Applying Attentional Mechanisms to Bi-modal 3D Laser Data

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

Abstract

In this paper we present experimental results on a novel application of visual attention mechanisms for the selection of points of interest in an arbitrary scene. The imaging sensor used is a multi-modal 3D laser scanner. In a single 3D scan pass, it is capable of providing range data as well as a gray-scale intensity image. The scanner is mounted on top of an autonomous mobile robot and serves control purposes. We present results achieved by applying the visual attention system of Itti et al. [8] to recorded scans of indoor and outdoor scenes. The vast majority of the primary attended locations pointed to scene objects of potential interest for navigation and object detection tasks. Moreover, both sensor modalities complement each other, resulting in a greater variety of points of interest than one modality alone can provide.

1. Introduction

Common tasks in the control of autonomous mobile robots are collision avoidance, navigation, and the manipulation of objects. In order to execute these tasks correctly, the robot needs to detect objects and free space in its environment fast and reliably. One method to find potential points of interest in the environment is to model human visual attention.

In human vision, attention helps to identify relevant data and so to efficiently select information from the broad sensory input. These effects are even more desired in computational applications like image processing. Our work is based on the model of visual attention by Itti et al. [8]. In this model, different features like intensity, color and orientation are evaluated in parallel and fused into a single saliency map that topographically codes salient locations in a visual scene. These locations can be analyzed later by object recognition modules and the information about found objects can help to accomplish robot control tasks.

This model, like many others, includes no depth feature, although in robotic applications depth is often employed for object detection tasks. Objects usually have range discontinuities at their borders which can help to detect them. Models comprising depth as a feature typically use stereo vision to compute it [10, 2]. But stereo vision is computationally expensive, and only a fraction of the image pixels contribute to the computed 3D point clouds.

As an alternative, 3D laser scanners are a class of sensors suitable for the fast acquisition of precise and dense depth or range information. The multi-modal 3D laser scanner used for our work [12] provides the technical means to acquire both range data as well as an intensity image of a scene in a single 3D scan pass. Since the data from the different sensor modalities result from the same measurement, we know exactly which remission or intensity value belongs to which range data. There is no need to establish correspondence by complex algorithms.

Such a multi-sensor offers new algorithmic possibilities. It is to be expected that range data and remission values complement each other, providing some redundancy that can be exploited. Contrasts in range and in intensity need not necessarily correspond for one object. That is, an object producing the same intensity like its background may not be detected in a gray-scale image, but probably in the range data. On the other hand a flat object on a flat background – e.g. a poster on a wall or a letter on a desk – that could be clearly distinguished in an intensity image, may be too flat to be detected in the range data.

In this paper we use the data from the 3D laser scanner as input to the attentional model by generating a special input image by combining range and intensity data in a suitable way. We show the applicability of the laser data for attentional mechanisms and compare these results to the ones from corresponding camera images. For our experiments, we used recorded scan data from indoor and outdoor scenes.

The 3D laser scanner is mounted on top of a robot (Fig. 1), and data acquisition is steered by the robot's CPU.

The remainder of this article is structured as follows. We start with analyzing the state of the art in robotic 3D scene imaging and in models of visual attention. Then we describe our system setup, that is, the multi-modal 3D scanner and our use of the visual attention system of Itti et al. In the main section we describe the acquisition and evaluation of the data and analyze the results. Finally, we summarize the arguments and give an outlook on future work.

2. State of the Art

Some groups have developed methods to build 3D volumetric representations of environments using 2D laser range finders. Thrun et al. [7], Früh et al. [6] and Zhao et al. [16] 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 [16]. 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 out edges. These obstacles are only detected while passing them.

A few other groups use true 3D laser scanners that are able to generate consistent 3D data points within a single 3D scan. The RESOLV project aimed at modelling interiors for virtual reality and tele-presence [11]. They employed a RIEGL laser range finder. The AVENUE project develops a robot for modelling urban environments [1]. This robot is equipped with an expensive CYRAX 3D laser scanner.

The multi-modal 3D laser range finder employed for this work [12] is a precise, fast scanning, reliable, and cost effective multi purpose sensor. Range data and remission images are acquired in one 3D scan pass. The interpretation of these data may require exhaustive time ressources. One approach to reduce these is to use attentional mechanisms that help to find regions of interest in the data.

