

# Global Context Extraction for Object Recognition Using a Combination of Range and Visual Features

We put a dummy line in here to make sure that your final manuscript does not exceed ten pages in length.

Same here, use up some space  
to make sure  
the final contribution will not exceed  
ten pages in length.

**Abstract.** It has been highlighted by many researchers, that the use of context information as an additional cue for high-level object recognition is important to close the gap between human and computer vision. We present an approach to context extraction in the form of global features for place recognition. Based on an uncalibrated combination of range data of a time-of-flight (ToF) camera and images obtained from a visual sensor, our system is able to classify the environment in predefined places (e.g. kitchen, corridor, office) by representing the sensor data with various global features. Besides state-of-the-art feature types, such as power spectrum models and Gabor filters, we introduce histograms of surface normals as a new representation of range images. An evaluation with different classifiers shows the potential of range data from a ToF camera as an additional cue for this task.

## 1 Introduction

The development of time-of-flight (ToF) cameras [1], which provide range information in realtime, has lead to a large number of applications. Most of them concentrate on the support of vision-based systems in tasks like 3d reconstruction and robot navigation [2]. Alternatively to geometric reconstruction techniques, we show how to utilize a classification based system for place recognition or rough self localization of a mobile robot.

Instead of describing the position of a robot in exact geometric terms, it is often beneficial to use a discretization of predefined places or scenes, e.g. kitchen, corridor or office. Especially for subsequent object detection tasks [3], information about the current place can be used as high-level contextual information [4]. Due to the large variability of scene appearances, the estimation of the most probable label is a challenging recognition task. For this reason we calculate a feature representation from ToF range data and from an image obtained using a standard visual sensor (Fig. 1). This allows to describe a scene using rough 3d information and visual appearance. Furthermore we present a simple method



**Fig. 1.** Setup of our place recognition system with a ToF sensor and a visual sensor mounted on a mobile robot. Data is obtained from both uncalibrated cameras in order to build the combined feature representation of the current view.

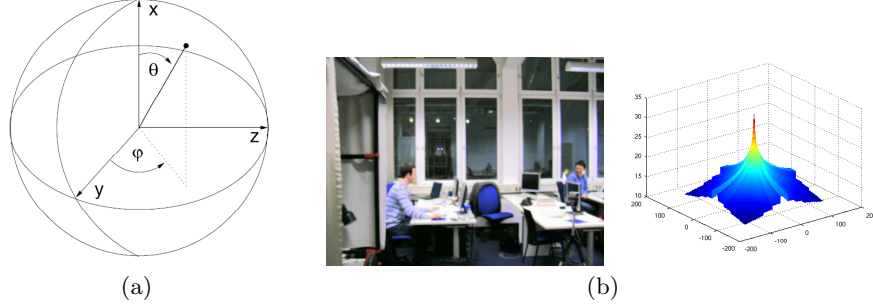
for feature calculation in range images which describes the image as a collection of planar patches. It can be seen as an instance of the bag-of-features concept, which has been shown to be well suited for scene recognition [5]. Features from visual images are calculated using two state-of-the-art approaches often used for the task of scene recognition. Our work extends the scene recognition approach of [4] to multiple sensors and range data.

The remainder of the paper is organized as follows: First of all, we present histograms of surface normals as a feature type for range images which is well suited for the place recognition task. In Sect. 3 we describe state-of-the-art global feature representations that can be applied to data from the visual and the range sensor. Classification techniques and details of the feature combination are explained in Section 4. Experiments in Sect. 5 compare feature types and different classifiers and show the performance benefit of feature combination from different sensors. A summary of our findings and a discussion of future research directions conclude the paper.

## 2 Histogram of Surface Normals

Range images captured by ToF sensors consist of dense distance measurements of scene elements in the field of view of the camera. Using a simple histogram representation of all depth values would be a typical global representation of the scene. However, for scene and place recognition with standard cameras, feature types that use aggregated local statistics of pixel neighborhoods showed to be successful. A simple but efficient approach to incorporate information from a small environment of a pixel is the representation of a range image as a collection of small planar patches or patchlets [6]. A statistic of the orientation of such planar patches then corresponds to local surface characteristics.

