Multi-modal Semantic Place Classification

A. Pronobis
Centre for Autonomous Systems,
Royal Institute of Technology,
SE-100 44 Stockholm, Sweden
pronobis@csc.kth.se

O. Martínez Mozos
Department of Computer Science,
University of Freiburg,
D-79110, Freiburg, Germany
omartine@informatik.uni-freiburg.de

B. Caputo
Idiap Research Institute,
CH-1920 Martigny, Switzerland
bcaputo@idiap.ch

P. Jensfelt
Centre for Autonomous Systems,
Royal Institute of Technology,
SE-100 44 Stockholm, Sweden
patric@csc.kth.se

Abstract

The ability to represent knowledge about space and its position therein is crucial for a mobile robot. To this end, topological and semantic descriptions are gaining popularity for augmenting purely metric space representations. In this paper we present a multi-modal place classification system that allows a mobile robot to identify places and recognize semantic categories in an indoor environment. The system effectively utilizes information from different robotic sensors by fusing multiple visual cues and laser range data. This is achieved using a high-level cue integration scheme based on a Support Vector Machine (SVM) that learns how to optimally combine and weight each cue. Our multi-modal place classification approach can be used to obtain a real-time semantic space labeling system which integrates information over time and space. We perform an extensive experimental evaluation of the method for two different platforms and environments, on a realistic off-line database and in a live experiment on an autonomous robot. The results clearly demonstrate the effectiveness of our cue integration scheme and its value for robust place classification under varying conditions.

KEY WORDS—recognition, sensor fusion, localization, multi-modal place classification, sensor and cue integration, semantic annotation of space

The International Journal of Robotics Research
Vol. 00, No. 00, Xxxxxxxx 2009, pp. 000–000
DOI: 10.1177/0278364909356483

Figures 1–15, 17, 18 appear in color online: http://ijr.sagepub.com

1. Introduction

The most fundamental competence for an autonomous mobile agent is to know its position in the world. This can be represented in terms of raw metric coordinates, topological location, or even semantic description. Recently, there has been a growing interest in augmenting (or even replacing) purely metric space representations with topological and semantic place information. Several attempts have been made to build autonomous cognitive agents able to perform human-like tasks[1]. Enhancing the space representation to be more meaningful from the point of view of spatial reasoning and human–robot interaction have been at the forefront of the issues being addressed (Kuipers 2006; Topp and Christensen 2006; Zender et

[1] See, e.g., CoSy (Cognitive Systems for Cognitive Assistants) http://www.cognitivesystems.org/ or COGNIRON (the cognitive robot companion) http://www.cogniron.org.

al. 2008). Indeed, in the concrete case of indoor environments, the ability to understand the existing topological relations and associate semantic terms such as “corridor” or “office” with places, gives a much more intuitive idea of the position of the robot than global metric coordinates. In addition, the semantic information about places can extend the capabilities of a robot in other tasks such as localization (Rottmann et al. 2005), exploration (Stachniss et al. 2006), or navigation (Galindo et al. 2005).

Nowadays, robots are usually equipped with several sensors providing both geometrical and visual information about the environment. Previous work on place classification relied on sonar and/or laser range data as robust sensory modalities (Mozos et al. 2005). However, the advantages of geometric solutions, such as invariance to visual variations and low dimensionality of the processed information, quickly became their weaknesses. The inability to capture many aspects of complex environments leads to the problem of perceptual aliasing (Kuipers and Beeson 2002) and can limit the usefulness of such methods for topological and semantic mapping. Recent advances in vision have made this modality emerge as a natural and viable alternative. Vision provides richer sensory input allowing for better discrimination. Moreover, a large share of the semantic description of a place is encoded in its visual appearance. However, visual information tends to be noisy and difficult to interpret as the appearance of places varies over time due to changing illumination and human activity. At the same time, the visual variability within place classes is huge, making the semantic place classification a challenging problem. Clearly, each modality has its own characteristics. Interestingly, the weaknesses of one often correspond to the strengths of the other.

In this paper, we propose an approach to semantic place classification which combines the stability of geometrical solutions with the versatility of vision. First, we present a recognition system implemented on a mobile robot platform integrating multiple cues and modalities. The system is able to perform robust place classification under different types of variations that occur in indoor environments over a span of time of several months. This comprises variations in illumination conditions and in configuration of furniture and small objects. The system relies on different types of visual information provided by global and local descriptors and on geometric cues derived from laser range scans. For the vision channel we apply the Scale-Invariant Feature Transform (SIFT) (Lowe 2004) and Composed Receptive Field Histograms (CRFH) (Linde and Lindeberg 2004). For the laser channel we use the features proposed in Mozos et al. (2005, 2007).

We combine the cues using a new high-level accumulation scheme, which builds on our previous work (Nilsback and Caputo 2004; Pronobis and Caputo 2007). We train for each cue a large margin classifier which outputs a set of scores encoding confidence of the decision. Integration is then achieved by feeding the scores to a Support Vector Machine (SVM) (Cristianini and Shawe-Taylor 2000). Such an approach allows to optimally combine cues, even obtained using different types of models, with a complex, possibly non-linear function. We call this algorithm the SVM-based Discriminative Accumulation Scheme (SVM-DAS).

Finally, we show how to build a self-contained semantic space labeling system, which relies on multi-modal place classification as one of its components. The system is implemented as a part of an integrated cognitive robotic architecture[2] and runs on-line on a mobile robot platform. While the robot explores the environment, the system acquires evidence about the semantic category of the current area produced by the place classification component and accumulates them both over time and space. As soon as the system is confident about its decision, the area is assigned a semantic label. We integrate the system with a Simultaneous Localization and Mapping (SLAM) algorithm and show how a metric and topological space representation can be augmented with a semantic description.

We evaluated the robustness of the presented methods in several sets of extensive experiments. We conducted experiments on two different robot platforms, in two different environments and for two different scenarios. First, we run a series of off-line experiments of increasing difficulty on the IDOL2 database (Luo et al. 2006) to precisely measure the performance of the place classification algorithm in presence of different types of variations. These ranged from short-term visual variations caused by changing illumination to long-term changes which occurred in the office environment over several months. Second, we run a live experiment where a robot performs SLAM and semantic labeling in a new environment using prebuilt models of place categories. Results show that integrating different visual cues, as well as different modalities, allows to greatly increase the robustness of the recognition system, achieving high accuracy under severe dynamic variations. Moreover, the place classification system, when used in the framework of semantic space labeling, can yield a fully correct semantic representation even for a new, unknown environment.

The rest of the paper is organized as follows. After a review of the related literature (Section 2), Section 3 presents the main principle behind our multi-modal place classification algorithm and describes the methods used to extract each cue. Then, Section 4 gives details about the new cue integration scheme and Section 5 describes the architecture of the semantic labeling system. Finally, Section 6 presents detailed experimental evaluation of the place classification system and Section 7 reports results of the live experiment with semantic labeling of space. The paper concludes with a summary and possible avenues for future research.

[2] See CoSy (Cognitive Systems for Cognitive Assistants) http://www.cognitivesystems.org/ and CAST (The CoSy Architecture Schema Toolkit) http://www.cs.bham.ac.uk/research/projects/cosy/cast/.

2. Related Work

Place classification is a vastly researched topic in the robotic community. Purely geometric solutions based on laser range data have proven to be successful for certain tasks and several approaches were proposed using laser scanners as the only sensors. Koenig and Simmons (1998) used a pre-programmed routine to detect doorways from range data. In addition, Althaus and Christensen (2003) used line features to detect corridors and doorways. In their work, Buschka and Saffiotti (2002) partitioned grid maps of indoor environments into two different classes of open spaces, i.e. rooms and corridors. The division of the open spaces was done incrementally on local submaps. Finally, Mozos et al. (2005) applied boosting to create a classifier based on a set of geometrical features extracted from range data to classify different places in indoor environments. A similar idea was used by Topp and Christensen (2006) to describe regions from laser readings.

The limitations of geometric solutions inspired many researchers to turn towards vision which nowadays becomes tractable in real-time applications. The proposed methods employed either perspective (Torralba et al. 2003; Tamimi and Zell 2004; Filliat 2007) or omnidirectional cameras (Gaspar et al. 2000; Ulrich and Nourbakhsh 2000; Blaer and Allen 2002; Menegatti et al. 2004; Andreasson et al. 2005; Murillo et al. 2007; Valgren and Lilienthal 2008). The main differences between the approaches relate to the way the scene is perceived, and thus the method used to extract characteristic features from the scene. Landmark-based techniques make use of either artificial or natural landmarks in order to extract information about a place. Siagian and Itti (2007) relied on visually distinctive image regions as landmarks. Many other solutions employed local image features, with SIFT being the most frequently applied (Se et al. 2001; Lowe 2004; Andreasson et al. 2005; Pronobis and Caputo 2007). Zivkovic et al. (2005) used the SIFT descriptor to build a topological representation by clustering a graph representing relations between images. Other approaches used the bag-of-words technique (Filliat 2007; Fraundorfer et al. 2007), the SURF features (Bay et al. 2006; Murillo et al. 2007; Valgren and Lilienthal 2008), or representation based on information extracted from local patches using Kernel PCA (Tamimi and Zell 2004). Global features are also commonly used for place recognition. Torralba et al. (Torralba and Sinha 2001; Torralba et al. 2003; Torralba 2003) suggested to use a representation called the "gist" of a scene, which is a vector of principal components of outputs of a bank of spatially organized filters applied to the image. Other approaches use color histograms (Ulrich and Nourbakhsh 2000; Blaer and Allen 2002), gradient orientation histograms (Bradley et al. 2005), eigenspace representation of images (Gaspar et al. 2000), or Fourier coefficients of low-frequency image components (Menegatti et al. 2004).

In all of the previous approaches only one modality is used for the recognition of places. Recently, several authors observed that robustness and efficiency of the recognition system can be improved by combining information provided by different visual cues. Siagian and Itti (2007) and Weiss et al. (2007) used a global representation of the images together with local visual landmarks to localize a robot in outdoor environments. Pronobis and Caputo (2007) used two cues composed of global and local image features to recognize places in indoor environments. The cues were combined using discriminative accumulation. Here, we extend this approach by integrating information provided by a laser range sensor using a more sophisticated algorithm.

Other approaches also employed a combination of different sensors, mainly laser and vision. Tapus and Siegwart (2005) used an omnidirectional camera and two lasers covering 360° field of view to extract fingerprints of places for topological mapping. This approach was not used for extracting semantic information about the environment. Posner et al. (2007) and Douillard et al. (2007) relied on range data and vision for recognition of objects in outdoor environments (e.g. grass, walls, or cars). Finally, Rottmann et al. (2005) used a combination of both modalities to categorize places in indoor environments. Each observation was composed of a set of geometrical features and a set of objects found in the scene. The geometrical features were calculated from laser scans and the objects were detected using Haar-like features from images. The extracted information was integrated at the feature level. In contrast, the method presented in this work learns how to combine and weigh outputs of several classifiers, keeping features and therefore the information from different modalities separate.

Various cue integration methods have been proposed in the robotics and machine learning community (Poggio et al. 1985; Matas et al. 1995; Triesch and Eckes 1998; Nilsback and Caputo 2004; Tapus and Siegwart 2005; Pronobis and Caputo 2007). These approaches can be described according to various criteria. For instance, Clark and Yuille (1990) suggest to classify them into two main groups, *weak coupling* and *strong coupling*. Assuming that each cue is used as input of a different classifier, weak coupling is when the output of two or more independent classifiers are combined. Strong coupling is when the output of one classifier is affected by the output of another classifier, so that their outputs are no longer independent. Another possible classification is into *low-level* and *high-level* integration methods, where the emphasis is on the level at which integration happens. We call *low-level integration methods* those algorithms where cues are combined together at the feature level, and then used as input to a single classifier. This approach has been used successfully for object recognition using multiple visual cues (Matas et al. 1995), and for topological mapping using multiple sensor modalities (Tapus and Siegwart 2005). In spite of remarkable performances for specific tasks, there are several drawbacks of the low-level methods. First, if one of the cues gives misleading information, it is quite probable that the new feature vector will be adversely affected influencing the whole performance. Second, we can

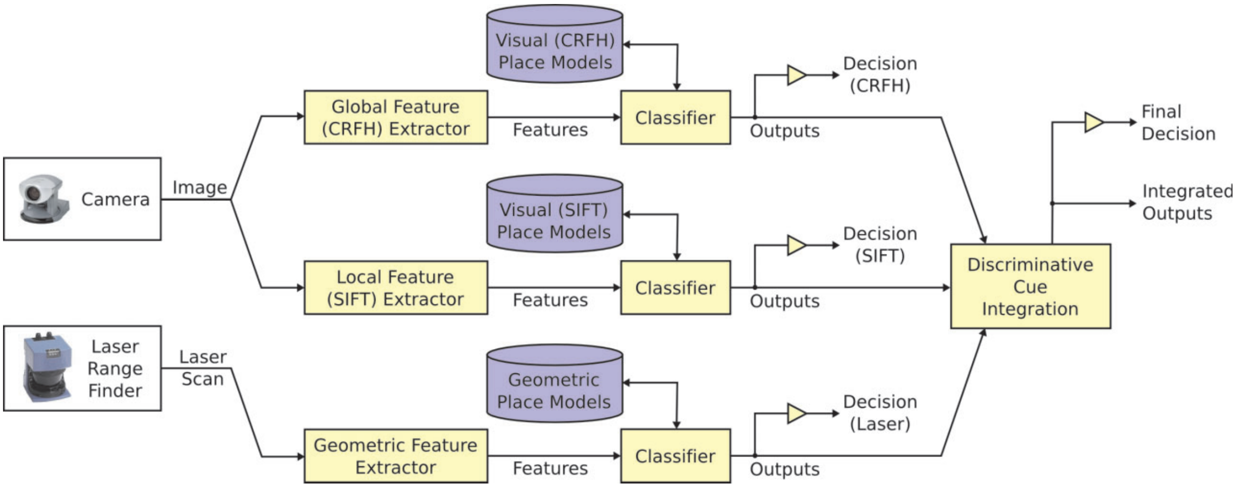


Fig. 1. Architecture of the multi-modal place classification system.

expect the dimension of such a feature vector to increase as the number of cues grow, and each of the cues needs to be used even if one would allow for correct classification. This implies longer learning and recognition times, greater memory requirements, and possible curse of dimensionality effects. Another strategy is to keep the cues separated and to integrate the outputs of individual classifiers, each trained on a different cue (Poggio et al. 1985; Nilsback and Caputo 2004; Pronobis and Caputo 2007). We call such algorithms *high-level integration methods*, of which voting is the most popular (Duda et al. 2001). These techniques are more robust with respect to noisy cues or sensory channels, allow the use of different classifiers adapted to the characteristics of each single cue and decide on the number of cues that should be extracted and used for each particular classification task (Pronobis and Caputo 2007). In this paper, we focus on a weak coupling, high-level integration method called *accumulation*. The underlying idea is that information from different cues can be summed together, thus accumulated. The idea was first proposed in probabilistic framework by Poggio et al. (1985) and further explored by Aloimonos and Shulman (1989). The method was then extended to discriminative methods in Nilsback and Caputo (2004) and Pronobis and Caputo (2007).

