# A Nonparametric Learning Approach to Range Sensing from Omnidirectional Vision

Christian Plagemann[a], Cyrill Stachniss[b], Jürgen Hess[b],
Felix Endres[b], Nathan Franklin[b]

[a] *Stanford University, Computer Science Dept., 353 Serra Mall, Stanford, CA 94305-9010*
[b] *University of Freiburg, Dept. of CS, Georges-Koehler-Allee 79, 79110 Freiburg, Germany*

**Abstract**

We present a novel approach to estimating depth from single omnidirectional camera images by learning the relationship between visual features and range measurements available during a training phase. Our model not only yields the most likely distance to obstacles in all directions, but also the predictive uncertainties for these estimates. This information can be utilized by a mobile robot to build an occupancy grid map of the environment or to avoid obstacles during exploration—tasks that typically require dedicated proximity sensors such as laser range finders or sonars. We show in this paper how an omnidirectional camera can be used as an alternative to such range sensors. As the learning engine, we apply Gaussian processes, a nonparametric approach to function regression, as well as a recently developed extension for dealing with input-dependent noise. In practical experiments carried out in different indoor environments with a mobile robot equipped with an omnidirectional camera system, we demonstrate that our system is able to estimate range with an accuracy comparable to that of dedicated sensors based on sonar or infrared light.

*Key words:* omnidirectional vision, learning, range sensing, Gaussian processes

*Preprint submitted to RAS - Special Issue on Omnidirectional Robot Vision December 18, 2009*

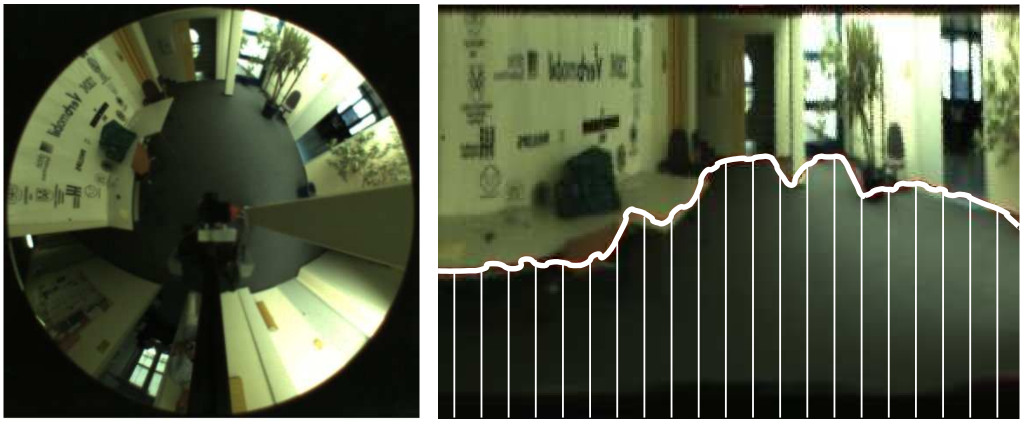

Figure 1: Our system records intensity images (left) and estimates the distances to nearby obstacles (right) after having learned how visual appearance is related to depth.

## 1. Introduction

The major role of perception, in humans as well as in robotic systems, is to discover geometric properties of the current scene in order to act in it reasonably and safely. For artificial systems, omnidirectional vision provides a rich source of information about the local environment, since it captures the entire scene—or at least the most relevant part of it—in a single image. Much research has thus concentrated on the question of how to extract geometric scene properties, such as distances to nearby objects, from such images.

This task is complicated by the fact that only a projection of the scene is recorded and, thus, it is not possible to sense depth information directly. From a geometric point of view, one needs at least two images taken from different locations to recover the depth information analytically. An alternative approach that requires just one monocular camera image and that we follow here, is to learn from previous experience how visual appearance is related to depth. Such an ability is also highly developed in humans, who are able to utilize monocular cues for depth perception [1]. As a motivating example, consider the right image in Figure 1, which shows the image of an office environment (180° of the omnidirectional image on the left warped to a panoramic view). Overlayed in white, we visualize the most likely area of free space that is predicted by our approach. We explicitly do not try to estimate a depth map for the whole image, as for example done by Saxena *et al.* [2]. Rather, we aim at learning the function that, given an image, maps measurement directions to their corresponding distances to the

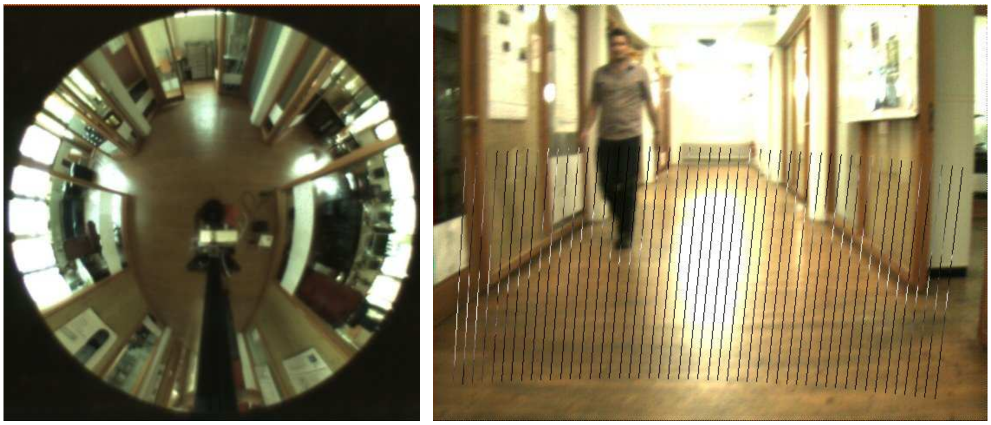

Figure 2: Reflections, glass walls and inhomogeneous surfaces make the relationship between visual appearance and depth hard to model. One of the test environments at the University of Freiburg (left) exhibits many of these factors. Our approach was also tested using a standard perspective camera in this challenging environment (right).

closest obstacles. Such a function can be utilized to solve various tasks of mobile robots including local obstacle avoidance, localization, mapping, exploration, or place classification.

The contribution of this paper is a new approach to range estimation based on omnidirectional images. The task is formulated as a supervised regression problem in which the training set is acquired by combining image date with proximity information provided by a laser range finder. We explain how to extract appropriate visual features from the images using algorithms for edge detection as well as for supervised and unsupervised dimensionality reduction. As a learning framework in our proposed system, we apply Gaussian processes since this technique is able to model non-linear functions, offers a direct way of estimating uncertainties for its predictions, and it has proven successful in previous work involving range functions [3].

The paper is organized as follows. First, we discuss related work in Section 2. Section 3 introduces the used visual features and how they can be extracted from images. We then formalize the problem of predicting range from these features and introduce the proposed learning framework in Section 4. In Section 5, we present the experimental evaluation of our algorithm as well as an application to the mapping problem.

## 2. Related Work

Estimating the geometry of a scene based on visual input is one of the fundamental problems in computer vision and is also frequently addressed in the robotics literature. Monocular cameras do not directly provide 3D information and therefore stereo systems are widely used to estimate the missing depth information. Stereo systems either require a careful calibration to analytically calculate depth using geometric constraints or, as Sinz *et al.* [4] demonstrated, can be used in combination with non-linear, supervised learning approaches to recover depth information. Often, sets of features such as SIFT [5] or SURF [6] are extracted from two images and matched against each other. Then, the feature pairs are used to constrain the poses of the two camera locations and/or the point in the scene that corresponds to the image feature. In this spirit, the motion of a single camera has been used by Davidson *et al.* [7] and Strasdat *et al.* [8] to estimate the location of landmarks in the environment. In their work, Mikusic and Padjla [9] have proposed a similar approach for recovering 3D structure from sequences of omnidirectional images.