Let  $\mathbf{x}$  be a three dimensional point obtained from the range image and  $N(\mathbf{x})$  the set of all points in the (rectangular) image neighborhood of size  $P \times P$  with center  $(\mathbf{x}_1, \mathbf{x}_2)^T$ . By assuming orthogonal projection, each plane that does not intersect the camera center can be described by  $\mathbf{n}^T \mathbf{x} = 1$ , where  $\mathbf{n} = (n_x, n_y, n_z)^T$  denotes the surface normal. We estimate the parameters of the



**Fig. 2.** a) Representation of surface normals as angles in sphere coordinates [7]. b) Sample image and its power spectrum representation with 16 sectors.

planar patch in each point  $\mathbf{x}^i$  with Iteratively Reweighted Least Squares (IRLS) applied to the resulting optimization problem:  $\mathbf{n}^i = \arg \min_{\mathbf{n}} \sum_{\mathbf{x} \in N(\mathbf{x}^i)} |\mathbf{n}^T \mathbf{x} - 1|$ .

Instead of absolute depth values, we use local surface characteristics as a feature. Therefore we utilize the normal representation of Hetzel et al. [7], which transforms  $\mathbf{n}^i$  in a pair of angles  $(\varphi^i, \theta^i)^T$  in sphere coordinates, where  $\varphi = \arctan\left(\frac{n_z}{n_y}\right)$  and  $\theta = \arctan\left(\frac{n_y^2 + n_z^2}{n_x}\right)$ , as illustrated in Fig. 2. Thus, the resulting representation is a two dimensional histogram with  $B_\varphi$  and  $B_\theta$  bins for  $\phi^i$  and  $\theta^i$ , and  $B_\varphi \times B_\theta$  entries.

### 3 Visual Features

In the subsequent sections low-level visual features are described. In addition to its originally motivated purpose, which is the representation of visual images, we also use the following features to extract second order and structure information from range images.

#### 3.1 Power Spectrum Features

One famous approach, which was first described by Mezrich et al. [8] in the late seventies, is to fit the Fourier power spectrum to an isotropic model. Empirical studies on natural images [8, 9] show that the average power spectrum approximately obeys the power law  $M(\mathbf{f}) = A \cdot \|\mathbf{f}\|_2^{-\alpha}$ , with parameter  $A$  and  $\alpha$ , where  $\mathbf{f}$  denotes frequency. Straightforward linear least squares optimization can be used to estimate the model parameters.

However, Oliva and Torralba [9] empirically show that the power law does not hold for artificial images. Thus, since we concentrate on indoor environments and want to calculate features from a single image, it is unlikely that an isotropic representation is sufficient to properly describe present second order statistics. We therefore use an extended representation [9], where the power spectrum is

radially divided in  $\Omega$  non-overlapping sectors. Each sector  $\omega$  is then assumed to obey a power law:

$$M_\omega(\mathbf{f}) = \frac{A_\omega}{\|\mathbf{f}\|_2^{\alpha_\omega}} \quad 1 \leq \omega \leq \Omega . \quad (1)$$

However, this anisotropic power spectrum model, which is illustrated in Fig. 2, is a weak representation, since all phase information is lost. In the remainder of this paper, a 16-sector model is used which results in a 32-dimensional feature vector  $(\alpha_1, \dots, \alpha_{16}, A_1, \dots, A_{16})$ .

### 3.2 Gabor Features

Phase preserving representations can be computed using properties of the amplitude spectra. Gabor filters are selective filters that respond to structures of a specific range of frequencies and orientations. A bank of Gabor filters, therefore, can be used as a global image representation. Since the collection of responses is very high dimensional, we follow the approach of [10], where subsampled squared response images are used. This results in substantially reduced feature vectors. Prior to gabor filtering, the image is preprocessed by a whitening step, followed by divisive normalization in order to increase contrast and, thus, amplify higher order structures.

