Visual Attention for Object Recognition in Spatial 3D Data

Simone Frintrop, Andreas Nüchter, Hartmut Surmann
Fraunhofer Institut für Autonome Intelligente Systeme
Schloss Birlinghoven
53754 Sankt Augustin, Germany
{*firstname.surname*}@ais.fraunhofer.de
http://www.ais.fraunhofer.de

Abstract

In this paper, we present a new recognition system for the fast detection and classification of objects in spatial 3D data. The system consists of two main components: A biologically motivated attention system and a fast classifier. Input is provided by a 3D laser scanner, mounted on an autonomous mobile robot, that acquires illumination independent range and reflectance data. These are rendered into images and fed into the attention system that detects regions of potential interest. The classifier is applied only to a region of interest, yielding a significantly faster classification that requires only 30% of the time of an exhaustive search. Furthermore, both the attention and the classification system benefit from the fusion of the bi-modal data, considering more object properties for the detection of regions of interest and a lower false detection rate in classification.

1 Introduction

Object recognition and classification belong to the hardest problems in computer vision and have been intensely researched [2]. Generally, for a given domain a large number of different object classes has to be considered. Although fast classifiers have been built recently [22], it is time consuming to apply many classifiers to an image. To preserve high quality of recognition despite of limited time, the input set has to be reduced. One approach is to confine classification to image regions of potential interest found by attentional mechanisms.

In human vision, attention helps identify relevant parts of a scene and focus processing on corresponding sensory input. Psychological work shows evidence that different features, like color, orientations, and motion, are determined in parallel, coding the saliency of different regions [20]. Many computational models of attention are inspired by these findings [7, 21, 1].

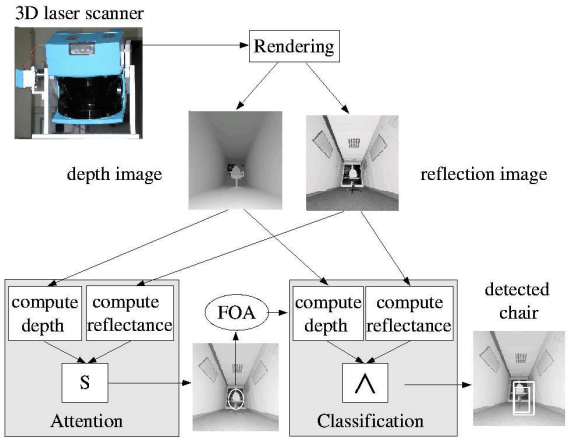

Figure 1. The recognition system. Two laser modes are provided by a 3D laser scanner, rendered into images and fed into an attention system which computes a focus of attention (FOA) from the data of both modes. The classifier searches for objects only in the FOA-region in depth and reflectance image and combines the results by an appropriate connection. The rectangles in the result image (right) depict a detected object.

In this paper, we present a new system for the fast detection and recognition of objects in spatial 3D data, using attentional mechanisms as a front end for object recognition (Fig. 1). Input is provided by a 3D laser scanner, mounted on an autonomous mobile robot. The scanner yields range as well as reflectance data in a single 3D scan pass [19]. Both data modalities are transformed into images and fed into a visual attention system based on one of the standard models of visual attention by Koch & Ullman [9]. In both laser images, the system detects regions which are salient according to intensity and orientations. Finally, the focus of attention is sequentially directed to the most salient regions.

A focus region is searched for objects by a cascade of classifiers built originally for face detection by Viola et al. [22]. Each classifier is composed of several simple classifiers containing edge, line or center surround features. The classifier is applied to both laser modes. We show how the classification is significantly sped up by concentrating on regions of interest with a time saving that increases proportionally with the number of object classes. The performance of the system is investigated on the example of finding chairs in an office environment. The future goal will be a flexible vision system that is able to find different objects in order of their saliency. The recognized objects will be registered in semantic 3D maps, automatically created by the mobile robot.

The fusion of two sensor modalities is performed in analogy to humans who use information from all senses. Different qualities of the modes enable to utilize their respective advantages, e.g., there is a high probability that discontinuities in range data correspond to object boundaries what facilitates the detection of objects: an object producing the same intensity like its background is difficult to detect in an intensity image, but easily in the range data. Additionally, misclassification of shadows, mirrored objects and wall paintings is avoided. On the other hand, a flat object, e.g., a sign on a wall, is likely not to be detected in the range but in the reflectance image. The respective qualities of the modes significantly improve the performance of both systems by considering more object properties for focus computation and a lower rate of false detections in classification. Furthermore, the scanner modalities are illumination independent, i.e. they are the same in sunshine as in complete darkness and no reflection artifacts confuse the recognition.