Sim and Little [10] present a stereo-vision based approach to the SLAM problem, which includes the recovery of depth information. Their approach contains both the matching of discrete landmarks and dense grid mapping using vision.

An active way of sensing depth using a single monocular camera is known as *depth from defocus* [11] or *depth from blur*. Such approaches typically adjust the focal length of the camera and analyze the resulting local changes in image sharpness. Torralba and Oliva [12] present an approach for estimating the mean depth of full scenes from single images using spectral signatures. While their approach is likely to improve a large number of recognition algorithms by providing a rough scale estimate, the spatial resolution of their depth estimates does not appear to be sufficient for the problem studied in this paper. Dahlkamp *et al.* [13] learn a mapping from visual input to road traversability in a self-supervised manner. They use the information from laser range finders to estimate the terrain traversability locally and then use visual data to extend the prediction to areas outside the field of view of the laser range scanners. In contrast to our method, the laser range data is used at all times since learning is not a separated process as in this paper. Furthermore, different learning techniques and different features have been applied.

The problem addressed by Saxena *et al.* [2] is closely related to our paper. They utilize Markov random fields (MRFs) for reconstructing dense depth maps from single monocular images. Compared to these methods, our Gaussian process formulation provides the predictive uncertainties for the depth estimates directly, which is not straightforward to achieve in an MRF model. An alternative approach that predicts 2D range scans using reinforcement learning techniques has been presented by Michels *et al.* [14]. Menegatti *et al.* [15] proposed to simulate range scans from detected color transitions in omnidirectional images and to apply scan-matching and Monte-Carlo methods for localizing a mobile robot. Such color transitions are comparable to our set of edge-based features described in Section 3.3, which form the low-level input to the learning approach described in this paper.

Hoiem *et al.* [16] developed an approach to monocular scene reconstruction based on local features combined with global reasoning. Whereas Han and Zhu [17] presented a Bayesian method for reconstructing the 3D geometry of wire-like objects in simple scenes, Delage *et al.* [18] introduced an MRF model on orthogonal plane segments to recover the 3D structure of indoor scenes. Ewert *et al.* [19] extract depth cues from monocular image sequences in order to facilitate image retrieval from video sequences. Their major cue for feature extraction is the motion parallax. Thus, their approach assumes translational camera motion and a rigid scene.

In own previous work [3], we applied Gaussian processes to improve sensor models for laser range finders. In contrast to that, the goal here is to exchange the highly accurate and reliable laser measurements by noisy and ambiguous vision features.

As mentioned above, one potential application of the approach described in this paper is to learn occupancy grid maps. This type of maps and an algorithm to update such maps based on ultrasound data has been introduced by Moravec and Elfes [20]. In the past, different approaches to learn occupancy grid maps from stereo vision have been proposed [21, 22]. If the positions of the robot are unknown during the mapping process, the entire task turns into the so-called simultaneous localization and mapping (SLAM) problem. Vision-based techniques have been proposed by Elinas *et al.* [23] and Davison *et al.* [7] to solve this problem. In contrast to the mapping approach presented in this paper, these techniques mostly focus on landmark-based representations.

The contribution of this paper is a novel approach to estimating the proximity to nearby obstacles in indoor environments from a single camera image. It is an extension of our recent conference paper [24] that first presented the idea of estimating depth from camera images using GP regression. The work presented here additionally considers supervised dimensionality reduction, namely LDA, which allows us to find a low dimensional space in which feature vectors corresponding

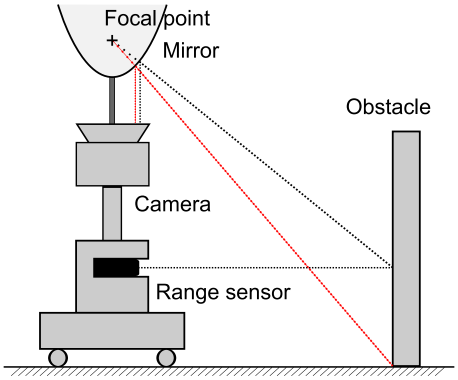

Figure 3: Our experimental setup. The training set was recorded using a mobile robot equipped with an omnidirectional camera (monocular camera with a parabolic mirror) as well as a laser range finder.

to different range measurements are better separated. In this way, the Gaussian process is able to provide better estimates about predicted ranges.

## 3. Omnidirectional Vision and Feature Extraction

The task of estimating range information from images requires us to learn the relationship between visual input and the extent of free space around the robot. Figure 3 depicts the configuration of our robot used for data acquisition. An omnidirectional camera system (catadioptric with a parabolic mirror) is mounted on top of a SICK laser range finder. This setup allows the robot to perceive the whole surrounding area at every time step as the two example images in Figure 2 illustrate. It furthermore enables the robot to collect proximity data from the laser range finders and relate them to the image data. As a result, our robot can easily acquire training data used in the regression task. The left images in Figure 1 and Figure 2 show typical situations from the two benchmark data sets used in this paper. They have been recorded at the University of Freiburg (Figure 1) and at the German Research Center for Artificial Intelligence (DFKI) in Saarbrücken (Figure 2). By considering these example images, it is clear that the part of an omnidirectional image which is most strongly correlated with the distance to the nearest obstacle in a certain direction $\alpha$ is the strip of pixels oriented in the same direction covering the area from the center of the image to its margins. With the type of camera used in our experiments, such strips have a dimensionality of 420 (140 pixels, each having a *hue*, *saturation*, and a *value* component). To make these strips easily accessible to filter operators, we warp the omnidirectional images into panoramic views (e.g., the right image in Figure 2) so that angles in the polar representation now correspond to column indices in the panoramic one. This transformation allows us to replace complicated image operations in the polar domain by easier and more robust ones in a Cartesian coordinate system.

In the following, we denote with $\mathbf{x}_i \in \mathbb{R}^{420}$ the individual pixel columns of an image and with $y_i \in \mathbb{R}$ the range values in the corresponding direction, that is, the distances to the closest obstacles, respectively. Before describing how to learn the relationship between the variables $\mathbf{x}$ and $y$, we discuss three alternative ways of extracting meaningful low-dimensional features $\mathbf{v}$ from $\mathbf{x}$ which can be utilized by the learning algorithm. The first approach applies unsupervised dimensionality reduction (PCA) to compute low-dimensional features. As an alternative, we also consider the linear discriminant analysis (LDA) as an supervised alternative to obtain low-dimensional features. Finally, we discuss the use of manually designed features extracted from the images that can be used for range prediction.

### *3.1. Unsupervised Dimensionality Reduction*

Principal component analysis (PCA) is arguably the most common approach to dimensionality reduction. We apply PCA for reducing the complexity of the data to the raw 420-dimensional pixel vectors that are contained in the columns of the panoramic images. In our approach, the PCA is implemented using eigenvalue decomposition of the covariance matrix of the 420-dimensional training vectors. PCA computes a linear transformation that maps the input vectors onto a new basis so that their dimensions are ordered by the amount of variance of the data set they carry. By selecting only the first $k$ vectors of this basis representing the dimensions with the highest variance in the data, one obtains a low-dimensional representation without losing a large amount of information. The left diagram in Figure 4 shows the remaining fraction of variance after truncating the transformed data vectors after a certain number of components. The right diagram in the same figure shows the 420 components of the first eigenvector for the Freiburg data set grouped by hue, saturation, and value. Our experiments revealed that the value channel of the visual input is more important than hue and saturation for our task.

For the experiments reported on in Section 5, we trained the PCA on 600 input images and retained the first six principal components. This results in a reduction from 420-dimensional input vectors to 6-dimensional ones. The GP model, described in the Section 4, is then learned with these 6D features and is named *PCA-GP* in the experimental section.