3. Multi-modal Place Classification

The ability to integrate multiple cues, possibly extracted from different sensors, is an important skill for a mobile robot. Different sensors usually capture different aspects of the environment. Therefore using multiple cues leads to obtaining a more descriptive representation. The visual sensor is an irreplaceable source of distinctive information about a place. However, this information tends to be noisy and difficult to analyze due to the susceptibility to variations introduced by changing illumination and everyday activities in the environment. At the same time, most recent robotic platforms are equipped with a laser range scanner which provides much more stable and robust geometric cues. These cues however, are unable to uniquely represent the properties of different places (perceptual aliasing) (Kuipers and Beeson 2002). Clearly performance could increase if different cues were combined effectively. Note that even alternative interpretations of the information obtained by the same sensor can be valuable, as we will show experimentally in Section 6.

This section describes our approach to multi-modal place classification. Our method is fully supervised and assumes that during training, each place (room) is represented by a collection of labeled data which captures its intrinsic visual and geometric properties under various viewpoints, at a fixed time and illumination setting. During testing, the algorithm is presented with data samples acquired in the same places, under roughly similar viewpoints but possibly under different conditions (e.g. illumination), and after some time (where the time range goes from some minutes to several months). The goal is to recognize correctly each single data sample provided to the system.

The architecture of the system is illustrated in Figure 1. We see that there is a separate path for each cue. We use two different visual cues corresponding to two types of image features (local and global) as well as simple geometrical features extracted from laser range scans. Each path consists of two main building blocks: a feature extractor and a classifier. Thus, separate decisions can be obtained for each of the cues in case only one cue is available. Alternatively, our method could decide when to acquire additional information (e.g. only in difficult cases) (Pronobis and Caputo 2007). In cases when several cues are available, the outputs encoding the confidence of the single-cue classifiers are combined using an efficient discriminative accumulation scheme.

The rest of this section gives details about the algorithms used to extract and classify each of the cues for the vision-based paths (Section 3.1) and laser-based path (Section 3.2). A comprehensive description of the algorithms used for cue integration is given in Section 4.

3.1. Vision-based Place Classification

As a basis for the vision-based channel, we used the place recognition system presented in Pronobis et al. (2006) and Pronobis and Caputo (2007), which is built around a SVM classifier (Cristianini and Shawe-Taylor 2000) and two types of visual features, global and local, extracted from the same image frame. We used CRFH (Linde and Lindeberg 2004) as global features, and SIFT (Lowe 2004) as local features. Both have already been proved successful in the domain of vision-based place recognition (Pronobis et al. 2006; Pronobis and Caputo 2007) and localization and mapping (Se et al. 2001; Andreasson et al. 2005).

CRFHs are a sparse multi-dimensional statistical representation of responses of several image filters applied to the input image. Following Pronobis et al. (2006), we used histograms of six dimensions, with 28 bins per dimension, computed from second-order normalized Gaussian derivative filters applied to the illumination channel at two scales. The SIFT descriptor instead represents local image patches around interest points characterized by coordinates in the scalespace in the form of histograms of gradient directions. To find the coordinates of the interest points, we used a scale and affine invariant region detector based on the difference-of-Gaussians (DoG) operator (Rothganger et al. 2006).

We used SVMs for creating models from both visual cues. A review of the theory behind SVMs can be found in Section 4.1. In case of SVMs, special care must be taken in choosing an appropriate kernel function. Here we used the χ^2 kernel (Chapelle et al. 1999) for CRFH, and the match kernel proposed by Wallraven et al. (2003) for SIFT. Both have been used in our previous work on SVM-based place recognition, obtaining good performances.

3.2. Laser-based Place Classification

In addition to the visual channel, we used a laser range sensor. A single two-dimensional (2D) laser scan covered a field of view of 180° in front of the robot. A laser observation $z = \{b_0, \ldots, b_{M-1}\}$ contains a set of beams b_i, in which each beam b_i consists of a tuple (α_i, d_i), where α_i is the angle of the beam relative to the robot and d_i is the length of the beam.

For each laser observation, we calculated a set of simple geometric features represented by single real values. The features were introduced for place classification by Mozos et al. (2005) where laser observations covering a 360° field of view were used. The complete set of features consists of two subsets. The first subset contains geometrical features calculated directly from the laser beams. The second subset comprises geometrical features extracted from a polygon approximation of the laser observation. This polygon is created by connecting the end points of the beams. The selection of features is based on the results presented in Mozos et al. (2005, 2007).

As classifiers for the laser-based channel, we tried both AdaBoost (Freund and Schapire 1995), following the work in Mozos et al. (2007), and SVMs. In the rest of the paper, we will refer to the two laser-based models as L-AB and L-SVM, respectively. For the geometric features, we used a Radial Basis Function (RBF) kernel (Cristianini and Shawe-Taylor 2000) with SVMs, chosen through a set of reference experiments[3]. Both classifiers were benchmarked on the laser-based place classification task. Results presented in Section 6.2 show an advantage of the more complex SVM classifier.

4. Discriminative Cue Integration

This section describes our approach to cue integration from one or multiple modalities. We propose an SVM-DAS, a technique performing non-linear cue integration by discriminative accumulation. For each cue, we train a separate large margin classifier which outputs a set of scores (outputs), encoding the confidence of the decision. We achieve integration by feeding the scores to an SVM. Compared to previous accumulation methods (Poggio et al. 1985; Caputo and Dorko 2002; Nilsback and Caputo 2004; Pronobis and Caputo 2007), SVM-DAS gives several advantages: (a) discriminative accumulation schemes achieve consistently better performances than probabilistic ones (Poggio et al. 1985; Caputo and Dorko 2002), as shown in Nilsback and Caputo (2004); (b) compared with previous discriminative accumulation schemes (Nilsback and Caputo 2004; Pronobis and Caputo 2007), our approach accumulates cues with a more complex, possibly non-linear function, by using the SVM framework and kernels (Cristianini and Shawe-Taylor 2000). Such an approach makes it possible to integrate outputs of different classifiers such as SVM and AdaBoost. At the same time, it learns the weight for each cue very efficiently, therefore making it possible to accumulate large numbers of cues without computational problems.

In the rest of the section we first sketch the theory behind SVMs (Section 4.1), a crucial component in our approach. We then describe the Generalized Discriminative Accumulation Scheme (G-DAS; see Pronobis and Caputo (2007) and Section 4.2) on which to a large extent we build. Finally, we introduce the new algorithm and discuss its advantages in Section 4.3.

[3] In the case of AdaBoost, we constructed a multi-class classifier by arranging several binary classifiers into a decision list in which each element corresponded to one specific class.

4.1. Support Vector Machines

Consider the problem of separating the set of labeled training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n), \ldots, (\boldsymbol{x}_N, y_N)$ into two classes, where $\boldsymbol{x}_n \in \Re^L$ is a feature vector and $y_n \in \{-1, +1\}$ its class label. Assuming that the two classes can be separated by a hyperplane in some Hilbert space $\mathcal{H}$, then the optimal separating hyperplane is the one which has maximum distance to the closest points in the training set resulting in a discriminant function

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n y_n \mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}) + \beta.$$

The classification result is then given by the sign of $f(\boldsymbol{x})$. The values of α_n and β are found by solving a constrained minimization problem, which can be done efficiently using the SMO algorithm (Platt 1999). Most of the α_n's take the value of zero; those $\boldsymbol{x}_n$ with non-zero α_n are the "support vectors". In case where the two classes are non-separable, the optimization is formulated in such a way that the classification error is minimized and the final solution remains identical. The mapping between the input space and the usually high-dimensional feature space $\mathcal{H}$ is done using kernels $\mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x})$.

The extension of SVM to multi-class problems can be done in several ways. Here we mention three approaches used throughout the paper:

1. *Standard one-against-all (OaA) strategy*. If M is the number of classes, M SVMs are trained, each separating a single class from all other classes. The decision is then based on the distance of the classified sample to each hyperplane, and the sample is assigned to the class corresponding to the hyperplane for which the distance is largest.

2. *Modified OoA strategy*. In Pronobis and Caputo (2007), a modified version of the OaA principle was proposed. The authors suggested to use distances to precomputed average distances of training samples to the hyperplanes (separately for each of the classes), instead of the distances to the hyperplanes directly. In this case, the sample is assigned to the class corresponding to the hyperplane for which the distance is smallest. Experiments presented in this paper and in Pronobis and Caputo (2007) show that in many applications this approach outperforms the standard OaA technique.

3. *One-against-one (OaO) strategy*. In this case, $M(M-1)/2$ two-class SVMs are trained for each pair of classes. The final decision can then be taken in different ways, based on the $M(M-1)/2$ outputs. A popular choice is to consider as output of each classifier the class label and count votes for each class; the test image is then assigned to the class that received more votes.

SVMs do not provide any out-of-the-box solution for estimating confidence of the decision; however, it is possible to derive confidence information and hypotheses ranking from the distances between the samples and the hyperplanes. In the work presented in this paper, we applied the distance-based methods proposed by Pronobis and Caputo (2007), which define confidence as a measure of unambiguity of the final decision related to the differences between the distances calculated for each of the binary classifiers.

4.2. Generalized Discriminative Accumulation Scheme

G-DAS was first proposed by Pronobis and Caputo (2007), as a more effective generalization of the algorithm presented in Nilsback and Caputo (2004). It accumulates multiple cues, possibly from different modalities, by turning classifiers into experts. The basic idea is to consider real-valued outputs of a multi-class discriminative classifier (e.g. SVM) as an indication of a soft decision for each class. Then, all of the outputs obtained from the various cues are summed together, therefore linearly accumulated. Specifically, suppose we are given M classes and, for each class, a set of N_m training samples $\{\{\boldsymbol{s}_{m,n}\}_{n=1}^{N_m}\}_{m=1}^{M}$. Suppose also that, from each sample, we extract a set of T different cues $\{\mathcal{T}_t(\boldsymbol{s}_{m,n})\}_{t=1}^{T}$. The goal is to perform recognition using all of them. The G-DAS algorithm consists of two steps:

1. *Single-cue Models.* From the original training set $\{\{\boldsymbol{s}_{m,n}\}_{n=1}^{N_m}\}_{m=1}^{M}$, containing samples belonging to all M classes, define T new training sets $\{\{\mathcal{T}_t(\boldsymbol{s}_{m,n})\}_{n=1}^{N_m}\}_{m=1}^{M}$, $t = 1, \ldots, T$, each relative to a single cue. For each new training set train a multi-class classifier. Then, given a test sample $\boldsymbol{s}$, for each of the T single-cue classifiers estimate a set of outputs $\{\mathcal{V}_{t,\upsilon}(\mathcal{T}_t(\boldsymbol{s}))\}_{\upsilon=1}^{V}$ reflecting the relation of the sample to the model. In the case of the SVMs with standard OaO and OaA multi-class extensions, the outputs would be values of the discriminant functions learned by the SVM algorithm during training, i.e. $\mathcal{V}_{t,\upsilon}(\mathcal{T}_t(\boldsymbol{s})) = f_{t,\upsilon}(\mathcal{T}_t(\boldsymbol{s}))$, $\upsilon = 1, \ldots, V$, and $V = M(M-1)/2$ for OaO or $V = M$ for OaA.

2. *Discriminative Accumulation.* After all the outputs are computed for all the cues, combine them with different weights by a linear function:

$$\mathcal{V}_\upsilon(\boldsymbol{s}) = \sum_{t=1}^{T} \sigma_t \mathcal{V}_{t,\upsilon}(\mathcal{T}_t(\boldsymbol{s})),\ \sigma_t \in \Re^+,\ \upsilon = 1, \ldots, V.$$

   The final decision can be estimated with any method commonly used for multi-class, single-cue SVM.