The presented architecture introduces an new approach for object recognition, however, parts of it have already been investigated. Pessoa and Exel combine attention and classification, but whereas we detect salient objects in complex scenes, they focus attention on discriminative parts of pre-segmented objects [17]. Miau, Papageorgiou and Itti detect pedestrians on attentionally focused image regions using a support vector machine algorithm [12]; however, their approach is computationally much more expensive and lacks real-time abilities. Object recognition in range data has been considered by Johnson and Hebert using an ICP algorithm for registration of 3D shapes [8], an approach extended in [18]; in contrast to our method, both use local, memory consuming surface signatures based on prior created mesh representations of the objects.

The paper is organized as follows: Section 2 describes the 3D laser scanner. In section 3 we introduce the attentional system and in 4 the object classification. Section 5 presents the experiments performed by the combination of attention and classification and discusses the results. Finally, section 6 concludes the paper.

2 The Multi-modal 3D Laser Scanner

The data acquisition in our experiments was performed with a 3D laser range finder (top of Fig. 1, [19]). It is built on the basis of a 2D range finder by extension with a mount and a small servomotor. The scanner works according to the time-of-flight principle: It sends out a laser beam and measures the returning reflected light. This yields two kinds of data: The distance of the scanned object (range data) and the intensity of the reflected light (reflectance data).

One horizontal slice is scanned by serially sending out laser beams using a rotating mirror. A 3D scan is performed by step-rotating the 2D scanner around a horizontal axis scanning one horizontal slice after the other. The area of $180°(\text{h}) \times 120°(\text{v})$ is scanned with different horizontal (181, 361, 721 pts) and vertical (210, 420 pts) resolutions. To visualize the 3D data, a viewer program based on OPENGL has been implemented. The program projects a 3D scene to the image plane, such that the data can be drawn and inspected from every perspective. Typical images have a size of 300×300 pixels. The depth information of the 3D data is visualized as a gray-scale image: small depth values are represented as bright intensities and large depth values as dark ones.

3 The Laser-Based Attention System

The Bimodal Laser-Based Attention System (BILAS) detects salient regions in laser data by simulating eye movements. Inspired by the psychological work of Treisman et al. [20], we determine conspicuities of different features in a bottom-up, data-driven manner. These conspicuities are fused into a saliency map and the focus of attention is directed to the brightest, most salient point in this map. Finally, the region surrounding this point is inhibited, allowing the computation of the next FOA.

The attention system is shown in Fig. 2 (cf. [5]); it is built on principles of one of the standard models of visual attention by Koch & Ullman [9] that is used by many computational attention systems [7, 1, 3, 10]. The implementation of the system is influenced by the Neuromorphic Vision Toolkit (NVT) by Itti et al. [7] that is publicly available and can be used for comparative experiments (cf. [5]). BILAS contains several major differences as compared to the NVT. In the following, we will describe our system in detail emphasizing the differences between both approaches.

The main difference to existing models is the capability of BILAS to process data of different sensor modalities simultaneously, an ability not available in any other attention system the authors know about. In humans, eye movements are not only influenced by vision but also by other senses and the fusion of different cues competing for attention is an essential part of human attention. The sensor modalities

