Saliency-based Object Recognition in 3D Data

Simone Frintrop, Andreas Nüchter, Hartmut Surmann, and Joachim Hertzberg
Fraunhofer Institute for Autonomous Intelligent Systems (AIS)
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
E-mail: {frintrop|nuechter|surmann|hertzberg}@ais.fraunhofer.de

***Abstract*— This paper presents a robust and real-time capable recognition system for the fast detection and classification of objects in spatial 3D data. Depth and reflection data from a 3D laser scanner are rendered into images and fed into a saliency-based visual attention system that detects regions of potential interest. Only these regions are examined by a fast classifier. The time saving of classifying objects in salient regions rather than in complete images is linear with the number of trained object classes. Robustness is achieved by the fusion of the bi-modal scanner data; in contrast to camera images, this data is completely illumination independent. The recognition system is trained for two different object classes and evaluated on real indoor data.**

I. Introduction

The interpretation of sensor data in real-time is one of the most important tasks in robotic applications. One approach to achieve a robust interpretation is to fuse different sensor modalities, e.g. depth and reflectance data from a 3D laser scanner. This enables to utilize the respective advantages of the modes, e.g., there is a high probability that discontinuities in range data correspond to object boundaries. This facilitates the detection of objects: An object producing a similar intensity like its background is difficult to detect in an intensity image, but easily in the range data. Additionally, misclassifications of shadows, mirrored objects and wall paintings are avoided (cf. Fig. 5, right). 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. 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.

In computer vision, classifiers are a common approach for object detection and recently, fast classifiers have been developed, e.g. by Viola & Jones [1]. However, the recognition time increases linearly with the number of different object classes. To preserve high quality of recognition despite of limited time and computation power, the input set has to be reduced. One approach is to restrict classification to image regions of potential interest found by a saliency-based attention system. Similar to human vision, such systems identify salient parts of a scene by computing feature contrasts according to different features [2], [3], [4].

A combination of attention and classification was done by Pessoa and Exel [5]; they focus attention on discriminative parts of pre-segmented objects. Miau, Papageorgiou and Itti detect pedestrians on attentionally focused image regions using a support vector machine algorithm [6]; however, their approach is computationally expensive and lacks real-time abilities. Object recognition in range data has been considered by Johnson and Hebert [7] using an ICP algorithm for registration of 3D shapes, an approach extended in [8]. Both use local, memory consuming surface signatures based on prior created mesh representations of the objects.

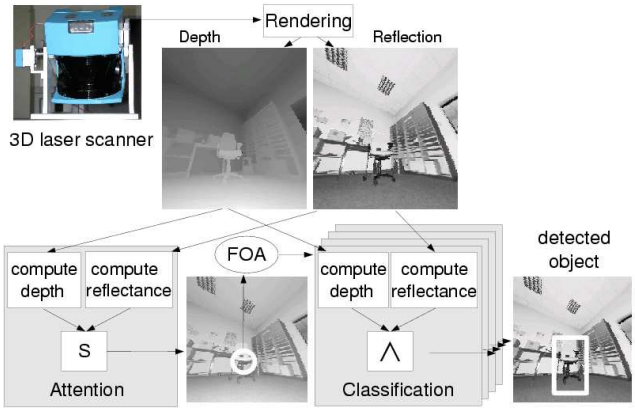

Fig. 1. The recognition system. Two laser modes, depth and reflection, are provided by a 3D laser scanner, rendered into images and fed into an attention and a classification system. The attention system fuses conspicuities of both modes in one saliency map (S). A focus of attention (FOA) is generated and fed into the classification system. The classifier searches for objects of predefined classes in the neighborhood of the FOA in both laser images and combines the results by an appropriate connection. The rectangle in the result image (right) depicts 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 the AIS 3D laser scanner [9], mounted on the autonomous mobile robot Kurt3D. The scanner yields range as well as reflectance data in a single 3D scan pass. Both data modalities are transformed into 2D images and fed into a visual attention system. In the depth as well as in the reflectance image, the system detects regions that 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. [1]. Each classifier is composed of several simple classifiers containing edge, line or center surround features. The classifier is applied to both laser modes. It is shown how the classification is significantly sped up by concentrating on

regions of interest. In this paper, we show the performance of the system for two object classes: office chairs and a mobile robot. For each object class, the same set of salient regions is considered, i.e., salient regions are computed only once for a scene.