An important property of accumulation is the ability to perform correct classification even when each of the single cues

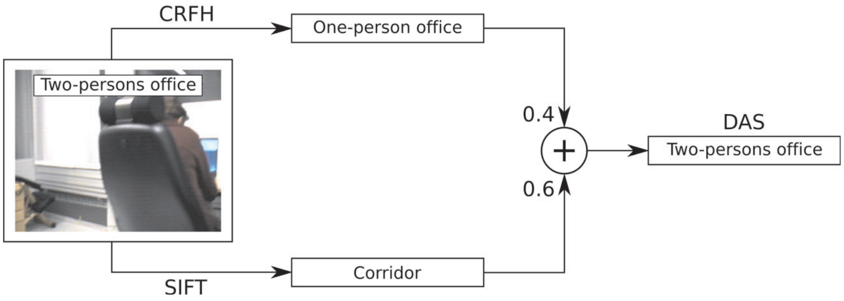


Fig. 2. A real example of test image misclassified by each of the single cues, but classified correctly using G-DAS.

gives misleading information. This behavior is illustrated on a real example in Figure 2. Despite these advantages, G-DAS presents some potential limitations: First, it uses only one weight for all outputs of each cue. This simplifies the parameter estimation process (usually, an extensive search is performed to find the coefficients $\{\sigma_t\}_{t=1}^{T}$), but also constrains the ability of the algorithm to adapt to the properties of each single cue. Second, accumulation is obtained via a linear function, which might not be sufficient in case of complex problems. The next section shows how our new accumulation scheme, SVM-DAS, addresses these issues.

4.3. SVM-based Discriminative Accumulation Scheme

The SVM-DAS accumulates the outputs generated by single-cue classifiers by using a more complex, possibly non-linear function. The outputs are used as an input to an SVM, and the parameters of the integration function are learned during the optimization process, for instance using the SMO algorithm (Platt 1999). These characteristics address the potential drawbacks of G-DAS discussed in the previous section.

More specifically, the new SVM-DAS accumulation function is given by

$$\mathcal{V}_u(\boldsymbol{s}) = \sum_{n=1}^{N} \alpha_{u,n} y_n \mathcal{K}(\boldsymbol{v}_n, \boldsymbol{v}) + \beta_u, \quad u = 1, \ldots, U,$$

where $\boldsymbol{v}$ is a vector containing all the outputs for all T cues:

$$\boldsymbol{v} = \left[\{\mathcal{V}_{1,\upsilon}(\mathcal{T}_1(\boldsymbol{s}))\}_{\upsilon=1}^{V_1}, \ldots, \{\mathcal{V}_{T,\upsilon}(\mathcal{T}_T(\boldsymbol{s}))\}_{\upsilon=1}^{V_T} \right].$$

The parameters $\alpha_{u,n}$, y_n, β_u, and the support vectors $\boldsymbol{v}_n$ are inferred from the training data either directly or efficiently during the optimization process. The number of the final outputs U and the way of obtaining the final decision depends on the multi-class extension used with SVM-DAS. We use the OaO extension throughout the paper for which $U = M(M-1)/2$.

The non-linearity is given by the choice of the kernel function $\mathcal{K}$, thus in the case of the linear kernel the method is still linear. In this sense, SVM-DAS is more general than G-DAS, while it preserves all of its important properties (e.g. the ability to give correct results for two misleading cues, see Figure 2). Also, for SVM-DAS each of the integrated outputs depend on all the outputs from single-cue classifiers, and the coefficients are learned optimally. Note that the outputs $\mathcal{V}_{t,\upsilon}(\mathcal{T}_t(\boldsymbol{s}))$ can be derived from a combination of different large margin classifiers, and not only from SVM[4].

5. Place Classification for Semantic Space Labeling

One of the applications of a place classification system is semantic labeling of space. This section provides a brief overview of the problem and describes how we employed our multi-modal place classification method to build a semantic labeling system. We evaluated the system in a live experiment described in Section 7.

5.1. Semantic Labeling of Space

The problem of semantic labeling can be described as assigning meaningful semantic descriptions (e.g. “corridor” or “kitchen”) to areas in the environment. Typically, semantic labeling is used as a way of augmenting the internal space representation with additional information. This can be used by the agent to reason about space and to enhance communication with a human user. In case of most typical environments, it is sufficient to distinguish between semantic categories which are usually associated with rooms (Zender et al. 2007), such as “office”, “meeting room” or “corridor”. It is labeling at this level that we will discuss in this paper.

[4] SVM-DAS can be seen as a variation of ensemble learning methods that employ multiple models to improve the recognition performance. The key reason why ensemble algorithms obtain better results is because the individual classifiers make errors on different data points. Typically, different training data is used for each classifier (Polikar 2006). In our experiments, we use data representing different types of information, e.g. obtained using different sensors.

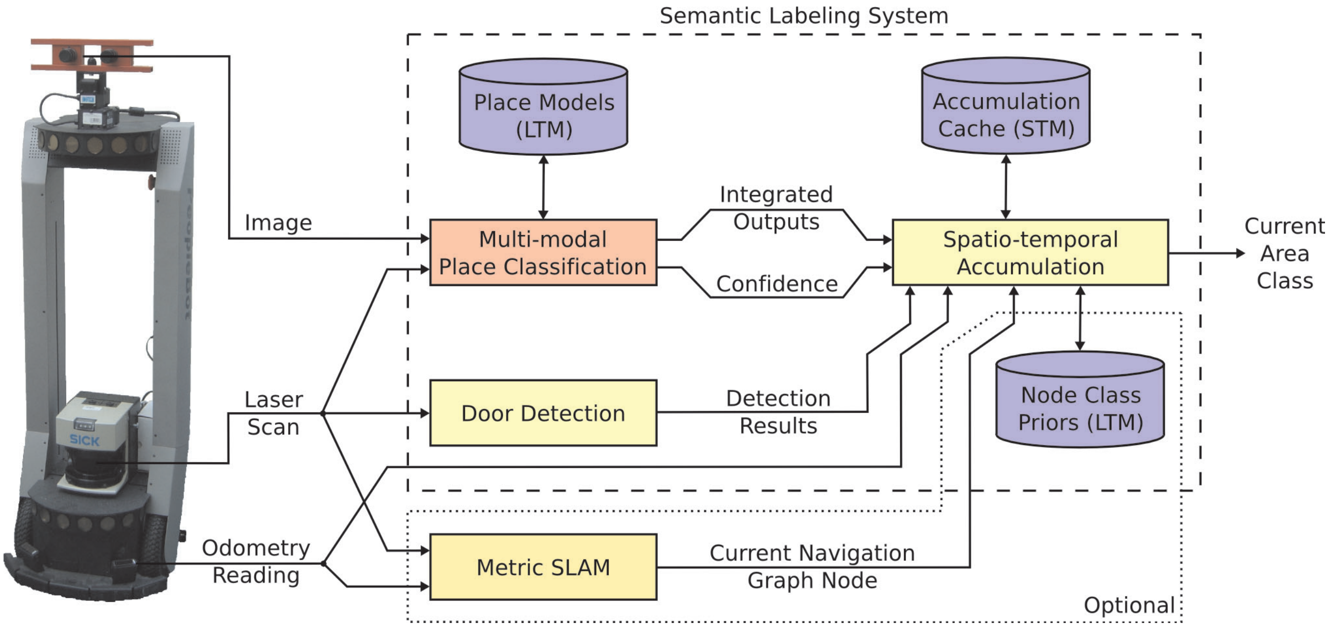


Fig. 3. Architecture of the semantic space labeling system based on place classification (LTM: Long-Term Memory; STM: Short-Term Memory).

As will be shown through experiments in Sections 6 and 7, the place classification system described in this paper can yield a place class with high accuracy given a single sample of multi-modal data (e.g. one image and a laser scan). However, when used for semantic labeling, the algorithm is requested to provide a label for the whole area under exploration. At the same time, the system must be resilient and able to deal with such problems as temporary lack of informative cues, continuous stream of similar information or long-term occlusions. Given that the system is operating on a mobile robot, crude information about its movement is available from wheel encoders. This information can be used to ensure robustness to the typical variations that occur in the environment but also to the problems mentioned above. Finally, the system should be able to measure its own confidence and restrain from making a decision until some confidence level is reached. All of these assumptions and requirements have been taken into consideration while designing the system described in the following section.

5.2. Architecture of the System

The architecture of our system is presented in Figure 3. The system relies on three sensor modalities typically found on a mobile robot platform: a monocular camera, a single 2D laser scanner, and wheel encoders. The images from the camera, together with the laser scans are used as an input for the multi-modal place classification component. For each pair of data samples, place classification provides its beliefs about the semantic category to which the samples belong. These beliefs are encoded in the integrated outputs as discussed in Section 4. Moreover, the confidence of the decision is also measured and provided by the classification component.

A labeling system should provide a robust and stable output over the whole area. Since the sensors employed are not omni-directional, it is necessary to accumulate and fuse information over time. Moreover, the data that the robot gathers are not evenly spread over different viewpoints. As a possible solution, we propose to use a confidence-based spatio-temporal accumulation method. The principle behind the method is illustrated in Figure 4. As the robot explores the environment, it moves with a varying speed. The robot has information about its own movement (odometry) provided by the wheel encoders. As errors accumulate over time, this information can only be used to estimate relative movement rather than absolute position. This is sufficient for our application. The spatio-temporal accumulation process creates a sparse histogram along the robot pose trajectory given by the odometry and described by the metric position (x, y) and heading (θ) as shown in Figure 4. The size of the histogram bins are adjusted so that each bin roughly corresponds to a single viewpoint. Then, as the robot moves, the beliefs about the current semantic category accumulate within the bins as in the case of G-DAS (with equal weights). This is what we call the temporal accumulation. It prevents a single viewpoint from becoming dominant due to long-term observation. Since each viewpoint observed by the robot will correspond to a different bin, performing accumula-

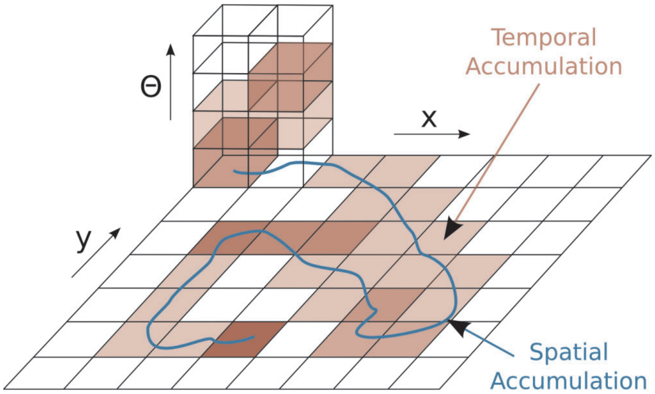


Fig. 4. Illustration of the spatio-temporal accumulation process. As the robot explores the environment, the beliefs collected on the way accumulate over time within the bin corresponding to the current pose (x, y, θ) and over space in different bins.

tion across the bins (this time spatially) allows to generate the final outputs to which each viewpoint contributes equally. In order to exclude most of the misclassifications before they get accumulated, we filter the decisions based on the confidence value provided by the place classification component. Moreover, as the odometry information is unreliable in the long term, the contents of bins visited a certain amount of viewpoints ago, are invalidated. Note that semantic labeling is an application of the method presented in this paper and not the main focus of the paper. The accumulation scheme we present here builds on the ideas of discriminative accumulation and confidence estimation to further illustrate their usefulness. If the emphasis was on labeling, more advanced methods based on Hidden Markov Models (Rottmann et al. 2005), probabilistic relaxation (Stachniss et al. 2007) or Conditional Random Fields (Douillard et al. 2007) should be taken into consideration. The advantages of our method are seamless integration with other components of the system and simplicity (the method does not require training or making assumptions on the transition probabilities between locations or areas).

The accumulation process ensures robustness and stability of the generated label for a single area. However, another mechanism is required to provide the system with information about area boundaries. This is required for the accumulation process not to fuse the beliefs across different areas. Here, we propose two solutions to that problem. As described in the previous sections, we can assume that each room of the environment should be assigned one semantic label. In the case of indoor environments, rooms are usually separated by a door or other narrow openings. Thus, as one solution, we propose to use a simple laser-based door detector which generates hypotheses about doors based on the width of the opening which the robot passes. Such a simple algorithm will surely generate a lot of false positives. However, this does not cause problems in the presented architecture as false positives only lead to oversegmentation. This is a problem mainly for other components relying on precise segmentation rather than for the labeling process itself. In fact, the labeling system could be used to identify false doors and improve the segmentation by looking for directly connected areas classified as being of the same category.

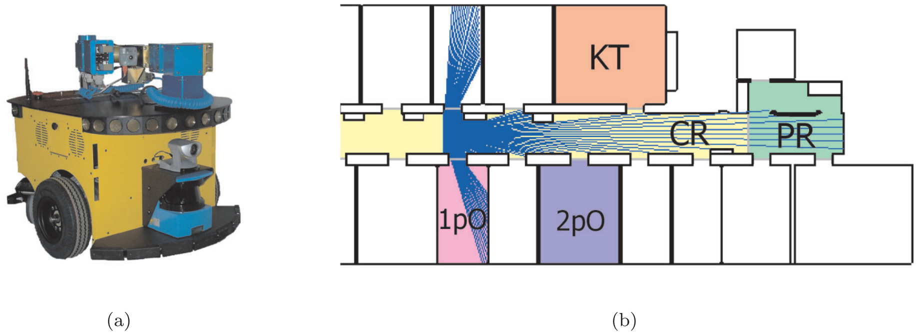
As a second solution, we propose to use another localization and mapping system in order to generate the space representation which will then be augmented with semantic labels. Here we take the multi-layered approach proposed in Zender et al. (2008). The method presented by Zender et al. (2008) builds a global metric map as the first layer and a navigation graph as the second. As the robot navigates through the environment, a marker or navigation node is dropped whenever the robot has traveled a certain distance from the closest existing node. Nodes are connected following the order in which they were generated. If information about the current node is provided to the spatio-temporal accumulation process, labels can be generated for each of the nodes separately. Moreover, as it is possible to detect whether the robot revisited an existing node, the accumulated information can be saved and used as a prior the next time the node is visited. For the live experiment described in this paper, we used the detected doors to bound the areas and navigation graph nodes to keep the priors. We then propagated the current area label to all the nodes in the area.