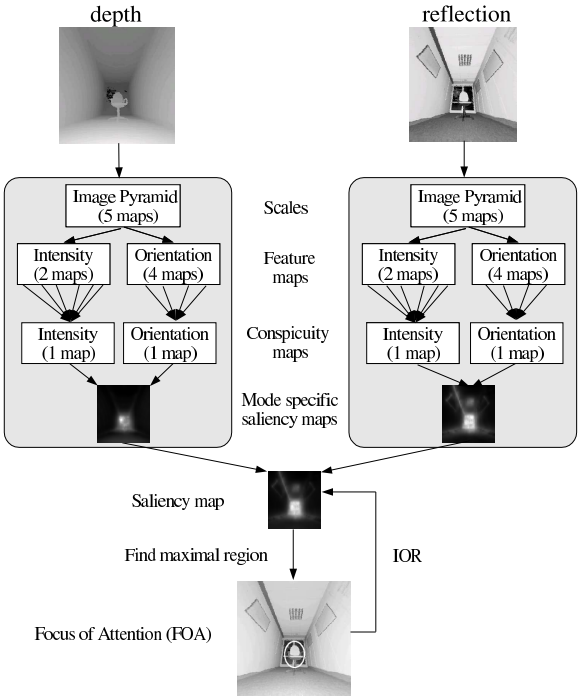

Figure 2. The Bimodal Laser-Based Attention System (BILAS). Depth and reflectance images rendered from the laser data are processed independently. Conspicuities according to intensity and orientations are determined and fused into a mode-specific saliency map. After combining both of these maps, a focus of attention (FOA) is directed to the most salient region.

used in this work are depth and reflectance values provided by the 3D laser scanner; in future work, we will use camera data additionally. The system computes saliencies for every mode in parallel and finally fuses them into a single saliency map.

Feature Computations

On images of both laser modalities, five different scales (0–4) are computed by Gaussian pyramids, which successively low-pass filter and subsample the input image; so scale $i+1$ has half the width and height of scale i. Feature computations on different scales enable the detection of salient regions with different sizes. In the NVT, 9 scales are used but the scales 5 to 8 are only used for implementation details (see below) so our approach yields the same performance with fewer scales. As features, we consider intensity and orientation.

The intensity feature maps are created by center-surround mechanisms which compute the intensity differences between image regions and their surroundings. These

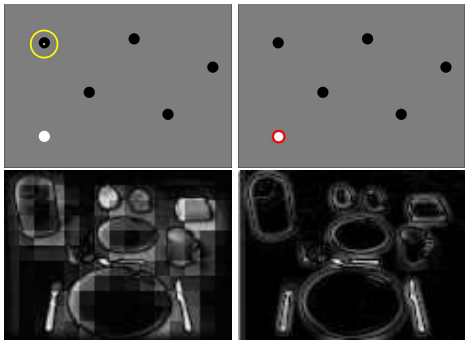

Figure 3. Differences of NVT (left) and BILAS (right). See text for details.

mechanisms simulate cells of the human visual system responding to intensity contrasts. The center c is given by a pixel in one of the scales $2-4$, the surround s is determined by computing the average of the surrounding pixels for two different sizes of surrounds with a radius of 3 resp. 7 pixels. According to the human system, we determine two kinds of center-surround differences: the on-center-off-surround difference $d_{\text{(on-off)}}(c,s) = c - s$, responding strongly to bright regions on a dark background, and the off-center-on-surround difference $d_{\text{(off-on)}}(c,s) = s - c$, responding strongly to dark regions on a bright background. This yields $2 \times 6 = 12$ intensity feature maps. The six maps for each center-surround variation are summed up by inter-scale addition, i.e. all maps are resized to scale 2 and then added up pixel by pixel. This yields 2 intensity maps.

The computations differ from these in the NVT, since we compute on-center-off-surround and off-center-on-surround differences separately. In the NVT, these computations are combined by taking the absolute value $|c - s|$. This approach is a faster approximation of the above solution but yields some problems. Firstly, a correct intensity pop-out is not warranted as is depicted in Fig. 3, top. The white object pops out in the computation with BILAS but not with the NVT. The reason is the amplification of maps with few peaks (see below). Secondly, if top-down influences are integrated into the system, a bias for dark-on-bright or bright-on-dark is not possible in the combined approach but in the separated one. This is for instance an important aspect if the robots searches for an open door, visible as a dark region in the depth image. The two approaches vary also in the computation of the differences themselves. In the NVT, the differences are determined by subtracting two scales at a time, e.g. $I_6 = scale(4) - scale(8)$. Our approach results in a slightly slower computation but is much more accurate (cf. fig. 3, bottom) and needs fewer scales.

The orientation maps are obtained by creating four oriented Gabor pyramids detecting bar-like features of orien-

tations $\{0°, 45°, 90°, 135°\}$. In contrast to Itti et al., we do not use the center-surround technique for computing the orientation maps. The Gabor Filters already provide maps, showing strong responses in regions of the preferred orientation and weak ones elsewhere, which is exactly the information needed. Finally, the maps $2-4$ of each pyramid are summed up by inter-scale addition. This yields four orientation feature maps of scale 2, one for each orientation.

