue ball has a higher saliency. In contrast, top-down cues depend on a predefined target. The features of a target are learned in a training phase and used during the test phase for assigning weights to the feature computations. Both bottom-up and top-down saliencies compete for global attention, resulting in a single region of interest. In a second step, this region is investigated by a fast classifier created by the Adaboost learning technique [66]. The classifier is trained off-line on different objects using large positive sample sets. During testing, it classifies the objects online within the selected region of interest. So the potentially compute-intensive classification process is applied only to the relatively small data section which VOCUS has deemed salient earlier in an efficient process working on the full data.

We investigate the performance of the combined system in different robotics scenarios: In a first experimental setting, the system operates in exploration mode. In this case, no target is given and VOCUS determines regions of interest purely based on bottom-up saliency. Intuition tells that the savings from the two-step process should increase with the computational demands of the classification, since the compute-intensive classification is applied only to salient data, rather than to the full set. We show that this is indeed the case, using different numbers of trained object classes. We achieve a time saving of 40% for a single object class and the time saving increased linearly with the number of classes. To illustrate the possibility to apply the system to data from different sensors and exploiting their advantages, these experiments are performed on data from a 3D laser scanner.

In a second experimental setting, VOCUS operates in top-down mode and searches for a predefined target. The classifier then verifies or falsifies the position hypothesis that VOCUS proposed. We investigate the factors leading to an increase in detection accuracy. It turned out that for simple objects as balls, which are often most difficult to classify correctly because they are hard to distinguish from their surroundings, the amount of false detections is significantly reduced by the use of attentive classification.

While other groups have already investigated the combination of visual attention and object recognition (we present and discuss the relevant state of the art in section 2), these approaches compute only bottom-up saliency and are therefore only able to recognize the most salient objects. New in our work is that the system can both operate in bottom-up, exploration mode, and, if pre-knowledge is available, perform visual search for a target with the top-down part of VOCUS. Thus, the pre-knowledge can be used to highlight target-relevant features and to inhibit target-irrelevant features. This leads to a considerably better recognition rate, less false detections and a faster system. Additionally, we have investigated the application of visual attention to other sensor data (laser scanner data). This paper is partly based upon the experience and results of previously published work [15, 33, 11]. Here, we combine, update, and extend the former results leading to a complete, robust, and flexible attentive classification system.

In the following, we first give an overview on the state of the art concerning visual attention systems and their combination with object recognition (section 2). Then, we introduce the attention system VOCUS in section 3 and the object recognition method in section 4. Next, the attentive classification method is presented in section 5, and in section 6, we show several experimental results. Section 7 discusses the time savings which are achieved and, finally, section 8 concludes the paper.

2 Related Work

2.1 Visual attention

Concerning visual attention, most research has so far been done in the field of *bottom-up* processing (in psychology [58, 69], neuro-biology [6, 46] and computer vision [3, 22, 24, 54]). Koch & Ullman [24] described the first explicit computational architecture for bottom-up visual attention; it is strongly influenced by Treisman's *feature-integration theory* [58]. Many computational systems have been presented meanwhile [22, 3, 54, 61], and most concentrate on computing the features intensity, color, and orientation. But there are also some considering other features such as curvature [32], spatial resolution [19], optical flow [61, 64], corners [9, 20, 44], or motion [21, 29, 43, 62]. Several systems also compute higher-level features that use approved techniques of computer vision to extract useful image information. Examples for such features are entropy [20], ellipses [26], eccentricity [3], or symmetry [3, 20, 26, 25].

Another important aspect in human perception that is rarely considered is depth. In the literature it is not clear whether depth is simply a visual feature. However, it clearly has some unusual properties distinguishing it from other features: if one of the dimensions in a conjunctive search is depth, a second feature can be searched in parallel [35], a property that does not exist for the other features. Computing depth for an attention system is usually solved with stereo vision [2, 29, 4]. The data obtained from stereo vision has the drawback that it is usually not very accurate and contains large regions without depth information. Another approach is to use special 3D sensors, such as 3D cameras [42]. The application of attentional mechanisms to data from several sensors and their fusion has to our knowledge not been done before.

There is also strong neuro-biological and psychophysical evidence for top-down influences modifying early visual processing in the brain due to pre-knowledge, motivations, and goals [72, 6, 71]. However, only a few computational attention models integrate top-down information. The earliest approach is the *guided search* model by Wolfe [69], a result of his psychological investigations of human visual search. Some systems inhibit target-irrelevant regions [61, 5], others prefer to excite target-relevant regions [18]. Newer findings suggest that inhibition and excitation both play an important rule [37]; this is realized in [38] and [12]. Navalpakkam and Itti [36] present an interesting approach in which not only knowledge about a target but also about distractors influence the search. Vincent et al. [65] learn the optimal top-down weighting with multiple linear regression. The here used system VOCUS combines bottom-up saliency detection with top-down visual search based on inhibition and excitation and is one of the few systems which is real-time capable [14]. Additionally, it is to our knowledge the only system extensively evaluated on real-world data, including evaluations of image transformations, viewpoint changes and variations in illumination [11].

2.2 Combining attention and classification

The combination of visual attention with object recognition has often been suggested and has recently gained interest in the field of computer vision. Several groups have investigated this using different recognition modules, but the combination has been restricted to bottom-up attention systems. One example of a combination of a bottom-up attentional front-end with a classifying object recognizer is presented in [30]. The recognizer is the biologically motivated system HMAX [51]. Since this system focuses on simulating processes in the human cortex, it is restricted in its capabilities and can only recognize simple artificial objects like circles or rectangles. In [31], the authors replace the HMAX system by a more powerful support vector machine algorithm to detect pedestrians in natural images. Walther and colleagues combine an attention system with an object recognizer based on SIFT features [28] and show that the recognition results are improved by the attentional front-end [68, 67]. Their approach is successful since they use highly textured and fixed views of objects which are pasted into real-world scenes. In [52] an attention system is combined with neural networks and an observable Markov model to do handwritten digit recognition and face recognition. In [41], an attention-based traffic sign recognition system is presented.

A different view on attention for object recognition is presented in [17, 45]: an information-theoretic saliency measure is used to determine discriminative regions of interest in objects. The saliency measure is computed by the conditional entropy of estimated posteriors of the local appearance patterns. In essence, regions of an object are considered to be salient if they discriminate the object well from other objects collected in an object data base. A similar approach is presented in [49].

Recently, some interesting work has been done to use visual attention for scene classification: Siagian and Itti [53] compute the *gist* of a scene — its overall appearance — from the features of the saliency computations. The resulting gist vector is used to differentiate outdoor scenes from various sites on a college campus.

