A Bimodal Laser-Based Attention System

Simone Frintrop, Erich Rome, Andreas Nüchter, Hartmut Surmann

Fraunhofer Institut für Autonome Intelligente Systeme, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Abstract

In this paper, we present a new visual attention system for robotic applications capable of processing data from different sensor modes simultaneously. The consideration of several sensor modalities is an obvious approach to regard a variety of object properties. Nevertheless, conventional attention systems only consider the processing of camera images. We present a bimodal system that processes two sensor modes simultaneously and is easily extensible to additional modes. In contrast to other systems, the input data to our system is provided by a bimodal 3D laser scanner, mounted on top of an autonomous mobile robot. In a single 3D scan pass, the scanner yields range as well as reflectance data. Both data modes are illumination independent, yielding a robust approach that enables all day operation. Data from both laser modes are fed into our attention system built on principles of one of the standard models of visual attention by Koch & Ullman. The system computes conspicuities of both modes in parallel and fuses them into one saliency map. The focus of attention is directed to the most salient points in this map sequentially. We present results on recorded scans of indoor and outdoor scenes showing the respective advantages of the sensor modalities enabling the mode-specific detection of different object properties. Furthermore, we show as an application of the attention system the recognition of objects for building semantic 3D maps of the robot's environment.

Key words: visual attention, saliency detection, bimodal sensor fusion, 3D laser scanner

Email addresses: simone.frintrop@ais.fraunhofer.de (Simone Frintrop), erich.rome@ais.fraunhofer.de (Erich Rome), andreas.nuechter@ais.fraunhofer.de (Andreas Nüchter), hartmut.surmann@ais.fraunhofer.de (Hartmut Surmann).

URLs: http://www.ais.fraunhofer.de/~frintrop (Simone Frintrop), http://www.ais.fraunhofer.de/~eric (Erich Rome), http://www.ais.fraunhofer.de/~nuechter (Andreas Nüchter), http://www.ais.fraunhofer.de/~surmann (Hartmut Surmann).

1 Introduction

Human visual attention scans a scene sequentially by directing a focus of attention to regions of potential interest. This helps identify relevant data and thus efficiently select information from the broad sensory input. These effects are even more desired in computational applications like image processing and robotics, since the interpretation of the complex sensor data is often computationally expensive. One approach to restrict the amount of processing is to concentrate on regions of potential interest, detected by a computational attention system.

The visual attention model by Koch and Ullman [1] is a popular basis for many computational attention systems [2–4]. This model is related to the psychological work of Treisman [5], the so-called "feature integration theory". In the Koch & Ullman model, conspicuities concerning different features like intensity, color, and orientation are determined in parallel and fused into a single saliency map that topographically codes salient locations in a visual scene. A winner-take-all network finds the most salient point in this map. Finally, the focus of attention is directed to this point and an inhibition of return mechanism enables directing the focus to the next salient location.

In humans, eye movements are not only influenced by vision but also by other senses, e.g., the gaze may be directed into the direction of a sound, a smell or even a touch, and the fusion of different cues competing for attention is an essential part of human attention. In robotics, attentional mechanisms might also profit from additional sensor modalities, since they yield a richer set of data that enable the detection of more object properties, resulting in more useful and interesting foci of attention. Nevertheless, existing attention models usually concentrate on camera data.

This paper presents a new approach to fuse salient regions of different sensor modes. The modes provided to the attention system are depth and reflection data acquired by a 3D laser scanner in a single 3D scan pass [6]. The attention system takes the data from both laser modes as input. Both modes are searched for saliencies according to principles of the Koch & Ullman model [1]: saliencies of different features, here intensity and orientation, are computed in parallel and fused into one map. As claimed above, the saliencies of the two laser modes correspond to different object properties: saliencies in the range mode imply a depth contrast whereas saliencies in the reflection mode imply a change of object materials. The saliencies of the two modes compete for attention. This is done by weighting the data according to their importance (given by the uniqueness of the features) and fusing them into a single saliency map. The Focus of Attention (FOA) is directed to the most salient region in this map.

Since the data from the different sensor modalities result from the same measurement, we know exactly which reflection value belongs to which range value. There is no need to establish correspondences and to perform costly calibration by complex algorithms. The laser data are illumination independent, i.e., the data is the same in sunshine as in complete darkness and no reflection artifacts occur. This yields a robust approach that enables all day operation.

We demonstrate the applicability of the laser data for attentional mechanisms on real-world indoor and outdoor scenes and elaborate on the different advantages of range data and reflectance values. It is shown that these data modes complement each other: contrasts in range and in intensity need not necessarily correspond for one scene element, i.e., an object of similar material as its background may not be detected in the reflection 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 – that could be distinguished in the reflection image, will likely not be detected in the range data. The results indicate that the combination of different modes enables considering a larger variety of object properties.

Besides investigating the strengths of the different laser modes, we compare the performance of attentional mechanisms on laser data with that of classical camera based approaches. For that purpose, we apply the system of Itti et al. [2] to camera images taken at the same position as the laser scans. The comparison reveals the respective advantages of the two kinds of sensors.

Finally, we present an application of our system in robotics: the recognition of objects for building semantically labeled 3D maps. The salient regions serve as input for a fast classifier, recognizing previously learned objects in these regions. Restricting classification to regions of interest enables the detection of objects in order of their relevance and speeds up the classification process significantly with a time saving that increases proportionally with the number of object classes. The autonomous mobile robot Kurt3D registers successive 3D scans to form a coherent 3D representation of an indoor or outdoor scene via fast scan matching algorithms. The semantic labeling of objects within these compound maps based on the output of the combined attention and classification system is ongoing work.

The remainder of this article is structured as follows. We start with a brief overview of the state of the art followed by a description of the bimodal 3D laser scanner. The next section introduces the Bimodal Laser-Based Attention System BILAS. Thereafter, we present our results in detail and describe the application scenario of object recognition. Finally, we summarize and give an outlook on future work.

2 State of the Art

2.1 Visual attention models

Many computational models of human visual attention are based on psychological and neuro-biological findings. The most relevant psychological theories include the work of Treisman et al., known as *feature-integration theory* [5], and the *guided search* model by Wolfe [7]. These theories introduced the ideas of several feature maps locally coding basic features and a master map of attention integrating them. They also describe the so-called pop-out effect, i.e., the fast detection of targets defined by a single feature. This enables for instance the immediate detection of a person wearing a red suit between many people wearing black ones. These psychological theories are supported by neuro-biological evidence indicating that different features as color, orientation and motion are processed in parallel in different brain areas [8].

The first explicit computational architecture for controlling visual attention was proposed by Koch and Ullman [1]. It already contains the main properties of many current models of visual attention, including parallel feature computation, a saliency map, a winner-take-all network, and an inhibition of return mechanism. Current systems based on this model include the well-known model of Itti et al. [2]. Additional systems are described in [4,9,3]. All these systems use classical linear filter operations for feature extraction, what makes them especially useful for the application to real-world scenes. Another approach is provided by models consisting of a pyramidal neural processing architecture, e.g., the *selective tuning model* by Tsotsos et al. [10].

Typically, these models use features like intensity, color, and orientation. Depth is rarely considered although it plays a special role in deploying attention. It is not clear from the literature whether depth is simply a feature, like color or motion, or something else. Definitely, it 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 [11], a property that does not exist for the other features. Two groups that include depth are Backer et al. [12] and Maki et al. [13]. They obtain depth data from stereo vision and regard it as another feature. The data obtained from stereo vision is usually not very accurate and contains large regions without depth information. This may justify the integration of the depth values as a feature in the above mentioned models; in our approach the range data come from a special sensor and yield dense and accurate range information, so we regard depth as an additional sensor mode. Ouerhani et al. [14] suggest the use of a 3D range camera to get depth information. They also regard depth as an additional feature.

Applications of attentional mechanisms in computer vision can be found, e.g., in the area of object recognition [15,16] or in robotics [17–20]. In the field of object recognition, Pessoa and Exel combine attention and classification by focusing attention on discriminative parts of pre-segmented objects [15]. Miau, Papageorgiou and Itti detect pedestrians in attentionally focused image regions using a support vector machine algorithm [16]; however, their approach is computationally very expensive and lacks real-time abilities.

In robotics, attentional mechanisms are often used to direct a camera to interesting points in the environment or to steer the robot to these regions. For example, Tsotsos et al. present a robot for disabled children that detects toys by the help of attention, moves to a toy and grasps it [17] and Breazeal introduces a robot that shall look at people or toys [21]. Bollmann et al. present a robot that uses attention to play at dominoes [18]. In this kind of applications it depends on the environment and the objects whether the introduced bottom-up method of attention is sufficient. If the environment is crowded and the objects to be grasped are not extremely salient by themselves, top-down information would be needed to enable the system to focus on the desired objects. Although a promising approach, this has rarely been considered in existing models so far.

Another application scenario is the detection of landmarks for localization. Especially in outdoor environments and open areas, the standard methods for localization like matching 2D laser range and sonar scans are likely to fail. Instead, localization by detection of visual landmarks with a known position can be used. Attentional mechanisms can facilitate the search of landmarks during operation by selecting interesting regions in the sensor data. By focusing on these regions and comparing the candidates with trained landmarks the most probable location can be determined. A project that follows this approach is the ARK project [20]. It relies on hand-coded maps, including the locations of known static obstacles as well as the locations of natural visual landmarks.