## 4 Classification and Feature Combination

In this paper, three different classifiers were used in order to learn the mapping between features and scene labels: multi-layer Perceptron (MLP), Parzen classifier, and Randomized Decision Forests (RDF). However, for the sake of brevity, only the latter two classifiers are described here.

### 4.1 Parzen Classifier Using Kernel Density Estimation

Core of the generative Parzen classifier for Gaussian kernel densities is the estimation of empirical likelihoods for each class  $\kappa$ :

$$p(\mathbf{f} \mid S_\kappa) = \frac{1}{M_\kappa} \sum_{i=1}^{M_\kappa} \mathcal{K}_\kappa(\mathbf{f} - \mathbf{f}_i) , \quad (2)$$

where  $\mathcal{K}_\kappa$  is a zero-mean normal density with covariance matrix  $\Sigma_\kappa$  and the set  $S_\kappa = \{\mathbf{f}_1, \dots, \mathbf{f}_{M_\kappa}\}$  denotes the  $n$ -dimensional training data labeled with class  $\kappa$ . An unseen feature  $\mathbf{f}$  is then classified using maximum likelihood estimation.

Although the shape of the empirical density is determined by the observed data  $S_\kappa$ , the smoothness depends solely on the kernel bandwidth parameter  $\Sigma_\kappa$ . The appropriate choice of a bandwidth is the most critical step in kernel density estimation, since small bandwidths lead to over-fitting, whereas huge bandwidths

125 result in oversmooth densities. In this paper, we use an ad-hoc method for band- 125  
 126 width selection known as generalized *Scott's rule* [11] for kernel densities: 126

$$\Sigma_{\kappa} \approx M_{\kappa}^{-\frac{1}{n+4}} \hat{\Sigma}_{\kappa}^{\frac{1}{2}}, \quad (3)$$

127 where  $\hat{\Sigma}_{\kappa}$  is the sample covariance with respect to  $S_{\kappa}$ . 127

## 128 4.2 Randomized Decision Forest 128

129 A Randomized Decision Forest (RDF) is a discriminative classifier that can 129  
 130 handle a large set of features without issues due to the curse of dimensionality. 130  
 131 Standard decision tree approaches suffer from severe over-fitting problems. A 131  
 132 RDF overcomes these problems by generating an ensemble (forest) of  $T$  decision 132  
 133 trees. During the classification, the overall probability of a class  $\kappa$  given a feature 133  
 134 vector  $\mathbf{f}$  can be obtained by simple averaging of the posterior probabilities  $p_{\tau}(\cdot)$  134  
 135 estimated by each tree of the ensemble: 135

$$p(\kappa | \mathbf{f}) = \frac{1}{T} \sum_{\tau=1}^T p_{\tau}(\kappa | \mathbf{f}) . \quad (4)$$

136 In contrast to Boosting, the RDF approach uses two types of randomization 136  
 137 to learn the ensemble. The first type of randomization is Bootstrap Aggregat- 137  
 138 ing [12], where each tree is trained with a random fraction of the training data. 138  
 139 Additionally, to reduce training time and to incorporate randomization into the 139  
 140 building process of a tree, the search for the most informative split function in 140  
 141 each inner node is done using only a random fraction of all features [13]. 141

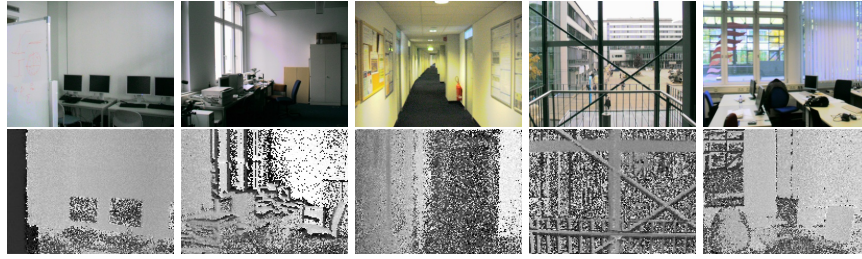
## 142 4.3 Feature Combination and Temporal Context 142

