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

Michael Kemmler, Erik Rodner, and Joachim Denzler

Chair for Computer Vision
Friedrich Schiller University of Jena
{Michael.Kemmler,Erik.Rodner,Joachim.Denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de

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 led 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 for feature calculation in range images which describes the image as a collection

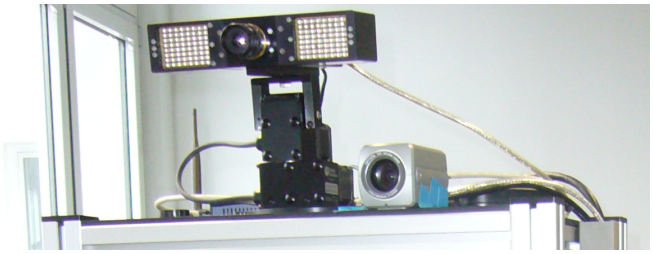


Fig. 1. Setup of our place recognition system with a ToF sensor and a visual sensor mounted on a mobile robot. Data is obtained from both uncalibrated cameras in order to build the combined feature representation of the current view.

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 $\boldsymbol{x}$ be a three dimensional point obtained from the range image and $N(\boldsymbol{x})$ the set of all points in the (rectangular) image neighborhood of size $P \times P$ with center $(\boldsymbol{x}_1, \boldsymbol{x}_2)^T$.

In the following we assume orthogonal projection. Note that we will show that our scene recognition system achieves a suitable performance without the need for an intrinsic camera calibration. With given camera parameters one can

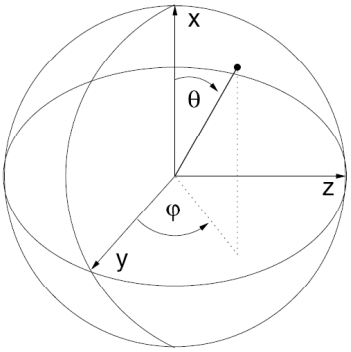


Fig. 2. Representation of surface normals in sphere coordinates [7]

easily undo the perspective projection, which might result in a better recognition performance. Nevertheless this influence is not investigated in this paper, because our results show that despite our severe assumption a histogram of surface normals can be a useful feature representation (cf. Sect. 5.2).

Each plane that does not intersect the camera center can be described by $\boldsymbol{n}^T\boldsymbol{x} = 1$, where $\boldsymbol{n} = (n_x, n_y, n_z)^T$ denotes the surface normal. We estimate the parameters of the planar patch in each point $\boldsymbol{x}^i$ with Iteratively Reweighted Least Squares (IRLS) applied to the resulting optimization problem:

$$\boldsymbol{n}^i = \arg\min_{\boldsymbol{n}} \sum_{\boldsymbol{x} \in N(\boldsymbol{x}^i)} |\boldsymbol{n}^T\boldsymbol{x} - 1| \ . \tag{1}$$

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 $\boldsymbol{n}^i$ into a pair of angles $(\varphi^i, \theta^i)^T$ in sphere coordinates, where:

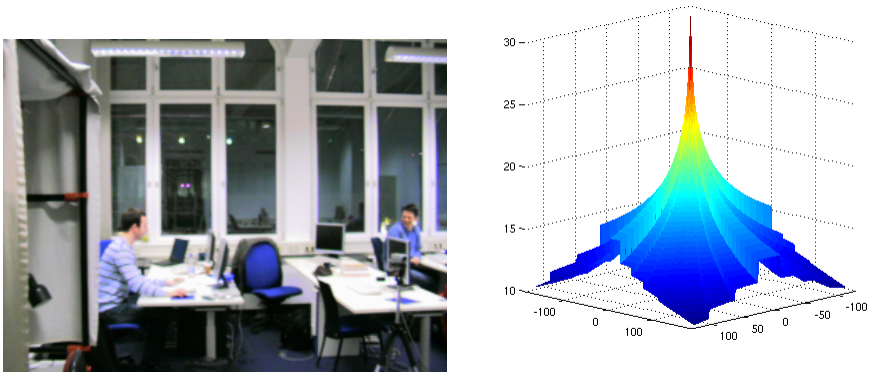
$$\varphi = \arctan\left(\frac{n_z}{n_y}\right) \tag{2}$$

$$\theta = \arctan\left(\frac{\sqrt{n_y^2 + n_z^2}}{n_x}\right) \tag{3}$$

as illustrated in Fig. 2. Thus, the resulting representation is a two dimensional histogram with B_φ and B_θ bins for ϕ^i and θ^i, and $B_\varphi \times B_\theta$ entries.