Many computational models of human visual attention are based on the psychological work of Treisman et al., known as *feature-integration theory* [13], and on the *guided search* model by Wolfe [15]. The first explicit computational architecture for controlling visual attention was proposed by Koch and Ullman [9]. It already contains the main properties of the more elaborated model of Itti et al. [8] which forms the basis of our work. This model belongs to the group of *feature-based models* that use classical linear filter operations for feature extraction, what makes them especially useful to real-world scenes. Another approach provide the connectionist models, e.g. the *selective tuning model* by Tsotsos et al. [14].

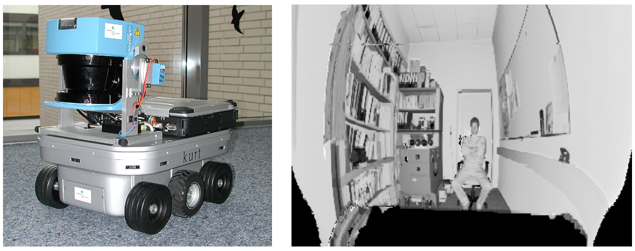

Figure 1. Left: The custom 3D range finder mounted on top of the mobile robot KURT 2. Right: An office scene imaged with the 3D scanner in remission value mode, medium resolution, 256 × 360 pixels.

Attentional systems using depth information can be found in [10] and [2], where stereo vision is applied to retrieve depth information. In robotics, attentional mechanisms are often used to direct the gaze (i.e. a camera) to interesting points in the environment. In [4], a robot shall look at people or toys and in [3] it uses attention to play at dominoes. However, the use of attentional mechanisms for robot control tasks is rarely considered.

3. Experimental Setup

3.1. The Multi-modal Custom 3D Laser Scanner

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 [12], the custom scanner setup is described in detail. The paper also describes reconstruction and scan matching algorithms and their use for robot applications. Here, we provide only a brief overview of the device.

A 3D scan is performed by step-rotating the 2D scanner around a horizontal axis by a total of up to 105 degrees. The 2D scanner is very fast (processing time about 13 ms for a 180° scan with 181 measurements) and precise (typical range error ~ 1 cm). A typical medium resolution 3D scan producing 256 layers of 362 values each, with an angular resolution of 0.5 degrees both horizontally and vertically, takes about 7.5 seconds and yields 92 672 data points. In lowest resolution mode, a 3D scan measures 128 × 180 points in about 1.75 seconds, in highest resolution mode, it yields 256 × 720 measurements in about 15 seconds.

The scanner can operate in two modes. In the usual mode, it returns range data in a predefined resolution. In an alternative mode, it is able to yield two different kinds of data for each measured point in a single scan pass: A distance value and a remission value that quantifies the intensity of the reflected laser light. The latter data type can directly be converted into a gray scale intensity image of the scanned parts of the scene. Compared to a normal camera image, the remission value image is spherically distorted (Fig. 1, right). This effect is due to the different measuring principle. The scanner has only one light sensitive element, which receives a beam reflected by a rotating mirror (for horizontal scanning), and is then rotated around a horizontal axis (for vertical scanning).

The current scanner software processes only the range data yielding, e.g., a rough 3D surface approximation of the environment [12]. For robot control tasks this needs to be processed further in order to identify objects and free space for navigation. The attention system shall be used to quickly identify points of interest as starting points for further segmentation and object detection tasks.

3.2. The Artificial Visual Attention System

Our experiments are based on an available implementation of the attentional model of Itti et al. [8] (large central box in Fig. 2). In their model, input is provided in the form of static color images taken from any kind of camera. The input is first decomposed into a set of topographic feature maps usually using intensity, color and orientation as feature dimensions. Each feature is computed by a set of linear *center-surround* operations, which are particularly well-suited to detecting locations which locally stand out from their surroundings. The feature maps are fed into a master *saliency map* (SM) which combines salient points from all feature maps. The most salient point in this map is found by a winner take all network (WTA). After shifting the focus of attention (FOA) to this location, local inhibition is activated in the SM, in the area of the FOA. This mechanism models the *inhibition of return* (IOR) phenomenon observed in humans. It yields dynamical shifts of the FOA and prevents the FOA from immediately returning to a previously attended location.

In our system (Fig. 2), input images are generated from the range data and remission values provided by the 3D scanner. The latter ones can immediately be used as a grayscale image in a straightforward manner. The depth values from the range data require a more elaborate transformation. The basic approach is to interpret the depth values of the range data as intensity values, representing small depth values as light 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:

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

with $I = 255$ denoting the maximum intensity value and $max = 1000$ the maximum distance in cm. This measure leads to a finer distinction of range discontinuities in the vicinity of the robot and works better than a linear function.