Fusing Saliencies

All feature maps of one feature are combined into a conspicuity map, using inter-scale addition for the intensity maps and pure pixel addition for the orientation maps. The intensity and the orientation conspicuity maps are summed up to a mode-specific saliency map, one representing depth and one reflection mode. These are finally summed up to the single saliency map S.

If the summation of maps was done in a straightforward manner, all maps would have the same influence. That means, that if there are many maps, the influence of each map is very small and its values do not contribute much to the summed map. To prevent this effect, we have to determine the most important maps and give them a higher influence. To enable pop-out effects, i.e. immediate detection of regions that differ in one feature, important maps are those that have few popping-out salient regions. These maps are determined by counting the number of local maxima in a map that exceed a certain threshold. To weight maps according to the number of peaks, each map is divided by the square root of the number of local maxima m: $w(\text{map}) = \text{map}/\sqrt{m}$.

The Focus of Attention

To determine the most salient location in S, the brightest point is located in a straightforward way instead of using a winner-take all network as proposed by Itti et al. While losing biological plausibility, the maximum is found even though with less computational resources. Starting from the most salient point, region growing finds recursively all neighbors with similar values within a certain threshold. The width and height of this region yield an elliptic FOA, considering size and shape of the salient region in contrast to the circular fixed-sized foci of most other systems.

Finally, an inhibition of return mechanism (IOR) is applied to the focused region by resetting the corresponding values in the saliency map, enabling the computation of the next FOA. Fig. 4 shows an example run of the system; to depict the output of the system, we present a trajectory for a single camera image instead of two laser images.

If two laser images are supplied as input, the attention system benefits from the depth as well as from the reflectance data, since these data modes complement each other: An object producing the same intensity like its background may not be detected in a gray-scale image, but in the range data. On the other hand, a flat object – e.g. a poster on a wall or a letter on a desk – is likely not to be detected in the depth but in the reflectance image (cf. [6]).

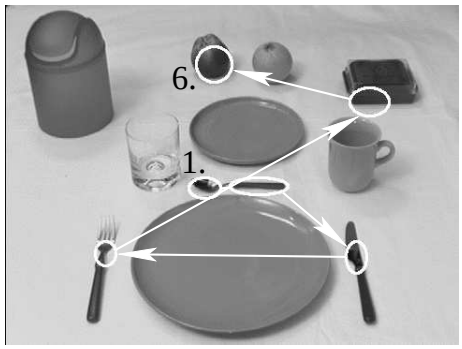

Figure 4. A trajectory of the first 6 foci of attention, generated by the attention system.

4 Object Classification

Recently, Viola and Jones have proposed a boosted cascade of simple classifiers for fast face detection [22]. Inspired by these ideas, we detect office chairs in 3D range and reflectance data using a cascade of classifiers composed of several simple classifiers.

Feature Detection using Integral Images

The six basic features used for classification are shown in Fig. 5 (left); they have the same structure as the Haar basis functions also considered in [15, 22]. The base resolution of the object detector is 20×40 pixels, thus the set of possible features in this area is very large (361760 features). A single feature is effectively computed on input images using integral images [22], also known as summed area tables [11]. An integral image I is an intermediate representation for the image and contains the sum of gray-scale pixel values of an image N:

$$I(x,y) = \sum_{x'=0}^{x} \sum_{y'=0}^{y} N(x',y').$$

The integral image is computed recursively by the formula: $I(x,y) = I(x,y-1) + I(x-1,y) + N(x,y) - I(x-1,y-1)$ with $I(-1,y) = I(x,-1) = 0$, requiring only one scan over the input data. This representation allows the computation of a feature value using several lookups and weighted subtractions (Fig. 5 right). To detect a feature, a threshold is required which is automatically determined

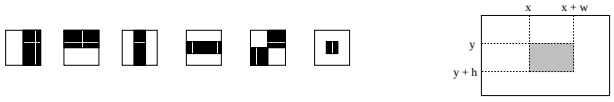

Figure 5. Left: Edge, line, diagonal and center surround features used for classification. Right: The computation of the sum of pixels in the shaded region is based on four integral image lookups: $F(x, y, h, w) = I(x+w, y+h) - I(x, y+h) - I(x+w, y) + I(x, y)$. Feature values are calculated by subtractions of these values weighted with the areas of the black and white parts.

during a fitting process, such that a minimum number of examples are misclassified.