3 Visual Features

In the subsequent sections low-level visual features are described, which we utilize to calculate a feature representation of the data of our visual sensor. Additionally, we use the following features to extract second order and structure information from range images.



(a) (b)

Fig. 3. Sample image (a) and its (logarithmed) power spectrum representation with 16 sectors (b).

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 $M(\boldsymbol{f}) = A \cdot ||\boldsymbol{f}||_2^{-\alpha}$, with parameter A and α, where $\boldsymbol{f}$ 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 Ω non-overlapping sectors. Each sector ω is then assumed to obey a power law:

$$M_\omega(\boldsymbol{f}) = \frac{A_\omega}{||\boldsymbol{f}||_2^{\alpha_\omega}} \quad 1 \leq \omega \leq \Omega \ . \tag{4}$$

In order to reduce noise, radially averaging [10] is employed for each sector prior to model fitting. Note that this anisotropic power spectrum model, which is illustrated in Fig. 3 does not incorporate phase information.

In the remainder of this paper, a 16-sector model is used which results in a 32-dimensional feature vector $(\alpha_1, \ldots, \alpha_{16}, A_1, \ldots, A_{16})$.

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 [11], 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 [12] in order to increase contrast and, thus, amplify higher-order structures.

4 Classification and Feature Combination

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

4.1 Parzen Classifier Using Kernel Density Estimation

Core of the generative Parzen classifier for Gaussian kernel densities [13, 14] is the estimation of empirical likelihoods for each class $\kappa \in \{1, \ldots, K\}$:

$$p(\mathbf{f} \mid S_\kappa) = \frac{1}{M_\kappa} \sum_{i=1}^{M_\kappa} \mathcal{K}_\kappa(\mathbf{f} - \mathbf{f}_i) \ , \tag{5}$$

where $\mathcal{K}_\kappa$ is a zero-mean normal density with covariance matrix $\mathbf{\Sigma}_\kappa$ and the set $S_\kappa = \{\mathbf{f}_1, \ldots, \mathbf{f}_{M_\kappa}\}$ denotes the n-dimensional training data labeled with class κ. 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_κ, the smoothness depends solely on the kernel bandwidth parameter $\mathbf{\Sigma}_\kappa$. The appropriate choice of a bandwidth is the most critical step in kernel density estimation, since small bandwidths lead to over-fitting, whereas too large bandwidths result in oversmooth densities. In this paper, we use an ad-hoc method for bandwidth selection known as generalized *Scott's rule* [14] for kernel densities:

$$\boldsymbol{\Sigma}_\kappa \approx M_\kappa^{-\frac{2}{n+4}} \widehat{\boldsymbol{\Sigma}}_\kappa \ , \tag{6}$$

where $\widehat{\boldsymbol{\Sigma}}_\kappa$ is the sample covariance with respect to S_κ.

4.2 Randomized Decision Forest

A Randomized Decision Forest (RDF) is a discriminative classifier that can handle a large set of features without issues due to the curse of dimensionality. Standard decision tree approaches suffer from severe over-fitting problems. A RDF overcomes these problems by generating an ensemble (forest) of T decision trees. During the classification, the overall probability of a class κ given a feature

vector $\mathbf{f}$ can be obtained by simple averaging of the posterior probabilities $p_\tau(\cdot)$ estimated by each tree of the ensemble:

$$p(\kappa \mid \mathbf{f}) = \frac{1}{T} \sum_{\tau=1}^{T} p_\tau(\kappa \mid \mathbf{f}) \ . \tag{7}$$

In contrast to Boosting, the RDF approach uses two types of randomization to learn the ensemble. The first type of randomization is Bootstrap Aggregating [15], where each tree is trained with a random fraction of the training data. Additionally, to reduce training time and to incorporate randomization into the building process of a tree, the search for the most informative split function in each inner node is done using only a random fraction of all features [16].

4.3 Support Vector Machines

In the last years, Support Vector Machines (SVM) have emerged to one of the most popular machine learning techniques. For a basic introduction we refer the reader to the textbook of Bishop [13] and concentrate on the detailed setup used for our evaluation.