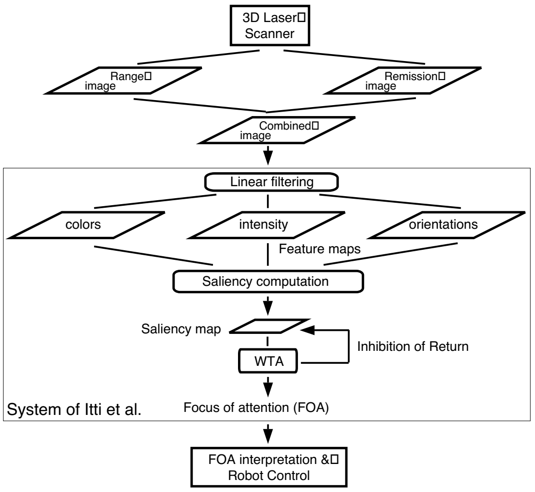

Figure 2. Simplified system architecture diagram

In order to generate salient locations from both sets of scanner data in one processing stage, we fuse a range and a remission image into one colorized image which serves as input to the model of Itti et al. To accomplish this, the transformed range data are treated as intensity values of the new input image, but the remission values are coded as color (hue) information of the new input image, i.e. we utilize the so far unused color feature dimension. High intensities are coded in red hues, low intensities in greens (Fig. 3). This results in suitable color images because the color feature map takes into account blue-yellow contrasts as well as red-green contrasts. This mapping enables us to utilize the attentional system as is without the need to adapt it. It would be also possible to code remission as intensity and depth as color but for implementation reasons our versions yields slightly better results. Of course, it is also still possible to use only either the range or the remission image as input to the system.

4. Results

We have tested our approach on scans of indoor as well as outdoor scenes. In Fig 3, we show a remission image, a

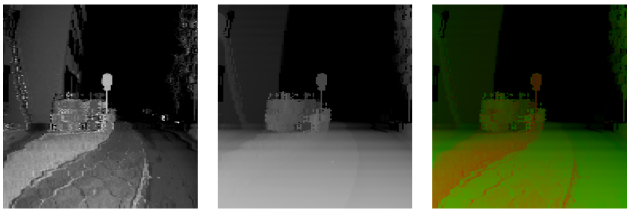

Figure 3. From left to right: A remission image, the corresponding range image and the combined colorized image. In gray-scale its hardly possible to see the red-green components from the remission image, so the colorized image on the right is very similar to the range image in the middle.

range image, and the combined color image. Since the color image does not print well in gray-scale, we have decided to display only the separately processed range and remission images in later figures (Fig. 4, 5). The circles indicate the FOAs, showing only the first resp. the first two FOAs. If there is more than one FOA shown, the arrow indicates their order of detection.

In a first test set we show the general applicability of the 3D laser scanner as a sensor for attentional mechanisms (Fig. 4). In the remission images (left column) the presented objects are detected at the first try in almost all cases except in the first row, where the cable reel is not regarded as an attentional point. In the range images (middle column), the FOAs always find the presented objects, because they are well separated from the background. These results show, that data from the 3D-laser scanner is generally well-suited to find attentional points as long as objects have a popping-out intensity or a certain distance to their background. More difficult is the detection of objects in crowded scenes like office environments (cf. Fig. 1) where these conditions are often not met.

The second test set shows a comparison of attentional points in either laser sensor mode and in corresponding camera images (Fig. 5). The columns show, from left to right, the remission, range and camera images each with some FOAs. The first row shows an example where all modes obtain the same results: the traffic sign is always detected by the first FOA.

An advantage of the laser data over the camera image can be found in row 2, where the traffic sign is easily detected by remission and range image but missed in the camera image. However the fire extinguisher in row 4 is only detected in the camera image because of its red color (1. FOA).

The complementary effect of the two laser modes is illustrated in rows 3 and 4. In row 3, the 2nd FOA in the remission image is placed on the icon on the ground marking a parking space for disabled people, which is impossible to be detected in the depth image. On the other hand, range shows its value in the last row, where the cable reel is detected immediately but missed completely in the remission image. The presented examples show in a convincing way the advantage of having different sensor modes at hand.