All of these approaches are based on cameras as input devices for the attentional systems. The only approach the authors are aware of which combines the use of laser data with attentional mechanisms is described in [20]. In contrast to our work, the attentional mechanisms are only applied to the camera data and a single laser beam is merely used to measure the distance to detected objects. Instead, the approach presented in this paper is to apply attentional mechanisms to the bimodal data of a 3D laser scanner. According to the knowledge of the authors, the approach of combining several sensor modalities for attentional mechanisms has not been considered before.

Laser scanners are common sensors in robotics. They usually retrieve 2D range data of their environment for a single horizontal plane, and use it to perform tasks like obstacle avoidance and localization. Recently, laser scanners have been employed to retrieve 3D information which is usually used to build 3D volumetric representations of environments [6].

There are different approaches to get 3D data from a laser scanner. Recently, some groups have developed methods to build 3D volumetric representations of environments using 2D laser range finders: several approaches [22–25] combine two 2D laser scanners for acquiring 3D data. One scanner is mounted horizontally, one vertically. Since the vertical scanner is not able to scan lateral surfaces of objects, Zhao et al. use two additional vertically mounted 2D scanners shifted by 45° to reduce occlusions [25]. The horizontal scanner is employed to compute the robot pose. The precision of 3D data points depends on that pose and on the precision of the scanners. In all of these approaches the robots have difficulties to navigate around 3D obstacles with jutting edges. These obstacles are only detected while passing them.

Another option to obtain 3D range data is to use true 3D laser scanners that are able to generate consistent 3D data points within a single scan [6,26–29]. The RESOLV project aimed at modeling interiors for virtual reality and tele-presence [29]. They employed a RIEGL laser range finder. The AVENUE project develops a robot for modeling urban environments [26,27]. This robot is equipped with a CYRAX 3D laser scanner. The research group of M. Hebert reconstructs environments using the expensive Zoller+Fröhlich laser scanner and aims to build 3D models without initial position estimates, i.e., without odometry information [28].

The bimodal 3D laser range finder employed for this work [6] is a precise, fast scanning, reliable, and cost effective multi purpose sensor that acquires range and reflectance data in a single 3D scan pass. The interpretation of these data may require exhaustive time resources. In this paper we will describe how attentional mechanisms can help to deal with this high amount of data by finding regions of potential interest.

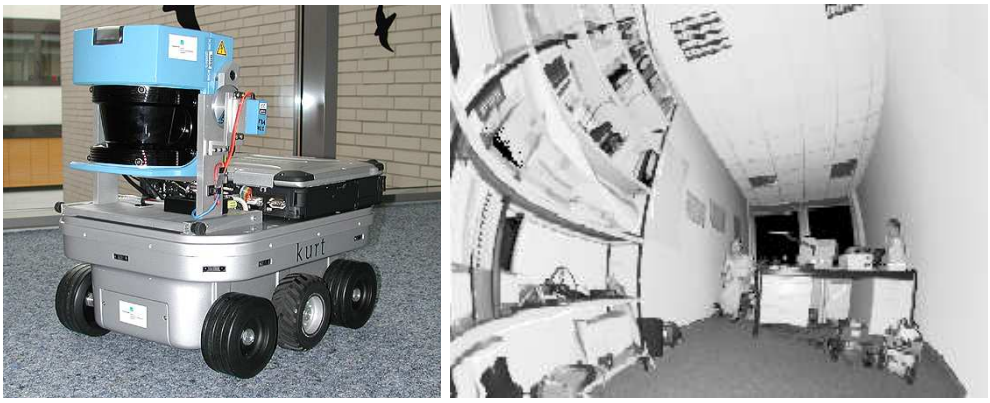

Fig. 1. Left: The custom 3D range finder mounted on top of the mobile robot Kurt3D. Right: An office scene imaged with the 3D scanner in remission value mode, medium resolution (361 × 210 pixels).

3 The Bimodal 3D Laser Scanner

3.1 Rendering Images from Laser Data

For the data acquisition in our experiments, we used a custom 3D laser range finder (Fig. 1, left). The scanner is based on a commercial SICK 2D laser range finder. In [30], the custom scanner setup is described in detail. The paper also describes reconstruction algorithms and their use for robot applications. Here, we provide only a brief overview of the device.

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 time the laser beam needs to come back gives us the distance of the scanned object (range data) and the intensity of the reflected light provides information about the reflection properties of the object (reflection data). This reflectance measurement is the result of the light measurement by the receiver diode. It measures the amount of infrared light that is returned from the object to the scanner and thus describes the surface properties concerning non-human visible light.

The 2D scanner serially sends out laser beams in one horizontal slice using a rotating mirror (LIDAR: LIght Detection And Ranging). The scanner is very fast and precise: the processing time is about 13 ms for a 180° scan with 181 measurements and the typical range error is about 1 cm. A 3D scan is performed by step-rotating the 2D scanner around a horizontal axis, i.e., the 3D scan is obtained by scanning one horizontal slice after the other. The area of 180°(h) × 120°(v) is scanned with different horizontal (181, 361, 721 pts) and vertical (210, 420 pts) resolutions.

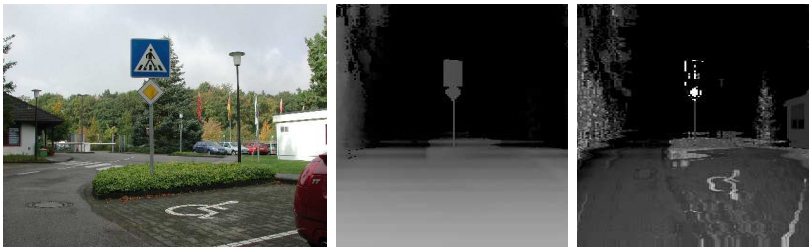

Fig. 2. Visualized laser data. Left: scene from camera image, middle: visualized depth data, right: visualized remission data. Depending on the sensor, the presented images have slightly different extensions, the laser scanner getting a wider angle than the camera in all directions.

The scanner is able to operate in two data modes. In the default mode, it returns only the range data in a predefined resolution. In an alternative mode, it is able to yield the range as well as the reflection data in a single scan pass. The reflection data can directly be converted into a gray scale intensity image (Fig. 1, right). The visualization of the depth values from the range data requires some transformation. The basic approach is to interpret the depth values as intensity values, representing small depth values as bright intensity values and large depth values as dark ones. Since close objects are considered more important for robot applications, we introduced an additional double proximity bias. Firstly, we consider only objects within a radius of 10 m of the robot's location. Secondly, we code the depth values by using their square roots, so pixel p computes from depth value d by:

$$p = \begin{cases} I - (\sqrt{d/max} * I) & : \quad d \leq max \\ 0 & : \quad d > max \end{cases} \tag{1}$$

with the maximal intensity value I and the maximal distance $max = 1000\,cm$. This measure leads to a finer distinction of range discontinuities in the vicinity of the robot and works better than a linear function. If the robot works outdoors and distant objects should be detected, the maximal distance can be increased. Fig. 2 shows an example of the visualized laser data.

3.2 Laser Data versus Stereo Vision

In current attention systems integrating depth information, the range data is usually extracted from stereo vision. With today's available computing power and advanced stereo algorithms, even real-time stereo vision at frame rate is possible. But for a number of reasons, stereo vision is not our sensor of choice.

Our 3D scanner's scan pass (between 1.2 and 15 seconds, with typically 7.5 s) is slow as compared to the frame rates of CCD cameras. However, for our target

application, the automatic 3D map building, high frame rates are not needed. Other possible robotic applications, like 3D obstacle avoidance, would benefit from a higher frame rate. This could be achieved by employing 3D cameras which yield range and image data in one snapshot. Such cameras are under development and about to enter the market. The algorithms developed for the 3D scanner can be applied directly to the data of 3D cameras.

For 3D map building, 3D laser range scanning has some considerable advantages over 3D stereo reconstruction. Firstly, range scanning yields very dense depth information. Only in rare cases the laser beam may be completely absorbed or reflected away, resulting in missing data for a few measured points. On the other hand, most 3D stereo vision algorithms rely on matching grey level values for finding pixel correspondences. This is often not possible:

- correspondences can only be found in overlapping parts of the stereo images, so that large image regions yield no depth data at all,
- ambiguous grey values that cannot be disambiguated result in false matches and
- shading may prevent finding matches at all.

Hence the generated depth maps are sparse, often containing large coherent regions without depth information.

Secondly, the precision of the depth measurement of a laser range scanner relies only on the tolerance that its construction foresees. Industry standard scanners like the SICK scanner that we use have an average depth (Z axis) error of 1 cm. The precision error of the Z axis measurement in 3D stereo reconstruction is dependent on a number of parameters, namely the width of the stereo base, the focal lengths of the lenses, the physical width of the CCD pixel, the object distance and the precision of the matching algorithm. The error increases by increased squared object distance, and decreases with increasing focal length (narrowing the field of view). For small robots like Kurt3D, the width of the stereo base is limited to small values ($\leq$ 20 cm), resulting in a typical Z axis error of about 78 cm for objects at the scanner's maximum ranging distance of about 8 m (stereo base $b = 200\ mm$, $f = 4\ mm$, distance $d = 8000\ mm$, pixel width w=0,0098 mm, precision 1 pixel, $error = d * (d * w)/(b * f)$).