143 In order to combine a set of features  $\mathbf{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_{|\mathbf{F}|}\}$ , simple concatenation 143  
 144 is performed. To avoid facing the curse of dimensionality, which often occurs 144  
 145 with generative classifier, a different scheme is used for the Parzen classifier. 145  
 146 In addition to subspace reduction via PCA, we choose a soft voting approach, 146  
 147 where each feature type  $\mathbf{f}_i$  is classified separately. The overall class probability 147  
 148  $p(\kappa|\mathbf{F})$  is then computed by averaging the separate class probabilities  $p(\kappa|\mathbf{f}_i)$ . 148

149 To further improve the classification performance, a hidden Markov model is 149  
 150 used to exploit temporal contextual properties. We use the approach from Tor- 150  
 151 ralba et al. [4], but instead of a sparse Parzen classifier, we utilize the classifiers 151  
 152 listed above. 152

## 153 5 Experiments 153

154 We experimentally evaluated our approach to illustrate the benefits of the com- 154  
 155 bination of range and visual features for the task of place recognition. In the 155  
 156 next sections the following hypotheses are empirically validated: 156



**Fig. 3.** Example images from one sequence, i.e. intensity images (upper row) and corresponding range images (lower row).

1. Incorporation of range features substantially improves the recognition performance.
2. A Randomized Decision Forest possesses the highest potential in combining a set of different feature types.
3. The use of temporal context information by means of hidden Markov models leads to an important gain in performance.

Our empirical evaluation is based on a place recognition scenario with seven different rooms (classes). The final dataset consists of eight sequences, where each sequence was captured by navigating a mobile robot through a subset of the rooms. Each second, a PMD[vision] 19k camera and a standard CCD camera obtained range and visual images (Fig. 1). As can be seen in Fig. 3, the images do not contain exactly the same image sections, which is due to the different angle of view of the cameras. Note that a calibration of the cameras was not necessary, because features are calculated from the different sensor images independently.

Training is done on two chosen sequences, which together cover all classes of the dataset. The remaining six sequences were then used to test the recognition performance. To measure recognition performance, unbiased average recognition rate was computed. Since more than one scene is used for testing, the mean of all average recognition rates (one for each sequence) is used to evaluate our system.

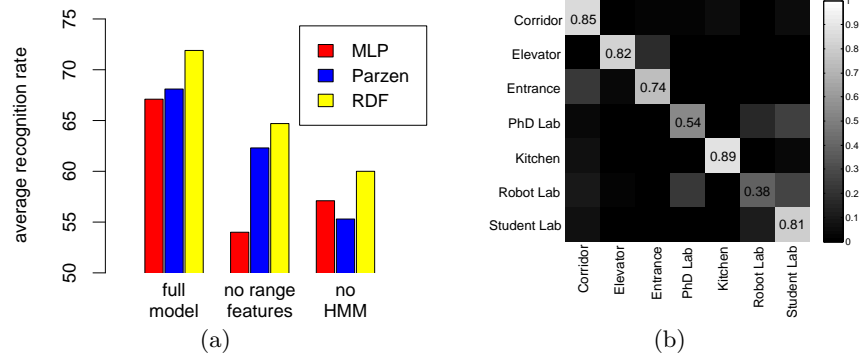
### 5.1 Evaluation of Feature Types and Combinations

In order to evaluate the effects of combined features, we first analyzed the classification performance on each feature type separately. The recognition results are illustrated in Table 1, whereby only the best (out of three) classifier result is shown. Regarding the range features, our experiments show that the surface normal histogram ( $B_\varphi = B_\theta = 10$ ,  $P = 3$ ) achieves the best place recognition result. However, gabor features computed using the data from the visual sensor yield a higher recognition performance.

As can be seen in Table 1, feature combination leads to a substantial performance gain over single feature types. The best combination scheme achieved hereby a recognition rate of 71.9%.

**Table 1.** Evaluation of different features types (incl. computation time). Features computed on the range image of the ToF sensor are tagged with a preceding *r*–.