We train K SVM classifiers using the one-vs.-all principle. All scores are converted to suitable probabilities using the logistic regression method of Platt et al. [17]. The classification result is the class with the highest probability (score of the corresponding binary SVM classifier). Each single classifier uses a radial basis function kernel with parameter γ and trade-off parameter C [13] optimized with cross-validation. Instead of simple grid search, we apply cyclic coordinate search which is faster and yields in our experiments to similar optimal parameters.

4.4 Feature Combination and Temporal Context

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

To further improve the classification performance, a hidden Markov model (HMM) is used to exploit temporally contextual properties. We use the approach from Torralba et al. [4], but instead of a sparse Parzen classifier, we utilize the classifiers listed above.

5 Experiments

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

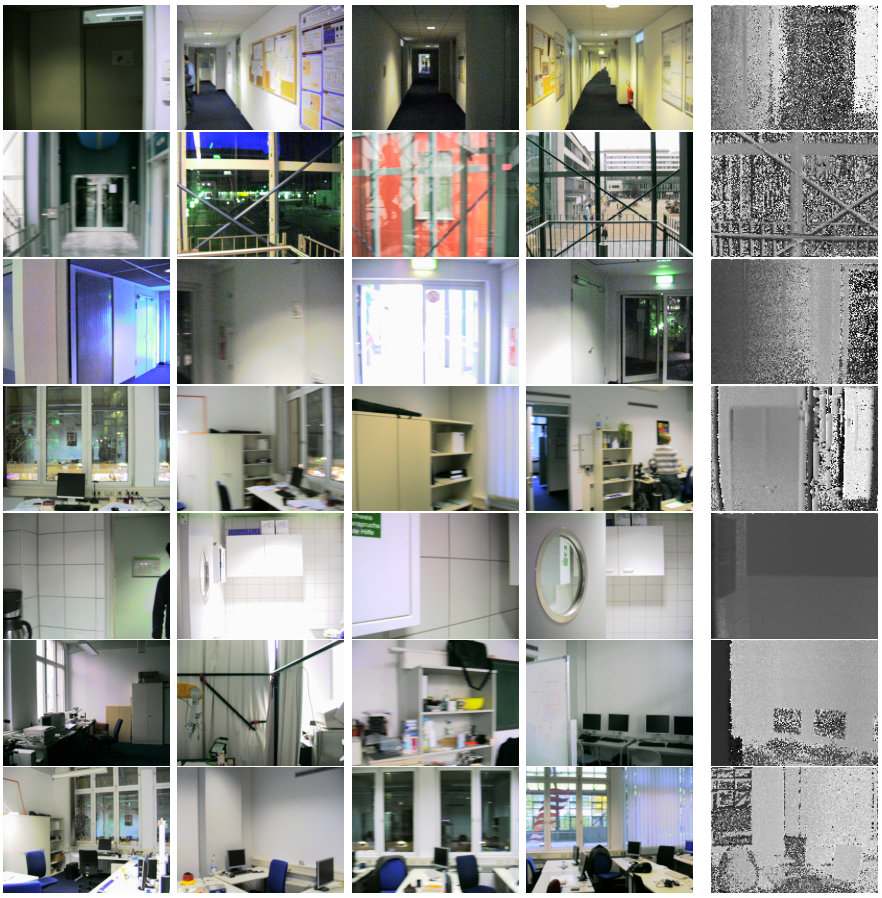


Fig. 4. Example images from different sequences, where each row comprises images from one scene. In addition to four visual example images, the range image which corresponds to the rightmost visual image is shown. The scene categories in our setting are (listed from top to bottom) *Corridor*, *Elevator Area*, *Entrance Area*, *PhD Lab*, *Kitchen*, *Robot Lab*, and *Student Lab*.

Table 1. Evaluation of different features types (incl. computation time) with the best classifier result and HMM integration. Features computed on the range image of the ToF sensor are tagged with a preceding $r-$.