And finally, our 3D laser scanner provides a very large field of view and the data of the laser scanner are illumination independent. This enables all-day operation and yields robust data. The named strengths make the 3D laser scanner the sensor of choice for the acquisition of 3D range and reflectance data and for the generation of input data for the attention system.

4 The Bimodal, Laser-Based Attention System (BILAS)

The Bimodal Laser-Based Attention System (BILAS) simulates human eye movements by generating saccades. In contrast to other systems, data from different sensor modalities are considered which compete for attention. Inspired by the psychological work of Treisman and Gelade [5], we determine conspicuities of different features, intensity and orientations, in a bottom-up, data-driven manner. The conspicuities are fused into a mode-specific saliency map which contains the saliencies according to the specific sensor mode. The saliencies of each mode are weighted according to their importance and fused into a saliency map. The focus of attention (FOA) is directed to the most salient point in this map and finally, the region surrounding this point is inhibited, allowing the computation of the next FOA. The model graphic in Fig. 3 illustrates this procedure and the algorithm in Fig. 4 elaborates on the computational details also described in the following.

The attention system is built on principles of one of the standard models of visual attention by Koch & Ullman [1] that is used by many computational attention systems [2,4,3,9]. The implementation of the system is influenced by the Neuromorphic Vision Toolkit (NVT) by Itti et al. [2] that is publicly available and can be used for comparative experiments (cf. section 5.3). The NVT has also been adopted by several other groups for their research on attention [3,9]. Our system 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. All existing models the authors know about concentrate on camera data. 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 used in this work are depth and reflectance values provided by the 3D laser scanner. The system computes saliencies for every mode in parallel and finally fuses them into a single saliency map. The system is easily extensible in a straightforward way to other sensor modalities. All sensor data that are representable in a 2D map might be used as input to the system.

The input to the attention system consists of two images representing the data of the two laser modes depth and reflection. On both images, 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 that our

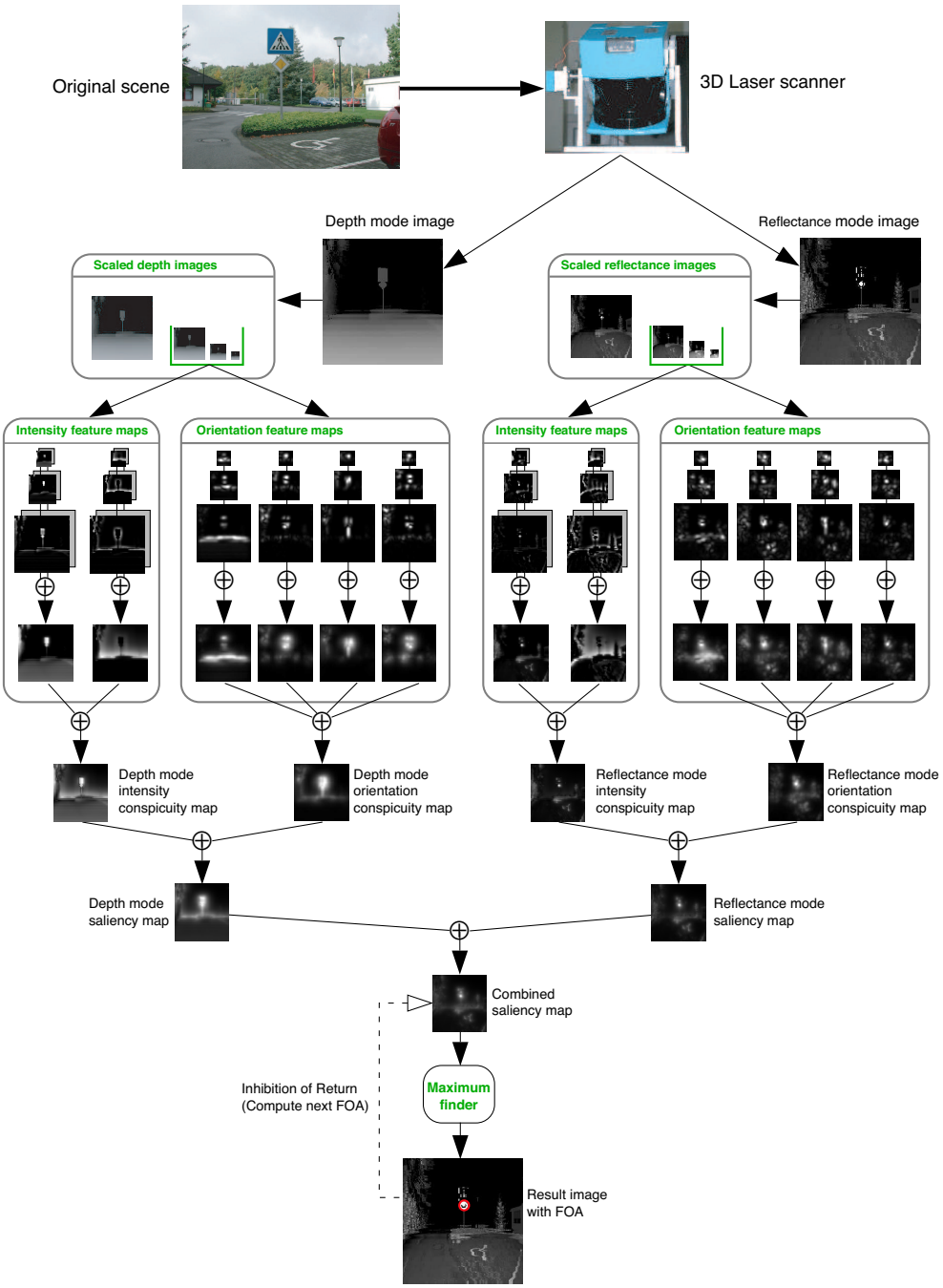

Fig. 3. The Bimodal Laser-Based Attention System (BILAS). The images from the two laser modes "depth" and "reflectance" are computed independently. Saliencies according to intensity and orientations are determined and fused into a mode-specific saliency map. After combining both of those maps, the focus of attention is directed to the most salient region.

approach yields the same performance with fewer scales.

for each mode $i \,\epsilon\, \{depth, reflectance\}$
- **compute 5 scales:** $s_0, ..., s_4$
- **compute feature maps:**
  - compute 2 intensity maps I_i, $i \,\epsilon\, \{1, 2\}$
    - compute 12 intensity maps $I_{i,c,s}$, $c \,\epsilon\, \{2, 3, 4\}, s \,\epsilon\, \{3, 7\}$
      - $I_{1,c,s} = d_{(\text{on-off})}(c, s)$
      - $I_{2,c,s} = d_{(\text{off-on})}(c, s)$
    - add per intensity channel: $I_i = \bigoplus_{c,s} I_{i,c,s}$
    - divide intensity maps: $I_i = I_i/6$
  - compute 4 orientation maps O_σ, $\sigma \,\epsilon\, \{0°, 45°, 90°, 135°\}$
    - compute 12 orientation maps $O_{s,\sigma}$
      - $O_{s,\sigma}$ = gabor filter σ applied to s
    - add per orientation: $O_\sigma = \bigoplus_s O_{s,\sigma}$
    - divide orientation maps: $O_\sigma = O_\sigma/3$
- **compute conspicuity maps:**
  - intensity map I:
    - get maximum of intensity maps: $max_I = max(I_{c,s})$
    - weight and add maps: $I = \bigoplus_i w(I_i)$
    - normalize I: $I = norm_{(0,max_I)}(I)$
  - orientation map O:
    - get maximum of orientation maps: $max_O = max(O_\sigma)$
    - weight and add maps: $O = \bigoplus_\sigma w(O_\sigma)$
    - normalize O: $O = norm_{(0,max_O)}(O)$
- **compute mode-specific saliency map:**
  - get maximum of conspicuity maps: $max_C = max(I, O)$
  - weight and add maps: $S_i = w(I) + w(O)$
  - normalize S_i: $S_i = norm_{(0,max_C)}(S_i)$

end for each mode

compute global saliency map:
- get maximum of S_i: $max_{S_i} = max(S_{depth}, S_{reflectance})$
- weight and add maps: $S = w(S_{depth}) + w(S_{reflectance})$
- normalize S: $S = norm_{(0,S_i)}(S)$

repeat until maximum number of FOAs is reached:
- **find most salient region:** find global maximum m in S,
  - determine salient region r surrounding m with region growing
- **direct FOA** to most salient region r
- **IOR:** inhibit region r in S

end repeat

Fig. 4. Algorithm of the Bimodal Laser-baser Attention System (BILAS)

4.1 Feature Computations

The features considered for the system are 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 mechanisms simulate cells of the human visual system responding to intensity contrasts (on-center-off-surround cells and off-center-on-surround cells). 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 $d_{(\text{on-off})}$, responding strongly to bright regions on a dark background, and the off-center-on-surround difference $d_{(\text{off-on})}$, responding strongly to dark regions on a bright background:

$$d_{(\text{on-off})}(c,s) = c - s, \qquad c \epsilon \{2,3,4\}, s \epsilon \{3,7\} \tag{2}$$

$$d_{(\text{off-on})}(c,s) = s - c, \qquad c \epsilon \{2,3,4\}, s \epsilon \{3,7\} \tag{3}$$

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. Imagine an image with a gray background and white and black objects on it, both producing the same intensity contrast to the background. If there is only one white object, but several black ones, the white object pops out in our approach but not in the NVT (cf. Fig. 5, top). The reason is the amplification of maps with few peaks (cf. section 4.2); in BILAS the on-off-intensity-map contains only one peak and gets a stronger weighting than the off-on-intensity-map, in the NVT the combined intensity-map contains six equally strong peaks. 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 seperated 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 (cf. Fig. 8).