In all of the presented systems which combine attention and classification, the saliency computations are only based on bottom-up information. Therefore, non-salient objects are not detected. This is a crucial drawback since in most application it is important to not only recognize objects which pop out of a scene but to find an object among others which are equally or more salient. Here, we investigate instead both approaches, object recognition in combination with bottom-up as well as with top-down attention. Which part shall be chosen depends on the available pre-knowledge and the task of the system — e.g. exploration or object search. The ability to deal with pre-knowledge enhances the system performance considerably.

3 The Visual Attention System VOCUS

The computational attention system VOCUS (Visual Object detection with a CompUtational attention System) consists of a bottom-up part, computing data-driven saliency, and a top-down part, enabling goal-directed search for a target. Global saliency is determined from both cues (cf. Fig. 2).

3.1 Bottom-up saliency

VOCUS' bottom-up part detects salient image regions by using image contrasts and uniqueness of a feature, e.g., a red ball on green grass. It was inspired by Itti et al. [22] but differs in several aspects, resulting in an improved system performance (see [11]). The feature computations are performed on 3 different scales using image pyramids. The feature intensity is computed by *center-surround mechanisms* extracting intensity differences between image regions and their surroundings, similar to cells in the human visual system [46]. In contrast to [22], we compute on-off and off-on contrasts separately; after summing up the scales, this yields 2 intensity maps. This is an important difference to existing models since it enables the detection of intensity pop-outs and top-down guidance to dark or bright regions [11, 13]. Similarly, 4 orientation maps $(0^\circ, 45^\circ, 90^\circ, 135^\circ)$ are computed by Gabor filters and 4 color maps (green, blue, red, yellow) by first converting the RGB image into the Lab color space, secondly determining the distance of the pixel color to the prototype color (the red map shows high activation for red regions and low activation for green regions) and thirdly,

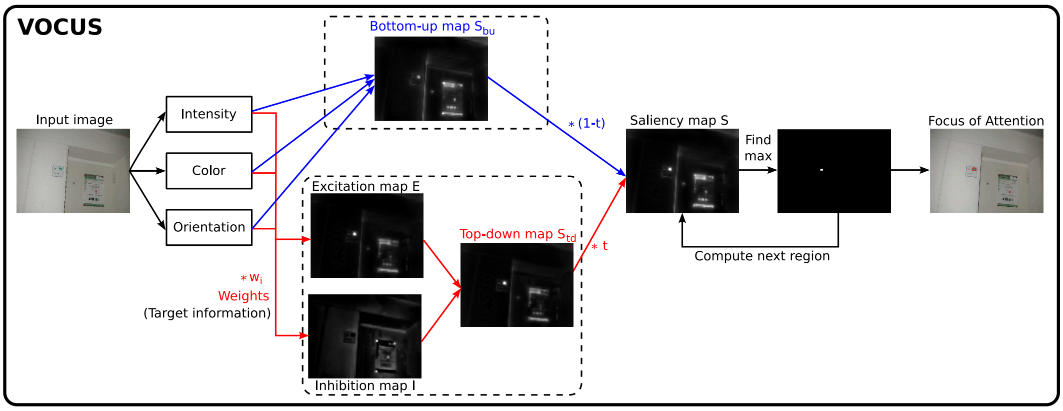


Fig. 2. The attention system VOCUS. On the input image (left), the features intensity, color, and orientation are computed and fused to a *Bottom-up Saliency Map* S_{bu}. If target information is available, this is used to strengthen important and inhibit unimportant features, resulting in an *Excitation* and an *Inhibition Map*. These are combined to a *Top-down Saliency Map* S_{td} which competes for global saliency with the Bottom-up Map. In the final *Saliency Map* S, the most salient region is selected to form the *Focus of Attention (FOA)*.

applying center-surround mechanisms (details in [11]). Each feature map X is weighted with the *uniqueness weight*

$$W(X) = X/\sqrt{m} \tag{1}$$

where m is the number of local maxima that exceed a threshold t. This weighting is essential since it emphasizes important maps with few peaks, enabling the detection of *pop-outs* (outliers). After weighting, the maps are summed up to the *bottom-up saliency map* S_{bu}.

3.2 Top-down saliency

To perform visual search for a *region of interest (ROI)* in an example image, VOCUS first computes target-specific weights (learning mode) and then uses these weights to adjust the saliency computations according to the target (search mode). We call this target-specific saliency *top-down saliency*.

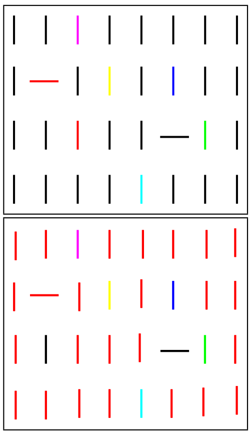
In **learning mode**, VOCUS is provided with a training image and coordinates of a ROI that includes the target. The region might be previously determined by VOCUS or be the output of a classifier specifying the target or determined manually by the user. Then, the system computes the bottom-up saliency map and the *most salient region (MSR)* inside the ROI. So, VOCUS is able to decide autonomously what is important in a ROI, concentrating on parts that are most salient and disregarding the background or less salient parts. Note that this makes VOCUS also robust to small changes of the ROI coordinates.

Next, weights are determined for the feature and conspicuity maps, indicating how important a feature is for the target. The weight w_i for map X_i is the ratio of the mean saliency in the target region $m_{(MSR)}$ and in the background $m_{(image-MSR)}$:

$$w_i = m_{(MSR)}/m_{(image-MSR)} \qquad \text{where} \qquad i \in \{1, ..., 13\}. \tag{2}$$

Two examples of weight vectors are shown in Figure 3. This computation does not only consider which features are the strongest in the target region, it also regards which features separate the region best from the rest of the image (cf. Fig. 3).

The learning of weights from one single training image yields good results if the target object occurs in all test images in a similar way, i.e., on a similar background and in a similar



Feature	weights (top)	weights (bottom)
intensity on/off	0.01	0.01
intensity off/on	9.13	13.17
orientation 0 °	20.64	**29.84**
orientation 45 °	1.65	1.96
orientation 90 °	0.31	0.31
orientation 135 °	1.65	1.96
color green	0.00	0.00
color blue	0.00	0.01
color red	**47.60**	10.29
color yellow	**36.25**	9.43
conspicuity I	4.83	6.12
conspicuity O	7.90	**11.31**
conspicuity C	**17.06**	2.44

Fig. 3. Effect of background information on the weight values. Left: the same target (red horizontal bar, 2nd in 2nd row) in different environments: all vertical bars are black (top) resp. red (bottom). Right: the learned weights (most important values for distinguishing the examples printed in bold face). In the upper image, the red color is the most important feature. In the lower image, surrounded by red distractors, red is no longer the prime feature to detect the bar but orientation is.

orientation. These conditions occur if the objects are fixed elements of the environment, e.g., fire extinguishers. Nevertheless, for movable objects, it is necessary to learn from several training images which features are stable and which are not. This is done by determining the average weights from n training images using the geometric mean of the weights, i.e.,

$$w_{i,(1..n)} = \sqrt[n]{\prod_{j=1}^{n} w_{i,j}}. \tag{3}$$

Instead of using all images from the training set, we choose the most suitable ones: first, the weights from one training image are applied to the training set, next, the image with the worst detection result is taken and the average weights from both images are computed. This procedure is repeated iteratively as long as the performance increases (details in [11, 13]).