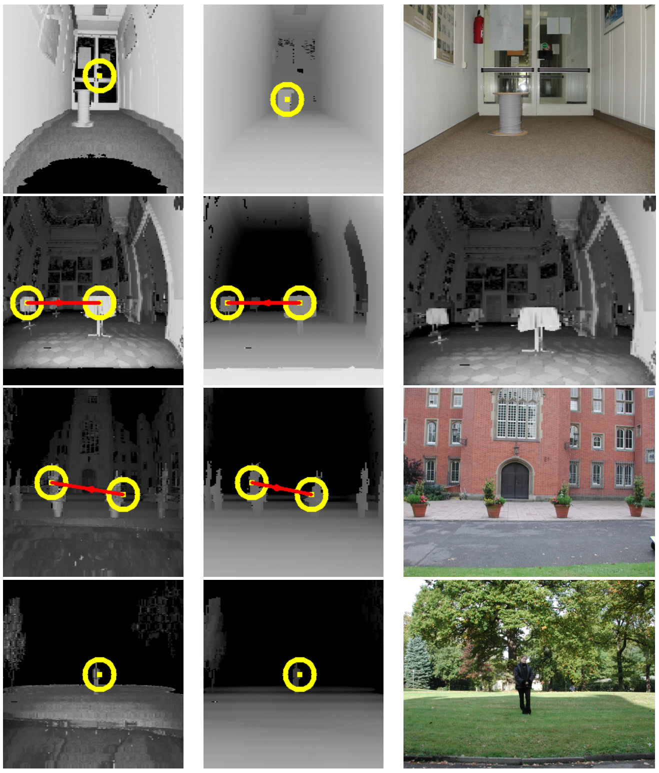

Figure 4. Test result set 1: Attentional points in laser data. Columns from left to right: Laser remission images with attentional points, laser range images with attentional points, reference scene images.

To evaluate the performance of our system, we generated the first 5 FOAs of 15 scenes and tested whether they showed an object of potential interest (OPI).

An OPI is an object the robot could derive benefit from. The evaluation whether an object is an OPI is highly task dependent and can only really be determined after a task was specified. If the task is the recognition of traffic signs, all traffic signs are defined as OPIs, if it is obstacle detection, possible obstacles like the cable reel are OPIs. If a task is specified the detection of the OPIs can be improved by increasing the features relevant to the current task. In the general case we are only able to define those objects as OPIs that are important for robot navigation, e.g. obstacles like the cable reel or landmarks like the fire extinguisher.

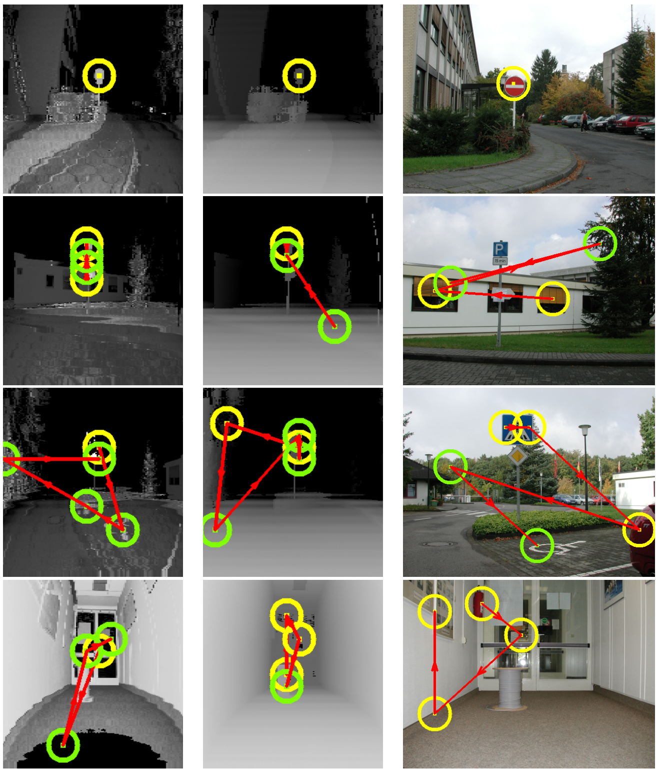

Figure 5. Test result set 2: Comparison of attentional points in camera images and in laser data. Columns from left to right: Laser remission images, laser range images, camera images. 1st row: same results in all modes. 2nd row: traffic sign not found in camera image. 3rd row: lower traffic sign not found in camera image, handicapped person sign not found in range image. 4th row: fire extinguisher only found in camera image, cable reel only found in range image.

The FOAs were generated for remission and range images yielding $5 * (2 * 15) = 150$ FOAs. Of these FOAs, 81 (54%) pointed to OPIs. Regarding this results one should consider that in most of our scenes the number of salient spots that generate FOAs is greater than the number of OPIs, so usually not all foci point to OPIs. Furthermore, if there is only one OPI in a scene, at most three of the first five FOAs can lie on an OPI, because IOR forces the focus to go away from the OPI before it can return again.