Feature type	Avg. Recognition Rate	Time (in sec)
$r-$hist	51.8	0.024
$r-$power	48.5	0.031
$r-$gabor	45.5	0.140
$r-$surface	**53.8**	0.303
power	55.4	0.040
gabor	**64.6**	0.512
feature combination	**67.0**	0.839

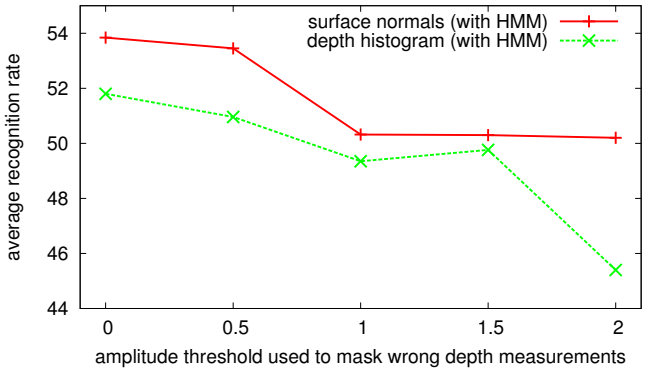


Fig. 5. Influence of additional preprocessing of the ToF data: all depth measurements below a given amplitude threshold are discarded in the computation of range features. A zero threshold corresponds to raw ToF data.

1. Incorporation of range features improves the recognition performance.
2. The Randomized Decision Forest classifier and the SVM classifier achieve the best recognition rates with a combination 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. Roughly 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. 4, visual and range 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 Preprocessing Techniques

Due to the severe noise of the ToF range data, one often has to mask outliers using the amplitude image. All depth measurements with corresponding amplitude value below a predefined threshold are discarded. Nevertheless, we do not apply this preprocessing technique prior to feature computation because it would decrease the recognition performance in our setting.

Table 2. Table of feature type combinations (among the tested subsets), which lead to the best recognition performances with HMM integration.

classifier	Gabor	power	$r-$Gabor	$r-$power	$r-$hist	$r-$surface	result
Parzen	×	×		×			65.4
MLP	×		×	×			65.5
RDF	×				×	×	**67.0**
SVM	×	×			×	×	65.6

We analyze this surprising effect in the following experiment. The recognition performance is evaluated for the surface normal feature and the range histogram feature using the RDF classifier with several values of the amplitude threshold. A threshold of zero corresponds to raw data without preprocessing.

The results are illustrated in Fig. 5. and show that the recognition rate decreases if we discard more and more measurements, even erroneous ones. Our place recognition system, therefore, seems to benefit also from wrong measurements which are possible cues of black or critical surfaces.

5.2 Evaluation of Feature Types and Combinations

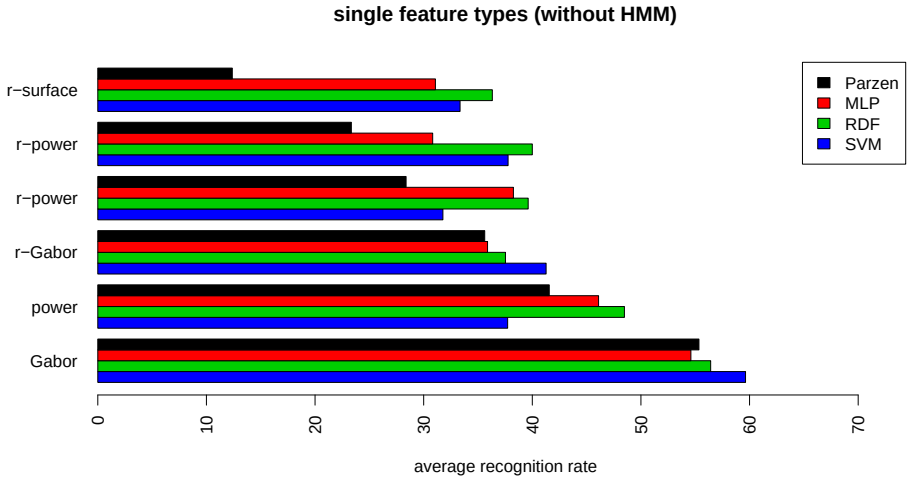
In order to evaluate the effects of combined features, we first analyzed the classification performance on each feature type separately. The recognition results are illustrated in detail in Fig. 6 and summarized in Table 1, where only the best (out of four) classifier result is shown. Regarding the range features, our experiments show that the surface normal histogram ($B_\varphi = B_\theta = 10,\ P = 3$) achieves the best place recognition result. However, Gabor and power spectrum 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 is a recognition rate of 67.0%.

