2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

SVM-based Discriminative Accumulation Scheme for Place Recognition

A. Pronobis*, O. Martínez Mozos†, B. Caputo‡§

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

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

‡IDIAP Research Institute
1920 Martigny, Switzerland
§EPFL, 1015 Lausanne, Switzerland
bcaputo@idiap.ch

Abstract— **Integrating information coming from different sensors is a fundamental capability for autonomous robots. For complex tasks like topological localization, it would be desirable to use multiple cues, possibly from different modalities, so to achieve robust performance. This paper proposes a new method for integrating multiple cues. For each cue we train a large margin classifier which outputs a set of scores indicating the confidence of the decision. These scores are then used as input to a Support Vector Machine, that learns how to weight each cue, for each class, optimally during training. We call this algorithm SVM-based Discriminative Accumulation Scheme (SVM-DAS). We applied our method to the topological localization task, using vision and laser-based cues. Experimental results clearly show the value of our approach.**

I. INTRODUCTION

The capability to integrate effectively multiple cues is fundamental for autonomous systems. Robots are usually equipped with several sensors used to acquire as much information about the external world as possible. In general, each sensor captures a different aspect of the environment; however, alternative interpretations of the information obtained by the same sensor can also be valuable. Consider, for instance, a mobile robot equipped with a laser range sensor and a camera, which performs topological localization in an indoor environment. The two dimensional range information extracted from the laser scans is robust to visual variations introduced e.g. by illumination, but suffers from perceptual aliasing (different places might look the same [1]). At the same time, the visual sensor provides much more descriptive, but also noisy data. Integrating these two sensory modalities allows to take the best of both worlds.

In this paper we propose a new high-level accumulation scheme for multiple cues. Our method builds on previous work [2], [3]: for each cue we train a large margin classifier which outputs a set of scores indicating the confidence of the decision. Integration is achieved by feeding the scores to a Support Vector Machine [4]. Compared to previous accumulation methods [5], [6], [3], [2] our algorithm offers several advantages: *(a)* as shown in [3], discriminative accumulation schemes achieve consistently better performances than probabilistic ones [5], [6]; *(b)* compared to previous discriminative accumulation schemes [3], [2] our new approach gives the possibility to accumulate cues with a much more complex, possibly non-linear function, by using the SVM framework and kernels [4]. Such approach makes it possible to integrate together outputs of different classifiers such as SVM and AdaBoost. We call the new algorithm SVM-based Discriminative Accumulation Scheme (SVM-DAS).

We applied SVM-DAS to the domain of mobile robot topological localization in indoor environment under dynamic changes. This is a particularly challenging task: recognizing rooms under varying illumination conditions, and with variations in the configuration of furniture and small objects, is a hard recognition problem. We tested SVM-DAS using multiple visual cues, and using cues derived from two different modalities, vision and laser. We used SIFT [7] and Composed Receptive Fields Histograms (CRFH, [8]) for the vision channel, and the features proposed in [9] for the laser channel.

We conducted several sets of experiments of increasing difficulty on the IDOL2 database [10], and we benchmarked against Generalized-DAS (G-DAS) framework presented in [2]. Results show that integrating different visual cues, or better, different modalities allows to greatly increase the robustness of a recognition system, achieving accuracy of more than 94% under severe dynamic variations. Moreover, the new integration framework consistently outperforms G-DAS on both types of integration problems, with increase in recognition rate of up to 8%.

The rest of the paper is organized as follows: after a review of the relevant literature (Section II), Section III gives a brief description of G-DAS and presents our new cue integration scheme. Section IV describes the experimental setup, and Section V reports the experimental results showing the effectiveness of the proposed approach. The paper concludes with a summary and possible avenues for future research.

II. RELATED WORK

Several cue integration methods have been proposed in the robotics and machine learning community [11], [3], [2], [5], [12], [13]. These approaches can be described according to various criteria. For instance, Clark and Yuille [14] 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 instead when the output of one classifier is

This work was sponsored by the EU integrated projects CoSy (AP, OMM, FP6-004250-IP) and DIRAC (BC, IST-027787, www.diracproject.org) and the Swedish Research Council contract 2005-3600-Complex (AP). The support is gratefully acknowledged.

affected by the output of another classifier, so that their outputs are not anymore 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 [13], and for topological mapping and place recognition using multiple sensor modalities [11], [15]. In spite of remarkable performances for specic tasks, there are two main 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 expect the dimension of such a feature vector to increase as the number of cues grows, 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 possibly 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 [5], [3], [2]. We call such algorithms *high level integration methods*, of which voting is the most popular [16]. These techniques are more robust with respect to noisy cues or sensory channels and allow to decide on the number of cues that should be extracted and used for each particular classification task [2].