The paper is organized as follows: Section II describes the 3D laser scanner. In section III we introduce the attention system and in IV the object classification. Section V presents the experiments performed by the combination of attention and classification and discusses the results. Finally, section VI concludes the paper.

II. The Multi-modal 3D Laser Scanner

The data acquisition in our experiments was performed with the AIS 3D laser range finder (top left of Fig. 1, [9]), mounted on the autonomous mobile robot Kurt3D. It is built on the basis of a 2D range finder by extension with a mount and a small servomotor step-rotating the scanner around a horizontal axis. The area of $180°(\text{h}) \times 120°(\text{v})$ is scanned with different horizontal (181, 361, 721 pts) and vertical (250, 500 pts) resolutions. The scanner yields two kinds of data: The distance of the scanned object (range data) and the intensity of the reflected light (reflectance data). 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.

III. The Laser-Based Attention System

The laser-based attention system detects salient regions in laser data. Rendering the laser data into images allows the investigation by computer vision methods. Saliencies are determined by computing conspicuities of the features intensity and orientation in a bottom-up, data-driven manner. These conspicuities are fused into a saliency map and, finally, the focus of attention is sequentially directed to the most salient points in this map. The system is shown in Fig. 2 (cf. [10]); it is built on principles of the standard model of visual attention by Koch & Ullman [11] used by many computational attention systems, e.g., [2], [4].

Since our sensor data consists of two modalities, depth and reflection, the attention system has to process several input images independently, an ability not available in any other attention systems the authors know about. Our system computes saliencies for every mode in parallel and finally fuses them into a single saliency map. This approach enables a straight-forward extension to additional sensor modes, e.g., camera data.

A. Feature Computations

Firstly, five different scales (0–4) are computed on images of both laser modalities by Gaussian pyramids, which successively low-pass filter and subsample the input image; i.e., 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. Two kinds of features are considered, intensities and orientations, and represented in different feature maps. The intensity feature maps are created by center-surround mechanisms which compute the intensity differences between image regions and their surroundings. 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. The center-surround difference $d = \mid c - s \mid$ is a measure for the intensity contrast in the specified region. This yields six intensity feature maps I_1 to I_6.

To obtain the orientation maps, four oriented Gabor pyramids are created, detecting bar-like features of the orientations $\{0°, 45°, 90°, 135°\}$. The maps $2-4$ of each pyramid 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 four orientation feature maps of scale 2, one for each orientation.

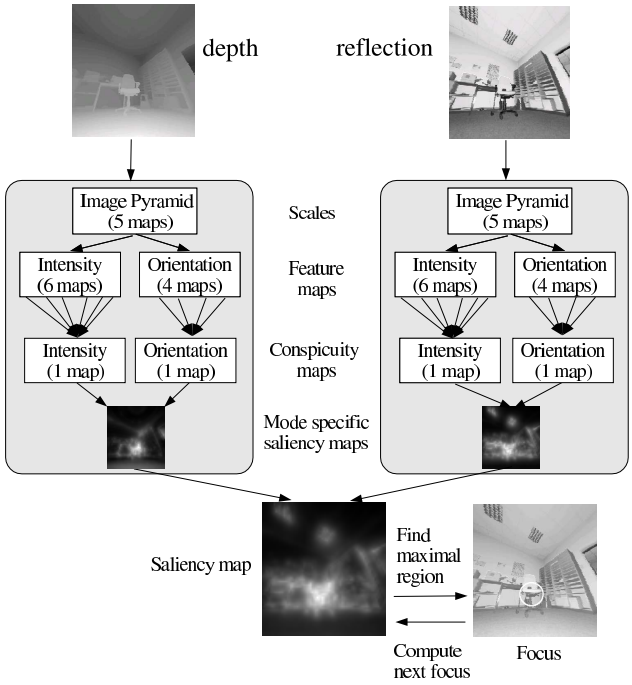

Fig. 2. The Laser-Based Attention System. Depth and reflectance images rendered from the laser data are processed independently. Conspicuities according to intensity and orientations are determined and fused into a mode-specific saliency map. After combining both of these maps, a focus of attention (FOA)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

B. Fusing Saliencies

All feature maps of one feature are combined into a conspicuity map. 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. The saliency map as well as some of the other maps are shown in Fig. 3.