Tab. 1 shows the distribution of the FOAs. Of the first attended locations, 86% show a potential object of interest. The 2nd and 3rd ranked FOAs find less objects, because there are frequently only one or two objects in the scene. The 4th and 5th FOAs have higher values than the 3rd one, because an inhibited object often attracts the focus again after a while.

Table 1. Distribution of the 5 most salient FOAs on a test set of 30 images.

Number of the FOA (decreasing saliency)	Attended objects of potential interest [%]
1.	86
2.	56
3.	30
4.	53
5.	43

5. Conclusion and Outlook

We have introduced a new application of visual attention algorithms for robot control purposes. The input images for the attentional system were provided by a 3D laser scanner yielding range as well as remission data. We have shown that the attention system is able to generate a high number of salient locations, both in indoor and outdoor real-world scenes. Furthermore, it has been demonstrated that under certain conditions range and remission values complement each other.

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. On the other hand, the 3D scanner is well-suited for applications in low dynamics environments, like security inspection tasks in facility maintenance and interior survey of buildings.

The duration of a 3D scan is determined by the scan mechanics, because only one point at a time is scanned and the data is acquired serially. A much faster way to acquire range and remission values are 3D laser “cameras”, that use a sensor array to measure these values in parallel. Several research prototypes of 3D laser “cameras” are known, e.g. at KTH [5] and at DaimlerChrysler. These devices are currently very expensive, are mostly restricted to shorter ranges and extremely low resolutions, and usually yield results that are less precise than those of a laser scanner.

The multi-modality of the 3D laser scanner offers new algorithmic possibilities, like the one shown in this paper. The remission values are dependent only on the laser energy, i.e. they are independent of current lighting conditions of the scene, which is another advantage over camera systems. More expensive laser range finders are able to yield up to two more data qualities, namely color and temperature. These data could also be fed into the attention model, and it is to be expected that this setup would yield even more complementary regions of interest.

As 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 intensity information. The extended data modes will be searched in parallel for interesting regions, which would then be fused into a single saliency map. Due to the distortions of the laser images and the different fields of view of laser and camera this fusion is not a trivial task and has to be examined carefully.

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 [13]. 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 the extension of the system to cope with dynamic image sequences.

A major application issue will be the decision on what to do with the detected image parts. There are roughly two possible answers: Firstly, an object recognition module could be used to analyze the found regions and compare the extracted data with an object data base (landmarks, things that can be grasped). Secondly, the information about interesting parts could be used to perform active vision. For example, the regions of potential interest could be zoomed in by a camera or the robot could be steered into the direction of a region that has been identified as a goal object. These methods could support robot control tasks like collision avoidance, navigation and manipulation of objects.

Future work will also concentrate on the inclusion of top-down mechanisms into the model. This approach would allow for adapting the attentional mechanisms to the robot's tasks. For robot control, bottom-up attention is well-suited for exploring unknown environments; in contrast top-down modulation adapts the generation of FOAs in order to effectively search for expected objects with known features.

Acknowledgements The authors wish to thank Laurent Itti for providing access to his attention modelling software. This support is gratefully acknowledged.

References

[1] P. Allen, I. Stamos, A. Gueorguiev, E. Gold, and 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, May 2001.

[2] G. Backer and 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)*, pages 69–74, 2000.

[3] M. Bollmann. *Entwicklung einer Aufmerksamkeitssteuerung für ein aktives Sehsystem*. PhD thesis, Universität Hamburg, 1999.

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

[5] T. E. Carlsson, J. Gustafsson, and B. Nilsson. Development of a 3d camera. In S. A. Benton, editor, *Practical Holography XIII*, volume 3637 of *Proc. SPIE*, pages 218–224, 1999.

[6] C. Früh and 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, December 2001.

[7] D. Hähnel, W. Burgard, and S. Thrun. Learning Compact 3D Models of Indoor and Outdoor Environments with a Mobile Robot. In *Proc. 4th European Workshop on Advanced Mobile Robots (EUROBOT '01)*, Lund, Sweden, September 2001.

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

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

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

[11] V. Sequeira, K. Ng, E. Wolfart, J. Goncalves, and 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.

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

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

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

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

[16] H. Zhao and 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)*, pages 284 – 295, Vancouver, Canada, July 2001.