In **search mode**, we determine a top-down saliency map that is integrated with the bottom-up map to yield global saliency. The top-down map itself is composed of an excitation and an inhibition map. The excitation map E is the weighted sum of all feature and conspicuity maps X_i that are important for the learned region, i.e., of all maps with $w_i > 1$. The inhibition map I shows the features more present in the background than in the target region, i.e., those with $w_i < 1$:

$$\begin{array}{ll} E = \sum_i (w_i * X_i) & \forall i : w_i > 1, \\ I \ = \sum_i ((1/w_i) * X_i) & \forall i : w_i < 1. \end{array} \tag{4}$$

The top-down saliency map S_{td} is obtained by

$$S_{td} = E - I \tag{5}$$

and a clipping of negative values. To make S_{td} comparable to S_{bu}, it is normalized to the same range.

3.3 Global Saliency

The global saliency map S is the weighted sum of S_{bu} and S_{td}. The contribution of each map is adjusted by the top-down factor $t \in [0..1]$:

$$S = (1-t) * S_{bu} + t * S_{td}. \tag{6}$$

For $t = 1$, VOCUS considers only target-relevant features (pure top-down). For a lower t, salient bottom-up cues may divert the focus of attention, an important mechanism in human attention: a person suddenly entering a room immediately catches our attention. Also, colored cues divert the search for non-colored objects as shown in [57]. Determining appropriate values for t depends on the system state, the environment and the current task; this is beyond the scope of this article and will be considered for future work.

After the computation of the global saliency map S, the most salient region is determined by *seeded region growing* [1] starting with the maximum of S. Finally, the focus of attention (FOA) is directed to this region. To compute the next FOA, this region is inhibited and the selection process is repeated.

The computations to determine the center-surround differences are done using integral images [14]. With this method, features can be computed at any scale in constant time, only 8 table look-ups are needed. In contrast to previous versions of the system, as well as to other attention systems, this makes the system fast and real-time capable: it needs only 60 ms to compute a saliency map for a 400 × 300 pixel image (2,8GHz).

4 Object Recognition

General object recognition is far from being solved in computer vision [8]. To illustrate this point, it is necessary to compare current algorithms in computer vision to the human visual system, which is impressively good at recognizing objects. Humans can name many thousands of different objects, categorize them spontaneously into groups, add new objects into these groups, re-detect them in arbitrary orientations, from different viewpoints, under most difficult illumination conditions and even if they are partially occluded.

Humans are also able to recognize objects on different hierarchy levels, such as recognizing a poodle as a poodle but also as a dog, a mammal, an animal, and a creature. Which level is appropriate in a particular situation seems to be intuitively clear to us. Furthermore, we are able to generalize, recognizing different kinds of chairs, for example, whether they have one or four legs, have armrests or not, and regardless of whether they have been previously observed in their present form or not. Finally, we are able to form new object categories from a small number of examples.

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*, a method that searches 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 a classifier based on the Adaboost learning technique [66], since it is currently one of the most accurate and efficient classifiers available, showing high detection rates in short amounts of time. It will be introduced in the following section.

4.1 Classification with Adaboost

In this section, we briefly introduce the classifier of Viola and Jones which is based on the Adaboost learning technique and was originally built for face detection [66]. It is publically available from the OpenCV library[1]. 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.

Learning Features: The idea of Viola and Jones' 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.

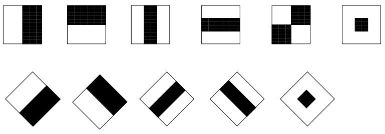


Fig. 4. Haar-like feature detection masks used by the Viola-Jones classifier for the detection of edge, line, and blob features. [66]

Fig. 4 shows the basic features the classifier considers. The 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. Haar-like features are also used in [27, 47, 60].

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* [66] or using *rotated integral images* [27]. After having created one 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 Adaboost Algorithm [10], 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 Trees) [27]. These binary trees enable the system to learn objects with different characteristics, e.g., objects from different viewpoints or with different patterns (see Fig. 5).

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 and the classifier returns a negative result. This enables 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 rate, multiplied in each stage, approaches zero.

[1] http://opencvlibrary.sourceforge.net/

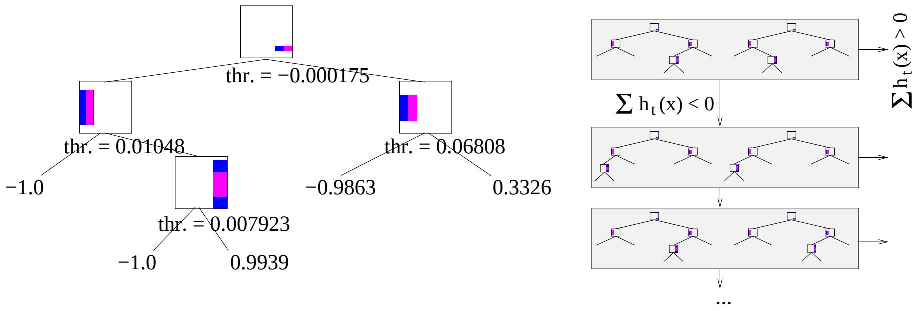


Fig. 5. Left: A Classification and Regression Tree (CART) with 4 splits. According to the specific filter applied to the image input section x, the output of the tree, $h_t(x)$ is calculated, depending on the threshold values. Right: A cascade of CARTs [34]. $h_t(x)$ is determined depending on the path through the tree.

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. This is done for the whole image, beginning with a search window of a specified small size. 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; details follow in the next section.

5 Attentive Classification

Attentive classification is the combination of a fast attention system with a powerful classifier. Whereas the attention system is applied to the whole scene, the classifier is restricted to a region of interest which is provided by the attention system (cf. Fig. 1). This is an effective way to improve the performance of visual systems: the attention system points to a region of interest but is not able to determine which objects are in this region (bottom-up) or whether a searched target is actually present (top-down). On the other hand, a general classifier needs much time if applied to the whole image. Restricting the classification to the region of interest is more effective and also improves the quality of recognition in certain cases by eliminating false positives (cf. section 6). The more complex and general a recognition system is, the more useful is its 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 where no special target is given. The system either favors salient objects or recognizes 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 searches 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 strategy that is usually used 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 scarce, 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 more effective 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 if this is inevitable due to a lack of time, the missing of non-salient objects is the lesser evil.