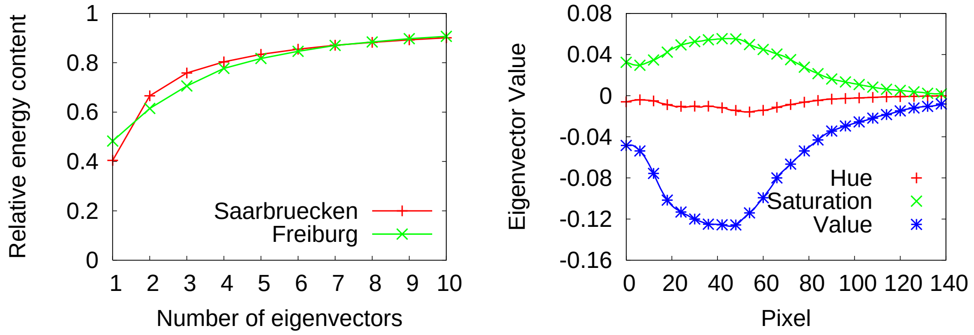

Figure 4: Left: The amount of variance explained by the first principal components (eigenvectors) of the pixel columns in the two data sets. Right: The 420 components of the first eigenvector of the Freiburg data set.

### *3.2. Supervised Dimensionality Reduction*

A drawback of PCA in our regression task is the fact that it does not consider the range values $y_i$ when reducing the dimensionality of the input vectors $\mathbf{x}_i$. In this way, it treats all components of the input vectors equally—no matter how much information they actually carry about the range to be predicted. It can thus be expected that *supervised* dimensionality reduction, where external, dependent variables are considered explicitly, can lead to more accurate predictions. See Alpaydin [25] for an overview of approaches and comparisons. One such technique is the linear discriminant analysis (LDA). LDA is related to PCA in that it also assumes a linear transformation between the original space and the reduced one. But in contrast to PCA, it allows each data point to be given a class label. LDA seeks a low-dimensional space in which the classes of the dataset are separated best as illustrated in Figure 5 for a reduction from $\mathbb{R}^2$ to $\mathbb{R}$.

Let $K$ be the number of classes $\mathcal{C}_i$ and $\mathbf{x}_i$ the $d$-dimensional inputs. The objective is to find a $d \times k$ matrix $W$ so that $\mathbf{v}_i = W^T\mathbf{x}_i$ with $\mathbf{v}_i \in \mathbb{R}^k$ and so that the classes $\mathcal{C}_i$ are separated best in terms of distances between the $\mathbf{v}_i$. Let $r_{i,t}$ be an indicator variable with $r_{i,t} = 1$ if $\mathbf{x}_t \in \mathcal{C}_i$ and 0 otherwise. Let $\mathbf{m}_i$ be the mean of $d$-dimensional vectors $\mathbf{x}_i$. Then, the so-called *scatter matrix* of $\mathcal{C}_i$ is

$$S_i \;=\; \sum_t r_{i,t}(\mathbf{x}_t - \mathbf{m}_i)(\mathbf{x}_t - \mathbf{m}_i)^T\,,$$

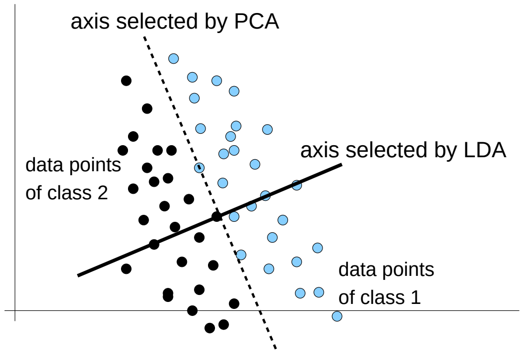

Figure 5: Reduction from $\mathbb{R}^2$ to $\mathbb{R}$ for PCA and LDA: PCA aims to keep the variance in the data while LDA seeks to separate the two classes (illustrated by black and blue) as well as possible.

the *total within-scatter matrix* becomes

$$S_W = \sum_{i=1}^{K} S_i = \sum_{i=1}^{K} \sum_{t} r_{i,t} (\mathbf{x}_t - \mathbf{m}_i)(\mathbf{x}_t - \mathbf{m}_i)^T ,$$

and the *between-class scatter matrix* is

$$S_B = \sum_{i=1}^{K} \left( \sum_{t} r_{i,t} \right) (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T ,$$

with $\mathbf{m} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{m}_i$. Now let us consider the scatter matrices after projecting using $W$. The between-class scatter matrix after projection is $W^T S_B W$, and the within-scatter matrix accordingly, both are $k \times k$ dimensional. To goal is to determine $W$ in a way that the means of the classes $W^T \mathbf{m}_i$ are as far apart from each other as possible while the spread of their individual projected class samples is small. Similarly to covariance matrices, the determinant of a scatter matrix characterizes the spread and it is computed as the product of the eigenvalues specifying the variance along the eigenvectors. Thus, we aim at finding the matrix $W$ that maximizes

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}.$$

The largest eigenvectors of $S_W^{-1} S_B$ are the solution to this problem.

Applied to the range-regression task, we selected the discretized laser range measurements as class label for each input data point. LDA then projects to a low-dimensional space so that data points corresponding to the discretized range measurements can be separated best. The GP model learned with the low-dimensional features is named *LDA-GP* in our experimental evaluation.

### *3.3. Edge-based Features*

Principal component analysis is an unsupervised method that does not take into account any prior information and also the linear discriminant analysis only uses information about class labels to perform dimensionality reduction to keep the data separated. However, there might be additional information available about the task to be solved—like the fact that distances to the closest obstacles are to be predicted in our case. Driven by the observation that there typically is a strong correlation between the extent of free space and the presence of horizontal edge features in the panoramic image, we also assessed the potential of edge-type features in our approach.

To compute such visual features from the warped images, we apply Laws' convolution masks [26]. They provide an easy way of constructing local feature extractors for discretized signals. The idea is to define three basic convolution masks

- $L_3 = (1, 2, 1)^T$ (Weighted Sum: Averaging),
- $E_3 = (-1, 0, 1)^T$ (First difference: Edges), and
- $S_3 = (-1, 2, -1)^T$ (Second difference: Spots),

each having a different effect on (1D) patterns, and to construct more complex filters by a combination of the basic masks. In our application domain, we obtained good results with the (2D) directed edge filter $E_5 L_5^\top$, which is the outer product of $E_5$ and $L_5$. Here, $E_5$ is a convolution of $E_3$ with $L_3$ and $L_5$ denotes $L_3$ convolved with itself. After filtering with this mask, we apply an optimized threshold to yield a binary response. This feature type is denoted as *Laws5* in the experimental section. As another well-known feature type, we applied the $E_3 L_3^\top$ filter, i.e. the Sobel operator, in conjunction with Canny's algorithm [27]. This filter yields binary responses at the image locations with maximal gray-value gradients in gradient direction. We denote this feature type as *Laws3+Canny* in Section 5. For both edge detectors, *Laws5* and *Laws3+Canny*, we search along each image

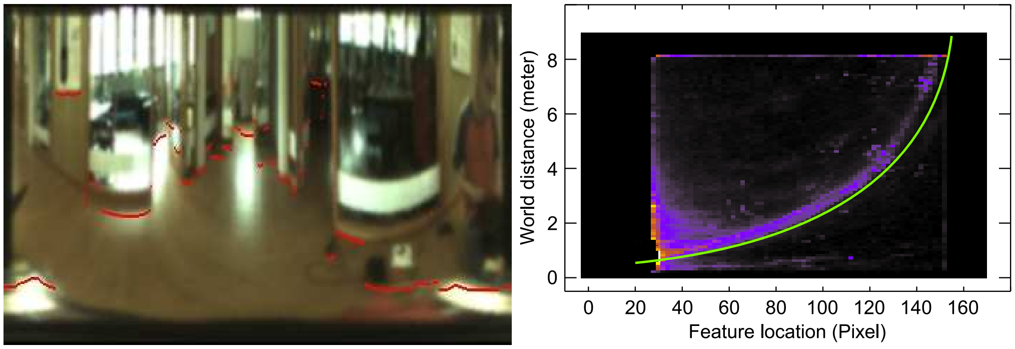

Figure 6: Left: Example *Laws5+LMD* feature extracted from one of the Freiburg images. Right: Histogram for *Laws5+LMD* edge features. Each cell in the histogram is indexed by the pixel location of the edge feature (x-axis) and the length of the corresponding laser beam (y-axis). The optimized (parametric) mapping function that is used as a benchmark in our experiments is overlaid in green.

column for the first detected edge. This pixel index then constitutes the feature value.