Learning Classification Functions

The Gentle Ada Boost Algorithm is a variant of the powerful boosting learning technique [4]. It is used to select a set of simple features to achieve a given detection and error rate. The various Ada Boost algorithms differ in the update scheme of the weights. According to Lienhart et al., the Gentle Ada Boost Algorithm is the most successful learning procedure for face detection applications [11].

Learning is based on N weighted training examples $(x_i, y_i), i \in \{1 \dots N\}$, where x_i are the images and $y_i \in \{-1, 1\}$ the supervised classified output. At the beginning, the weights w_i are initialized with $w_i = 1/N$. Three steps are repeated to select simple features until a given detection rate d is reached: First, every simple feature is fit to the data. Hereby, the error e is evaluated with respect to the weights w_i. Second, the best feature classifier h_t is chosen for the classification function and the counter t is increased. Finally, the weights are updated with $w_i := w_i \cdot e^{-y_i h_t(x_i)}$ and renormalized.

The final output of the classifier is $\text{sign}(\sum_{t=1}^{T} h_t(x))$, with $h(x) = \alpha$, if $x \geq$ thr. and $h(x) = \beta$ otherwise. 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. Next, a cascade based on these classifiers is built.

The Cascade of Classifiers

The performance of one classifier is not suitable for object classification, since it produces a high hit rate, e.g., 0.999, and error rate, e.g., 0.5. Nevertheless, the hit rate is much higher than the error rate. To construct an overall good classifier, several classifiers are arranged in a cascade, i.e., a degenerated decision tree. In every stage of the cascade, a decision is made whether the image contains the object or not. This computation reduces both rates. Since the hit rate is close to one, their multiplication results also in a value close to one, while the multiplication of the smaller error rates approaches zero. Furthermore, the whole classification process speeds up, because the whole cascade is rarely needed. Fig. 6 shows an example cascade of classifiers for detecting chairs in depth images.

To learn an effective cascade, the classification function $h(x)$ is learned for every stage until the required hit rate is reached. The process continues with the next stage using only the currently misclassified negative examples. The number of features used in each classifier increases with additional stages (cf. Fig. 6).

The detection of an object is done by laying a search window over several parts of the input image, usually running over the whole image from the upper left to the lower right corner. To find objects on larger scales, the detector is enlarged by rescaling the features. This is effectively done by several look-ups in the integral image. In our approach, the search windows are only applied in the neighborhood of the region of interest detected by the attentional system.

5 Experiments and Results

We investigate the performance of the system on the example of finding chairs in an office environment. However, the future goal will be the construction of a flexible vision system that is able to search for and detect different object classes while the robot drives through its environment. If the robot moves, the time for the recognition is limited and it is not possible to search for many objects in a scene. A naive approach to restrict processing is to search the whole image for the first object class, then for the second class and so on. The problem is that if there is not enough time to check all object classes, some of the classes of the data base would never be checked. In our approach, we restrict processing to the salient regions in the image recognizing objects in order of their saliency.

To show the performance of the system, we claim three points: Firstly, the attention system detects regions of interest. Secondly, the classifier has good detection and false alarm rates on laser data. And finally, the combination of both systems yields a significant speed up and reliably detects objects at regions of interest. These three points will be investigated in the following.

Firstly, the performance of attention systems on camera data was evaluated by Parkhurst et al. [16] and Ouerhani et al. [14]. They demonstrate that attention systems based on the Koch-Ullman model [9] detect salient regions with a performance comparable to humans. We showed in [6] and [5] that attentional mechanisms work also reliably on laser data and that the two laser modes complement each other, enabling the consideration of more object qualities. Two examples of these results are depicted in Fig. 7.