Another application scenario, in which recognizing only salient objects is even preferred to the recognition of all objects, comprises the following: if a system is very complex and knows about a wide variety of object classes, it might be useful not to consider everything in the environment. This is also true for humans: not every object in the environment is noticed but mainly salient and task-relevant objects. The socket in the corner will probably not be noticed if you look around in a newly entered room unless you need a power supply. Here, it is sensible to have an attention system narrowing down the choice of regions for recognition.

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. In our experiments it turns out that choosing a region which is four times as large as the expected size of the target object yields good results. The search windows were placed so that the middle of the search window lies inside the region selected by the attention system.

6 Experiments and Results

An attentive classification system is able to operate in two modes: in *bottom-up* and in *top-down* mode. Which mode is chosen depends on the available pre-knowledge and the goals. If there are no goals and the task is the exploration of the environment, the bottom-up mode is chosen. If a goal is known, this information is used in the top-down mode.

Additionally to these distinctions, we distinguish also between the types of input data for the system: we apply the system to camera data as well as to data from a 3D laser scanner. Although VOCUS works equally well, independently of which sensor the input image comes from, the sensors highlight different aspects of the world. For example, in depth images from a laser scanner, objects "pop out" when they have a certain distance from their background. This feature can be very useful for a robotic system. In [16], we have presented the complementary effects of different sensor data for the attentional system.

These distinctions lead us to four different test scenarios: (1) bottom-up attentive classification on camera data, (2) bottom-up attentive classification on laser data, (3) top-down attentive classification on camera data, and (4) top-down attentive classification on laser data:

	bottom-up	top-down
Camera images		sec. 6.2
Laser images	sec. 6.1	

To give a representative overview, we decided to pick one of the scenarios for each of the columns of the table for presentation. Since scenario 1, the application of bottom-up attentive classification to camera data, was already investigated in previous work (cf. section

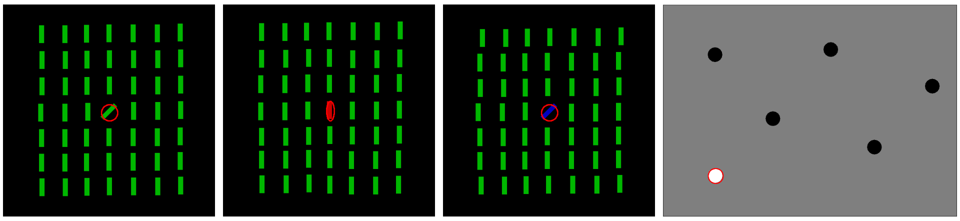


Fig. 6. Foci of attention (red ellipses) on some psychological pop-out images (images (1-3) provided by B. Schönwälder, psychological institute of LMU Munich)

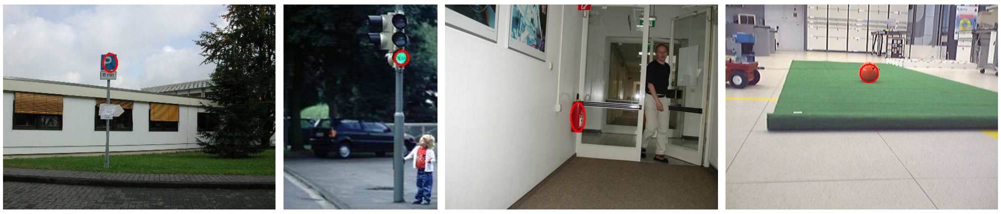


Fig. 7. Foci of attention (red ellipses) on some natural images

2), we chose to demonstrate scenario (2) instead, the application of bottom-up attentive classification to laser data. For illustrating the top-down attentive classification, we chose camera images instead. Experiments for these two scenarios are presented in the next two sections.

6.1 Experiment 1: Bottom-Up Attentive Classification

In a first step, we use the bottom-up mode of VOCUS for the attentive classification, i.e., VOCUS finds regions of interest and the classifier recognizes the content of the regions. This approach allows the recognition of the most salient objects in a scene, which is useful in complex systems that can recognize a large variety of objects but do not have the time to analyze all objects in a scene. The attention system provides the priority determining which region to analyze first.

The most common question at this point is whether the regions of interest are actually useful, i.e., whether they are at least better than random regions. In the case of biologically motivated attention systems, evolution has performed this evaluation for us: the human visual attention system picks those parts of the sensory input which are most useful for surviving. It is reasonable to assume that for a technical system the same regions are of interest, since it acts in the same world as humans and performs tasks provided by humans. It remains to show that the salient regions of the attention system indeed correspond to human eye movements. This is usually tested on artificial images like the ones in Fig. 6, since on such images human viewing behavior is definite and well investigated [69, 70, 59]. The foci of attention detected by VOCUS are the same as humans' first fixations (Fig. 6). There are also several psychological studies which investigated the coherence of computational saliency and human eye movements on natural scenes [48, 56, 7, 50]. They show that there is a correspondence between computationally salient regions and human fixation points and that the hot spots in the saliency map predict the locations of objects significantly above chance. Fig. 7 illustrates the behavior of VOCUS on natural scenes. More experimental data of VOCUS as well as a comparison to other attention systems can be found in [11].

The detection of salient regions is usually not sufficient, it is also necessary to know what the object of interest actually is. That is when the classifier comes into play. As motivated before, we show the combined attentive classification approach on images from a 3D laser scanner.

Our experimental setup was as follows: The laser images are rendered from range and reflection data from a 3D laser scanner. The scanner works according to the time-of-flight principle: it sends out a laser beam and measures the returning reflected light. To achieve 3D data, it scans subsequently horizontal lines of the environment. The scanner returns two kinds of data: the time the laser pulse needs to come back gives the distance (range data) and the intensity of the reflected light provides information about the reflection properties (reflection data). The reflection data are directly transformed into images: the range data are rendered into images by interpreting small depth values as bright intensity values and large depth values as dark ones. Both, VOCUS and the classifier, are applied to both laser modes and the results are combined to a single image (details in [16, 15, 11]). This helps to exploit different object properties (VOCUS) and to eliminate false detections (classifier).

First, we applied VOCUS in bottom-up mode to the laser images. It finds the most salient regions, without any further information about current objects of interest. Second, the classifier was applied to the salient regions. In an ideal setting, the classifier would be able to recognize a large amount of objects, in the order of at least several dozens. In practice, this is currently almost impossible, because the training of objects with current classifiers is usually very time-consuming: training the Viola-Jones classifier for one object, including taking training pictures, marking the target objects and learning the classifier, usually takes several days. To illustrate the approach, we trained the classifier on two typical objects of our office environment: chairs and the robot Kurt3D.

Since for laser scanner data different images can be rendered automatically from a single scan [40], the amount of required scans is lower than the amount of required images. We rendered 200 training images with chairs from 46 scans and 1083 training images with the robot from 200 scans, the images had a size of 300 × 300 pixels. Additionally, we provided 738 negative example images to the classifier from which a multiple of sub-images is created automatically. The performance of the classifier on these images in terms of detection and false detection rates can be found in [15]. In Fig. 8 we present examples of focused and classified objects in laser data. Note that if there is no focus on an object, it is not detected. This conforms to our goal to detect only salient objects in the order of decreasing saliency. The time saving achieved by attentive classification compared with exhaustive classification is already 40% for a single object class and grows with the number of object classes. This will be discussed in detail in section 7.