To increase the robustness of the edge detectors described above, we applied *lightmap damping* as an optional preprocessing step to the raw images. This means that, in a first step, a copy of the image is converted to gray scale and strongly smoothed with a Gaussian filter, such that every pixel represents the brightness of its local environment. This is referred to as the *lightmap*. The brightness of the original image is then scaled with respect to the lightmap, such that the *value* component of the color is increased in dark areas and decreased in bright areas. In the experimental section, this operation is marked by adding *+LMD* to the feature descriptions. Figure 6 shows *Laws5+LMD* edge features extracted from an image of the Freiburg data set.

All parameters involved in the edge detection procedures described above were optimized to yield features that lie as close as possible to the laser end points projected onto the omnidirectional image using the acquired training set. For our regression model, we can now construct 4D feature vectors $\mathbf{v}$ consisting of the Canny-based feature, the *Laws5*-based feature, and both features with additional preprocessing using lightmap-damping. Since every one of these individual features captures slightly different aspects of the visual input, the combination of all, in what we call the *Feature-GP*, can be expected to yield more accurate predictions than any single one.

As a benchmark for predicting range information from edge features, we also evaluated the individual features directly. For doing so, we fitted a parametric function to training samples of feature-range pairs. This mapping function computes for each pixel location of an edge feature the length of the corresponding laser beam. The right diagram in Figure 6 shows the feature histogram for the *Laws5+LMD* features from one of our test runs that was used for the optimization. The color of a cell $(c_x, c_y)$ in this diagram encodes the relative amount of feature detections that were extracted at the pixel location $c_x$ (measured from the center of the omnidirectional image) and that have a corresponding laser beam with a length of $c_y$ in the training set. The optimized projection function is overlayed in green.

## 4. Learning Depth from Images

This section presents the learning method used in our approach to find the relationship between visual input and the free space around the robot. Given a training set of images and corresponding range scans acquired in a setting, we can treat the problem of predicting range in *new* situations as a supervised learning problem. The omnidirectional images can be mapped directly to the laser scans since both measurements can be represented in a common, polar coordinate system. Note that our approach is not restricted to omnidirectional cameras in principle. However, the correspondence between range measurements and omnidirectional images is a more direct one and the field of view is considerably larger compared to standard perspective optics.

### *4.1. Gaussian Processes for Range Predictions*

In the spirit of the Gaussian beam processes (GBP) model introduced in [3], we propose to put a Gaussian process (GP) prior on the range function, but in contrast, here we use the visual features $\mathbf{v}$ described in the previous section as indices for the range values rather than the bearing angles $\alpha$.

We extract for every viewing direction $\alpha$ a vector of visual features $\mathbf{v}$ from an image $\mathbf{c}$ and phrase the problem as learning the range function $f(\mathbf{v}) = y$ that maps the visual input $\mathbf{v}$ to distances $y$. We learn this function in a supervised manner using a training set $\mathcal{D} = \{\mathbf{v}_i, y_i\}_{i=1}^n$ of observed features $\mathbf{v}_i$ and corresponding laser range measurements $y_i$. If we place a GP prior (see, e.g., [28]) on the non-linear function $f$, i.e., we assume that all range samples $y$ indexed by their corresponding feature vectors $\mathbf{v}$ are jointly Gaussian distributed, we obtain

$$y_* \;=\; f(\mathbf{v}_*) \;\sim\; \mathcal{N}(\mu_*, \sigma_*^2) \tag{1}$$

for the noise-free range with

$$\mu_* \;=\; \mathbf{k}_{\mathbf{v}_*\mathbf{v}}^\top (\mathbf{K}_{\mathbf{vv}} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \tag{2}$$

$$\sigma_*^2 \;=\; k(\mathbf{v}_*, \mathbf{v}_*) - \mathbf{k}_{\mathbf{v}_*\mathbf{v}}^\top (\mathbf{K}_{\mathbf{vv}} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_{\mathbf{v}_*\mathbf{v}} \tag{3}$$

for a new query feature $\mathbf{v}^*$. Here, the matrix $\mathbf{K}_{\mathbf{vv}} \in \mathbb{R}^{n \times n}$ denotes the covariance matrix with $[\mathbf{K}_{\mathbf{vv}}]_{ij} = k(\mathbf{v}_i, \mathbf{v}_j)$. Furthermore, $\mathbf{k}_{\mathbf{v}_*\mathbf{v}} \in \mathbb{R}^n$ is given by $[\mathbf{k}_{\mathbf{v}_*\mathbf{v}}]_i = k(\mathbf{v}_*, \mathbf{v}_i)$, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, and $\mathbf{I}$ is the identity matrix. $\sigma_n$ denotes the global noise parameter. As covariance function, we apply the squared exponential

$$k(\mathbf{v}_p, \mathbf{v}_q) = \sigma_f^2 \cdot \exp\left(-\frac{1}{2\ell^2} |\mathbf{v}_p - \mathbf{v}_q|\right), \tag{4}$$

where $l$ and $\sigma_f$, as well as the global noise parameter $\sigma_n$, are the so-called hyperparameters. A standard way of learning these hyperparameters from data, which we applied in this work, is to maximize the log data likelihood of the training data using scaled conjugate gradients (see, e.g., [28] for details).

A particularly useful property of Gaussian processes for our application is the availability of the predictive uncertainty at every query point. This means that new features $\mathbf{v}_*$ which lie close to points $\mathbf{v}$ of the training set, result in more confident predictions than features which fall into a less densely sampled region of feature space.

### *4.2. Modeling Angular Dependencies*

So far, our model assumes *independent* range variables $y_i$ and it thus ignores dependencies that arise, for instance, because "neighboring" range variables and visual features are likely to correspond to the same object in the environment. Angular dependencies can be included, for example, by (a) explicitly considering the angle $\alpha$ as an additional index dimension in $\mathbf{v}$ or by (b) applying Gaussian beam processes (GBPs) as an independent post-processing step to the *predicted* range scan. While the first variant would require a large amount of additional training data—since it effectively couples the visual appearance and the angle of observation, the second alternative is relatively easy to realize and to implement. Figure 3 gives a graphical representation of the second approach. The gray bars group sets of variables that are fully connected and jointly distributed according to

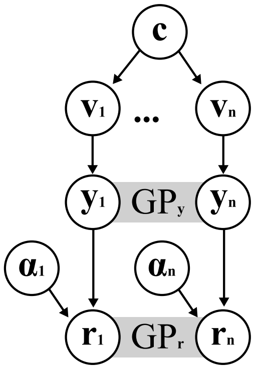

Figure 7: Graphical model for predicting ranges $\mathbf{r}$ from a camera image $\mathbf{c}$. The gray bars group sets of variables that are fully connected and that are jointly distributed according to a GP model.