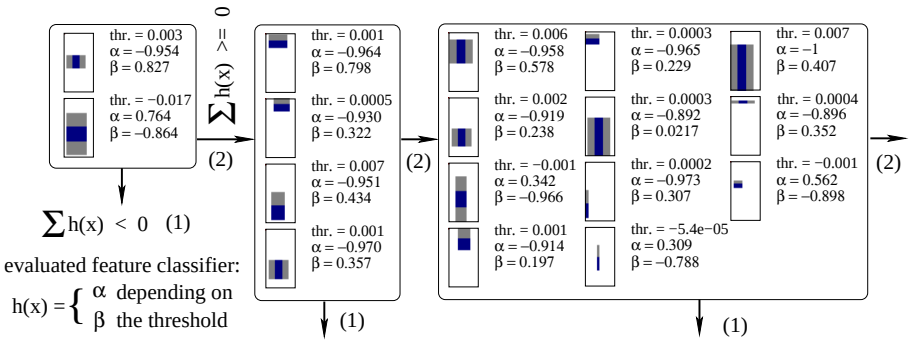

Figure 6. The first three stages of a cascade of classifiers to detect an office chair in depth data. Every stage contains several simple classifiers that use Haar-like features.

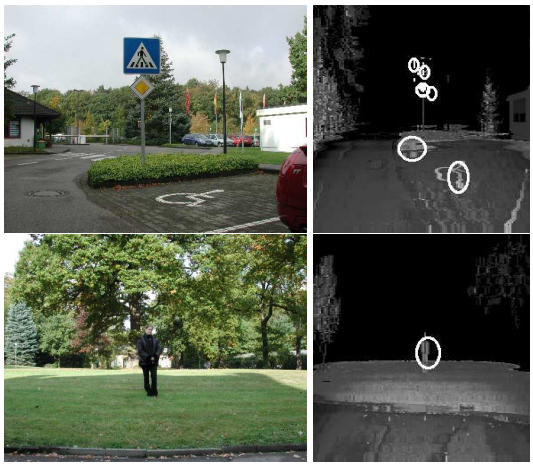

Figure 7. Two examples of foci found by the attention system on laser data. Left: Camera image of the scene. Right: The corresponding scene in laser data with foci of attention. The traffic sign, the handicapped person sign and the person were focused (taken from our results in [5]).

Secondly, we tested the performance of the classifier. Its high performance for face detection was shown in [22], here we show the performance on laser data. The classifier was trained on laser images (300×300 pixels) of office chairs. We rendered 200 training images with chairs from 46 scans. Additionally, we provided 738 negative example images to the classifier from which a multiple of sub-images is created automatically.

The cascade in Fig. 6 presents the first three stages of the classifier for the object class "office chair" using depth values. One main feature is the horizontal bar in the first stage representing the seat of the chair. The detection starts with a classifier of size 20×40 pixels. To test the general performance of the classifier, the image is searched from top left to bottom right by applying the cascade. To detect objects

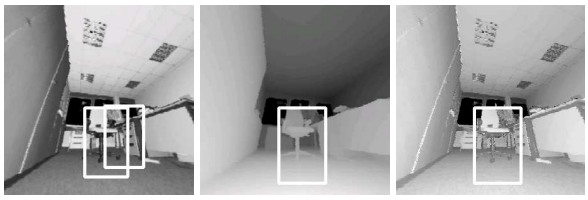

Figure 8. The combination of both laser modes for classification reduces the amount of false detections. The false detection in the reflectance image (left) does not exist in the depth image (middle) and therefore it is eliminated in the combined result (right).

Table 1. Detections and false detections of the classifier applied to 31 chair images.

object class	no. of obj.	detections			false detections		
		refl. im.	depth im.	**comb.**	refl. im.	depth im.	**comb.**
chair	33	30	29	**29**	2	2	**0**

at larger scales, the detector is rescaled. The classification is performed on a joint cascade of range and reflectance data. A logical "and" combines the results of both modes, yielding a reduction of false detections. Fig. 8 shows an example of a recognized chair in a region found by the attention system. The false detection in the reflectance image (left) does not exist in the depth image (middle) and therefore is eliminated in the combined result (right). Table 1 summarizes the results of exhaustive classification, i.e., searching the whole image, with a test data set of 31 scans that are not used for learning (see also [13]). It shows that the number of false detections is reduced to zero by the combination of the modes while the detection rates change only slightly.