6. Experiments with Place Classification

We conducted several series of experiments to evaluate the performance of our place classification system. We tested its robustness to different types of variations, such as those introduced by changing illumination or human activity over a long period of time. The evaluation was performed on data acquired using a mobile robot platform over a time span of six months, taken from the IDOL2 database (Image Database for rObot Localization 2, see Luo et al. (2007)). Details about the database and experimental setup are given in Section 6.1. The experiments were performed for single-cue models and models based on different combinations of cues and modalities. We present the results in Sections 6.2 and 6.3 respectively. In addition, we analyze performance and properties of different cue integration schemes in Section 6.4.

6.1. Experimental Setup

The IDOL2 database was introduced in Luo et al. (2007). It comprises of 24 image sequences accompanied by laser scans and odometry data acquired using two mobile robot platforms (PeopleBot and PowerBot). The images were captured with a Canon VC-C4 perspective camera using the resolution of



(a)

(b)

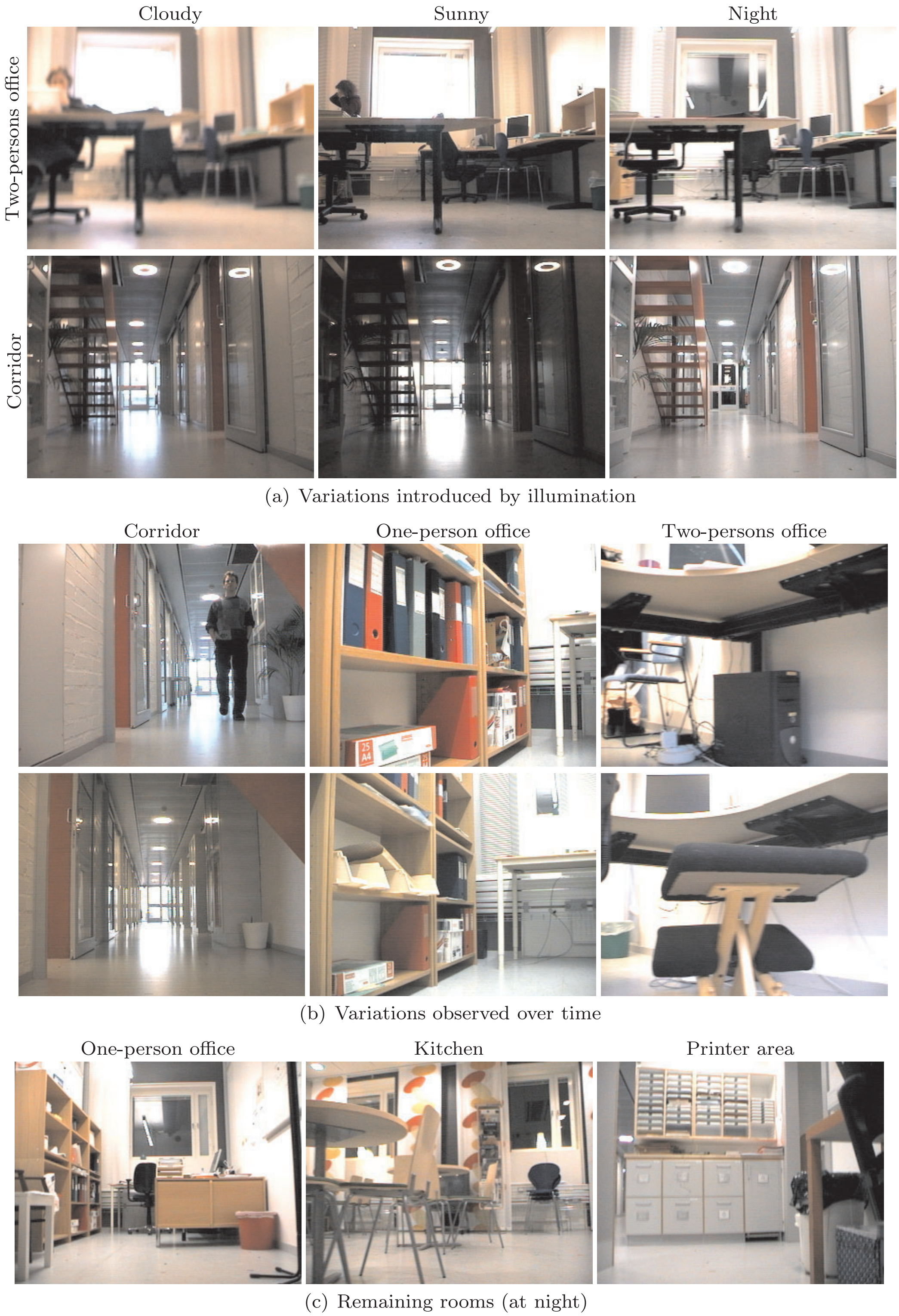
Fig. 5. (a) The mobile robot platform used in the experiments. (b) Map of the environment used during data acquisition and an example laser scan simulated in the corridor. The rooms used during the experiments are annotated.

320 × 240 pixels. In this paper, we will use only the 12 data sequences acquired with the PowerBot, shown in Figure 5(a).

The acquisition was performed in a five room subsection of a larger office environment, selected in such a way that each of the five rooms represented a different functional area: a one-person office (1pO), a two-persons office (2pO), a kitchen (KT), a corridor (CR), and a printer area (PR). The map of the environment and an example laser scan are shown in Figure 5(b). Example pictures showing interiors of the rooms are presented in Figure 6. The appearance of the rooms was captured under three different illumination conditions: in cloudy weather, in sunny weather, and at night. The robot was manually driven through each of the five rooms while continuously acquiring images and laser range scans at a rate of 5 fps. Each data sample was then labelled as belonging to one of the rooms according to the position of the robot during acquisition. Extension 1 contains a video illustrating the acquisition process of a typical data sequence in the database. The acquisition was conducted in two phases. Two sequences were acquired for each type of illumination conditions over the time span of more than two weeks, and another two sequences for each setting were recorded six months later (12 sequences in total). Thus, the sequences captured variability introduced not only by illumination but also natural activities in the environment (presence/absence of people, furniture/objects relocated etc.). Example images illustrating the captured variability are shown in Figure 6.

We conducted four sets of experiments, first for each cue separately and then for cues combined. In order to simplify the experiments with multiple cues, we matched images with closest laser scans on the basis of the acquisition timestamp. In case of each single experiment, both training and testing were performed on one data sequence. The first set consisted of 12 experiments, performed on different combinations of training and test data acquired closely in time and under similar illumination conditions. In this case, the variability comes from human activity and viewpoint differences. For the second set of experiments, we used 24 pairs of sequences captured still at relatively close times, but under different illumination conditions. In this way, we increased the complexity of the problem (Pronobis et al. 2006; Pronobis and Caputo 2007). In the third set of experiments, we tested the robustness of the system to long-term variations in the environment. Therefore, we conducted 12 experiments, where we tested on data acquired six months later, or earlier, than the training data, again under similar illumination conditions. Finally, we combined both types of variations and performed experiments on 24 pairs of training and test sets, obtained six months from each other and under different illumination settings. Note that in the last two sets of experiments described, the task becomes more and more challenging as the difference between training and test set increases. By doing this, we aim at testing the gain in robustness expected from cue integration in very difficult, but still realistic, scenarios.

For all experiments, model parameters were determined via cross validation. Since the datasets in the IDOL2 database are unbalanced (on average 443 samples for CR, 114 for 1pO, 129 for 2pO, 133 for KT and 135 for PR), as a measure of performance for the reported results and parameter selection, we used the average of classification rates obtained separately for each actual class (average per-class recall). For each single experiment, the percentage of properly classified samples was first calculated separately for each room and then averaged with equal weights independently of the number of samples acquired in the room. This allowed to eliminate the influence that large classes could have on the performance score. Statistical significance of the presented results was verified using the Wilcoxon signed-ranks test (when performance of two methods was compared) or Friedman and *post hoc* Nemenyi test (when multiple methods were compared) at a confidence level of $\alpha = 0.05$ as suggested in Demšar (2006). The results of the *post hoc* tests were visualized using critical difference diagrams. The diagrams show average ranks of the compared methods and the groups of methods that are not significantly different are connected (the difference is smaller than the critical difference presented above the main axis of the diagram).



(a) Variations introduced by illumination

(b) Variations observed over time

(c) Remaining rooms (at night)

Fig. 6. Examples of pictures taken from the IDOL2 database showing the interiors of the rooms, variations observed over time and caused by activity in the environment as well as introduced by changing illumination.

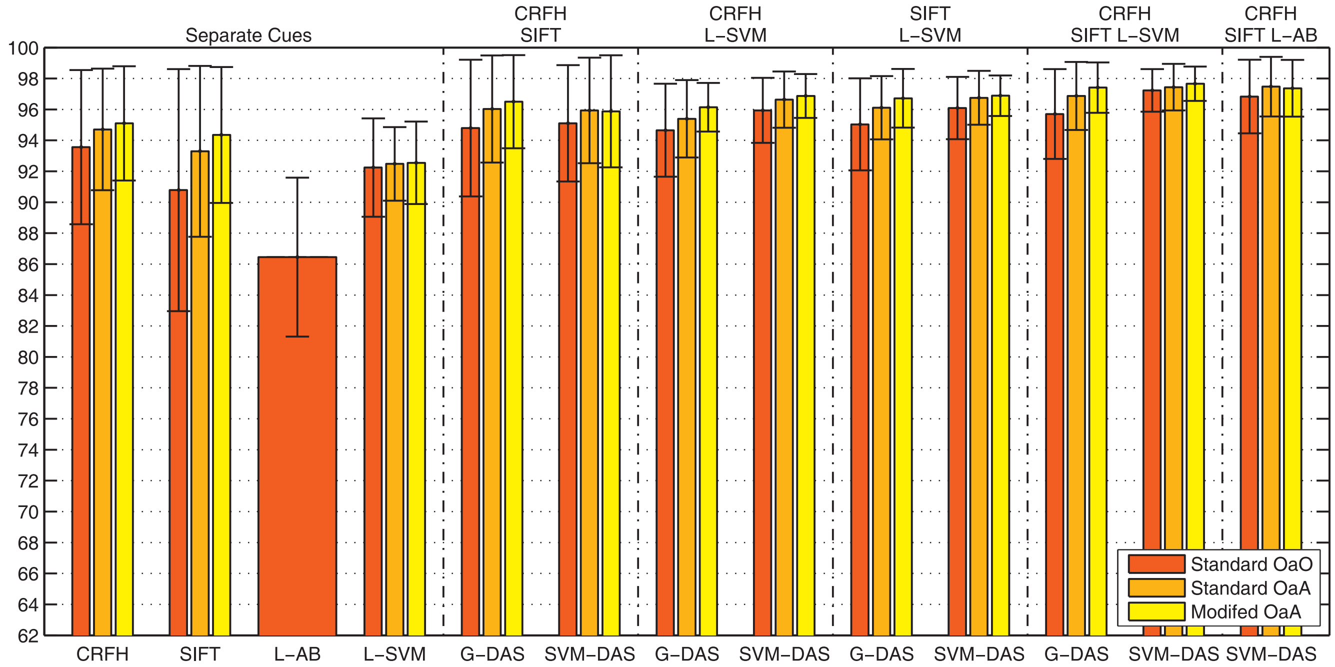


Fig. 7. Classification results for Experiment 1: stable illumination conditions, close in time.

The reader is referred to Demšar (2006) for details on the applied tests and the critical difference diagrams presented below.

6.2. Experiments with Separate Cues

We first evaluated the performance of all four types of single-cue models: the two SVM models based on visual features (CRFH, SIFT), the AdaBoost and the SVM models trained on the laser range cues (referred to as L-AB and L-SVM). For SVM, we tried the three multi-class extensions described in Section 4.1. The results of all four sets of experiments for these models are presented in Figures 7–10 (the first four bar groups). Moreover, the results of statistical significance tests comparing the models based on the combined results of all four experiments are illustrated in Figure 11. We first note that, as expected, CRFH and SIFT suffer from changes in illumination (−15.3% and −11.0% respectively), while the geometrical features do not (−1.9% for L-AB and −0.64% for L-SVM). Long-term variations pose a challenge for both modalities (from −7.0 to −10.2% for vision and −3.7 to −7.9% for laser). We also see differences in performance between the two visual cues: CRFH suffers more from changes in illumination, while SIFT is less robust to variations induced by human activities. It is also interesting to note that under stable conditions, the vision-based methods outperform the systems based on laser range cues (95.1% for CRFH and 92.5% for L-SVM; the difference is statistically significant). This illustrates the potential of visual cues, but also stresses the need for more robust solutions.

These experiments are also a comparison between two recognition algorithms using laser-range features, namely the boosting-based implementation (L-AB) presented in Mozos et al. (2005) and the current SVM-based implementation (L-SVM). Figures 7–10 and Figure 11 show the results. We can see that the difference in performance is statistically significant in favor of the SVM-based method for all three multi-class extensions (from 6.1% for Experiment 1 to 10.3% for Experiment 4 in average). The classification results for the L-AB are worse than the results of the original paper by Mozos et al. (2005). There are two main reasons for that. First, the number of classes is increased to five, while in Mozos et al. (2005) was of a maximum of four. Second, in these experiments, we used a restricted field of view of 180°, whilst in Mozos et al. (2005) the field of view was of 360°. This decreases the classification rate, as has been shown in previous work (Mozos et al. 2007).