a GP model. We denote with $\mathcal{GP}_y$ the Gaussian process that maps visual features to ranges and with $\mathcal{GP}_r$ the so-called heteroscedastic GP that is applied as a post-processing step to single, predicted range scans. For $\mathcal{GP}_r$, the task is to learn the mapping $\alpha \mapsto r$ using a training set of *predicted* range values $\mathbf{r}$. Since we do not want to constrain the model to learning from the *mean* predictions $\mu_*(\mathbf{x}_i)$ only, we need a way of incorporating the predictive uncertainties $\sigma_*^2(\mathbf{v}_i)$ for the feature-based range predictions $y_*$. This can be achieved by not using a fixed noise matrix $\sigma_n^2\mathbf{I}$ as in $\mathcal{GP}_y$ (compare Eq. (2) and Eq. (3)), but instead its heteroscedastic extension

$$\mathbf{R} = \text{diag}\left(\sigma_*^2(\mathbf{v}_1), \ldots, \sigma_*^2(\mathbf{v}_n)\right) \ , \tag{5}$$

see [3]. This matrix does not depend on a *global* noise parameter $\sigma_n$, but rather on the individual confidence estimates $\sigma_*^2(\mathbf{v}_i)$, with which $\mathcal{GP}_y$ estimated the corresponding range value. Note that this “trick” of gating out training points by artificially increasing their associated variance was also applied in recent work on modeling gas distributions [29] for deriving a GP mixture model. A more detailed discussion of the approach can be found there and in [30].

### *4.3. Summary of Our Approach*

The full approach that also considers the angular dependencies in a range scan is denoted by the postfix *+GBP* in the experimental evaluation. To obtain the prediction of a full range scan given one omnidirectional image, we proceed as follows:

1. Warp the omnidirectional image into a panoramic view.

2. Extract for every pixel column $i$ a vector of visual features $\mathbf{v}_i$.
3. Use $\mathcal{GP}_y$ to make independent range predictions about $y_i$.
4. Learn a heteroscedastic GBP $\mathcal{GP}_r$ for the set of predicted ranges $\{y_i\}_{i=1}^n$ indexed by their bearing angles $\alpha_i$ and make the final range predictions $r_i$ for the same bearing angles.

As the following experimental evaluation revealed, this additional GBP treatment (post-processing with $\mathcal{GP}_r$) further increases the accuracy of range predictions. The gain, however, is rather small compared to what the GP treatment $\mathcal{GP}_y$ adds to the accuracy achievable with the baseline feature mappings. This might be due to the fact that the extracted features—and the constellation of several feature types even more so—carry information of neighboring pixel strips, such that angular dependencies are incorporated at this early stage already.

## 5. Experimental Evaluation

The system for predicting range from single, omnidirectional images described in the previous sections was implemented in `C/C++` and `Python` and tested on two benchmark data sets for image-based localization. The data sets, named Freiburg and Saarbrücken have been acquired in the context of the EU project CoSy. They have been made publicly available at [31] under the names COLD-Freiburg and COLD-Saarbruecken. The data was recorded using a mobile robot equipped with a laser scanner, an omnidirectional camera, and Odometry sensors at the AIS lab of the University of Freiburg and at the German Research Center for Artificial Intelligence (DFKI) in Saarbrücken. The two environments have quite different characteristics—especially in the visual aspects. While the environment in Saarbrücken mainly consists of solid, regular structures and a homogeneously colored floor, the lab in Freiburg exhibits many glass panes, an irregular, wooden floor and challenging lighting conditions.

The goal of the experimental evaluation was to verify that the proposed system is able to make sensible range predictions from single omnidirectional camera images and to quantify the benefits of the GP approach in comparison to conceptually simpler approaches. We document a series of different experiments: First, we evaluate the accuracy of the estimated range scans using (a) the individual edge features directly, (b) the *PCA-GP*, (c) the *LDA-GP*, and (d) the *Feature-GP*, which constitutes our regression model with the four edge-based vision features as input dimensions. Then, we illustrate how these estimates can be used to build grid maps of the environment. We also evaluated whether applying the GBP model,

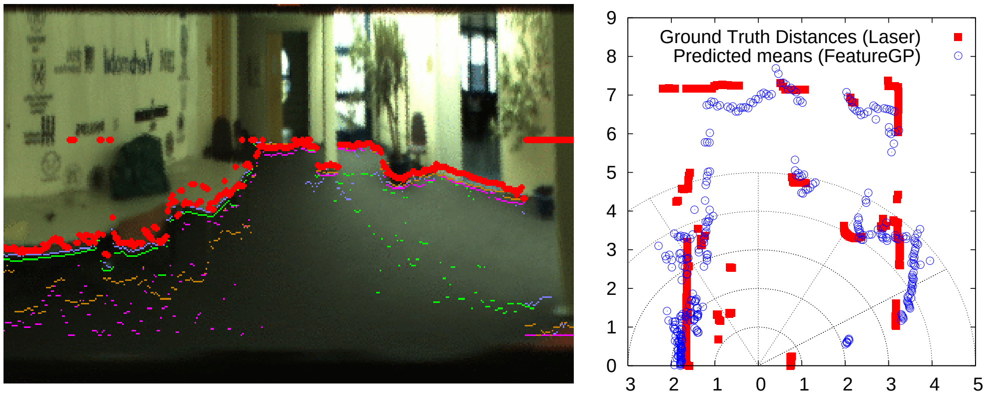

Figure 8: Left: Estimated ranges projected back onto the camera image using the feature detectors directly (small dots) and using the *Feature-GP* model (red points). Right: Prediction results and the true laser scan at one of the test locations visualized from a birds-eye view.

which was introduced in [3], as a post-processing step to the predicted range scans can further increase the prediction accuracy. The GBP model places a Gaussian process prior on the range function (rather than on the function that maps features to distances) and, thus, also models angular dependencies. We denote these models by *Feature-GP+GBP*, *PCA-GP+GBP*, and *LDA-GP+GBP*.

### *5.1. Quantitative Results*

Table 1 summarizes the average RMSE (root mean squared error) obtained for different system configurations, which are detailed in the following. The error is measured as the difference between *measured* laser ranges and ranges *predicted* using the visual input. The first four configurations, referred to as C01 to C04, apply the optimized mapping functions for the different edge features (see Figure 6). Depending on the data, the features provide estimates with an RMSE of between 1.7 m and 3 m. We then evaluated the configurations C05 and C06 which use the four edge-based features as inputs to a Gaussian process model as described in Section 4 to learn the mapping from the feature vectors to the distances. The learning algorithm was able to perform range estimation with an RMSE of around 1 m. Note that we measure the prediction error relative to the recorded laser beams rather than to the true geometry of the environment. Thus, we report a conservative error estimate that also includes mismatches due to reflected laser beams or due to imperfect calibration of the individual components. To give a visual impression of the prediction accuracy of the *Feature-GP*, we give a typical laser scan and the mean predictions in the right diagram in Figure 8.

Table 1: Average errors obtained with the different methods. The root mean squared errors (RMSE) are calculated relative to the mean predictions for the complete test sets.

| | RMSE on test set | |
|---|---|---|
| **Configuration** | **Saarbrücken** | **Freiburg** |
| C01: Laws5 | 1.70m | 2.87m |
| C02: Laws5+LMD | 2.01m | 2.08m |
| C03: Laws3+Canny | 1.74m | 2.87m |
| C04: Laws3+Canny+LMD | 2.06m | 2.59m |
| C07: PCA-GP | 1.24m | 1.40m |
| C09: LDA-GP | 1.20m | 1.31m |
| C05: Feature-GP | 1.04m | 1.04m |
| C08: PCA-GP+GBP | 1.22m | 1.41m |
| C10: LDA-GP+GBP | 1.17m | 1.29m |
| C06: Feature-GP+GBP | 1.03m | 0.94m |

The *PCA-GP* approach (denoted as C07) that does not require engineered features, but rather works on the low-dimensional representation of the raw visual input computed using the PCA. The resulting six-dimensional feature vector is used as input to the Gaussian process model. With an RMSE of 1.2 m to 1.4 m, the *PCA-GP* outperforms all four engineered features, but is not as accurate as the *Feature-GP*. When using LDA for dimensionality reduction (C09) instead of PCA, we observe a reduction of the prediction error by around 4-8 per cent. Also the LDA is outperformed by *Feature-GP* in terms of prediction accuracy. It should be stressed, however, that *PCA-GP* as well as *LDA-GP* do not require any manually defined features as they operate on the observed 420-dimensional pixel columns directly. For configurations C06, C08, and C10, we predicted the ranges per scan using the same methods as above, but additionally applying the GBP model [3] to incorporate angular dependencies between the predicted beams. This post-processing step yields slight improvements compared to the original variants C05, C07, and C09.