Feature type	Avg. Recognition Rate	Time (in sec)
<i>r</i> –histogram	44.2	0.024
<i>r</i> –power-spectrum	45.8	0.031
<i>r</i> –gabor	47.4	0.140
<i>r</i> –surface-normal	<b>49.1</b>	0.303
power-spectrum	49.2	0.040
gabor	<b>63.3</b>	0.512
feature combination	<b>71.9</b>	1.050



**Fig. 4.** a) Performances of features type combination, influence of range features, and impact of HMM. b) Average confusion matrix of best feature combination.

## 5.2 Evaluation of Different Classifiers

In the preceding section we showed that the combination of features can improve the classification performance. However, the amount of performance gain depends on the used classifier. Fig. 4(a) (full model) contains the average classification rate of RDF ( $T = 100$ ), MLP, and the Parzen ensemble. The RDF classifier was trained with all feature types listed in Table 1. We also observed that the MLP and the Parzen approach did not achieve comparable results when all feature types were used. We therefore searched for appropriate feature type combinations using a greedy strategy. Nevertheless, the RDF turned out to have the highest classification rate for our dataset. The detailed classification behavior is illustrated in the average confusion matrix shown in Fig.4(b).

In order to further evaluate the power of range information, we extracted all range features from the used feature type subsets mentioned above, i.e. only a combination of visual features remains. The average recognition rates in Fig. 4(a) (no range features) illustrates a drop in classification performance. These results clearly show the superiority of our multi-sensor approach. To finally analyze the impact of temporal contextual cues, we switched off the hidden Markov model (HMM) which leads to a substantial decrease in performance (cf. Fig. 4(a)).



## 6 Conclusion and Further Work

We presented an approach to place and scene recognition which combines information from both a ToF sensor and a standard visual sensor without calibration. We utilized state-of-the-art feature representations from the field of scene recognition [9, 4] and developed a novel description of the range image using planar patches. To show the applicability of our method, we performed experiments with multiple image sequences collected by a mobile robot. The resulting impressive performance gain of the combined feature representation highlights the usefulness of a ToF sensor for the task of place recognition.

As an interesting direction for future research, our feature description of the range image as a histogram of surface normals could be used in conjunction with the principle of spatial pyramid matching [5]. This approach has shown to lead to a significant performance gain by incorporating rough spatial information within images. The most interesting application of our place recognition system would be to use the probabilities of places as prior information in an object detection setting as proposed in [10].

## References

1. Lange, R.: 3D Time-of-Flight Distance Measurement with Custom Solid-State Image Sensors in CMOS/CCD-Technology. PhD thesis, University of Siegen (2000)
2. Prusak, A., Melnychuk, O., Roth, H., Schiller, I., Koch, R.: Pose estimation and map building with a time-of-flight camera for robot navigation. *Int. J. Intell. Syst. Technol. Appl.* **5** (2008) 355–364
3. Hegazy, D., Denzler, J.: Generic 3d object recognition from time-of-flight images using boosted combined shape features. In: *Proc. of VISAPP*. (2009) 321–326
4. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: *Proc. of ICCV*. (2003) 273–280
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *Proc. of CVPR*. (2006) 2169–2178
6. Murray, D.R.: Patchlets: a method of interpreting correlation stereo three-dimensional data. PhD thesis, The University of British Columbia (Canada) (2004)
7. Hetzel, G., Leibe, B., Levi, P., Schiele, B.: 3d object recognition from range images using local feature histograms. In: *Proc. of CVPR*. Volume 2. (2001) 394–399
8. Mezrich, J., Carlson, C., Cohen, R.: Image descriptors for displays. Technical Report PRRL-77-CR-7, Office of Naval Research (1977)
9. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV* **42** (2001) 145–175
10. Torralba, A.: Contextual priming for object detection. *International Journal of Computer Vision* **53** (2003) 169–191
11. Schimek, M.G.: Smoothing and Regression: Approaches, Computation, and Application. Series in Probability and Statistics. Wiley (1996)
12. Breiman, L.: Random forests. *Machine Learning* **45** (2001) 5–32
13. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. *Maching Learning* **63** (2006) 3–42