As already mentioned, all the experiments with SVMs were repeated for three different multi-class extensions: standard OaO and OaA as well as modified OaA algorithm. The obtained results are in agreement with those of Pronobis and Caputo (2007): in the case of single cue and G-DAS experiments, the modified version gives the best performance with a statistically significant difference independently of the modality on which the classifier was trained.

Figure 12 shows the distribution of errors for each actual class (room) made by the four models. It is apparent that each of the cues makes errors according to a different pattern. At the

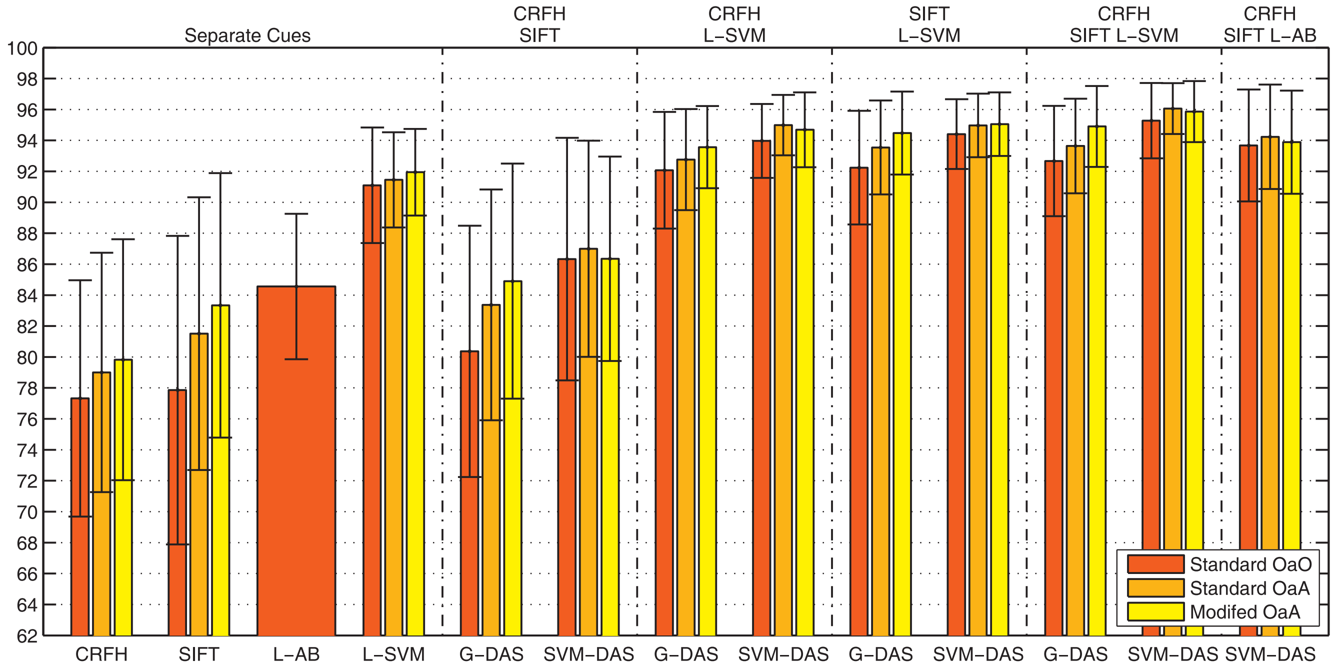


Fig. 8. Classification results for Experiment 2: varying illumination conditions, close in time.

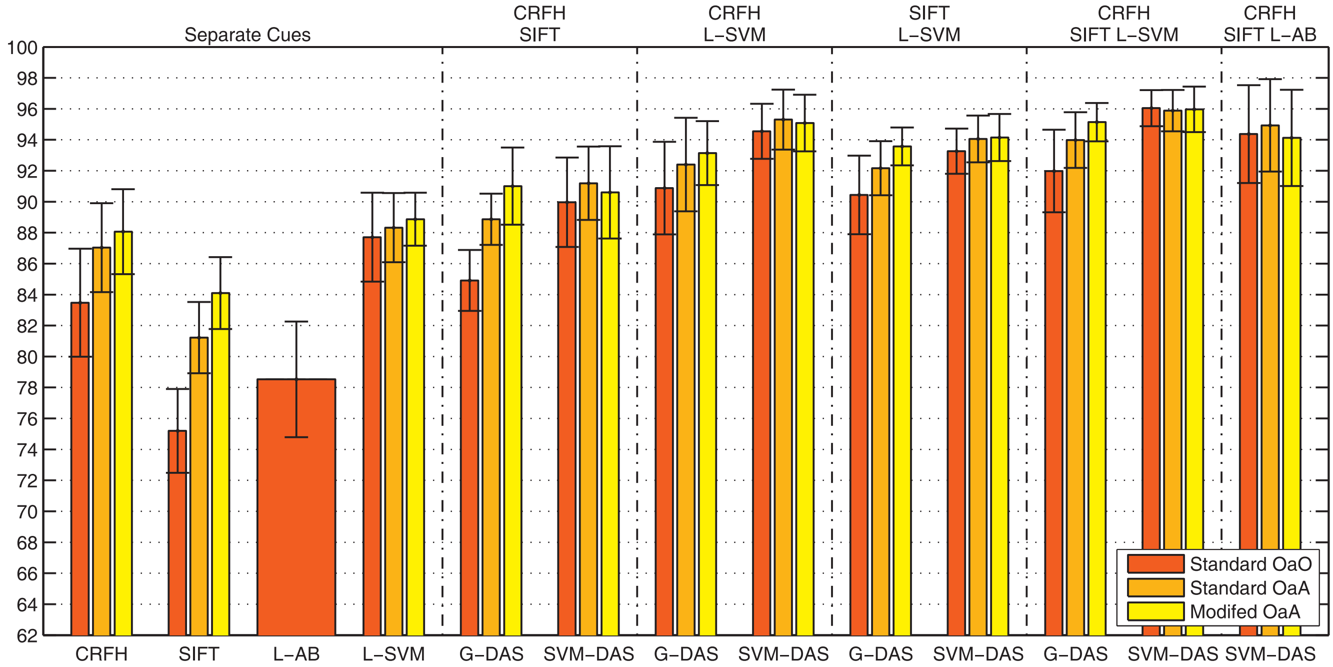


Fig. 9. Classification results for Experiment 3: stable illumination conditions, distant in time.

same time, similarities occur between the same modalities. We see that visual models are biased towards the corridor, while the geometrical models tend to misclassify places as the printer area. A possible explanation is that the vision-based models were trained on images acquired with perspective camera with constrained viewing angle. As a result, similar visual stimuli

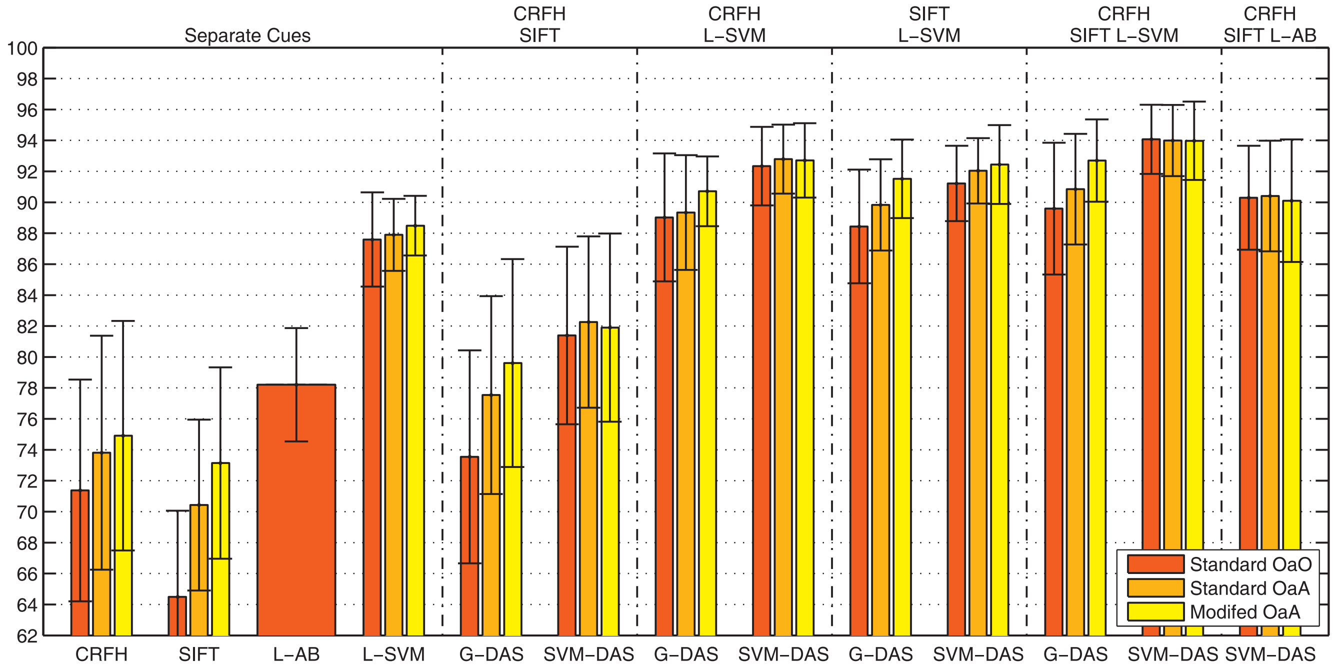


Fig. 10. Classification results for Experiment 4: varying illumination conditions, distant in time.

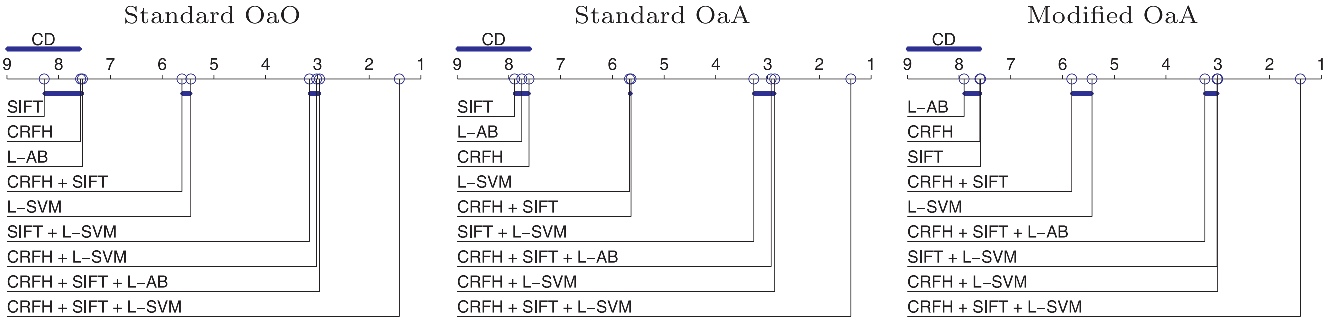
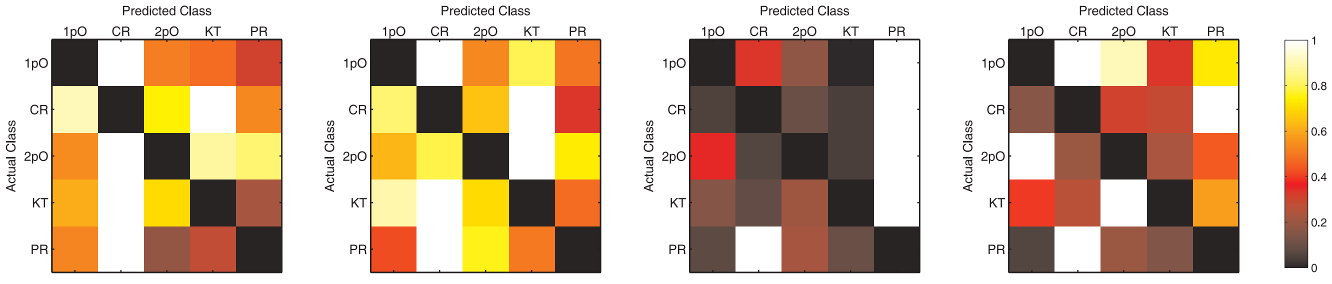


Fig. 11. Critical difference diagrams comparing four single-cue models and solutions based on multiple cues integrated using SVM-DAS with the Nemenyi test for a confidence level of $\alpha = 0.05$. The comparison is based on the combined results of Experiments 1–4 and presented separately for each multi-class extension. The average ranks of the methods are plotted on the axis and the groups of methods that are not significantly different are connected.

coming from the corridor are present in the images captured by the robot leaving each of the rooms. The same area close to a doorway, from the geometrical point of view, is similar to the narrow passage in the printer area. This analysis is a strong motivation to integrate these various cues with a stack of classifiers, as theory indicates that this is the ideal condition for exploiting the different informative content (Polikar 2006).

6.3. Experiments with Cue Integration

For the final experiments, we selected four different cue accumulation methods: G-DAS and SVM-DAS with three kernel types (linear, RBF, and histogram intersection (HI) kernel (Barla et al. 2003)). The parameters of the algorithms (weights in case of G-DAS and SVM model in case of SVM-DAS) were always adjusted on the basis of outputs generated during all experiments with single-cue models trained on one particular data sequence. Then, during testing, the previously obtained integration scheme was applied to all experiments with models trained on a different sequence, acquired under similar illumination and closely in time. This way, the generalization abilities of each of the methods were tested in a realistic scenario. In all experiments, we found that SVM-DAS with an RBF kernel outperforms the other methods and the



(a) CRFH (b) SIFT (c) L-AB (d) L-SVM

Fig. 12. Distribution of errors made by the four models for each actual class (bright colors indicate errors). The diagonal elements were removed.