The left image in Figure 8 depicts the predictions based on the individual vision features and the *Feature-GP*. It can be clearly seen from the image, that the different edge-based features model different parts of the range scan well. The *Feature-GP* fuses these unreliable estimates to achieve high accuracy on the whole scan. The result of the *Feature-GP+GBP* variant for the same situation is given in Figure 1. The right diagram in Figure 8 visualizes a typical prediction result and

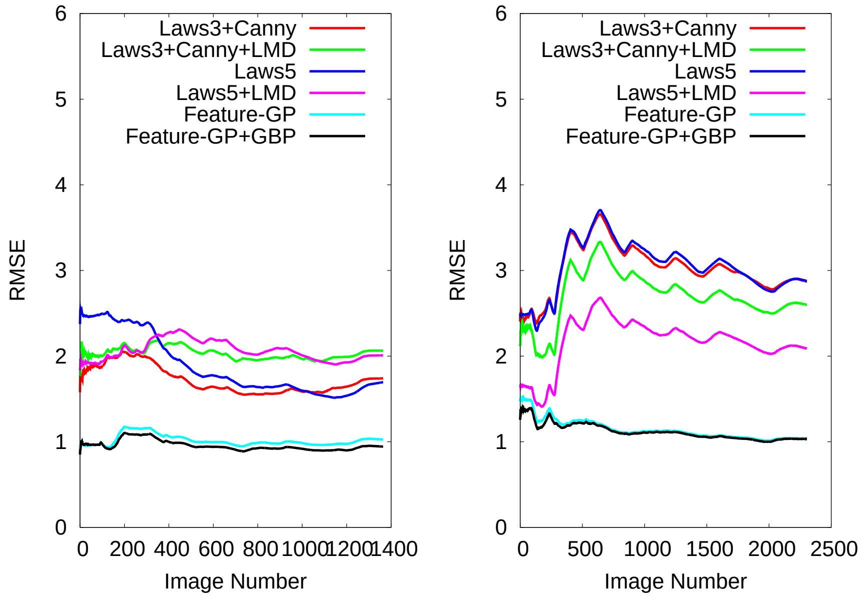

Figure 9: The evolution of the root mean squared error (RSME) for the individual images of the Saarbrücken (left) and Freiburg (right) data sets.

the corresponding laser scan—which can be regarded here as the ground truth—from a birds-eye view. The evolution of the RMSE for the different methods over time is given in Figure 9. As can be seen from the diagrams, the prediction using the *Feature-GP* model outperforms the other techniques and achieves a near-constant error rate.

In summary, our GP-based technique outperforms the individual, engineered features for range prediction. The smoothed approach (C06) yields the best predictions with an RMSE of around 1 m. One can obtain good results by a combination of LDA for dimensionality reduction and GP learning with an error that is only slightly larger (C10 versus C06), even though this unsupervised method does not have access to background information.

### *5.2. Error Analysis*

In this section, we analyze the prediction accuracy of our proposed method beyond the RMSE measure, that is, considering the entire distribution of prediction error in order to identify and document its different causes. The left diagram in Fig. 10 shows the histogram of prediction errors for a typical scan. It is clearly visible, that the reported RMSE values are strongly influenced by the heavy tails

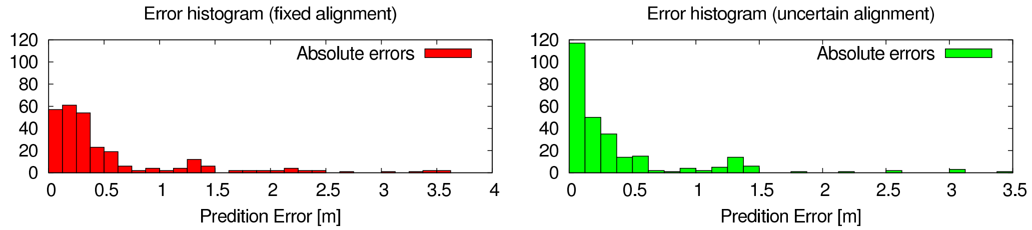

Figure 10: Error histograms for omnidirectional range prediction from single images. Left: Independent predictions for fixed beam orientations. Right: Accounting for uncertain beam orientations due to small angular miscalibration of the test and training setup.

of the error distribution. The large majority of the predictions is accurate (less than 30cm error), while very few predictions have a high error of up to 3m. Close inspection of the results reveals, that such isolated high errors are mostly caused by a small angular misalignment between the camera and the laser scanner which recorded the reference test set. This effect is visualized in the right diagram in Fig. 11. The diagram shows the "true distances" as measured by the laser scanner, the predicted distances and the respective absolute errors. It can be seen that most of the absolute prediction error accounts to beam 18520 (ID in the entire test set) which is located close to a depth discontinuity. Already a very small angular misalignment between laser scanner and camera can lead to such peaks in the error function.

As a result, the reported RMSE values have to be seen as a tool for comparing different approaches and settings rather than as a measure of precision for an actual application. From our experience, error histograms and concrete prediction examples deliver the best picture of the actual precision to be expected.

To show the influence of angular misalignment as well as long-range predictions quantitatively, we give a comparison of different error measures in Tab. 2. In the row labeled "fixed aligment", we give the errors for direct comparisons

Table 2: Comparison of error measures.

| | Root Mean Squared Error (RMSE) | Mean Absolute Error | Mean relative Error |
|---|---|---|---|
| Fixed alignment | 0.88 m | 0.55 m | 0.19 |
| Uncertain alignment | 0.67 m | 0.38 m | 0.14 |

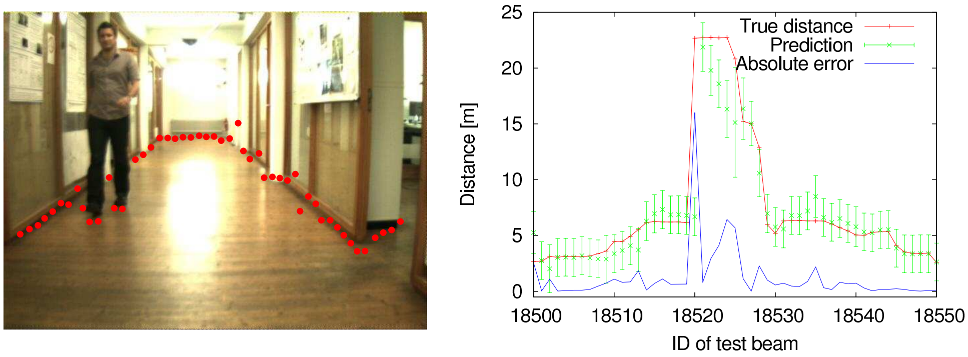

Figure 11: Left: Range predictions of method C05 (Feature-GP) for a single perspective camera image taken from a test sequence. Right: Visualization of the prediction errors for a typical scan from a test sequence. Most of the absolute prediction error is caused by small angular misalignments (here: beam 18520) and for long-range predictions (exceeding 10m).

between laser beams and prediction w.r.t. a fixed beam orientation. For “uncertain aligment”, we allow for a small angular misalignment of the laser beams and their projections to the camera images. The relative errors in the last column are computed by dividing the range predictions by the true distances.