Finally, we show the results of the combination of attention and classification system and analyze the time performance. The coordinates of the focus serve as input for

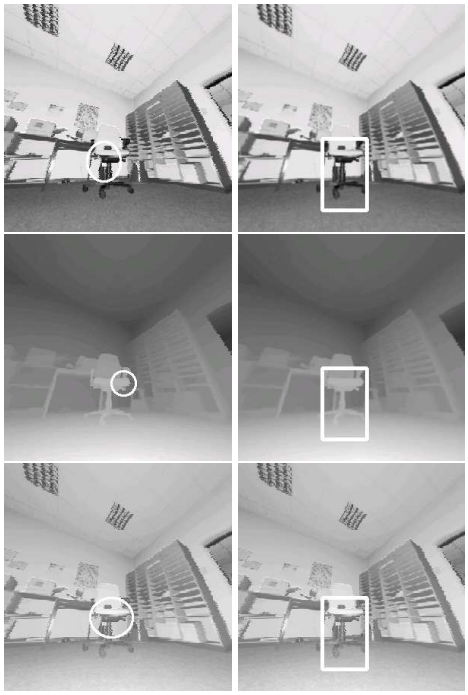

Figure 9. A complete run of the recognition system. Left column: Focus in reflectance, depth, and combined image. The last one is transferred to the classifier. Right column: chair detected by the classifier in reflectance image, depth image and combined image.

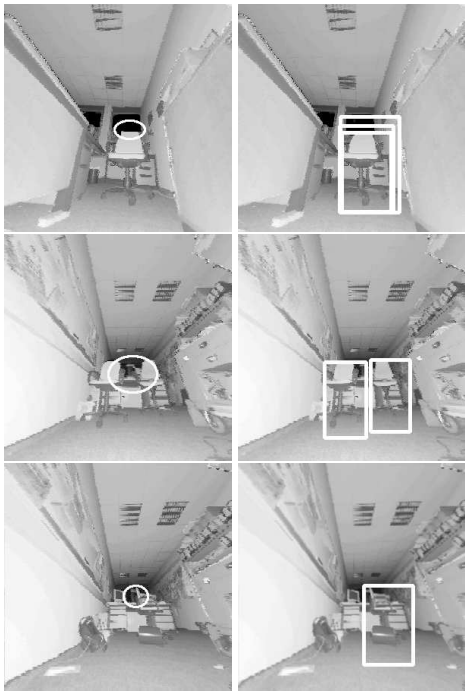

Figure 10. Three examples of chair detections. Left: Combined foci of attention. Right: corresponding detections by the classifier. Top: Chair is detected even if the focus is at its border. Middle: Detection of two chairs. Bottom: Chair is detected although it is presented sidewards and partially occluded.

the classifier. Since a focus is not always at the center of an object but often at the borders, the classifier searches for objects in a specified region around the focus (here: radius 20 pixels). In this region, the classifier begins its search for objects with a 20×40 search window. To find chairs at larger scales, the detector is enlarged by rescaling the simple features.

In all of our examples, the objects were detected if a focus of attention pointed to them and if the object was detected when searching the whole image. If no focus points to an object, this object is not detected. This is conform to our goal to detect only salient objects in the order of decreasing saliency.

Fig. 9 shows some images from a complete run of the recognition system. The left column depicts the most salient regions in the single reflectance (top) and depth image (middle) and the computed focus in the combination of both modes (bottom). The right column shows the rectangle that denotes a detected chair in reflectance and depth image and at the bottom the chair that is finally recognized by the joint cascade. Fig. 10 shows some more examples: the chairs are successfully detected even if the focus is at the object's border (top and middle) and if the object is partially occluded (bottom). However, severely occluded objects are not detected; the amount of occlusion still enabling detection has to be investigated further.

The classification needs 60 ms if a focus is provided as a starting point, 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. The attention system requires 230 ms to compute a focus for both modes, i.e., for m object classes the exhaustive search needs $m * 200$ ms, the attentive search needs $230 + m * 60$ ms. Therefore, already for two different object classes in the data base, the return of investment is reached and the time saving increases proportionally with the number of objects.

6 Conclusions

In this paper, we have presented a new architecture for combining biologically motivated attentional mechanisms with a fast method for object classification. Input data

are provided by a 3D laser scanner mounted on top of an autonomous robot. The scanner provides illumination-independent, bi-modal data that are transformed to depth and reflectance images. These serve as input to an attention system, directing the focus of attention sequentially to regions of potential interest. The foci determine starting regions for a cascade of classifiers which use Haar-like features. By concentrating classification on salient regions, the classifier has to consider only a fraction of the search windows than in the case of an exhaustive search over the whole image. This speeds up the classification part significantly.