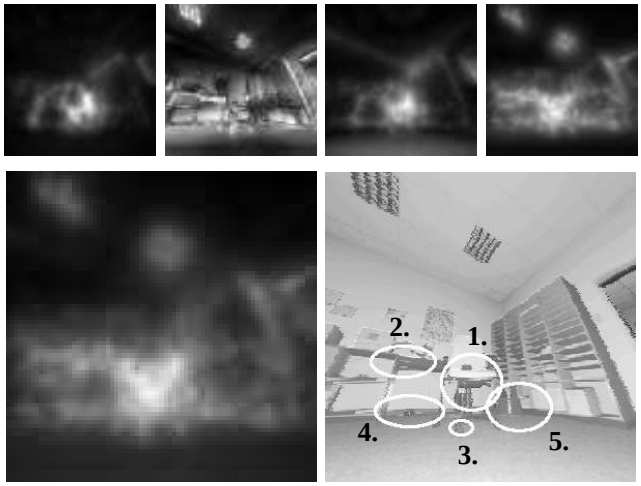

Fig. 3. First row: Orientation conspicuity map depth, intensity conspicuity map reflectance, saliency map depth, saliency map reflectance. Second row: Left: Combined saliency map. Right: The 5 most salient regions. The numbers denote the degree of saliency.

The summation of the maps is done by weighting them, resizing them to scale 2 and pixel-by-pixel addition. If there was no weighting, 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 weigh maps according to the number of peaks, each map is divided by the square root of the number of local maxima m: $w(\text{map}) = \text{map}/\sqrt{m}$.

C. The Focus of Attention

To determine the most salient location in S, the brightest point is located. Starting from this 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. Finally, the values in the focused region are reseted in the saliency map, enabling the computation of the next FOA. Fig. 3 (bottom, right) shows the five most salient locations in a test image.

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 letter on a desk, is likely not to be detected in the depth but in the reflectance image (cf. [12]).

IV. Object Classification

Recently, Viola and Jones have proposed a boosted cascade of simple classifiers for fast face detection [1].

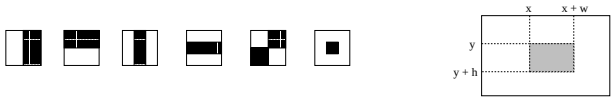

Fig. 4. Left: Edge, line, diagonal and center surround features used for classification. Right: The computation of the sum of pixels in the shaded region is based on four integral image lookups: $F(x, y, h, w) = I(x + w, y + h) - I(x, y + h) - I(x + w, y) + I(x, y)$. Feature values are calculated by subtractions of these values weighted with the areas of the black and white parts.

Inspired by these ideas, we detect objects in 3D range and reflectance data using a cascade of classifiers composed of several simple classifiers.

A. Feature Detection using Integral Images

The features used here have the same structure as the Haar basis functions also considered in [13], [1]. Fig. 4 (left) shows the six basis features, i.e., edge, line, and center surround features. The set of possible features in an object detection area is very large, e.g. 361760 features for an object detection area of 20×40 pixels. A single feature is effectively computed on input images using integral images [1], also known as summed area tables [14]. An integral image I is an intermediate representation for the image and contains the sum of gray-scale pixel values of an $x \times y$ image N, i.e.,

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

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. 4 right). To detect a feature, a threshold is required which is automatically determined during a fitting process, such that a minimum number of examples are misclassified.

B. Learning Classification Functions

The Gentle Ada Boost Algorithm is a variant of the powerful boosting learning technique [15]. 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 [14].

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.