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.* [5] and further explored by Aloimonos and Shulman [17]. The method was then extended to discriminative methods in [3], [2].

III. CUE INTEGRATION VIA ACCUMULATION

This section describes our cue integration scheme. We first briefly review the theory behind the Support Vector Machines (Section III-A), that constitute a key building block of our approach. Then, we describe the Generalized Discriminative Accumulation Scheme (G-DAS, [2]) on which to large extent we build (Section III-B). Finally, we introduce our new algorithm and discuss its advantages in Section III-C.

A. Support Vector Machines

Consider the problem of separating the set of training data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ into two classes, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label. If we assume 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

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

The classification result is then given by the sign of $f(\boldsymbol{x})$. The values of α_i and b are found by solving a constrained minimization problem, which can be done efficiently using the SMO algorithm [4]. Most of the α_i's take the value of zero; those $\boldsymbol{x}_i$ with nonzero α_i are the "support vectors". In case where the two classes are non-separable, the optimization is formulated in such 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 $K(\boldsymbol{x}_i, \boldsymbol{x})$.

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

1) *Standard one-against-all (OaA) strategy*. If M is the no. 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 one-against-all strategy*. In [2], 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. Experiments presented in this paper and in [2] 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 machines 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.

B. Generalized Discriminative Accumulation Scheme

The G-DAS algorithm was proposed in [2], and in a preliminary version in [3], as a way to integrate multiple visual cues using the principle of accumulation. The basic idea is to consider real-valued outputs of a multi-class discriminative classifier (e.g. SVM) as an indication of confidence of the decision for each class, and accumulate all the outputs obtained for various cues with a linear function. Specifically, suppose we are given M classes and, for each class, a set of n_j training samples $\{\boldsymbol{I}_i^j\}_{i=1}^{n_j}$, $j = 1, \ldots, M$. Suppose also that, from each sample, we extract a set of P different cues $\{T_p(\boldsymbol{I}_i^j)\}_{p=1}^{P}$. Note that the samples here could be images, and then the cues would be different visual features; but they could also be outputs from different sensory modalities, like vision and laser scans, in which case the cues would be features extracted from these different sensors. In both cases, the goal is to perform recognition using all the cues. The G-DAS algorithm consists of two steps:

1) *Single-cue Models.* From the original training set $\{\{\boldsymbol{I}_i^j\}_{i=1}^{n_j}\}_{j=1}^{M}$, containing images belonging to all M classes, define P new training sets $\{\{T_p(\boldsymbol{I}_i^j)\}_{i=1}^{n_j}\}_{j=1}^{M}$, $p = 1, \ldots, P$, each relative to a single cue. For each new

training set train a multi-class classifier. Model parameters can be estimated during the training step via cross validation. Then, given a test sample $\boldsymbol{I}$, for each single-cue classifier estimate a set of outputs $\{O_h^p(T_p(\boldsymbol{I}))\}_{h\in H}$ reflecting the relation of the sample to the model. In case of the SVMs with standard OaO and OaA multi-class extensions, the outputs would be values of the $M(M-1)/2$ or M discriminant functions $f_h^p(T_p(\boldsymbol{I}))$ learned by the SVM algorithm during training.

2) *Discriminative Accumulation.* After all the outputs are computed for all the cues, they are being combined with different weights by a linear function:

$$O_h^{\Sigma P}(\boldsymbol{I}) = \sum_{p=1}^{P} a_p O_h^p(T_p(\boldsymbol{I})),\ a_p \in \Re^+.$$

As a result, any method of estimating the final decision can be used within the G-DAS framework, the same way it would be used for a single-cue classifier.

It is important to note that only one weight is used for all outputs of each cue, which simplifies the parameter estimation process (usually, an extensive search is performed in order to find the coefficients $\{a_p\}_{p=1}^P$), but also constraints the ability of the algorithm to adopt to the properties of each single cue. For a more comprehensive discussion on the G-DAS algorithm we refer the reader to [2].