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,

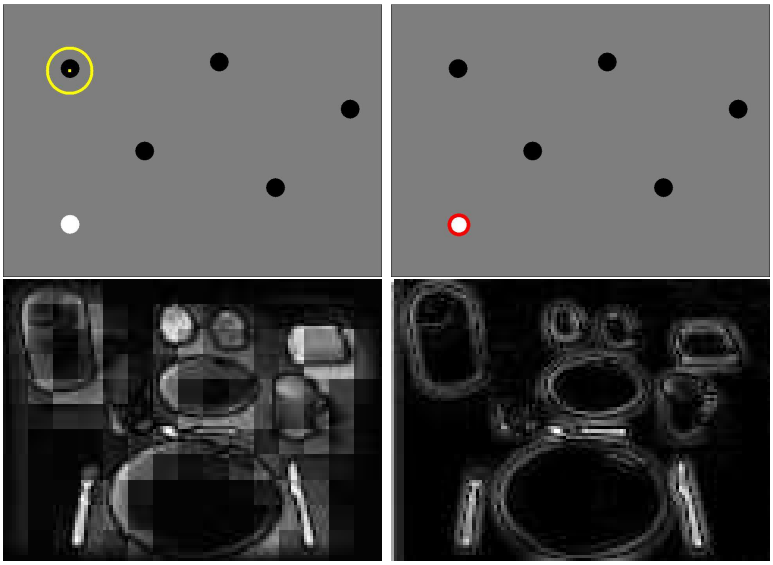

Fig. 5. Top: The white pop-out is not detected by the NVT (left) but by BILAS (right). Only separating the on-center-off-surround difference from the off-center-on-surround difference enables the pop-out. Bottom: Two intensity maps of a breakfast table scene, computed by the NVT (left) and by BILAS (right). The square-based structure in the left image resulting from taking the difference between two scales can be seen clearly, the right image shows a much more accurate solution.

e.g. $I_6 = scale(4) - scale(8)$. The problem with this approach is that it yields sort of "square-based" feature maps and uneven transitions at the borders of the coarser scale (cf. Fig. 5, bottom left). Our approach results in a slightly slower computation but is much more accurate (cf. Fig. 5, bottom right) and needs fewer scales. BILAS uses only five scales (0-4), instead of nine in the NVT, since Itti et al. need the four coarsest scales (5-8) merely to represent the surround.

The orientation maps are obtained by creating four oriented Gabor pyramids detecting bar-like features of orientations $\{0°, 45°, 90°, 135°\}$. 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. So we take the orientation maps as are, yielding $3 \times 4 = 12$ orientation maps $O_{s,\sigma}$ for scales $s \,\epsilon\, \{2, 3, 4\}$ and orientations $\sigma \,\epsilon\, \{0°, 45°, 90°, 135°\}$. The orientation maps are summed up by inter-scale addition for each orientation, yielding four feature maps O_σ of scale 2, one for each orientation.

If the summation of maps is done in a straightforward manner, all maps 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. An operator enabling this is the operator $N()$ presented by Itti et al. in [2]. It promotes maps with one strong peak and suppresses those which contain many almost equivalent peaks. This operator works by normalizing the maps to a fixed range and multiplying it by the squared difference of the global maximum M and the average of the local maxima $\bar{m}$: $N(map) = map * (M - \bar{m})^2$.

There are two problems with this approach. The first problem was already pointed out in [31]: Taking the difference of the global and the local maxima only yields the desired result if there is just one strong maximum. If there are two equally high maxima, the difference yields zero, ignoring the map completely, while humans would consider both maxima as salient (imagine the eyes of a wolf in the dark). In the same article a sophisticated complex iterative scheme is proposed to overcome this problem by local competition between neighboring salient locations. For simplicity reasons, we chose an alternative approach: we divide each map by the square root of the number of local maxima in a pre-specified range from the global maximum: $w(\text{map}) = \text{map}/\sqrt{\text{num-local-max}}$. This method yielded good results in first experiments, but it should be examined further whether it stands the test.

The second problem with $N()$ concerns the normalization of maps to a fixed range. This was done by Itti et al. to weed out the differences between a priori not comparable modalities with different extraction mechanisms. In addition, it prevents the higher weighting of channels that have more feature maps than others. However, there is a problem with this approach: normalizing maps to a fixed range removes important information about the magnitude of the maps. Assume that one intensity and one orientation map belonging to an image with high intensity but low orientation contrasts are to be fused into one saliency map. The intensity map will contain very bright regions, but the orientation map will show only some moderately bright regions. Normalizing both maps to a fixed range forces the values of the orientation maps to the same range as the intensity values, ignoring the fact that orientation is not an important feature in this case.

Since some normalization has to be done to make the maps comparable after they were summed up and weighted at least once, we propose the following normalization technique: we store the maximum m of the maps that have to be summed up and weighted. Then summation and weighting are performed

and finally, the normalization is done between 0 and m, expressed by the term $n_{(0,m)}(\text{map})$ in the algorithm in Fig. 4. This technique yielded much better results in our experiments than the normalization to a fixed range.

The next step in the feature computation is the generation of the conspicuity maps. All feature maps belonging to one feature are combined into one conspicuity map, yielding map I for intensity and O for orientation. In contrast to the NVT, we compute the interscale-addition by interpolating scales 3 and 4 to the finest scale 2 and perform point-by-point addition, instead of reducing all maps to the coarsest scale 4 before adding them. This enables us to keep all the information. Conspicuity maps I and O are summed up to the mode-specific saliency map S_i, $i \,\epsilon\, \{\text{depth}, \text{reflectance}\}$. Each of these maps represents the salient regions of the respective sensor mode, e.g., the depth map shows depth contrasts, the reflectance map shows strongly reflecting image regions. Both of these maps are weighted again with the weighting function $w()$ to determine how strong the salient regions pop out concerning this sensor mode. Finally, the maps are summed up to the single saliency map S.

4.3 The Focus of Attention

To determine the most salient location in S, a straightforward maximum-finding strategy is applied instead of the WTA network proposed by Itti et al. Although biologically less plausible, equivalent results are achieved with less computational resources. Starting from the most salient point, region growing recursively finds all neighbors with similar values within a certain range. 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 the NVT and most other systems. One of the few systems considering the size of the salient region is presented in [3]. In a final step, inhibition of return (IOR) is applied to the FOA region in the saliency map. After inhibition, the next salient region in S is determined and the focus jumps to that position.

5 Results

We have tested our approach on scans of both indoor and outdoor scenes. The laser scans were taken at two different resolutions: 152×256 and 360×211 data points. From these points, images of sizes 244×256 and 288×211 were generated. The pixel dimensions do not match exactly the number of data points, since some of the border pixels in horizontal direction are ignored due to distortion effects and in the lower resolution mode the pixels in the horizontal direction were duplicated to yield adequately dimensioned images.

The lower resolution proved to be sufficient for the application of attentional mechanisms. The computations of the first focus on both laser images took 230 ms on a Pentium IV, 2400 MHz. The computation of further foci was determined nearly at once (less than 10 ms).

The camera images depicted in this section represent the same scenes as the laser scans to facilitate the scene recognition for the reader and to enable comparison between the sensor modalities. It has to be remarked that camera and laser images do not show identical parts of the scene, since the sizes of their fields of view are different.

In all of the presented examples, it is hard to evaluate which foci are "good ones" and which are not. This decision depends highly on the current task, since the robot needs to detect obstacles if its task is obstacle avoidance, but open doors if the task is navigation into another room. If no such information is available, every strong cue might pop out and attract the focus of attention, the same way a human looks around when exploring an unknown scene. The only possibility to evaluate bottom-up foci is to decide whether the detected region is also considered salient by humans. A good hint is to regard if the focus is on something considered being an object. For further information on the evaluation of attentional systems in comparison to humans refer to [32] and [33]. However, the main point of this section is to show the diversity of the saliencies of the different sensor modes, enabling the consideration of more object properties, rather than to assess the quality of the foci.

In this section, we focus on three aspects. Firstly, we show the general performance of attentional mechanisms on laser data. Secondly, the different qualities of the two laser modes are shown, and finally, we compare the performance of attentional mechanisms on laser images with those on corresponding camera images.

5.1 General performance

Here, we briefly demonstrate the general performance of attentional mechanisms on laser data to indicate that the choice of a 3D laser scanner as a sensor for attentional mechanisms is a sensible one. Fig. 6 shows four scenes, a camera image as reference on the left and the laser image combined from both laser modes on the right.