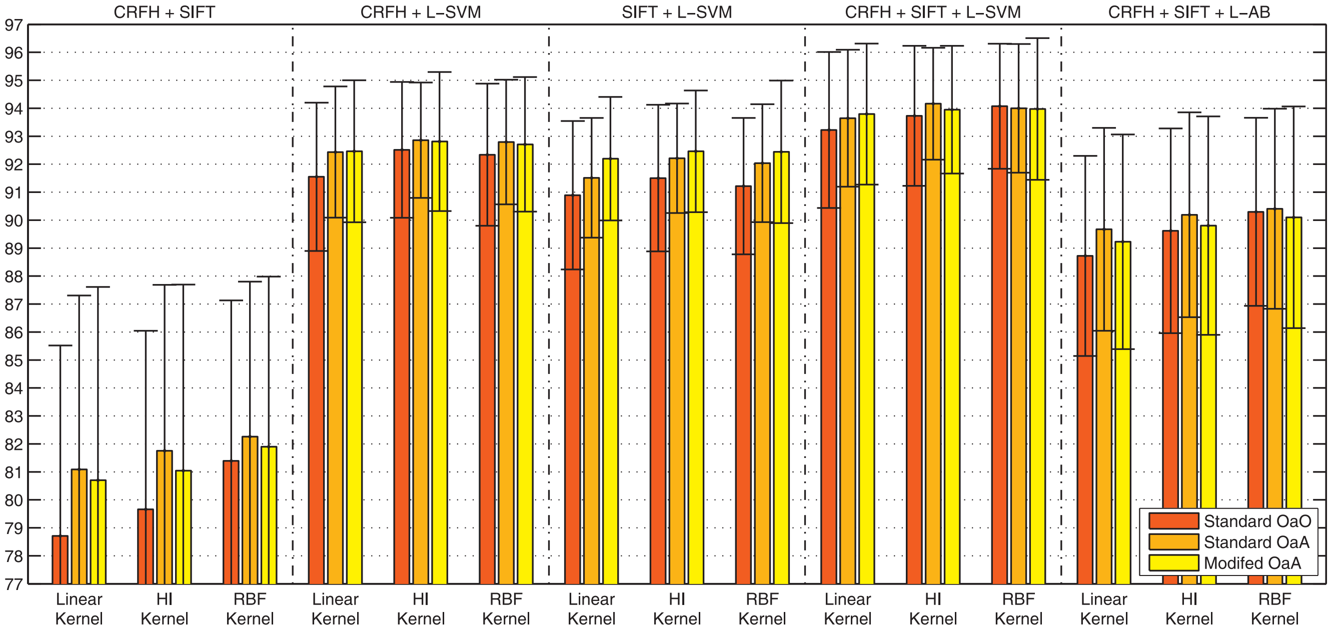


Fig. 13. Comparison of performance of SVM-DAS based on different kernel functions for the most complex problem (Experiment 4).

difference in performance with respect to G-DAS was statistically significant for all combinations of cues and multi-class extensions (Wilcoxon test). For space reasons, we report results of each of the experiments only using SVM-DAS based on the RBF kernel and G-DAS for comparison (Figures 7–10, last nine bar groups). A detailed comparison of all variants of SVM-DAS for the most complex problem (Experiment 4) is given in Figure 13. Results of statistical significance tests comparing the multi-cue solutions with single-cue models based on the combined results of all experiments are illustrated in Figure 11.

We tested the methods with several combinations of different cues and modalities. First, we combined the two visual cues. We see that the generalization of a purely visual recognition system can be significantly improved by integrating different types of cues, in this case local and global. This can be observed especially for Experiment 4, where the algorithms had to tackle the largest variability. Despite that, according to the error distributions in Figure 12, we should expect the largest gain when different modalities are combined. As we can see from Figures 7–10 this is indeed the case. By combining one visual cue and one laser range cue (e.g. CRFH + L-SVM), we exploit the descriptive power of vision in the case of stable illumination conditions and the invariance of geometrical features to the visual noise. Moreover, if the computational cost is not an issue, the performance can be further improved by using both visual cues instead of just one. As can be seen from Figure 11, by integrating single-cue models or adding another

Table 1. Confusion Matrix for the Multi-cue System Based on CRFH, SIFT and L-SVM Integrated Using SVM-DAS

	Predicted class				
Actual class	1pO	CR	2pO	KT	PR
1pO	**11.20 (93.71)**	0.36 (3.06)	0.16 (1.33)	0.11 (0.96)	0.11 (0.94)
CR	0.25 (0.53)	**45.36 (97.73)**	0.19 (0.42)	0.33 (0.70)	0.29 (0.62)
2pO	0.17 (1.22)	0.11 (0.8)	**13.26 (96.92)**	0.06 (0.46)	0.08 (0.60)
KT	0.17 (1.18)	0.35 (2.45)	0.08 (0.57)	**13.42 (95.12)**	0.09 (0.67)
PR	0.09 (0.65)	0.77 (5.59)	0.03 (0.19)	0.05 (0.33)	**12.90 (93.24)**

Normalized average values in percentage over all experiments are reported. The values in brackets were normalized separately for each actual class (row). The presented results are only for the standard OaO multi-class extension since the results for the remaining extensions were comparable.

cue to a multi-cue system, we always get an improvement statistically significant.

We performed a more detailed analysis of the best results. Table 1 contains the confusion matrix for the multi-cue system based on CRFH, SIFT and L-SVM integrated using SVM-DAS with an RBF kernel. We see that even if the corridor class contained on average four times more samples than each of the room classes and was visually and geometrically distinctive, the results are balanced and the recognition rates for each actual class are similar. In general, during our experiments, more balanced solutions were preferred due to the performance metric used (average of the diagonal values in brackets in Table 1).

As it was mentioned in Section 4.3, SVM-DAS can be applied for problems where outputs of different classifiers need to be integrated. To test this in practice, we combined the SVM models trained on visual cues with AdaBoost model based on geometrical features (L-AB)[5]. We present the results in Figures 7–10 (last bar group) and Figure 11. The method obtained large and statistically significant improvements compared to each of the individual cues. For instance for Experiment 4, the recognition rate increased by 12.2% in average. This proves the versatility of our approach.

6.4. Analysis of Cue Integration Schemes

Results presented so far clearly show that SVM-DAS performs significantly better than G-DAS and, by using more sophisticated kernel types for SVM-DAS, it is possible to perform non-linear cue accumulation. Moreover, the experiments (see Figure 13) show that we can expect better results with the RBF kernel (especially for the OaO multi-class extension), although there is no drastic improvement. We therefore suggest to choose the kernel according to constraints on the computational cost of the solution. Since there are fast implementations of linear SVMs, it might be beneficial to use a linear kernel in cases when the integration scheme must be trained on a very large number of samples. In applications where only the number of training parameters is an issue, the non-parametric HI kernel can be used instead of RBF.

We now further discuss differences between high-level (e.g. SVM-DAS) and low-level (feature-level) cue integration. There are several advantages in integrating multiple cues with a high-level strategy. First, different learning algorithms can be used for each single cue. In our experiments, this allowed to combine SVM-based models employing different kernel functions (e.g. the χ^2 kernel for CRFH and the match kernel for SIFT) or even different classifiers (AdaBoost and SVM). Moreover, parameters can be tuned separately for each of the cues. Second, both the training and recognition tasks can be divided into smaller subproblems that can be easily parallelized. Finally, it is possible to decide on the number of cues that should be extracted and used for each particular classification task. This is an important feature, since, in most cases, decisions based on a subset of cues are correct while extraction and classification of additional features introduces additional cost. For example, a solution based on global visual features, laser range cues and SVM-DAS runs in real-time at a rate of approximately 5 fps, which would not be possible if an additional visual cue like SIFT was used. The computational cost can be significantly reduced by taking the approach presented in Pronobis and Caputo (2007). By combining confidence estimation methods with cue integration, we can use additional sources of information only when necessary – when the decision based on one cue only is not confident enough. This scheme is referred to as Confidence-based Cue Integration. Table 2 presents the results of applying the scheme to the experiments presented in this section. We see that, in general, we can base our decision on the fastest model (marked with bold font in Table 2), such as the efficient and low-dimensional model based on simple laser-range features, and we can retain the maximal performance by using additional cues only in approximately 30% of cases. This greatly reduces the computational time required on average e.g. approximately three times for

[5] As usual, for SVM we used several multi-class extensions that in most cases produced outputs having different interpretation than those generated by the multi-class algorithm used for AdaBoost. In those cases G-DAS could not be applied.

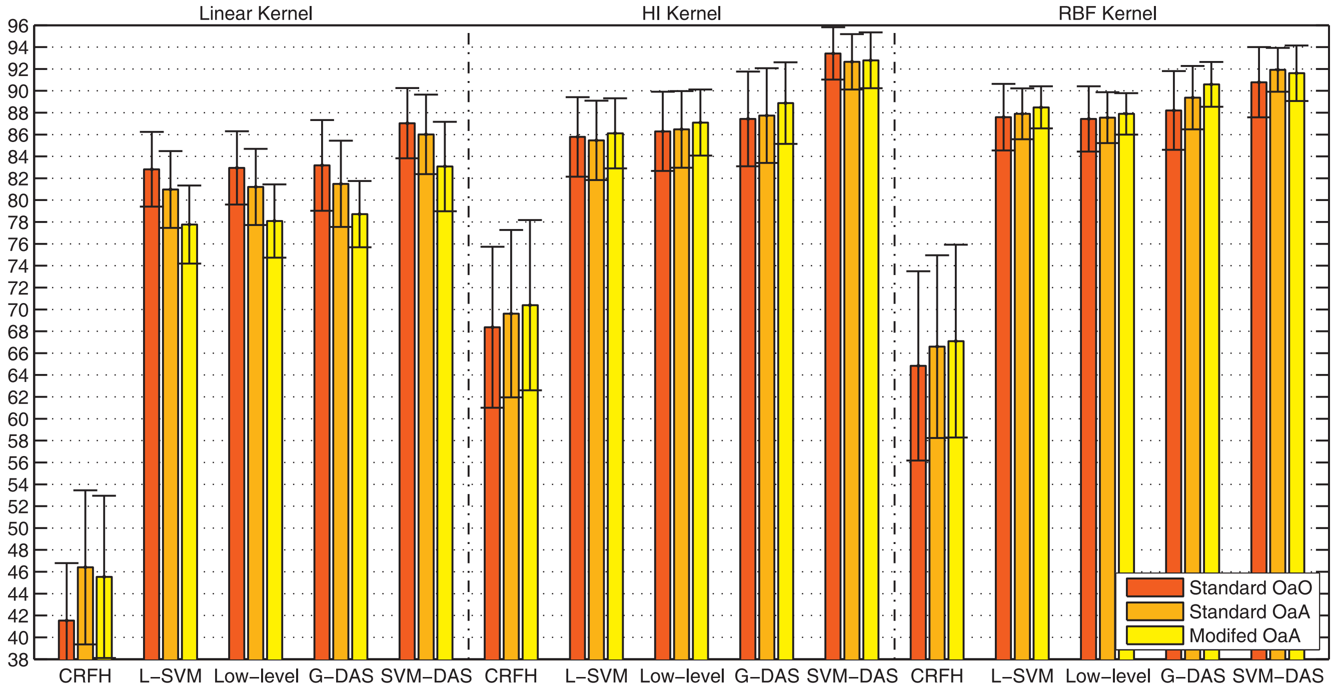


Fig. 14. Comparison of performance of two single-cue models and solutions based on the cues integrated on both low and high level for the most complex problem (Experiment 4).

Table 2. Average Percentages (with Standard Deviations) of Test Samples for which all Cues had to be Used in Order to Obtain the Maximal Recognition Rate

Cues (**Primary cue**)	Cue integration method G-DAS	 SVM-DAS RBF Kernel
CRFH + SIFT	25.971 ± 18.503	29.453 ± 22.139
CRFH + **L-SVM**	21.230 ± 20.199	32.736 ± 20.256
SIFT + **L-SVM**	28.820 ± 20.982	33.344 ± 22.425
SIFT + CRFH + **L-SVM**	31.858 ± 20.474	40.833 ± 21.916

CRFH, L-SVM and SVM-DAS. Additional cues will be used more often when the variability is large, and rarely for less difficult cases. This is not possible in the case of low-level integration where all the cues must be extracted and classified in order to obtain a decision.

Another important factor is performance. During our experiments, we compared the performance of G-DAS and SVM-DAS (with an RBF kernel) with models built on cues combined on the feature level. We performed three different sets of comparisons. In the first comparison, we built single-cue models and models based on features combined on the low level using SVM and the non-parametric linear kernel, using the same values of the SVM training parameters for all models. Then, we integrated the outputs of the single-cue models using G-DAS and SVM-DAS. In the case when G-DAS was used, the solution remained linear. In the second comparison, for building the models we used the non-linear, non-parametric HI kernel. In the final comparison, we used an RBF kernel and performed parameter selection for each of the models. All comparisons were based on CRFH and laser-range cues, since the dedicated kernel function required by SIFT could not be used with any of the other features for low-level integration.

The results for the most complex problem (Experiment 4) are given in Figure 14 and statistical significance tests comparing the solutions are illustrated in Figure 15. It can be observed that, in every case, the high-level integration significantly outperformed solutions based on features combined on the low level. In only one case there was no significant difference between G-DAS and low-level integration; however, SVM-DAS still performed better than the other solutions. This is in agreement with the results reported by Tommasi et al. (2008) and Nilsback and Caputo (2004) and can be explained by greater robustness of the high-level methods to noisy cues or sensory channels and the ability of different classifiers to adapt to the characteristics of each single cue.