6.2 Experiment 2: Top-down Attentive Classification

If the system is searching for a target instead of exploring the environment, it is clear which classifier needs to be applied. Therefore, in this approach, the attention system provides a hypothesis for the target location, which is then verified or falsified by the classification system. In this setting, the time saving is not as large as in the bottom-up approach, since the classifier has to search the image usually only for one known object. Details about the time saving in top-down mode are described in section 7. However, we found that for certain objects, the attentive classification increases the recognition accuracy considerably.

We investigated the quality of search performance for the example of detecting balls for the robot soccer scenario ROBOCUP [63]. The most common techniques for ball detection in this context rely on color information. In the last few years, fast color segmentation algorithms have been developed to detect and track balls in this scenario [23, 55]. The community has agreed that in the near future, visual cues like color coding will be removed to come to a more realistic setup with robots playing with a “normal” soccer ball [60]. A first approach

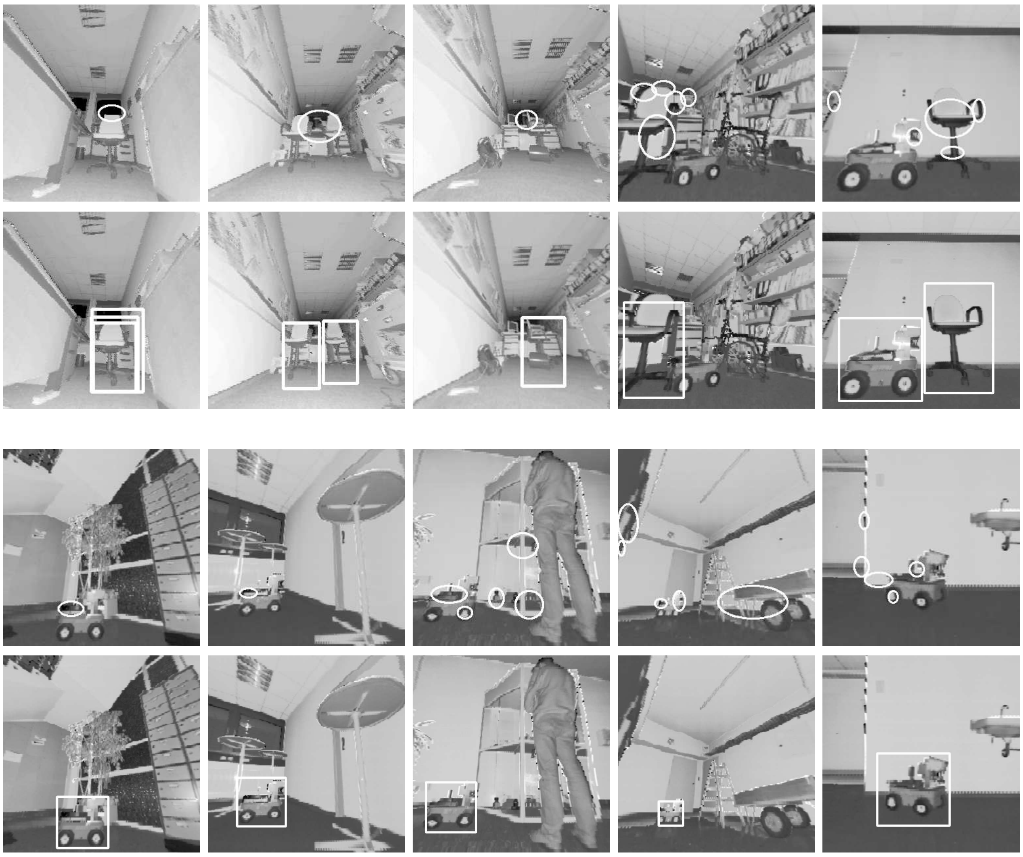


Fig. 8. Experiment 1: Attentive Classification. VOCUS was used in bottom-up mode and the Viola-Jones Classifier was trained for chairs and a robot in laser data. 1st/3rd row: the first resp. the first 5 foci of attention computed on depth and reflection data. 2nd/4th row: classified objects in the focus regions [15].

was presented by Treptow and Zell who learned balls with Adaboost conglomerations of Haar-like classifiers and arrange them in a cascade to recognize balls without color information [60]. This approach doesn't use color information at all. However, although color-coding can not be used to train the classifier, it is an important cue for detection as soon as the current type of ball is known. The same is true for other recognition tasks. Imagine the recognition of a cup. A general classifier should be color-independent, but when searching for a special cup, knowledge about the color is helpful to speed up the detection process.

Instead of training a classifier for all possible occurances of an object, we chose the more convenient way to train a color-independent classifier and combine it with the easily adaptable attention system. This 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 in general, by considering balls of different sizes, colors, and surface patterns. In the adaptation phase, VOCUS is quickly adapted to the actual ball that will be used in the game by learning ball-dependent features from a few training images. During the detection phase, the attentive classification takes place: first, VOCUS computes regions of interest; second, the classifier is applied to these regions, verifying the object hypotheses.

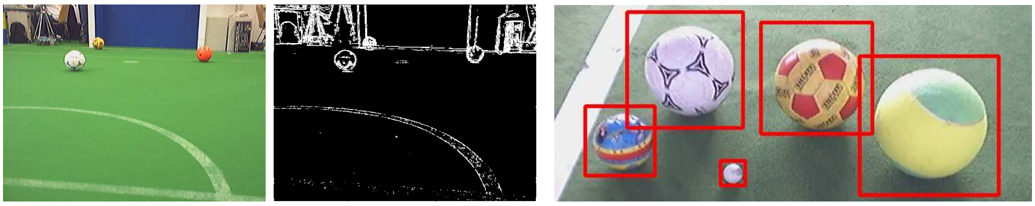


Fig. 9. Experiment 2: An image of a ROBOCUP scene showing the three kinds of balls that were used for training the classifier (left) and the corresponding edge image generated with a Sobel filter on which the training was actually performed (middle). Right: Five different kinds of balls are detected by the classifier.

Classifier (Training phase): According to Mitri et al. [34], we trained and tested the classifier for balls on edge images instead of the original image data. They showed that the classifier, when trained on different balls in the original image data, performs badly because the object is too simple and contains few features. In various experiments they explored how performance was significantly improved if edge filters were applied before training. Thus, a Sobel filter was used to obtain edge images as the one in Fig. 9 (middle) which was then fed 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 [34], this yielded much better edge images than the application of the filter to a gray-scale image.