In the first three laser images, the FOAs point to objects that also a human observer would consider as salient: a traffic sign, two flower pots and a statue with flowers. These objects are focused because they are highly salient in laser images: the traffic sign has strong reflection properties that yield high saliencies in the reflection image. Furthermore, it pops out in depth and shows

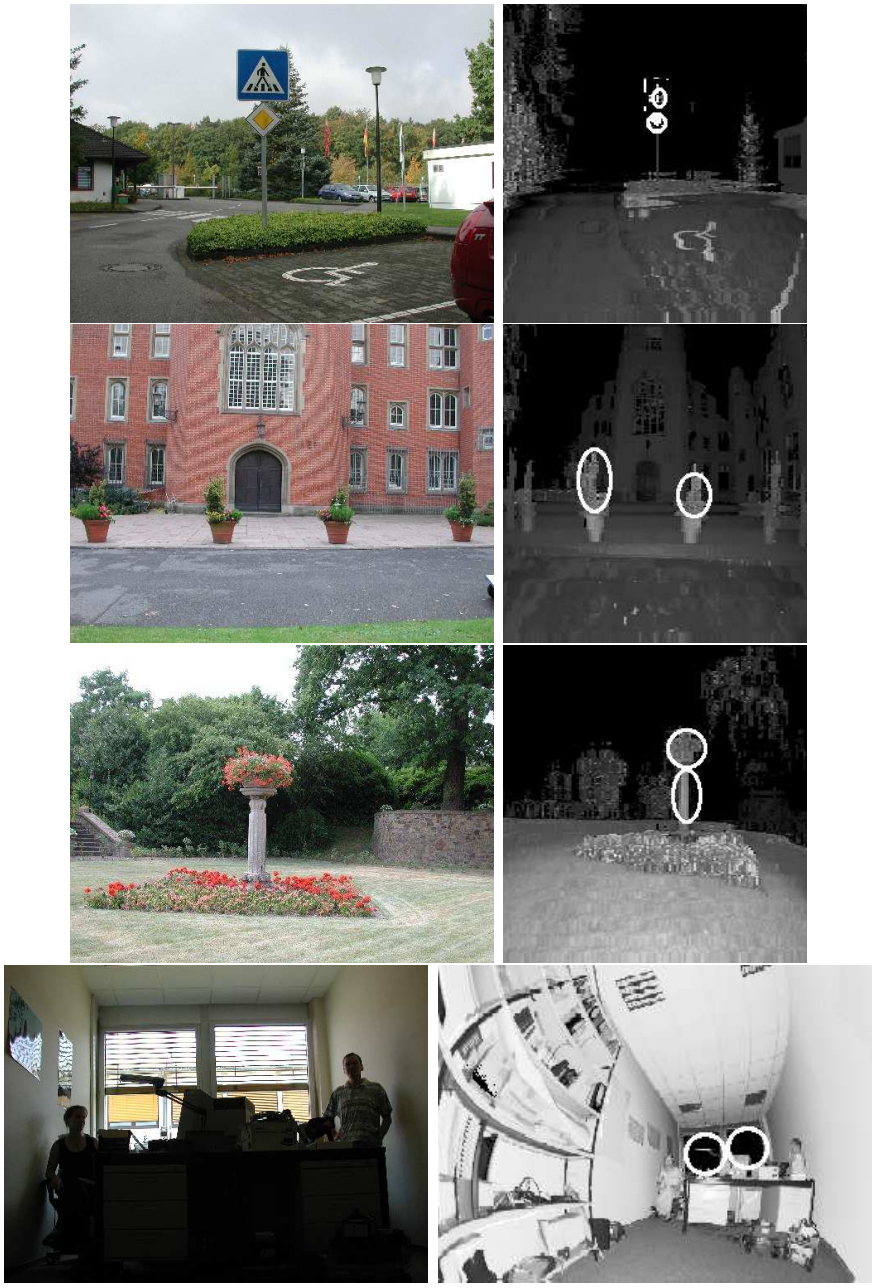

Fig. 6. The first two foci of attention computed by BILAS on laser scanner data. Left: the scene in a camera image. Right: foci on the combination of range and reflection data.

a vertical orientation (cf. Fig. 3). Similar effects are true for the objects in the next two images. The last row shows an example of a scene in which the foci point to regions, the windows, that most human observers would not consider as conspicuous, since they are not useful to most tasks. However, in a pure bottom-up approach the window region is highly salient in the laser data, because the glass is transparent for the laser scanner, yielding black regions in both laser modes. Note that similar effects would arise in the processing

of the camera image, which shows the window region much brighter than the rest of the image.

5.2 The two laser modes

This section concentrates on showing the different qualities of the two laser modes. For that purpose, we applied our system separately to range and reflection data. Additionally, we applied it to the simultaneous input of both modes, showing how their different properties influence the detection of salient regions. We start with the presentation of some scenes where certain saliencies are only detected in the range data and other saliencies only in the reflection data. The shown examples (Fig. 7–10) are presented in reading order as follows: depth image, reflection image, combined image, and camera image as a reference of the scene.

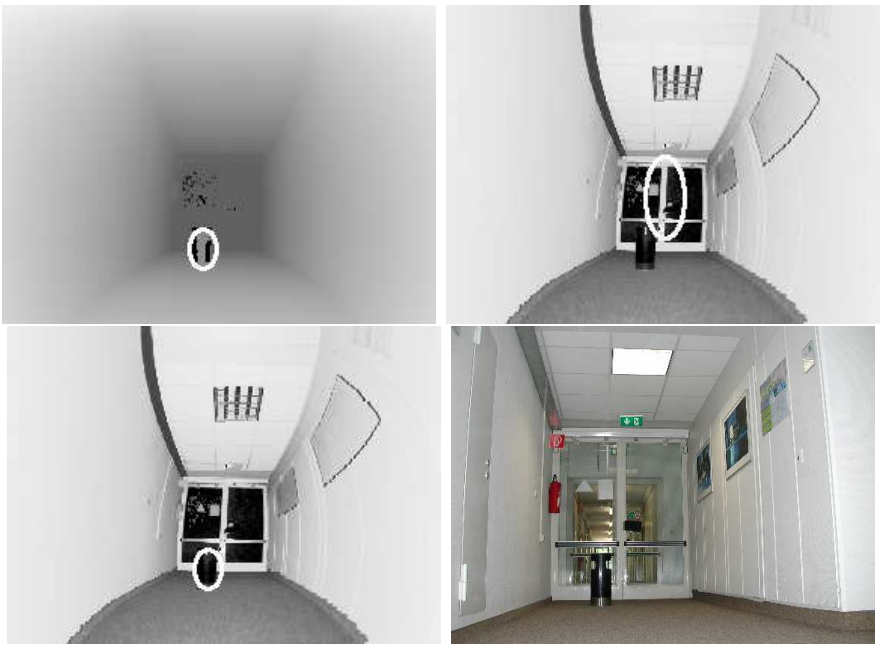

Fig. 7. The foci in laser data show some advantages of the depth mode. In reading order: depth image, reflection image, combined image, camera image. The rubbish bin is salient only in the range data. Here, the stronger influence of the depth image causes the first focus to point to the rubbish bin in the combined image, too.

The advantages of the depth mode are illustrated in Fig. 7 and 8. The example in Fig. 7 shows a rubbish bin in a corridor. The rubbish bin is highly salient in the depth image, but not in the reflectance image. Here, the vertical line of the door attracts the attention. In the combined image, the influence of the depth focus is stronger, resulting in a focus on the rubbish bin. Remember that the influence of the maps is determined by the weighting function w that strengthens maps with few salient regions (cf. sec. 4). Of course, the focus in

Fig. 8. The foci in laser data show some advantages of the depth mode. In reading order: depth image, reflection image, combined image, camera image. The open door is salient only in the range data.

the combined image is not always on the desired object since this is a task-dependent evaluation. The region with the highest bottom-up saliency wins and attracts the FOA.

The example in Fig. 8 shows a hallway scene. The depth image shows a FOA on an open door which could be interesting for a robot. In the reflection image the foci point to other regions. Here again, the influence of the depth image is stronger, resulting in FOAs on the open door in the combined image, too.

Please note that the foci in the combined image are not a union of the foci of both modes. In the combined image, the first focus might point to a region that is the most salient region neither in the depth nor in the reflection image. This might happen for a simple reason: if the depth image has its most salient point at location a and the reflection image at location b, whereas both images have a point with lower saliency at location c, then the saliency of location c can sum up to the highest saliency in the combined image, yielding the primary focus of attention.

The advantages of the reflection mode are shown in Fig. 9 and 10. Although the traffic sign in Fig. 9 attracts the first FOA in both laser modes, in the reflection image the 5th FOA is directed to the handicapped person sign on the floor. In the depth data this sign is completely invisible. In the combined data this detection occurs later: the 6th FOA is on the handicapped person sign.

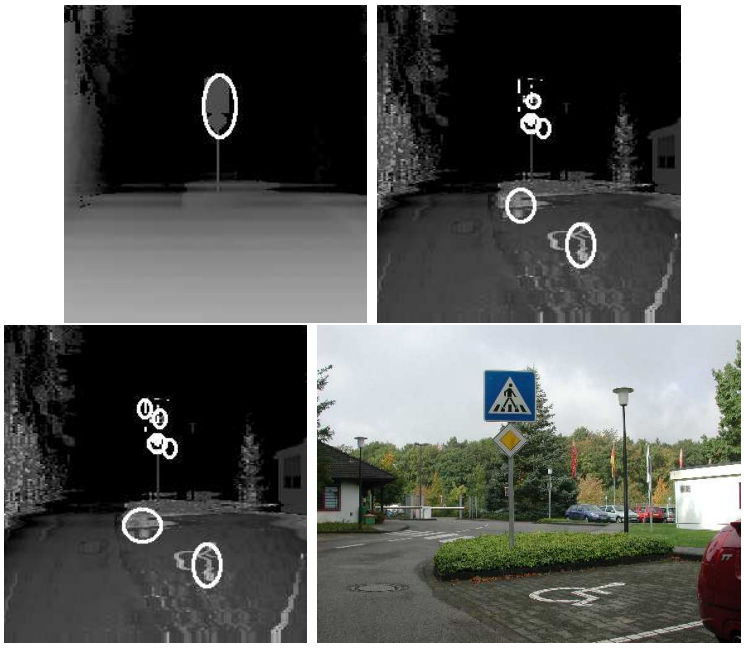

Fig. 9. The foci in laser data show some advantages of the reflection mode. In reading order: depth image, reflection image, combined image, camera image. The handicapped person sign is salient only in the reflection data.