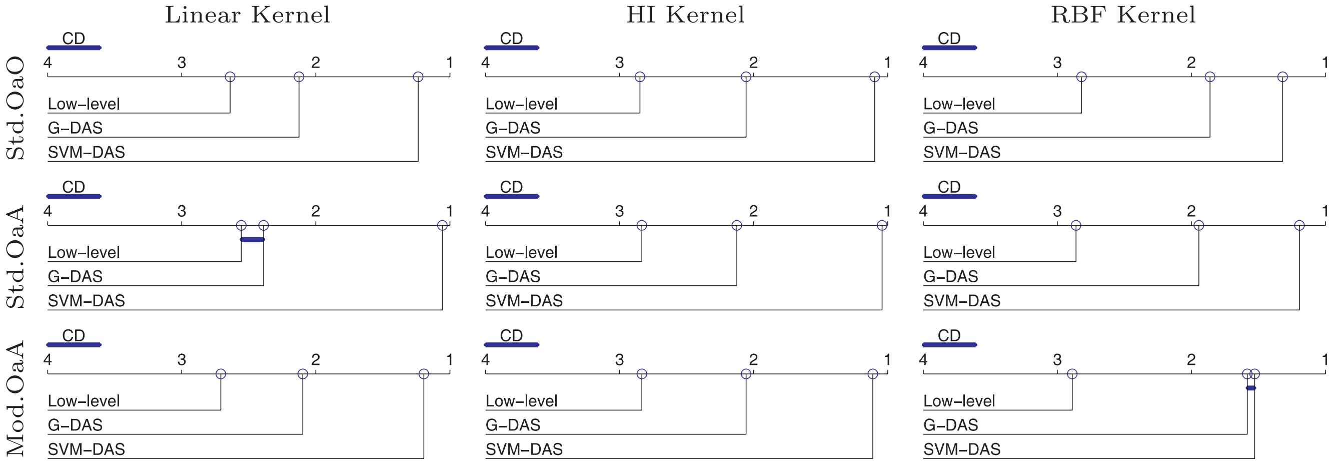


Fig. 15. Critical difference diagrams comparing two single-cue models and solutions based on the cues integrated at both the low and high level with the Nemenyi test for a confidence level of $\alpha = 0.05$. The comparison is based on the combined results of Experiments 1–4 and presented separately for three kernel functions and multi-class extensions used with SVM. The average ranks of the methods are plotted on the axis and the groups of methods that are not significantly different are connected.

7. Experiments with Semantic Space Labeling

We performed an independent live experiment to test our multi-modal semantic space labeling system running in real-time on a mobile robot platform. The experiment was performed during working hours in a typical office environment. Both the environment and the robot platform were different than in the case of the off-line evaluation described in Section 6. The whole experiment was videotaped and a video presenting the setup, experimental procedure, and visualization of the results can be found in Extension 2.

7.1. Experimental Setup

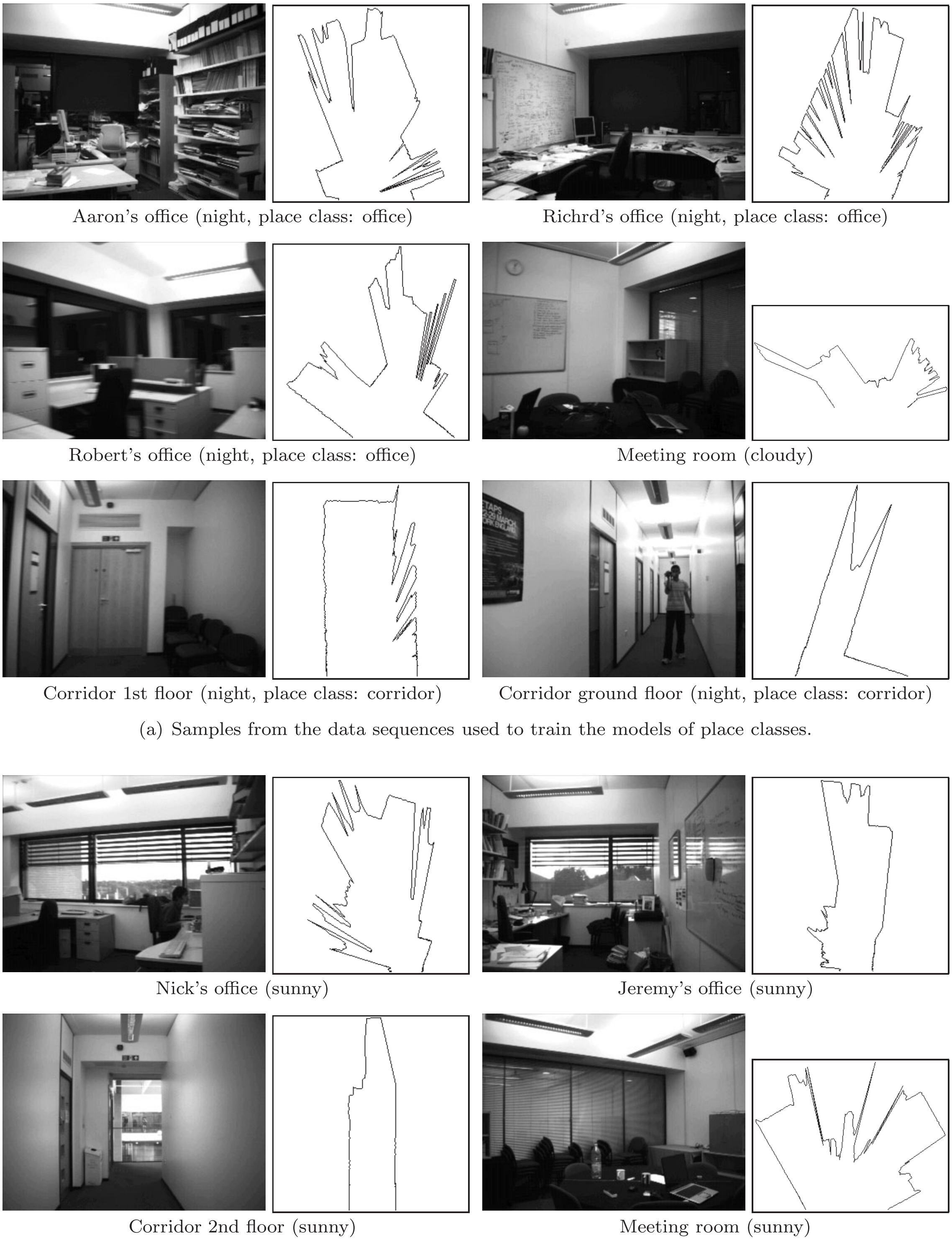
The experiment was performed between the 7th and 10th of September 2008 in the building of the School of Computer Science at the University of Birmingham, Birmingham, UK. The interior of the building consists of several office environments located on three floors. For our experiments, we selected three semantic categories of rooms that could be found in the building: a corridor, an office and a meeting room. To build the model of an office, we acquired data in three different offices: Aaron's office (first floor), Robert's office (first floor) and Richard's office (ground floor). To create a representation of the corridor class, we recorded data in two corridors, one on the ground floor and one on the first floor. The acquisition was performed at night. Finally, to train the model of a meeting room, we used an instance on the second floor. All training data except the one from the meeting room was acquired in another part of the building than the one used for testing. The data for this class were recorded during the day. A video illustrating the whole data acquisition process is available as Extension 3. The interiors of the rooms are presented in Figure 16(a), as seen by vision and laser. The robot was manually driven around each room and data samples were recorded at the rate of 5 fps. All the collected training data are available as Extension 4. In the case of the meeting room, the corridor on the first floor as well as Aaron's and Richard's offices, the acquisition was repeated twice.

For the real-time experiment, we built the system as described in Section 5. Following the findings of the off-line experiments, we used SVM-DAS with the RBF kernel to integrate the classifier outputs for vision and laser range data. For efficiency reasons, we used only global features (CRFH) for the vision channel. We used the OaA multi-class SVM extension for the place models. Other parameters were set as described in Section 6.

We trained the place models separately for each modality on a dataset created from one data sequence recorded in each of the rooms. One of the advantages of SVM-DAS is the ability to infer the integration function from the training data, after training the models. We used the additional data sequences acquired in some of the rooms and trained SVM-DAS on the outputs of the uni-modal models tested on these data.

The PeopleBot robot platform shown in Figure 3 was used for data acquisition and the final experiment. The robot was equipped with a SICK laser range finder and Videre STH-MDCS2 stereo head (only one of the cameras was used). The images were acquired at the resolution of 320 × 240 pixels. The whole system was implemented in the CAST (The CoSy Architecture Schema Toolkit)[6] framework and run on a standard

[6] See http://www.cs.bham.ac.uk/research/projects/cosy/cast/



Aaron's office (night, place class: office)

Richrd's office (night, place class: office)

Robert's office (night, place class: office)

Meeting room (cloudy)

Corridor 1st floor (night, place class: corridor)

Corridor ground floor (night, place class: corridor)

(a) Samples from the data sequences used to train the models of place classes.

Nick's office (sunny)

Jeremy's office (sunny)

Corridor 2nd floor (sunny)

Meeting room (sunny)

(b) Samples acquired during the test run.

Fig. 16. Examples of images and laser scans (synchronized) taken from the data sequences used for training the models of place classes (a) and acquired during the test run (b) in each of the rooms considered during the experiment. The within-category variations for corridors and offices are illustrated as well as other types of variability observed for each place class (e.g. different illumination conditions, activity in the environment).

2.5 GHz dual-core laptop. The processing for both modalities was executed in parallel using both of the CPU cores.

7.2. Experimental Procedure and Results

Three days after the training data were collected, we performed a live experiment in the lab on the second floor in the same building. The experiment was conducted during the day with sunny weather. The part of the environment that was explored by the robot consisted of two offices (Nick's office and Jeremy's office), a corridor and a meeting room. The interiors of the rooms and the influence of illumination can be seen in the images in Figure 16(b).

The SLAM system of the robot constructs a metric map and navigation graph. In this experiment, the task is to semantically label the navigation graph nodes and areas as the map is being built. The only knowledge given to the robot before the experiment consisted of the models of the three place classes: "office", "corridor" and "meeting room". As stated in Section 5, every time the robot created or revisited a node, the accumulated beliefs about the semantic category of the area were used to label the node and saved as a future prior. The label was also propagated to the whole area. We used detected doors to assign nodes to areas.

The whole experiment was videotaped and a video presenting the experimental setup, the test run and visualization of the obtained results can be found in Extension 2. The robot started in Nick's office, and was manually driven through the corridor to Jeremy's office. Then, it was taken to the meeting room where the autonomous exploration mode was turned on. The robot used a frontier-based algorithm based on Yamauchi (1997). Laser data was limited to 2 m distance in the exploration to make sure that the robot not just perceived how the environment looked but also covered it to build the navigation graph. After the meeting room was explored, the robot was manually driven back to Nick's office where the experiment finished. A video presenting visualization of the full test run is available in Extension 5. The labeling process was running online and the place classification was performed approximately at the rate of 5 times per second. The final semantic map build during the run is shown in Figure 17. We can see that the system correctly labeled all the areas in the environment.

The sensory data acquired during the test run are available as Extension 4. Moreover, a video presenting the sequence of images and laser scans is presented in Extension 6. The fact that the data were stored allowed for additional performance analysis of the multi-modal place classification system, similar to the one presented in Section 6. The results are displayed in Figure 18. When we look at the overall classification rate for all the data samples in the test sequence, we see that vision significantly outperformed laser in this experiment (66% versus 84%). Still, the performance of the system was boosted by an additional 8% compared with vision alone when the two modalities were integrated. The gain is even more apparent if we look at the detailed results for each of the classes (the first three charts in Figure 18). We see that the modalities achieved different performance, but also different error patterns, for each class. Clearly, the system based on laser range data is a very good corridor detector. On the other hand, vision was able to distinguish between the offices and the meeting room almost perfectly. Finally, the integrated system always achieved the performance of the more reliable modality and for two out of three classes outperformed the uni-modal systems. As can be seen in the video in Extensions 2 and 5, this provided stable performance for each of the classes and a robust base for the semantic labeling system.

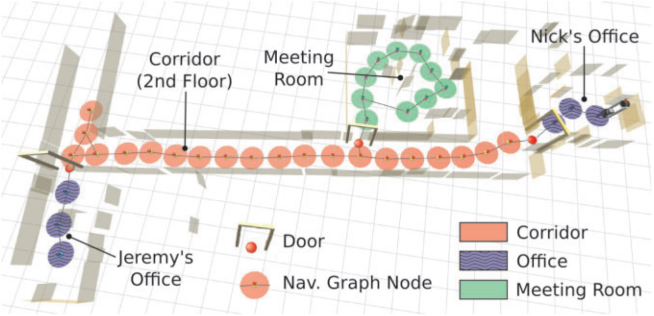


Fig. 17. Final map obtained after the test run. The navigation graph is overlaid on the metric map and the color of the circles around the graph nodes indicate the place class assigned to each area bounded by detected doors. The system correctly labeled all of the areas in the environment.