The architecture benefits from the fusion of the two laser modes resulting in more detected objects and a lower false classification rate. The range data enables the detection of objects with the same intensity like their background whereas the reflection data is able to detect flat objects. Moreover, misclassifications of shadows, mirroring objects and pictures of objects on the wall are avoided.

In future work, we will include top-down mechanisms in the attention model, enabling goal dependent search for objects. Furthermore, we plan to additionally integrate camera data into the system, allowing the simultaneous use of color, depth, and reflectance. The classifier will be trained for additional objects which compete for saliency. The overall goal will be a flexible vision system that recognizes salient objects first, guided by attentional mechanisms, and registers the recognized objects in semantic maps autonomously built by a mobile robot. The maps will serve as interface between robot and humans.

References

[1] G. Backer, B. Mertsching, and M. Bollmann. Data- and model-driven gaze control for an active-vision system. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 23(12):1415–1429, 2001.

[2] M. Bennamoun and G. Mamic. *Object Recognition: Fundamentals and Case Studies*. Springer, 2002.

[3] B. Draper and A. Lionelle. Evaluation of selective attention under similarity transforms. In *Proc. of the Int'l Workshop on Attention and Performance in Computer Vision (WAPCV'03)*, pages 31–38, Graz, Austria, April 3 2003.

[4] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proc. 13th International Conference*, pages 148–156, 1996.

[5] S. Frintrop, E. Rome, A. Nüchter, and H. Surmann. A bimodal laser-based attention system. submitted.

[6] S. Frintrop, E. Rome, A. Nüchter, and H. Surmann. An attentive, multi-modal laser "eye". In J. Crowley, J. Piater, M. Vincze, and L. Paletta, editors, *Proc. of 3rd Int'l Conf. on Computer Vision Systems (ICVS 2003)*, pages 202–211. Springer, Berlin, LNCS 2626, 2003.

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

[8] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 21(5):433–449, May 1999.

[9] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, pages 219–227, 1985.

[10] 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'03)*, pages 55–62, Graz, Austria, April 3 2003.

[11] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. In *Proc. 25th German Pattern Recognition Symposium (DAGM '03)*, Magdeburg, Germany, Sep 2003.

[12] F. Miau, C. Papageorgiou, and L. Itti. Neuromorphic algorithms for computer vision and attention. In *Proc. SPIE 46 Annual International Symposium on Optical Science and Technology*, volume 4479, pages 12–23, Nov 2001.

[13] A. Nüchter, H. Surmann, and J. Hertzberg. Automatic Classification of Objects in 3D Laser Range Scans. In *Proc. 8th Conf. on Intelligent Autonomous Systems (IAS '04)*, pages 963–970, Amsterdam, The Netherlands, March 2004. IOS Press.

[14] N. Ouerhani, R. von Wartburg, H. Hügli, and R. Müri. Empirical validation of the saliency-based model of visual attention. In *Elec. Letters on Computer Vision and Image Analysis*, volume 3, pages 13–24. Computer Vision Center, 2004.

[15] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. 6th Int'l Conf. on Computer Vision (ICCV '98)*, pages 555–562, Bombay, India, January 1998.

[16] 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.

[17] L. Pessoa and S. Exel. Attentional strategies for object recognition. In J. Mira and J. Sachez-Andres, editors, *Proc. of the IWANN, Alicante, Spain 1999*, volume 1606 of *Lecture Notes in Computer Science*, pages 850–859. Springer, 1999.

[18] S. Ruiz-Correa, L. G. Shapiro, and M. Meila. A New Paradigm for Recognizing 3-D Object Shapes from Range Data. In *Proc. Int'l Conf. on Computer Vision (ICCV '03)*, Nice, France, Oct 2003.

[19] H. Surmann, K. Lingemann, A. Nüchter, and J. Hertzberg. A 3d laser range finder for autonomous mobile robots. In *Proc. 32nd Intl. Symp. on Robotics (ISR 2001) (April 19–21, 2001, Seoul, South Korea)*, pages 153–158, April 2001.

[20] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[21] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *AI*, 78(1-2):507–545, 1995.

[22] P. Viola and M. Jones. Robust Real-time Object Detection. In *Proc. 2nd Int'l Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing and Sampling*, Vancouver, Canada, July 2001.