Fig. 10. The foci in laser data show some advantages of the reflection mode. In reading order: depth image, reflection image, combined image, camera image. All of the four cars are among the first six focus regions in the reflection data.

Fig. 11. Foci showing the same region in camera and in laser data. Left: a camera image with a focus on a red traffic sign generated by the NVT. Right: a laser image, combined from depth and reflection data, with a focus generated by BILAS.

Another example is shown in Fig. 10. Three of the four cars in the scene are among the first four FOAs in the reflection image and within the first seven FOAs in the combined data. Obviously, the strongly reflecting license plates are the reason for high saliency in these regions. In the depth image, the cars are not focused, because the saliency of the nearer tree is stronger.

These examples show the different advantages of the two laser modes. Note again that the decision which results are better depends highly on the task. The point of these experiments is to show the complementary effect of the two modes and the possibility to concentrate on different object properties. In an application scenario with a special task, top-down influence should strengthen the influence of one mode to enable focusing on special object properties.

5.3 Camera versus laser

Usually, computational visual attention systems take camera images as input. In this section, we compare this approach to our one, considering the respective advantages of the sensors. The computations for the camera images were performed by the NVT by Itti et al. [2], the computations for the laser images by BILAS. Note that some differences in performance may not only result from the different sensors, but also from the differences in the implementation of the two systems.

We present three different cases: FOAs that are similar in both kinds of sensor data, those that are unique in camera images and those being unique in laser data. Fig. 11 shows an example of a scene where both sensor modalities yield the same results: the focus is immediately attracted by the traffic sign in both images. We remark that this is due to different reasons: the camera FOA is attracted by the color of the traffic sign, the laser FOA by its depth and reflection properties. Obviously, the design of traffic signs is carefully examined, attracting bottom-up attention of different kinds.

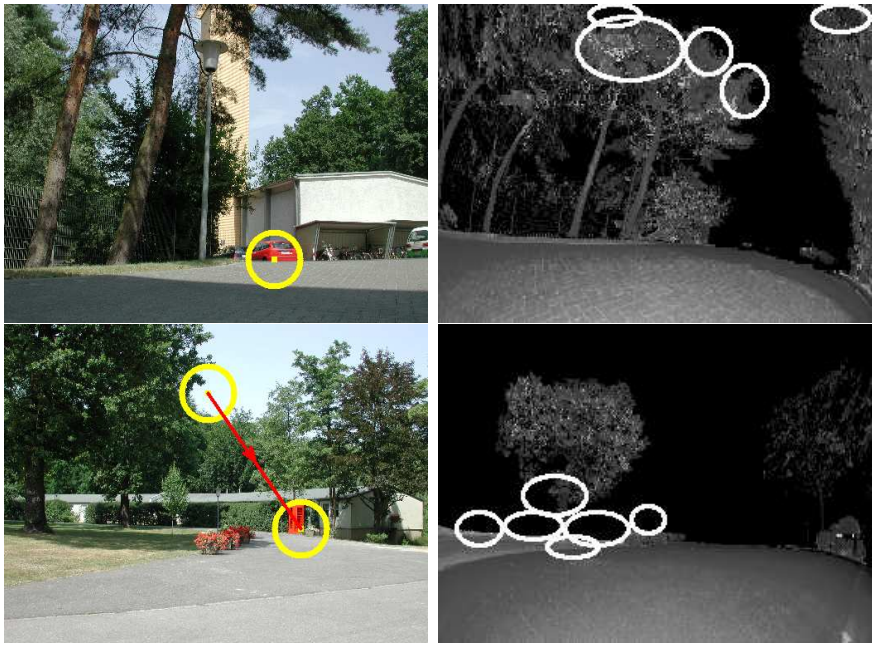

Fig. 12. The foci show some advantages of camera images: the red car and the red telephone box are only focused in the camera images (left), but not in laser data (right).

One of the advantages of a camera is its ability to obtain color information. Although laser scanners exist that are able to record color and even temperature information, ours is not. Both scenes in Fig. 12 show cases in which color properties alone produced saliencies in image regions (the car in the upper image, the telephone box in the lower one) that would hardly be salient in the laser mode data.

On the other hand, Fig. 13 shows objects that are only focused in the laser images. The traffic sign pops out in the laser data, whereas the FOA in the camera image sticks to the yellow blinds. In the lower row of Fig. 13, the person is only focused in the laser image.

Of course, the decision which sensor yields the better results depends not only on the scenes but also on the task. The advantage of the laser scanner is that it is able to detect salient object attributes that can not or hardly be found in camera images. Best results should be achieved by a combination of both sensors, inducing a much richer variety of salient regions. Selection from these regions could be controlled by integrating top-down mechanisms to the system. These topics are subject to future work.

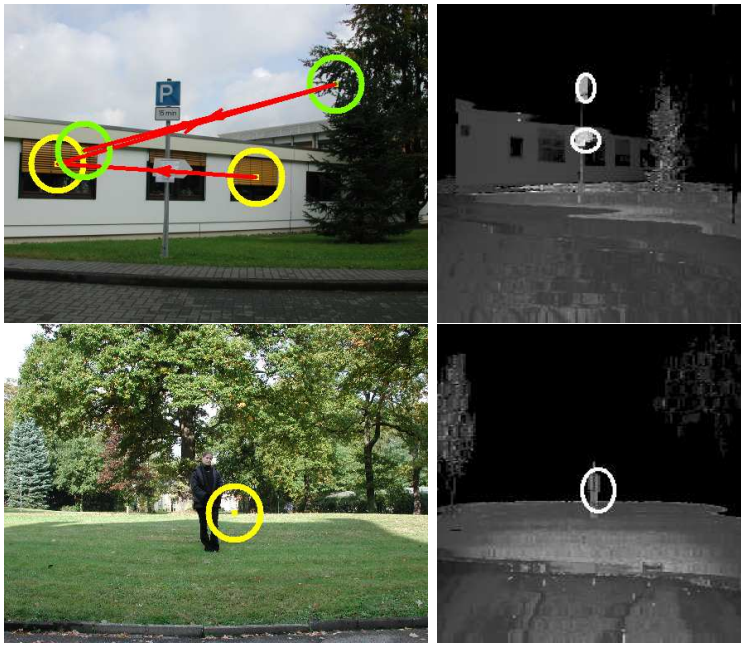

Fig. 13. The foci show some advantages of the laser data: the traffic sign (ahead) and the person (below) are only focused in the laser data (right), but not in camera images (left).

6 An Application Scenario: Semantic Map Building with Kurt3D

As discussed in section 2, one of the application fields of computational visual attention is robotics. Here, we describe the application of attentional mechanisms for a task that was not investigated before: the recognition of objects in bimodal laser scanner data and the registration of the objects in semantic 3D maps created autonomously by the mobile robot Kurt3D. In an exploration phase, 3D maps of a room or a building are generated and objects within these maps shall be labelled. The goal is to label not all objects in the maps but the ones that are most salient and at the same time belong to task specific classes that are predetermined and trained in advance. This prevents the maps from being overcrowded, preserves high quality of recognition despite of limited time and computation power, and enables the robot to detect objects in order of their relevance. The approach prefers objects near the robot, resulting in a detailed map representation in the explored regions. Later, the robot is able to navigate in an already known and mapped environment.

While the integration of the objects into the 3D maps is subject for future work, we already investigated the detection and recognition of objects in laser data supported by the visual attention system BILAS. This specific method combination is motivated by the fact that objects are hard to detect and to segment in 3D range data alone, when no 3D or CAD models of the objects are available. By applying object detection algorithms to the visualized bimodal

data, we can identify regions in these visualizations that contain searched objects. For each such object, we can determine exactly the 3D points that belong to it. In the remainder of this section, we concentrate on the first step.

The attention system searches the scene for salient regions and thus performs a preselection of locations. The salient regions detected by the attention system serve as input for a fast classifier by Viola & Jones [34] that was originally developed for face detection. To detect objects, their algorithm uses a cascade of simple classifiers, i.e., a linear decision tree. Each stage of the cascade consists of several primitive classifiers for detecting edge, line or center surround features. A fast computation of these primitive features is enabled by using an intermediate representation called integral image. For learning the desired object classes from a large set of sample images, the boosting technique *Ada Boost* is used [34].

We trained the classifier to recognize two object classes: office chairs and the autonomous mobile robot Kurt3D; for further information on object detection in 3D laser range data, please refer to [35]. The restriction of classification to salient regions enables a preference of salient objects against less salient ones. Furthermore, it significantly speeds up the classification part, since only about 30% of the image have to be investigated. This is especially useful if many object classes are considered: the time saving increases proportionally with the number of object classes. Some of the results of the object recognition are shown in Fig. 14 and 15. A detailed discussion of the combination of visual attention and classification in 3D laser data can be found in [36].