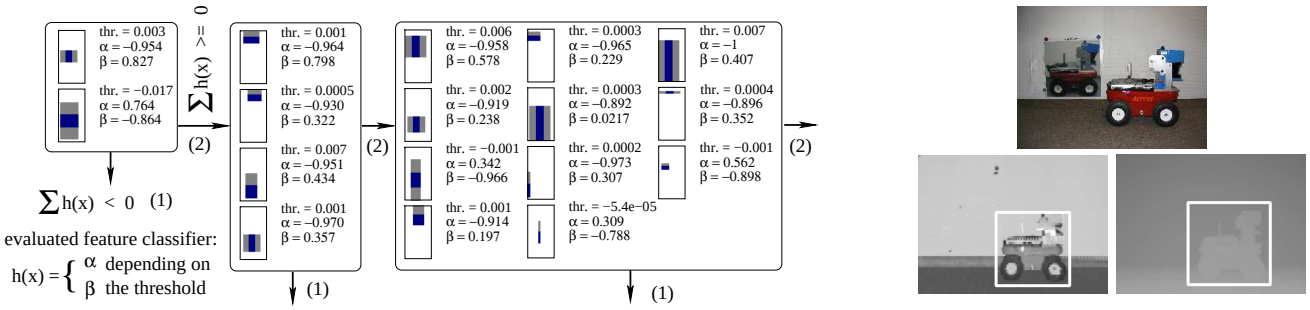

Fig. 5. Left: The first three stages of a cascade of classifiers for an office chair in depth data. Every stage contains several simple classifiers that use Haar-like features. Right: A camera image of the robot next to a poster showing a robot (top). In the laser data of the same scene, the poster is not visible due to the infrared light and the range information (bottom); this prevents misclassification: Only the real robot is detected.

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.

C. 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, but also a high 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. 5 left shows an example cascade of classifiers for detecting chairs in depth images.

An effective cascade is learned by a simple iterative method. For every stage, the classification function $h(x)$ is learned until the required hit rate is reached. The process continues with the next stage using only the currently misclassified examples. The number of features used in each classifier increases with additional stages (cf. Fig. 5, left).

An object is detected 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.

V. Experiments and Results

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. [17]. They demonstrate that attention systems based on the Koch-Ullman model [11] detect salient regions with a performance comparable to humans. We showed in [12] and [10] 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. 6.

Fig. 6. 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 arrows indicate the order of saliency of the foci. The traffic sign, the handicapped person sign and the person were focused (taken from our results in [10]).

Secondly, we tested the performance of the classifier. Its high performance for face detection was shown in [1], here we show the performance on laser data. The classifier was trained on laser images (300×300 pixels) of chairs and of the robot. We rendered 200 training images with chairs from 46 scans and 1083 training images with the robot from 200 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. 5 (left) 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. To test the general performance of the classifier, the image is searched from top left to bottom right by applying the cascade. The detection starts with a classifier of size 20×40 pixels for a chair and 24×24 for the robot. To detect objects at larger scales, the detector is rescaled. The classification is performed on a joint cascade of range and reflectance data. The detected results have to be combined by an appropriate connection, in this case we used a logical "and", yielding a reduction of false detections.

Table I summarizes the results of exhaustive classification with a test data set of 31 scenes with chairs (cf. our results in [18]) and of 33 scenes with the robot. 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.

TABLE I

DETECTIONS AND FALSE DETECTIONS OF THE CLASSIFIER APPLIED TO 31 CHAIR AND 33 ROBOT 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**
robot	33	29	29	**29**	10	1	**0**

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 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.

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. 7 and 8 show some examples of our results. The objects are successfully detected even if the focus is at the object's border (Fig. 7, left) and if the object is partially occluded (Fig. 7, middle). However, severely occluded objects are not detected; the amount of occlusion still enabling detection depends on the learned object class and has to be investigated further. We also tested the robustness of the classifier according to rotations for the object class robot. It showed that a robot rotated up to 40° is still recognized (Fig. 8, right).

The classification needs on average 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 vs. $230 + m * 60$ ms for the attentive search. Therefore, already for two different object classes the return of investment is reached: The exhaustive search needs 400 ms, whereas the attentive search requires only 350 ms. The time saving increases proportionally with the number of objects.

VI. CONCLUSIONS

In this paper, we have presented a new system for combining visual attention 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. By concentrating classification on salient regions, the classifier has to consider only a fraction of the search windows of those of an exhaustive search over the whole image. This speeds up the classification part significantly. The time saving of classifying objects in salient regions rather than in complete images is linear with the number of trained object classes. The saving is especially important in time critical robotic applications.

The architecture benefits from the fusion of the two laser modes resulting in more detected objects and a zero false classification rate. The range data facilitates 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 are avoided.

We have investigated the performance of the system with two different object classes: Office chairs and a mobile robot. In future work, we will integrate camera data into the system, allowing the simultaneous use of color, depth, and reflectance. Furthermore, the attention model will be extended by top-down mechanisms, enabling goal dependent search for objects. The classifier will be trained for additional objects 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 which are autonomously built by a mobile robot. The maps will serve as an interface between robot and humans.

REFERENCES

[1] 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.