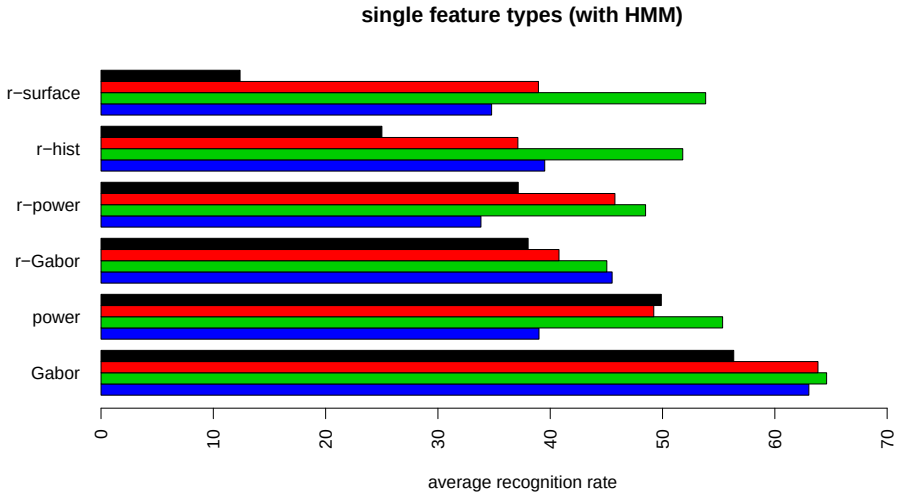
5.3 Evaluation of Different Classifiers

In the preceding section we showed that the combination of different feature types can improve the classification performance. However, the amount of performance gain depends on the used classifier. We also observed that the classifiers achieved best results when only a subset of all feature types were used. By analyzing either a manually chosen list of feature combinations (for RDF and SVM) or by applying a greedy search algorithm on the space of combinations (for Parzen and MLP), we obtained the results shown in Fig. 7 with corresponding combinations listed in Table 2. These average recognition rates suggest that the RDF is the appropriate classifier for our scene recognition task.

In order to further evaluate the power of range information, we removed all range features from the used feature type subsets mentioned above, i.e. only a



(a)



(b)

Fig. 6. Performances of single features types without hidden Markov model (a) and with hidden Markov model (b).

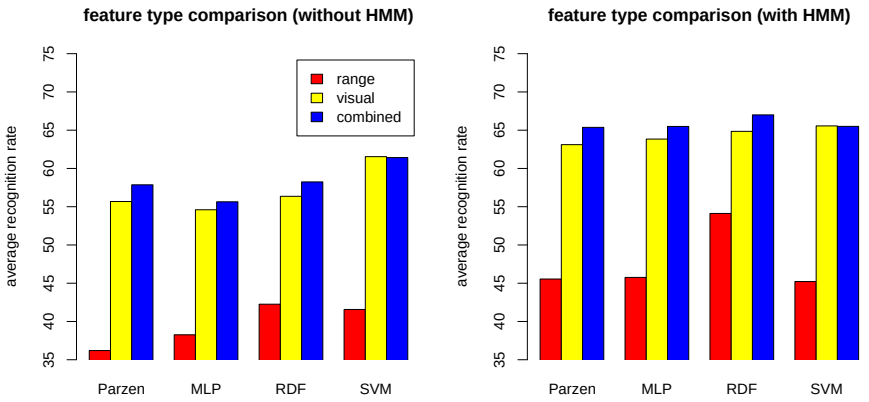


Fig. 7. Comparison of various feature type combinations (sensor-specific and mixed).

combination of visual features remains. The average recognition rates in Fig. 7 (visual) illustrates a drop in classification performance for all classifiers except SVM. These results clearly show the advantage of our multi-sensor approach.

It can be also seen that without the integration of the hidden Markov model the recognition performance decreases substantially. This observation highlights the importance of temporally contextual information in our scene recognition task.

Finally, in order to allow a more detailed analysis of the scene recognition result obtained by the best feature type combination, we computed the confusion matrix for this setting (averaged over 30 results). As can be seen in Fig. 9, the recognition rates for six out of eight rooms vary between 76.9% and 85.3%. The significantly lower overall recognition rate (67.0%) is thus directly related to the low recognition rates of the remaining two categories *PhD Lab* and *Robot Lab*, which tend to be recognized as *Student Lab*. This behavior stems from the close holistic similarity of these rooms and suggests that more locally receptive features could be promising in order to differentiate between these similar rooms.