The ball detection cascade was learned with 1000 positive example images (640×480 pixels) showing complex scenes with three soccer balls of different colors and patterns. The three balls used for training are shown in Fig. 9, left. To enable the detection of different kinds of balls, the training was done with CARTs. Fig. 9 (right) 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, resulting in 180 test images. Table 1 shows the detection and false detection rates for each kind of ball. The detection rate of the classifier is adjustable with the number of stages used in the cascade, where a lower number of stages increases the number of detections, but also the amount of false detections; with more stages, the false detection rate diminishes but there are also fewer 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 mainly learns the round shape of the balls and so it is difficult to differentiate between soccer balls and other round image regions. At least 12 stages are needed to decrease the number of false detections to 80, which is unacceptably high. Using such a large number of stages, though, 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.

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 described in [11, 13] 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 turns out that two training images were sufficient to yield a local optimum in performance.

In Table 2, 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 quite successful. The best results are obtained for 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 detected with the 1.7th focus on average. The reason we do not use only the attention system is that this approach 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 likely to look like a ball.

Attentive Classification (Detection phase): In the combined approach, first the balls are searched with the top-down modulated attention system, and then the first five FOA regions are investigated by the classifier. Therefore, the output is the intersection of both result sets.

The results of the attentive classification are shown in Table 3. It shows that the false detections are significantly reduced in the combined approach versus pure classification to 21 from 160 while the detections remained 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. Some of the results are depicted in Fig. 10.

Taking a closer look at the results of the different kinds of balls, it is noticeable how the performance differs 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 positives remain: here, the foci pointed to regions which were also misclassified by the classifier. Interestingly, this is different for the white ball: although 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 are disjoint, resulting in a false detection rate of zero.

7 Time Performance

For most applications in robotics, real-time capability is essential. 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 specific scene. The more complex and time-consuming the classifier, the higher the performance gain achieved with the approach of attentive classification.

The most current version of VOCUS is quite fast (60 ms to compute a saliency map for a 400 × 300 pixel image on a 2.8 GHz PC), since the feature computations are based on integral images (cf. sec. 3). This makes the system considerably faster than older versions of the system (2.7 sec. [11, 15]) and to most other attention systems (e.g. 30 sec. [31] or 30 sec. [3] on little older machines).

Although the classifier itself is fast, the combination of attention and classification pays off: applying the classifier to the whole image requires 200 ms, applying it only to a region of interest needs on average 60 ms, which constitutes 30% of the time of the exhaustive classifier. In total, attentive classification needs 120 ms, 80 ms less then the classifier-only approach. This corresponds to a time saving of 40% if the classifier knows only one single object class. But usually, it has a database with many objects. Then, the combination pays off even more, since the computation time increases with the number of object classes m: $60 + m * 60$ ms versus $m * 200$ ms for the conventional classifier-only approach. Already for 10 object classes, the conventional approach needs 2 seconds, whereas the attentive classification requires only 660 ms.

The top-down attention system is used if target information is available. This means it is clear which object is searched for and which classifier should be applied. If several objects have to be searched for in the same scene, the top-down attention system has to

Table 1. Experiment 2: Classifier: The Viola-Jones classifier was trained for balls. We examined different numbers of stages of the classification cascade. The cascade with 10 stages (bold face) was used for the attentive classification experiments.

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

Table 2. Experiment 2: Attention system: Detection results of VOCUS when searching in top-down mode for different balls. In each image, the first 5 focused regions are considered. The average hit number is the average rank of the foci that hit the target.

	# test im.	Detections	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	1.27

Table 3. Experiment 2: Exhaustive classification vs. attentive classification. We used the classification cascade with 10 stages. Column 2 (attention) shows the average hit number (the average rank of the foci that hit the target). It shows that the false detections are significantly reduced in the attentive approach while the detection rate remains nearly stable.

	# test im.	Attention	Exhaustive classification		Attentive classification	
		Av. hit nb.	Detections	False pos.	Detections	False pos.
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	1.27	146	160	142	21