[2] 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*, vol. 20, no. 11, pp. 1254–1259, 1998.

[3] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *AI*, vol. 78, no. 1-2, pp. 507–545, 1995.

[4] 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*, vol. 23(12), pp. 1415–1429, 2001.

[5] L. Pessoa and S. Exel, "Attentional strategies for object recognition," in *Proc. of the IWANN, Alicante, Spain 1999*, ser. Lecture Notes in Computer Science, J. Mira and J. Sachez-Andres, Eds., vol. 1606. Springer, 1999, pp. 850–859.

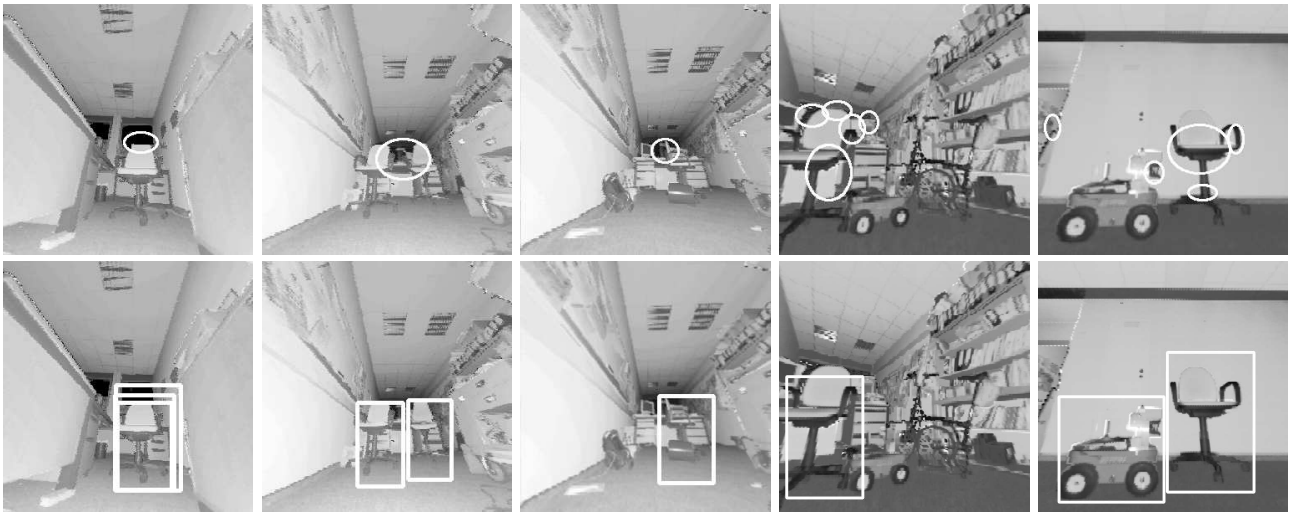

Fig. 7. Top row: The first resp. the first 5 foci of attention computed on depth and reflection data. Bottom row: Classified objects in the focus regions. Left to right: 1) Chair is detected even if the focus is at its border; 2) detection of two chairs; 3) chair is detected although it is presented sidewards and partially occluded; 4) only the chair is focused, therefore the chair but not the robot is classified; 5) both objects are focused and classified.

Fig. 8. Top row: The first resp. the first 5 foci of attention computed on depth and reflection data. Bottom row: Classified objects in the focus regions. Right: A robot rotated by 30° is still detected.

[6] 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*, vol. 4479, Nov 2001, pp. 12–23.

[7] A. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Trans. on Pattern Analysis & Machine Intelligence*, vol. 21, no. 5, pp. 433–449, May 1999.

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

[9] 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)*, April 2001, pp. 153–158.

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

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

[12] S. Frintrop, E. Rome, A. Nüchter, and H. Surmann, "An attentive, multi-modal laser "eye"," in *Proceedings of the 3rd International Conference on Computer Vision Systems (ICVS 2003)*, J. Crowley, J. Piater, M. Vincze, and L. Paletta, Eds. Springer, Berlin, LNCS 2626, 2003, pp. 202–211.

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

[14] 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.

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

[16] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[17] N. Ouerhani, R. von Wartburg, H. Hügli, and R. Müri, "Empirical validation of the saliency-based model of visual attention," in *Electronic Letters on Computer Vision and Image Analysis*, vol. 3, no. 1. Computer Vision Center, 2004, pp. 13–24.

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