8. Conclusions

In this paper we have addressed the problem of place classification and showed how it can be applied to semantic knowledge extraction in robotic systems. This is an important and challenging task, where multiple sensor modalities are necessary in order to achieve generality and robustness, and enable systems to work in realistic settings. To this end, we presented a new cue integration method able to combine multiple cues derived by a single modality, as well as cues obtained by multiple sensors. The method was thoroughly tested in off-line experiments on realistic data collected under varying conditions and as part of a real-time system running on a robotic platform. The results obtained using multiple visual cues alone, and combined with laser range features, clearly show the value of our approach. Finally, we showed that the system can successfully be applied for the space labeling problem where it can be used to augment the internal space representation with semantic place information. All of the data used in the paper are available as extensions to the paper and from the IDOL2 database (Luo et al. 2006).

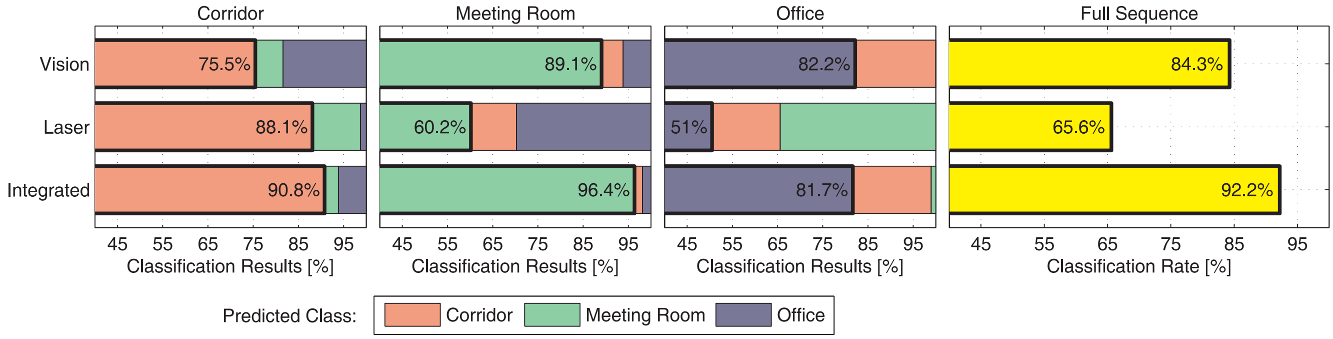


Fig. 18. Place classification results obtained on the dataset recorded during the test run. The first three bar charts show the results separately for each place class: "corridor", "meeting room" and "office". The charts show the percentage of the samples that were properly classified (most left bars marked with thick lines), but also how the misclassifications were distributed. The chart on the right presents the percentage of properly classified samples during the whole run. The two top rows give results for single modalities, while the bottom row shows results for the multi-modal system.

In the future, we plan to extend this method and attack the scalability issue, with particular attention to indoor office environments. These are usually characterized by a large number of rooms with very similar characteristics; we expect that in such a domain our approach will be particularly effective. Another important aspect of place classification is the intrinsic dynamics in the sensory information: as rooms are used daily, furniture is moved around, objects are taken in and out of drawers and people appear. All of this affects the sensor inputs of places in time. We plan to combine our approach with incremental extensions of the SVM algorithm (Luo et al. 2007; Orabona et al. 2007) and to extend these methods from fully supervised to semi-supervised learning, so to obtain a system able to learn continuously from multiple sensors.

Acknowledgments

Special thanks go to Sagar Behere for his great help with running the integrated system on the robotic platform, data acquisition and videotaping. This work was sponsored by the EU integrated projects FP6-004250-IP CoSy (A. Pronobis, O. Martínez Mozos, P. Jensfelt), ICT-215181-CogX (A. Pronobis and P. Jensfelt) and IST-027787 DIRAC (B. Caputo), and the Swedish Research Council contract 2005-3600-Complex (A. Pronobis). The support is gratefully acknowledged.

A preliminary version of part of the experimental evaluation reported in this work was presented in Pronobis et al. (2008).

Appendix: Index to Multimedia Extensions

The multimedia extension page is found at http://www.ijrr.org

Table of Multimedia Extensions

Extension	Type	Description
1	Video	The acquisition procedure of a typical data sequence in the IDOL2 database.
2	Video	The setup, procedure and visualization of the experiment with semantic space labeling based on multi-modal place classification.
3	Video	The process of acquiring data for training the models of places for the experiment with semantic space labeling.
4	Data	The dataset (sequences of images and laser scans) collected during the experiment with semantic labeling of space.
5	Video	Visualization of the complete test run and results obtained during the experiment with semantic space labeling.
6	Video	The complete sequence of images and laser scans acquired during the test run of the experiment with semantic space labeling.

References

Aloimonos, J. and Shulman, D. (1989). *Integration of Visual Modules: an Extension of the Marr Paradigm*. New York, Academic Press.

Althaus, P. and Christensen, H. I. (2003). Behaviour coordination in structured environments. *Advanced Robotics*, **17**(7): 657–674.

Andreasson, H., Treptow, A. and Duckett, T. (2005). Localization for mobile robots using panoramic vision, local features and particle filter. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Barcelona, Spain.

Barla, A., Odone, F. and Verri, A. (2003). Histogram intersection kernel for image classification. *Proceedings of the International Conference on Image Processing (ICIP)*, Barcelona, Spain.

Bay, H., Tuytelaars, T. and Van Gool, L. J. (2006). Surf: Speeded up robust features. *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, Graz, Austria.

Blaer, P. and Allen, P. (2002). Topological mobile robot localization using fast vision techniques. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Washington, DC.

Bradley, D. M., Patel, R., Vandapel, N. and Thayer, S. M. (2005). Real-time image-based topological localization in large outdoor environments. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, August, Edmonton, AB, Canada.

Buschka, P. and Saffiotti, A. (2002). A virtual sensor for room detection. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Lausanne, Switzerland.

Caputo, B. and Dorko, G. (2002). How to combine color and shape information for 3D object recognition: Kernels do the trick. *Neural Information Processing Systems*, Vancouver, BC, Canada.

Chapelle, O., Haffner, P. and Vapnik, V. (1999). Support vector machines for histogram-based image classification. *Transactions on Neural Networks*, **10**(5): 1055–1064.

Clark, J. and Yuille, A. (1990). *Data Fusion for Sensory Information Processing Systems*. Dordrecht, Kluwer.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge, Cambridge University Press.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, **7**: 1–30.

Douillard, B., Fox, D. and Ramos, F. (2007). A spatio-temporal probabilistic model for multi-sensor object recognition. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA.

Duda, R., Hart, P. and Stork, D. (2001). *Pattern Classification*, 2nd edn. New York, Wiley.

Filliat, D. (2007). A visual bag of words method for interactive qualitative localization and mapping. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Rome, Italy.

Fraundorfer, F., Engels, C. and Nistér, D. (2007). Topological mapping, localization and navigation using image collections. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, October, San Diego, CA.

Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. *Proceedings of the European Conference on Computational Learning Theory*, Barcelona, Spain.

Galindo, C., Saffiotti, A., Coradeschi, S., Buschka, P., Fernández-Madrigal, J. and González, J. (2005). Multi-hierarchical semantic maps for mobile robotics. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Edmonton, AB, Canada.

Gaspar, J., Winters, N. and Santos-Victor, J. (2000). Vision-based navigation and environmental representations with an omni-directional camera. *Transactions on Robotics and Automation*, **16**(6): 890–898.

Koenig, S. and Simmons, R. G. (1998). Xavier: A robot navigation architecture based on partially observable Markov decision process models. *Artificial Intelligence Based Mobile Robotics: Case Studies of Successful Robot Systems*, Kortenkamp, D., Bonasso, R. and Murphy, R. (eds). Cambridge, MA, MIT Press, pp. 91–122.

Kuipers, B. (2006). An intellectual history of the spatial semantic hierarchy. *Robot and Cognitive Approaches to Spatial Mapping*. Berlin, Springer.

Kuipers, B. and Beeson, P. (2002). Bootstrap learning for place recognition. *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI)*, Edmonton, AB, Canada.

Linde, O. and Lindeberg, T. (2004). Object recognition using composed receptive field histograms of higher dimensionality. *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Cambridge, UK.

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

Luo, J., Pronobis, A., Caputo, B. and Jensfelt, P. (2006). *The IDOL2 Database*. Technical Report, KTH, CAS/CVAP. Available at http://www.cas.kth.se/IDOL/.

Luo, J., Pronobis, A., Caputo, B. and Jensfelt, P. (2007). Incremental learning for place recognition in dynamic environments. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA.

Matas, J., Marik, R. and Kittler, J. (1995). On representation and matching of multi-coloured objects. *Proceedings of the International Conference on Computer Vision (ICCV)*, Boston, MA.

Menegatti, E., Zoccarato, M., Pagello, E. and Ishiguro, H. (2004). Image-based Monte-Carlo localisation with omnidirectional images. *Robotics and Autonomous Systems*, **48**(1): 17–30.

Mozos, O. M., Stachniss, C. and Burgard, W. (2005). Supervised learning of places from range data using AdaBoost. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Barcelona, Spain, pp. 1742–1747.

Mozos, O. M., Triebel, R., Jensfelt, P., Rottmann, A. and Burgard, W. (2007). Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems*, **55**(5): 391–402.

Murillo, A. C., Guerrero, J. J. and Sagues, C. (2007). Surf features for efficient robot localization with omnidirectional images. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Rome, Italy.

Nilsback, M. E. and Caputo, B. (2004). Cue integration through discriminative accumulation. *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC.

Orabona, F., Castellini, C., Caputo, B., Luo, J. and Sandini, G. (2007). Indoor place recognition using online independent support vector machines. *Proceedings of the British Machine Vision Conference (BMVC)*, Warwick, UK.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods: Support Vector Learning*, Schölkopf, B., Burges, C. and Smola, A. (eds). Cambridge, MA, MIT Press, pp. 185–208.

Poggio, T., Torre, V. and Koch, C. (1985). Computational vision and regularization theory. *Nature* **317**: 314–319.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, **6**: 21–45.

Posner, I., Schroeter, D. and Newman, P. M. (2007). Describing composite urban workspaces. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Rome, Italy.

Pronobis, A. and Caputo, B. (2007). Confidence-based cue integration for visual place recognition. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA.

Pronobis, A., Caputo, B., Jensfelt, P. and Christensen, H. I. (2006). A discriminative approach to robust visual place recognition. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, October, Beijing, China.

Pronobis, A., Mozos, O. M. and Caputo, B. (2008). SVM-based discriminative accumulation scheme for place recognition. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, May, Pasadena, CA.

Rothganger, F., Lazebnik, S., Schmid, C. and Ponce, J. (2006). 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, **66**(3): 231–259.

Rottmann, A., Mozos, O. M., Stachniss, C. and Burgard, W. (2005). Semantic place classification of indoor environments with mobile robots using boosting. *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*, Pittsburgh, PA.

Se, S., Lowe, D. G. and Little, J. (2001). Vision-based mobile robot localization and mapping using scale-invariant features. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Seoul, Korea.

Siagian, C. and Itti, L. (2007). Biologically-inspired robotics vision monte-carlo localization in the outdoor environment. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, October, San Diego, CA.

Stachniss, C., Mozos, O. M. and Burgard, W. (2006). Speeding-up multi-robot exploration by considering semantic place information. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Orlando, FL.

Stachniss, C., Grisetti, G., Mozos, O. and Burgard, W. (2007). Efficiently learning metric and topological maps with autonomous service robots. *Information Technology*, **49**: 232–237.

Tamimi, H. and Zell, A. (2004). Vision based localization of mobile robots using kernel approaches. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Sendai, Japan.

Tapus, A. and Siegwart, R. (2005). Incremental robot mapping with fingerprints of places. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Edmonton, AB, Canada.

Tommasi, T., Orabona, F. and Caputo, B. (2008). Discriminative cue integration for medical image annotation. *Pattern Recognition Letters, Special Issue on IMageCLEF Med Benchmark Evaluation*, **29**(15): 1996–2002.

Topp, E. A. and Christensen, H. I. (2006). Topological modelling for human augmented mapping. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, October, Beijing, China.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, **53**(2): 169–191.

Torralba, A. and Sinha, P. (2001). *Recognizing Indoor Scenes*. Technical Report 2001-015, AI Memo.

Torralba, A., Murphy, K. P., Freeman, W. T. and Rubin, M. A. (2003). Context-based vision system for place and object recognition. *Proceedings of the International Conference on Computer Vision (ICCV)*, Nice, France.

Triesch, J. and Eckes, C. (1998). Object recognition with multiple feature types. *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, Skövde, Sweden.

Ulrich, I. and Nourbakhsh, I. (2000). Appearance-based place recognition for topological localization. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, San Francisco, CA.

Valgren, C. and Lilienthal, A. J. (2008). Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Pasadena, CA.

Wallraven, C., Caputo, B. and Graf, A. (2003). Recognition with local features: the kernel recipe. *Proceedings of the In-*

ternational Conference on Computer Vision (ICCV), Nice, France.

Weiss, C., Tamimi, H., Masselli, A. and Zell, A. (2007). A hybrid approach for vision-based outdoor robot localization using global and local image features. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, October, San Diego, CA.

Yamauchi, B. (1997). A frontier-based approach for autonomous exploration. *Proceedings of the International Symposium on Computational Intelligence in Robotics and Automation*, Monterey, CA.

Zender, H., Jensfelt, P., Mozos, O. M., Kruijff, G.-J. M. and Burgard, W. (2007). An integrated robotic system for spatial understanding and situated interaction in indoor environments. *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI)*, Vancouver, BC, Canada.

Zender, H., Mozos, O. M., Jensfelt, P., Kruijff, G.-J. M. and Burgard, W. (2008). Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, **56**(6): 493–502.

Zivkovic, Z., Bakker, B. and Kröse, B. (2005). Hierarchical map building using visual landmarks and geometric constraints. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Edmonton, AB, Canada.