7 Discussion and Outlook

7.1 Summary

In this paper, we have introduced a new computational system of visual attention capable of processing data from different sensor modalities simultaneously. The bimodal input data for the attention system, depth and reflection, were provided by a 3D laser scanner, in contrast to other systems that usually only work on camera images. Our new Bimodal Laser-Based Attention System (BILAS) processes the depth and reflection data in parallel, determining conspicuities concerning intensity and orientations and fusing them in an appropriate way.

We have tested our system on both indoor and outdoor real-world scenes. The results show that BILAS is able to focus on a large number of regions in the environment that contain objects of potential interest for robot tasks.

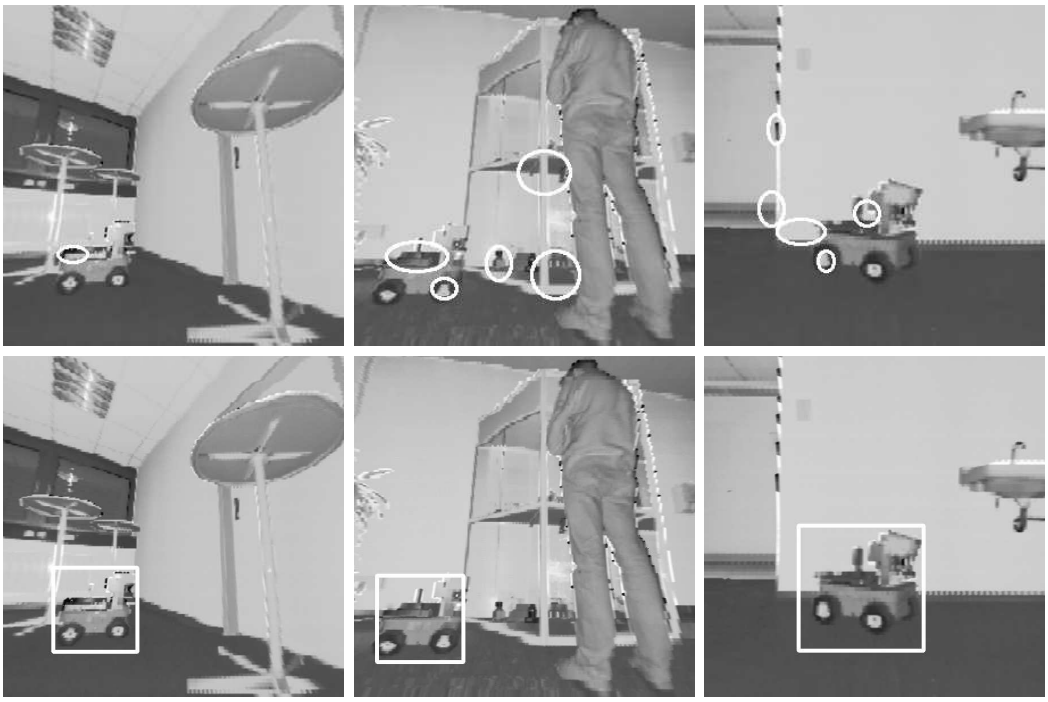

Fig. 14. Salient regions detected by the attention system serve as starting points for a classifier. Top row: The first resp. the first 5 foci of attention computed on depth and reflection data. Bottom row: Recognized objects in the focus regions.

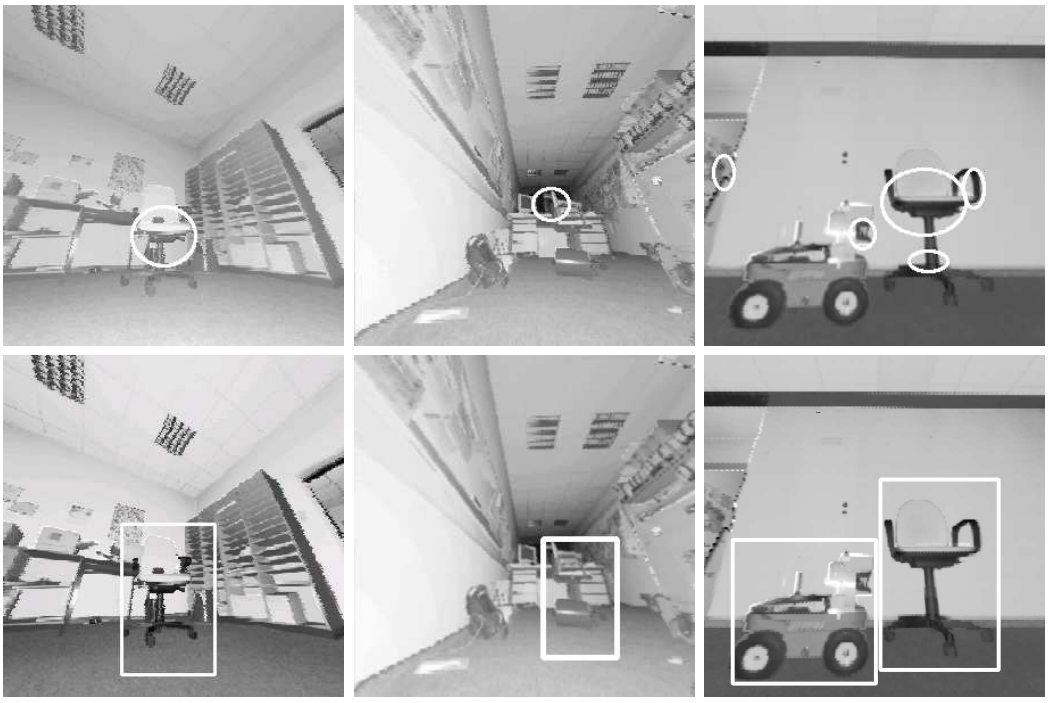

Fig. 15. Salient regions detected by the attention system serve as starting points for a classifier. Top row: The first resp. the first 5 foci of attention computed on depth and reflection data. Bottom row: Recognized objects in the focus regions.

An application scenario is presented, in which the attended regions serve as candidate regions for an object classification algorithm. This speeds up the object detection, since only a small part of the image needs to be processed by the classifier.

Furthermore, it has been demonstrated that range and reflection values complement each other: Some objects are salient in depth but not in reflection data and vice versa. The comparison between the 3D laser scanner and a camera as input sensors exhibited that their data also contain complementary features. In camera images, regions may be salient due to color contrast, which is not existent in laser data. On the other hand, laser data allow the detection of salient regions that cannot be identified in camera data. The ability to detect complementary features in the enviroment is considered crucial for robust and reliable object detection for robots.

7.2 Strengths and limitations

The approach presented in this paper offers an innovative idea to consider different sensor modalities for attentional mechanisms. This is the first step for the integration of multiple sensors for an attention system. The same way the two laser modes are fused, the system can be augmented to combine information of arbitrary sensors that provide the possibility to locate the sensor information in the environment. Not only camera data, even auditory information could be depicted in a map and searched for salient regions. However, the integration of different sensor information requires careful examination.

The laser scanner offers new possibilities for the detection of objects that can not or hardly be detected in camera images. This is done by taking into account so far unconsidered object properties. As the results show, there are situations in which a camera system fails to detect objects of low saliency. The 3D laser scanner is able to consider qualities like reflectance and depth discontinuities, enabling the detection of objects that are missed otherwise. Since the laser scanner does not provide color information, the best results could probably be achieved by combining the scanner data with those of a camera, utilizing the union of saliencies.

Furthermore, the scanner is independent of illumination variances. Different lighting conditions are a big problem in computer vision applications that rely on camera images. The laser scanner can be applied even in complete darkness, yielding the same results and providing a visual impression of the scene based on the reflection data. This can be an advantage in applications like surveillance in which the robot has to operate at night.

A limiting factor for the application of a scanning device in robot control is

the low scan speed. The minimum speed of the scanner is 1.7 seconds for a low resolution 3D scan. Therefore, data from other sensors have to be used for robot navigation in quickly changing environments. However, the attention system is not dependent on the sensor hardware and is equally able to process data from a 3D camera. On the other hand, the 3D scanner is well-suited for applications in low dynamics environments, like security inspection tasks in facility maintenance, interior survey of buildings and 3D digitalization.

As some of the laser images show, the raw data from the scanner is spherically distorted. The feature extraction methods are not designed for distorted images, so problems could arise. For example, straight lines in the environment are only mapped onto straight lines in the sensor data if they are located near the center of the image. At the borders, the distortion makes them curved, so they are not or only partly detected by the orientation filters. Two solutions are possible: Firstly, special filters could be designed for distorted images. This could be difficult, because the distortion is not evenly distributed. Secondly, the laser data could be rectified. The second approach was successfully examinated in later experiments as can be seen in Fig. 14 and 15.

One of the strengths of our attention system is that it is able to deal with real-world scenes (cf. sec. 5). Many attention systems are mainly applied to artificial images. This is much easier because these scenes usually do not contain as many details as real scenes, so the focus of attention is more likely to detect the desired objects.

Some limitations of the NVT of Itti et al. were shown by Draper et al. [3]. They point out that the system is not invariant to rotation, translation and reflection due to compromises in the implementation made for higher speed. Some of the suggestions of Draper are realized in our system too. Moreover, we have proposed some other improvements (cf. sec. 4). Currently, our system is still not completely invariant to the mentioned transformations, but it provides a robust solution for the detection of salient regions.