5.4 Influence of the number of trees

In our previous experiments we used $T = 100$ trees for the randomized decision forest. To investigate the influence of this parameter we perform tests with Gabor and combined features without HMM. The results can be seen in Fig. 8. To cope with the randomization, we average the results of 200 runs for each data point. As can be seen, the generalization performance increases with the number of trees even beyond $T = 100$. However, this effect levels out after a specific size of the forest.

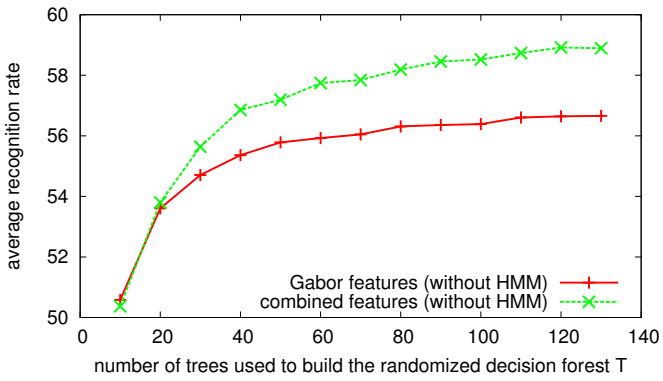


Fig. 8. Influence of the number of trees for the randomized decision forest with visual Gabor features or a combination of range and visual features.

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 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 been 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 [11].

Acknowledgements We would like to thank all four anonymous reviewers for their valuable comments, which really helped to improve the quality of the paper.

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

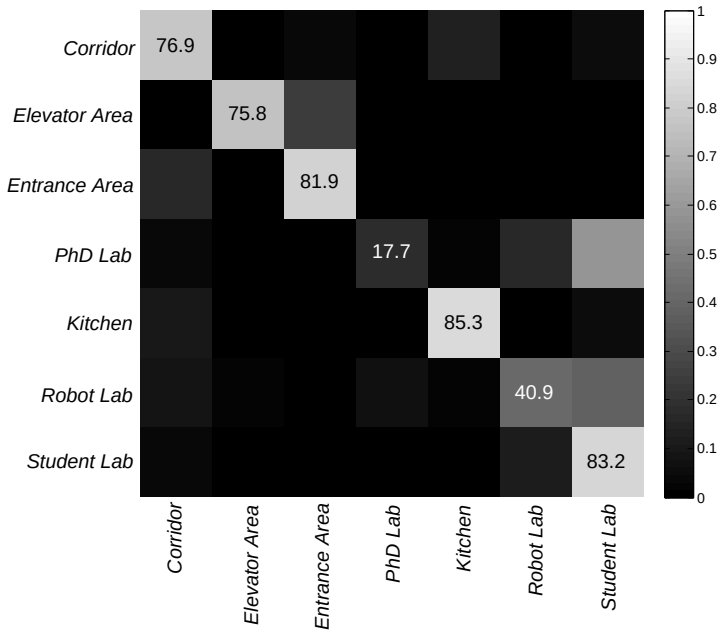


Fig. 9. Confusion matrix for the best feature combination setting (RDF+HMM).

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. Redies, C., Hasenstein, J., Denzler, J.: Fractal-like image statistics in visual art: similarity to natural scenes. Spatial Vision **21** (2007) 137–148
11. Torralba, A.: Contextual priming for object detection. International Journal of Computer Vision **53** (2003) 169–191
12. Wainwright, M.J., Schwartz, O., Simoncelli, E.P.: Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In Rao, R., Olshausen, B., Lewicki, M., eds.: Probabilistic Models of the Brain: Perception and Neural Function. MIT Press (2002) 203–222

13. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). 1 edn. Springer (2007)
14. Schimek, M.G.: Smoothing and Regression: Approaches, Computation, and Application. Series in Probability and Statistics. Wiley (1996)
15. Breiman, L.: Random forests. Machine Learning **45** (2001) 5–32
16. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Maching Learning **63** (2006) 3–42
17. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. In Smola, A., Bartlett, P., Schoelkopf, B., Schuurmans, D., eds.: Advances in Large Margin Classifiers. (2000) 61–74