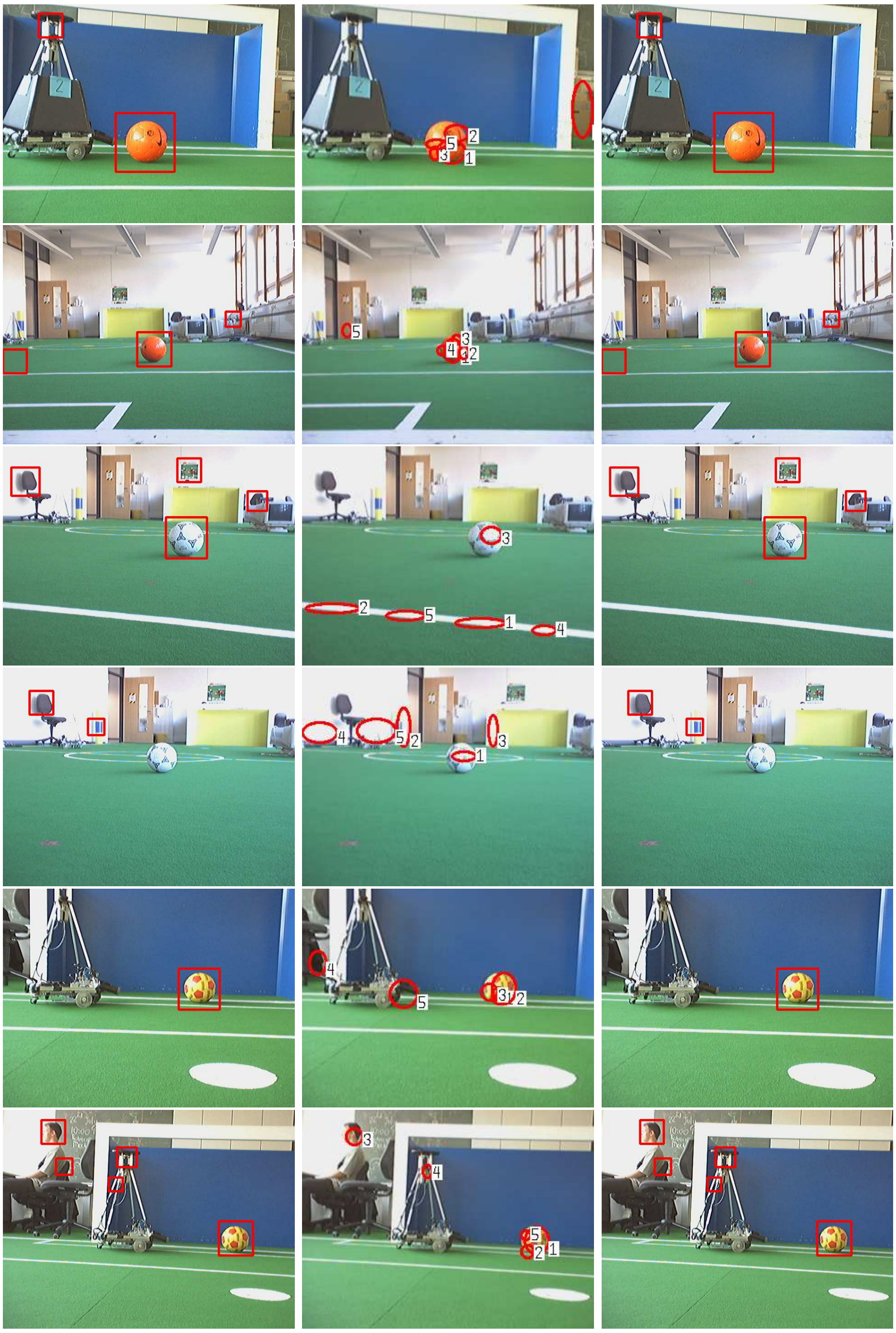


Fig. 10. Experiment 2: Detecting different kinds of balls. Left: classifier only. Middle: first 5 FOAs of VOCUS in top-down mode. Right: attentive classification; most false detections are eliminated.

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 maps must 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.

8 Conclusion

In this paper, we presented *Attentive Classification*, a biologically-motivated two-step approach for object recognition. The attention system VOCUS provides regions of interest, based on data-driven, bottom-up, as well as on target-dependent, top-down cues. These regions of interest were analyzed with a fast classifier. The method represents an important step towards effective general object recognition, since it constrains complex and time-consuming computations to restricted parts of the data.

Often, simple, low-textured objects cause problems in recognition, rather than complex ones. In general, the simpler the shape and texture of an object, the more difficult it is to distinguish it from other regions in a scene. We illustrate this for the example of detecting balls with the Viola-Jones classifier. We have shown on the example of ball detection how the combined approach of *Attentive Classification* enables a significant improvement in the detection results for simple objects, reducing the amount of false detections by 87%.

The approach of *Attentive Classification* is also superior to other detection and classification methods in terms of speed. In our experiments, we have shown an increase in speed of at least 40% to a standard classifier, which increases linearly with the number of objects to be detected.

The current system mimics the human visual attention system to a certain extent. However, it has to be noted that it can only be an approximation of human perception, since the human brain is highly complex and not yet well-understood. One difference to the current system is that in human perception, attention and recognition are more intertwined and share resources. In other words, the early extracted features give a first hint about the object, which is then verified successively. Approaches to simulate this behavior are presented in [19] and [38]. Although currently still preliminary in their recognition capabilities, these approaches might be a step into the right direction. In future work, we will investigate the interplay between attention and classification in more detail, coming up with a complete, more intertwined recognition system.

References

1. R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 16(6):641 – 647, 1994.
2. G. Backer and B. Mertsching. Integrating depth and motion into the attentional control of an active vision system. In G. Baratoff and H. Neumann, editors, *Dynamische Perzeption. Workshop der GI-Fachgruppe 1.0.4 Bildverstehen, Ulm, November 2000*, pages 69–74. Infix, 2000.
3. G. Backer, B. Mertsching, and M. Bollmann. Data- and model-driven gaze control for an active-vision system. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 23(12):1415–1429, 2001.
4. M. Björkman and J.-O. Eklundh. Vision in the real world: Finding, attending and recognizing objects. *Int'l Journal of Imaging Systems and Technology*, 16(2):189–208, 2007.
5. S.-B. Choi, S.-W. Ban, and M. Lee. Biologically motivated visual attention system using bottom-up saliency map and top-down inhibition. *Neural Information Processing-Letters and Reviews*, 2(1), 2004.

6. M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews*, 3(3):201–215, 2002.
7. L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3:3):1–15, Mar 2008.
8. D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach.* Pearson Prentice Hall, Berkeley, 2003.
9. F. Fraundorfer and H. Bischof. Utilizing saliency operators for image matching. In *Proc. of the Int'l Workshop on Attention and Performance in Computer Vision (WAPCV)*, pages 17–24, Graz, Austria, April 3 2003.
10. Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proc. of the 13th International Conference*, pages 148 – 156, 1996.
11. S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search.* PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany, July 2005. Published 2006 in Lecture Notes in Artificial Intelligence (LNAI), Vol. 3899, Springer Verlag.
12. S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *Proc. of the Annual meeting of the German Association for Pattern Recognition (DAGM)*, Lecture Notes in Computer Science (LNCS). Springer, Sept. 2005.
13. S. Frintrop, G. Backer, and E. Rome. Selecting what is important: Training visual attention. In *Proc. of the 28th German Conference on Artificial Intelligence (KI 2005)*, Lecture Notes in Computer Science (LNCS), pages 351–365, Conference: Koblenz, Germany, Sept. 2005. Springer.
14. S. Frintrop, M. Klodt, and E. Rome. A real-time visual attention system using integral images. In *Proc. of the 5th Int'l Conf. on Computer Vision Systems (ICVS)*, Bielefeld, Germany, March 2007.
15. S. Frintrop, A. Nüchter, H. Surmann, and J. Hertzberg. Saliency-based object recognition in 3D data. In *Proc. of the Int'l Conf. on Intelligent Robots and Systems (IROS)*, pages 2167 – 2172, Conference: Sendai, Japan, September 2004.
16. S. Frintrop, E. Rome, A. Nüchter, and H. Surmann. A bimodal laser-based attention system. *J. of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance in Computer Vision*, 100(1-2):124–151, Oct-Nov 2005.
17. G. Fritz, C. Seifert, and L. Paletta. Attentive object detection using an information theoretic saliency measure. In L. Paletta, J. K. Tsotsos, E. Rome, and G. W. Humphreys, editors, *Proc. of the 2nd Int'l Workshop on Attention and Performance in Computational Vision (WAPCV)*, pages 136–143, Conference: Prague, Czech Republic, May 2004.
18. F. H. Hamker. Modeling attention: From computational neuroscience to computer vision. In L. Paletta, J. K. Tsotsos, E. Rome, and G. W. Humphreys, editors, *Proc. of the 2nd international workshop on attention and performance in computational vision (WAPCV '04)*, pages 59–66, Conference: Prague, Czech Republic, May 2004.
19. F. H. Hamker. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance*, 100(1-2):64–106, 2005.
20. G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter. Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Machine Vision and Applications*, 16(1):64–73, 2004.
21. L. Itti. Real-time high-performance attention focusing in outdoors color video streams. In *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI)*, San Jose, CA, 2002.
22. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
23. T. B. J. Bruce and M. Veloso. Fast and inexpensive color image segmentation for interactive robots. In *Proceedings of the IEEE/RSL International Conference on Intelligent Robots and Systems*, volume 3, pages 2061 – 2066, 2000.
24. C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.
25. G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. In *Proc. of the British Machine Vision Conference (BMVC)*, Leeds, UK, September 1-4 2008.
26. K. Lee, H. Buxton, and J. Feng. Selective attention for cue-guided search using a spiking neural network. In *Proc. of the Int'l Workshop on Attention and Performance in Computer Vision (WAPCV)*, pages 55–62, Graz, Austria, April 3 2003.