As a reference for comparison, Saxena *et al.* [2] reports depth reconstruction errors in indoor environments of 0.084 on a log scale (base 10), which corresponds to 1.21m of mean absolute error. Including stereo information using a second camera, their error drops to 0.079 in log scale, that is, 1.19m on a regular scale.

### *5.3. Using Perspective Cameras*

To show the flexibility of our method, we conducted additional experiments using a single perspective camera (as opposed to an omnidirectional one). In this setting, the correspondence between range observations from the laser scanner (available only during training) and the camera image is not as direct as for omnidirectional, axis-aligned cameras. Nevertheless, the mapping function is bijective in the region observed by the camera and it can be computed analytically using projective geometry. The right image in Fig. 2 shows an example image and the laser beams transformed into image coordinates. The camera has a $50^\circ$ field of view and it covers approximately 2.2m to 30m in depth. The camera was tilted downwards by $10^\circ$. We recorded two data sets containing 600 images each. One set was used for training, the other one for testing.

In this experiment, we additionally evaluated a combination of PCA and LDA, for which we first reduced the dimensionality of the data to 50 using PCA and then applied LDA to further reduce to 6 dimensions. This is a common approach in face recognition, addressing the concern that LDA might perform poorly due to too little training data. PCA and LDA were learned from 100 images randomly drawn from the training set and the GP models used 300 random images. Test statistics were computed over the whole test set (50 beams per image).

Table 3 gives the quantitative results for this experiment. The observed trend is similar to the one described in Sec. 5.1: The feature-based GP performs best, while the unsupervised methods (LDA and PCA) follow up with a higher mean prediction error. The combination of LDA and PCA did not yield significant advantages in this setting. The absolute prediction errors are higher compared to the omnidirectional setting, mainly because the respective test set contains significantly more long-range predictions due to the heading of the camera towards the long corridor.

Referring to the discussion in the previous section, it should be noted that the distribution of errors is strongly biased towards errors caused by small angular misalignments (see the right diagram in Fig. 11). For most practical applications, for example obstacle avoidance, such small angular misalignments do not have a negative impact. The left image in Fig. 11 shows a typical image from the test sequence including the predictions made using C05 (Feature-GP).

Table 3: Average errors obtained with the different methods on single perspective camera images. The root mean squared errors (RMSE) are calculated relative to the mean predictions for the complete test sets.

| | Prediction errors (on single images of a perspective camera) | |
|---|---|---|
| **Configuration** | **Mean absolute error** | **RMSE** |
| C09: LDA-GP | 2.24m | 3.52m |
| C07: PCA-GP | 1.87m | 3.10m |
| C11: PCA-LDA-GP | 1.85m | 3.02m |
| C05: Feature-GP | 1.19m | 2.65m |

### *5.4. Non-Linear Dimensionality Reduction*

In addition to PCA and LDA, we also considered applying non-linear dimensionality reduction as a preprocessing step. Non-linear dimensionality reduction can be described as seeking a low-dimensional manifold (not necessarily a linear subspace) in which the observed data points can be represented well. Approaches to this problem include local linear embedding (LLE) [32] and ISOMAP [33].

We implemented LLE, due to our positive experience with this technique in the past. In this domain, however, LLE performed significantly worse than all other techniques evaluated in Table 1. An analysis of the constructed manifolds indicated that the low performance may be caused by the significant number of outliers present in our real-world data sets. Similar observations about LLE and the presence of outliers have also been reported by other researcher. Chang and Yeung [34], for example, report that adding between 5% and 10% outliers to perfect data can prevent LLE from finding an appropriate embedding.

### *5.5. Learning Occupancy Maps from Predicted Scans*

Our approach can be applied to a variety of robotics tasks such as obstacle avoidance, localization, or mapping. To illustrate this, we show how to learn a grid map of the environment from the predictive range distributions. Compared to occupancy grid mapping where one estimates for each cell the probability of being occupied or free, we use the so-called *reflection probability maps*. A cell of such a map models the probability that a laser beam passing this cell is reflected or not. Reflection probability maps, which are learned using the so-called *counting model*, have the advantage of requiring no hand-tuned sensor model such as occupancy grid maps (see [35] for further details). The reflection probability $m_i$ of a cell $i$ is given by

$$m_i \;=\; \frac{\alpha_i}{\alpha_i + \beta_i} \;, \tag{6}$$

where $\alpha_i$ is the number of times an observation hits the cell, i.e., ends in it, and $\beta_i$ is the number of misses, i.e., the number of times a beam has intercepted a cell without ending in it. Since our GP approach does not estimate a single laser end point, but rather a full (normal) distribution $p(z)$ of possible end points, we have to integrate over this distribution (see Figure 12). More precisely, for each grid cell $c_i$, we update the cell's reflectance values according to the predictive distribution $p(z)$ according to the following formulas:

$$\alpha_i \;\leftarrow\; \alpha_i + \int_{z \in c_i} p(z)\, dz \tag{7}$$

$$\beta_j \;\leftarrow\; \beta_i + \int_{z > c_i} p(z)\, dz \;. \tag{8}$$

Note that for perfectly accurate predictions, the extended update rule is equivalent to the standard formula stated above.

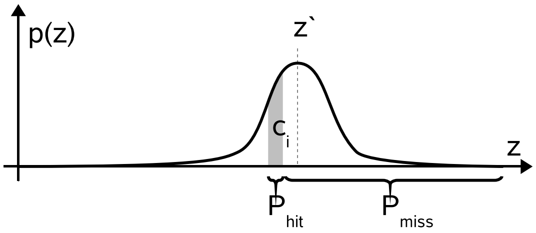

Figure 12: The counting model for reflectance grid maps in conjunction with sensor models that yield Gaussian predictive distributions over ranges.

We applied this extended reflection probability mapping approach to the trajectories and range predictions that resulted from the experiments reported above. Figure 13 presents the laser-based maps using a standard reflection probability mapping system (left column) and our extended variant using the predicted ranges (right column) for the two environments (Freiburg on top and Saarbrücken below). In both cases, it is possible to build an accurate map, which is comparable to maps obtained with infrared proximity sensors [36] or sonars [21].

## 6. Conclusion

This paper presents a new approach to estimating the free space around a robot based on single images recorded with an omnidirectional camera. The task of estimating the range to the closest obstacle is achieved by applying a Gaussian process model for regression, utilizing edge-based features extracted from the image or, alternatively, using PCA or LDA to find a low-dimensional representation of the visual input in an unsupervised manner. All learned models outperform the optimized individual features.

We furthermore showed in experiments with a real robot that the range predictions are accurate enough to feed them into a mapping algorithm considering predictive range distributions and that the resulting maps are comparable to maps obtained with infrared or sonar sensors.

## Acknowledgments

We would like to thank Andrzej Pronobis and Jie Luo for creating the CoSy data sets. This work has partly been supported by the EC under contract number FP7-231888-EUROPA and FP6-004250-CoSy, by the DFG under contract number SFB/TR-8, and by the German Ministry for Education and Research (BMBF) through the DESIRE project.