A general limitation of bottom-up attention systems is that they are only able to detect objects with high saliency. Our approach enables us to consider additional object properties enlarging the set of detectable objects. However, that does not help if the current interest is to detect an object with low saliency even concerning these properties or if there are other regions of very high saliency in the image, that attract the focus. This is not only true for laser data but also for camera images and even for human vision. Imagine driving a car at night. The attention is immediately attracted by the lights from the cars at the other side of the road. Only deliberately directing gaze and attention to the own side of the road enables safe driving. Similarly, top-down influences could enable the detection of desired objects in computational attention systems, a topic we consider for future work.

As the next step towards integrating more data modes, we plan to use the laser scanner together with an affordable ordinary camera to enable the simultaneous use of color, depth and reflectance information. The data modes will be searched in parallel for interesting regions and fused into a single saliency map. Due to the distortions of the laser data and the different fields of view of laser and camera, this fusion is not a trivial task and has to be examined carefully [29].

Concerning the visual attention system, there are many possible extensions and improvements. For example, there is psychological evidence that more than three features play an important part in the human attentional system [5]. Relevant features like motion, blob or region size, region shape etc. could be included in the model. The inclusion of motion would allow for tracking objects over time, but requires an extension of the system to cope with dynamic image sequences.

A major issue for future work will be the inclusion of top-down mechanisms into the model. This will help in several situations which are difficult for a bottom-up system: As mentioned above, the detection of objects with low saliency could be facilitated. Moreover, it would be possible to immediately focus on a relevant object, although there are several other salient regions in the image. For robot control, bottom-up attention is well-suited for exploring unknown environments; in contrast, top-down modulation utilizes knowledge of the environment in order to effectively search for expected objects or regions with known features.

The embedding of our attention system into the automatic building of semantically labelled 3D maps as described in section 6 will also be an important issue. To achieve this goal, the classifier has to be trained for further objects and the objects have to be localized not only in the 2D images, but also in the 3D point cloud.

Acknowledgements The authors wish to thank Laurent Itti for providing access to his *Neuromorphic Vision Toolkit (NVT).*

References

[1] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurobiology (1985) 219–227.

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

[3] B. Draper, A. Lionelle, Evaluation of selective attention under similarity transforms, in: Proceedings of the International Workshop on Attention and Performance in Computer Vision (WAPCV'03), Graz, Austria, 2003, pp. 31–38.

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

[5] A. Treisman, G. Gelade, A feature integration theory of attention, Cognitive Psychology 12 (1980) 97–136.

[6] H. Surmann, A. Nüchter, J. Hertzberg, An autonomous mobile robot with a 3D laser range finder for 3D exploration and digitalization of indoor environments, Journal Robotics and Autonomous Systems 45 (3-4) (2003) 181–198.

[7] J. Wolfe, K. Cave, S. Franzel, Guided search: An alternative to the feature integration model for visual search, J. of Experimental Psychology: Human Perception and Performance 15 (1989) 419–433.

[8] S. Zeki, A Vision of the Brain, Cambridge, Mass.: Blackwell Scientific., 1993.

[9] K. Lee, H. Buxton, J. Feng, Selective attention for cue-guided search using a spiking neural network, in: Proceedings of the International Workshop on Attention and Performance in Computer Vision (WAPCV'03), Graz, Austria, 2003, pp. 55–62.

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

[11] K. Kakayama, G. H. Silverman, Serial and parallel processing of visual feature conjunctions, Nature 320 (1986) 264–265.

[12] G. Backer, B. Mertsching, Integrating depth and motion into the attentional control of an active vision system, Baratoff, G.; Neumann, H. (eds.): Dynamische Perzeption. St. Augustin (Infix) (2000) 69–74.

[13] A. Maki, P. Nordlund, J.-O. Eklundh, Attentional scene segmentation: Integrating depth and motion, CVIU 78 (3) (2000) 351–373.

[14] N. Ouerhani, H. Hügli, Computing visual attention from scene depth, in: Proceedings of the 15th International Conference on Pattern Recognition, ICPR 2000, Vol. 1, IEEE Computer Society Press, 2000, pp. 375–378.

[15] L. Pessoa, S. Exel, Attentional strategies for object recognition, in: J. Mira, J. Sachez-Andres (Eds.), Proceedings of the IWANN, Alicante, Spain 1999, Vol. 1606 of Lecture Notes in Computer Science, Springer, 1999, pp. 850–859.

[16] F. Miau, C. Papageorgiou, L. Itti, Neuromorphic algorithms for computer vision and attention, in: Proc. SPIE 46 Annual International Symposium on Optical Science and Technology, Vol. 4479, 2001, pp. 12–23.

[17] J. Tsotsos, G. Verghese, S. Stevenson, M. Black, D. Metaxas, S. Culhane, S. Dickinson, M. Jenkin, A. Jepson, E. Milios, F. Nuflo, Y. Ye, R. Mann, Playbot: A visually-guided robot to assist physically disabled children in play, Image and Vision Computing 16, Special Issue on Vision for the Disabled (1998) 275–292.

[18] M. Bollmann, R. Hoischen, M. Jesikiewicz, C. Justkowski, B. Mertsching, Playing domino: A case study for an active vision system, in: H. Christensen (Ed.), Computer Vision Systems, Springer, 1999, pp. 392–411.

[19] L. Itti, The beobot platform for embedded real-time neuromorphic vision, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems, Vol. 15, Hardware Demo Track, MIT Press, Cambridge, MA, 2003.

[20] S. B. Nickerson, P. Jasiobedzki, D. Wilkes, M. Jenkin, E. Milios, J. Tsotsos, A. Jepson, O. N. Bains, The ARK project: Autonomous mobile robots for known industrial environments, Robotics and Autonomous Systems 25 (1-2) (1998) 83–104.

[21] C. Breazeal, A context-dependent attention system for a social robot, in: Proc. 16th Int'l Joint Conf. on Artifical Intelligence (IJCAI '99), Stockholm, Sweden, 1999, pp. 1146–1151.

[22] D. Hähnel, W. Burgard, S. Thrun, Learning compact 3D models of indoor and outdoor environments with a mobile robot, Robotics and Autonomous Systems 44 (1) (2003) 15–27.

[23] S. Thrun, D. Fox, W. Burgard, A real-time algorithm for mobile robot mapping with application to multi robot and 3D mapping, in: Proc. IEEE Int'l Conf. Robotics & Automation (ICRA '00), San Francisco, CA, USA, 2000, pp. 321–328.

[24] C. Früh, A. Zakhor, 3D model generation for cities using aerial photographs and ground level laser scans, in: Proc. Computer Vision & Pattern Recognition Conference (CVPR '01), Kauai, Hawaii, USA, 2001, p. 31ff.

[25] H. Zhao, R. Shibasaki, Reconstructing textured CAD model of urban environment using vehicle-borne laser range scanners and line cameras, in: 2nd Int'l Workshop on Computer Vision System (ICVS '01), Vancouver, Canada, 2001, pp. 284–295.

[26] P. Allen, I. Stamos, A. Gueorguiev, E. Gold, P. Blaer, AVENUE: automated site modelling in urban environments, in: Proc. 3rd Int'l Conf. on 3D Digital Imaging and Modeling (3DIM '01), Quebec City, Canada, 2001, pp. 357–364.

[27] A. Gueorguiev, P. Allen, E. Gold, P. Blaer, Design, architecture and control of a mobile site-modelling robot, in: Proc. IEEE Int'l Conf. on Robotics & Automation (ICRA '00), San Francisco, CA, USA, 2000, pp. 3266–3271.

[28] M. Hebert, M. Deans, D. Huber, B. Nabbe, N. Vandapel, Progress in 3-D mapping and localization, in: Proceedings of the 9th International Symposium on Intelligent Robotic Systems, (SIRS '01), Toulouse, France, 2001.

[29] V. Sequeira, K. Ng, E. Wolfart, J. Goncalves, D. Hogg, Automated 3D reconstruction of interiors with multiple scan–views, in: Proc. SPIE, Electronic Imaging '99, SPIE's 11th Annual Symposium, San Jose, CA, USA, 1999, pp. 106–117.

[30] H. Surmann, K. Lingemann, A. Nüchter, J. Hertzberg, A 3D laser range finder for autonomous mobile robots, in: Proc. 32nd Intl. Symp. on Robotics (ISR '01) (April 19–21, 2001, Seoul, South Korea), 2001, pp. 153–158.

[31] L. Itti, C. Koch, Feature combination strategies for saliency-based visual attention systems, Journal of Electronic Imaging 10 (1) (2001) 161–169.

[32] N. Ouerhani, R. von Wartburg, H. Hügli, R. Müri, Empirical validation of the saliency-based model of visual attention, in: Electronic Letters on Computer Vision and Image Analysis, Vol. 3, Computer Vision Center, 2004, pp. 13–24.

[33] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, Vision Research 42 (1) (2002) 107–123.

[34] P. Viola, M. J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.

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

[36] S. Frintrop, A. Nüchter, H. Surmann, Visual Attention for Object Recognition in Spatial 3D Data, in: accepted for International Workshop on Attention and Performance in Computer Vision (WAPCV 2004), Conference: Prag, Czech Republic, 2004.