27. R. Lienhart and J. Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. In *Proc. of the IEEE Conf. on Image Processing (ICIP '02)*, pages 155–162, New York, USA, Septmber 2002.
28. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. of Computer Vision (IJCV)*, 60(2):91–110, 2004.
29. A. Maki, P. Nordlund, and J.-O. Eklundh. Attentional scene segmentation: Integrating depth and motion. *Computer Vision and Image Understanding (CVIU)*, 78(3):351–373, 2000.
30. F. Miau and L. Itti. A neural model combining attentional orienting to object recognition: Preliminary explorations on the interplay between where and what. In *Proc. IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 789–792, Conference: Istanbul, Turkey, 2001.
31. F. Miau, C. Papageorgiou, and L. Itti. Neuromorphic algorithms for computer vision and attention. In *Proc. SPIE 46 Annual Int'l Symposium on Optical Science and Technology*, volume 4479, pages 12–23, Nov 2001.
32. R. Milanese. *Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation*. PhD thesis, University of Geneva, Switzerland, 1993.
33. S. Mitri, S. Frintrop, K. Pervölz, H. Surmann, and A. Nüchter. Robust object detection at regions of interest with an application in ball recognition. In *IEEE Proc. of the Int'l Conf. on Robotics and Automation (ICRA '05)*, pages 126–131, Conference: Barcelona, Spain, April 2005.
34. S. Mitri, K. Pervölz, A. Nüchter, and H. Surmann. Fast color-independent ball detection for mobile autonomous robots. In *Proc. of IEEE Mechatronics & Robotics (Mechrob '04)*, pages 900–905, Conference: Aachen, Germany, 2004.
35. K. Nakayama and G. H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320:264–265, 1986.
36. V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.
37. V. Navalpakkam, J. Rebesco, and L. Itti. Modeling the influence of knowledge of the target and distractors on visual search. *Journal of Vision*, 4(8):690, 2004.
38. V. Navalpakkam, J. Rebesco, and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, 2005.
39. U. Neisser. *Cognitive Psychology*. Appleton-Century-Crofts, New York, 1967.
40. A. Nüchter, K. Lingemann, J. Hertzberg, and H. Surmann. Accurate object localization in 3D laser range scans. In *Proceedings of the 12th International Conference on Advanced Robotics (ICAR '05)*, pages 665–672, 2005.
41. N. Ouerhani. *Visual Attention: From Bio-Inspired Modeling to Real-Time Implementation*. PhD thesis, Institut de Microtechnique Université de Neuchâtel, Switzerland, 2003.
42. N. Ouerhani and H. Hügli. Computing visual attention from scene depth. In *Proc. of Int'l Conf. on Pattern Recognition (ICPR 2000)*, volume 1, pages 375–378. IEEE Computer Society Press, September 2000.
43. N. Ouerhani and H. Hügli. A model of dynamic visual attention for object tracking in natural image sequences. In *International Conference on Artificial and Natural Neural Network (IWANN)*, volume 2686, pages 702–709. Springer Verlag, Lecture Notes in Computer Science (LNCS), 2003.
44. N. Ouerhani and H. Hügli. AttentiRobot: a visual attention-based landmark selection approach for mobile robot navigation. In L. Paletta, J. K. Tsotsos, E. Rome, and G. W. Humphreys, editors, *Proc. of the 2nd international workshop on attention and performance in computational vision (WAPCV '04)*, pages 83–89, Conference: Prague, Czech Republic, May 2004.
45. L. Paletta, G. Fritz, and C. Seifert. Q-learning of sequential attention for visual object recognition from informative local descriptors. In *Proc. 22nd International Conference on Machine Learning (ICML 2005)*, pages 649–656, 2005.
46. S. E. Palmer. *Vision Science, Photons to Phenomenology*. The MIT Press, Cambridge, MA, 1999.
47. C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. of the 6th International Conference on Computer Vision (ICCV '98)*, pages 555–562, Conference: Bombay, India, January 1998.
48. D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

49. L. Pessoa and S. Exel. Attentional strategies for object recognition. In J. Mira and J. Sachez-Andres, editors, *Proc. of the International Work-Conference on Artificial and Natural Neural Networks (IWANN '99)*, volume 1606 of *Lecture Notes in Computer Science (LNCS)*, pages 850–859, Alicante, Spain, 1999. Springer.
50. R. J. Peters and L. Itti. Applying computational tools to predict gaze direction in interactive visual environments. *ACM Trans. on Applied Perception*, 5(2):Article 8, 2008.
51. M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.
52. A. Salah, E. Alpaydin, and L. Akrun. A selective attention based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(3):420–425, 2002.
53. C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, 29(2):300–312, Feb. 2007.
54. Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77–123, 2003.
55. R. H. T. Bandlow, M. Klupsch and T. Schmitt. Fast image segmentation, object recognition and localization in a robocup scenario. In *3. RoboCup Workshop, IJCAI'99*, 1999.
56. B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45:643–659, 2005.
57. J. Theeuwes. Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review*, 11:65–70, 2004.
58. A. M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.
59. A. M. Treisman and S. Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95(1):15–48, 1988.
60. A. Treptow and A. Zell. Real-time object tracking for soccer-robots without color information. *Robotics and Autonomous Systems*, 48(1):41–48, August 2004.
61. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.
62. J. K. Tsotsos, Y. Liu, J. C. Martinez-Trujillo, M. Pomplun, E. Simine, and K. Zhou. Attenting to visual motion. *Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance*, 100(1-2):3–40, 2005.
63. Robot World Cup Soccer Games and Conferences. http://www.robocup.org, 02.
64. S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal. Overt visual attention for a humanoid robot. In *Proc. International Conference on Intelligence in Robotics and Autonomous Systems (IROS 2001)*, pages 2332–2337, Hawaii, 2001.
65. B. T. Vincent, T. Troscianko, and I. D. Gilchrist. Investigating a space-variant weighted salience account of visual selection. *Vision Research*, 47:1809–1820, 2007.
66. P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, May 2004.
67. D. Walther. *Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics.* PhD thesis, California Institute of Technology, Pasadena, CA, 2006.
68. D. Walther, U. Rutishauser, C. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding (CVIU)*, 100(1-2):41–63, 2005.
69. J. M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1(2):202–238, 1994.
70. J. M. Wolfe. What can 1,000,000 trials tell us about visual search? *Psychological Science*, 9(1):33–39, 1998.
71. J. M. Wolfe, T. Horowitz, N. Kenner, M. Hyle, and N. Vasan. How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research*, 44:1411–1426, 2004.
72. A. L. Yarbus. *Eye Movements and Vision.* Plenum Press (New York), 1967.