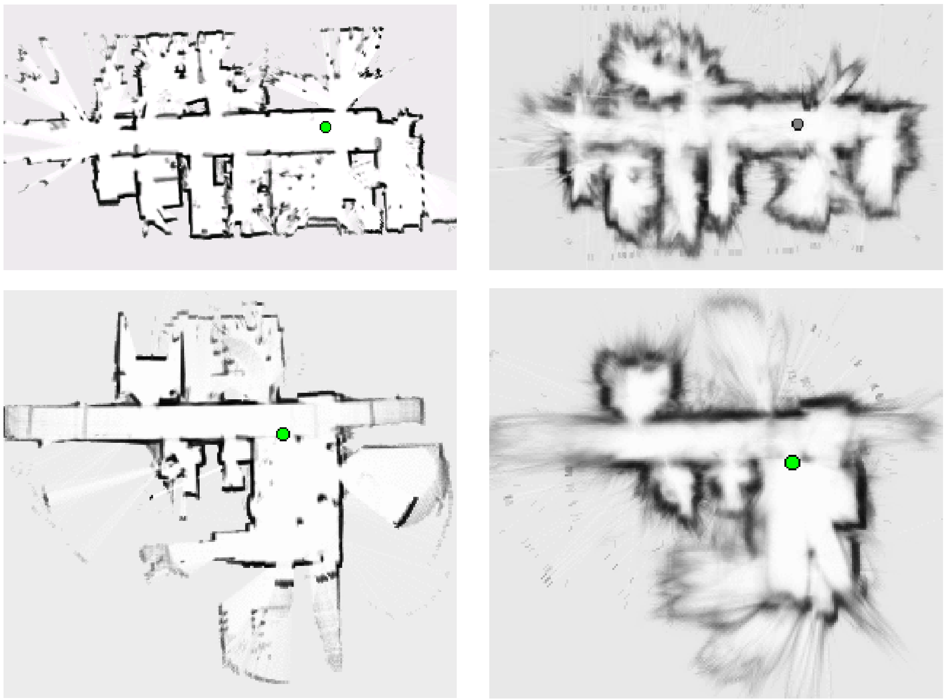

Figure 13: Maps of the Freiburg AIS lab (top row) and DFKI Saarbrücken (bottom row) using real laser data (left) and the predictions of the *Feature-GP* (right).

## References

[1] G. Swaminathan, S. Grossberg, Laminar cortical mechanisms for the perception of slanted and curved 3-D surfaces and their 2-D pictorical projections, Journal of Vision 2 (7) (2002) 79–79.

[2] A. Saxena, S. Chung, A. Ng., 3-d depth reconstruction from a single still image, Int. Journal of Computer Vision (IJCV).

[3] C. Plagemann, K. Kersting, P. Pfaff, W. Burgard, Gaussian beam processes: A nonparametric bayesian measurement model for range finders, in: Proc. of Robotics: Science and Systems (RSS), 2007.

[4] F. Sinz, J. Quinonero-Candela, G. Bakir, C. Rasmussen, M. Franz, Learning depth from stereo, in: 26th DAGM Symposium, 2004.

[5] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[6] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, SURF: Speeded up robust features, Computer Vision and Image Understanding (CVIU) 110 (3) (2008) 346–359.

[7] A. Davision, I. Reid, N. Molton, O. Stasse, Monoslam: Real-time single camera slam, IEEE Transaction on Pattern Analysis and Machine Intelligence 29 (6).

[8] H. Strasdat, C. Stachniss, M. Bennewitz, W. Burgard, Visual bearing-only simultaneous localization and mapping with improved feature matching, in: Fachgespräche Autonome Mobile Systeme (AMS), 2007.

[9] B. Micusik, T. Pajdla, Structure from motion with wide circular field of view cameras, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7) (2006) 1135–1149.

[10] R. Sim, J. J. Little, Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters, in: Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), 2006, pp. 2082–2089.

[11] P. Favaro, S. Soatto, A geometric approach to shape from defocus, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 406–417.

[12] A. Torralba, A. Oliva, Depth estimation from image structure, IEEE Transactions on Pattern Analysis and Machine Learning.

[13] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, G. Bradski, Self-supervised monocular road detection in desert terrain., in: Proc. of Robotics: Science and Systems (RSS), 2006.

[14] J. Michels, A. Saxena, A. Ng, High speed obstacle avoidance using monocular vision and reinforcement learning, in: Int. Conf. on Machine Learning (ICML), 2005, pp. 593–600.

[15] E. Menegatti, A. Pretto, A. Scarpa, E. Pagello, Omnidirectional vision scan matching for robot localization in dynamic environments, IEEE Transactions on Robotics 22 (3) (2006) 523–535.

[16] D. Hoiem, A. Efros, M. Herbert, Recovering surface layout from an image, Int. Journal of Computer Vision (IJCV) 75 (1).

[17] F. Han, S.-C. Zhu, Bayesian reconstruction of 3d shapes and scenes from a single image, in: IEEE Int. Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis, Washington, DC, USA, 2003, p. 12.

[18] E. Delage, H. Lee, A. Ng., Automatic single-image 3d reconstructions of indoor manhattan world scenes., in: Proceedings of the 12th International Symposium of Robotics Research (ISRR), 2005.

[19] R. Ewerth, M. Schwalb, B. Freisleben, Using depth features to retrieve monocular video shots, in: Proceedings of the 6th ACM international conference on Image and video retrieval (CIVR), ACM, New York, NY, USA, 2007, pp. 210–217.

[20] H. Moravec, A. Elfes, High resolution maps from wide angle sonar, in: Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA), St. Louis, MO, USA, 1985, pp. 116–121.

[21] S. Thrun, A. Bücken, W. Burgard, D. Fox, T. Fröhlinghaus, D. Hennig, T. Hofmann, M. Krell, T. Schimdt, Map learning and high-speed navigation in RHINO, in: AI-based Mobile Robots: Case studies of successful robot systems, MIT Press, Cambridge, MA, 1998.

[22] K. Sabe, M. Fukuchi, J.-S. Gutmann, T. Ohashi, K. Kawamoto, T. Yoshigahara, Obstacle avoidance and path planning for humanoid robots using stereo vision, in: Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA), New Orleans, LA, USA, 2004.

[23] P. Elinas, R. Sim, J. J. Little, $\sigma$SLAM: Stereo vision SLAM using the rao-blackwellised particle filter and a novel mixture proposal distribution, in: Proc. of ICRA, 2006.

[24] C. Plagemann, F. Endres, J. Hess, C. Stachniss, W. Burgard, Monocular range sensing: A non-parametric learning approach, in: Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA), Pasadena, CA, USA, 2008.

[25] E. Alpaydin, Introduction To Machine Learning, MIT Press, 2004.

[26] E. R. Davies, Laws texture energy in texture, in: Machine Vision: Theory, Algorithms, Practicalities, Acedemic Press, 1997.

[27] F. Canny, A computational approach to edge detection, IEEE Trans. Pattern Analysis and Machine Intelligence (1986) 679–714.

[28] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[29] C. Stachniss, C. Plagemann, A. Lilienthal, W. Burgard, Gas distribution modeling using sparse gaussian process mixture models, in: Proc. of Robotics: Science and Systems (RSS), Zurich, Switzerland, 2008.

[30] V. Tresp, Mixtures of gaussian processes, in: Proc. of the Conf. on Neural Information Processing Systems (NIPS), 2000.

[31] EU Project CoSy. [link].
URL `http://www.cognitivesystems.org/`

[32] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[33] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction., Science 290 (5500) (2000) 2319–2323.

[34] H. Chang, D.-Y. Yeung, Robust locally linear embedding, Pattern Recognition 39 (6) (2006) 1053–1065.

[35] W. Burgard, C. Stachniss, D. Haehnel, Autonomous Navigation in Dynamic Environments, Vol. 35 of STAR Springer tracts in advanced robotics, Springer Verlag, 2007, Ch. Mobile Robot Map Learning from Range Data in Dynamic Environments.

[36] Y. Ha, H. Kim, Environmental map building for a mobile robot using infrared range-finder sensors, Advanced Robotics 18 (4) (2004) 437–450.