C. SVM-based Discriminative Accumulation Scheme

There are several drawbacks of the G-DAS algorithm. First of all, the accumulation function is simple and linear, thus the algorithm is only able to weight the whole cues and not adopt to the characteristics of the models. This might be a limiting factor for complex tasks like robot localization. Moreover, there is no straightforward way to infer the weights from the training data. This is a problem in case of large number of cues, when exhaustive search becomes intractable.

What we propose here is to accumulate the outputs generated by single-cue classifiers using a more complex, possibly non-linear function, namely to use them as input to an SVM. As a result, the new accumulation function will be given as:

$$O_k^{\Sigma P}(\boldsymbol{I}) = \sum_{i=1}^{m} \alpha_i^k y_i K(\boldsymbol{O}_i, \boldsymbol{O}) + b^k,\ k = 1, \ldots, K,$$

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

$$\boldsymbol{O} = \left[\{O_h^1(T_1(\boldsymbol{I}))\}_{h\in H_1}, \ldots, \{O_h^P(T_P(\boldsymbol{I}))\}_{h\in H_P}\right].$$

The parameters α_i^k, y_i, and the support vectors $\boldsymbol{O}_i$ are inferred from the training data either directly or efficiently during the optimization process (e.g. by means of SMO [4]). The number of the final outputs K and the way of obtaining the final decision depends on the multi-class extension used with SVM-DAS. We use the one-against-one extension throughout the paper for which $K = M(M-1)/2$.

We call this new accumulation scheme SVM-DAS. The nonlinearity is given by the choice of the kernel function, thus in the case of the linear kernel the method is still linear. In this sense, SVM-DAS is more general that G-DAS. Also, for SVM-DAS each of the final outputs depends on all the outputs from single-cue classifiers, and the coefficients are learned optimally. Note that the outputs O_h^p can be derived from any large margin classifier, and not only from SVM. When SVM-DAS is used on outputs derived all from the same type of classifier, such as SVM or AdaBoost [18], then it can be seen as a variation of the stacking learning methods. In the case when the outputs are derived by different classifiers, for instance visual data outputs from SVM and laser range data outputs from AdaBoost, then SVM-DAS is a variation of the ensemble learning methods.

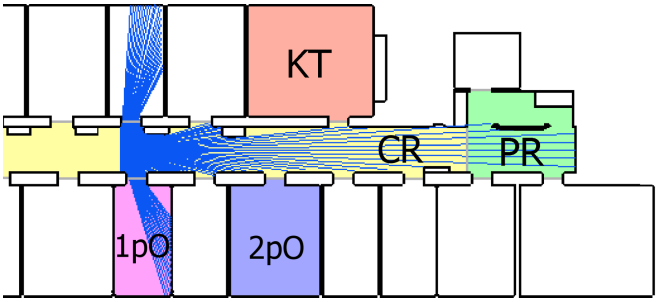


Fig. 1. Map of the environment used during data acquisition and an example laser scan simulated in the corridor. The rooms used during the experiments are annotated.

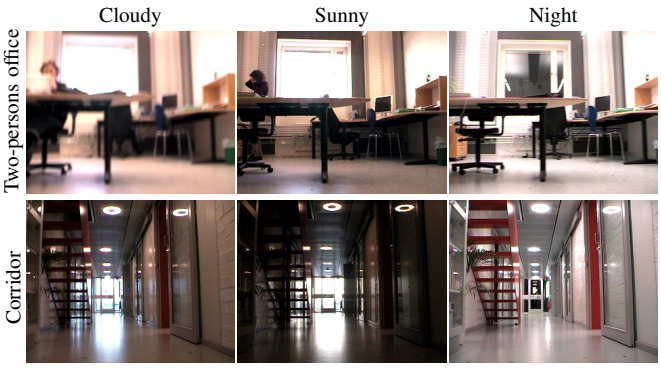
IV. EXPERIMENTAL SETUP

This section describes the setup used for the experiments reported in this paper. First, we describe the common scenario in which the evaluation took place (Section IV-A). Then, we present the overall architecture of our single-cue place recognition system as well as the methodology we followed during experiments (Sections IV-B). Finally, we briefly review the main building blocks of the system: the feature extractors and classifiers that were used to generate the cues benchmarked in the paper (Section IV-C).

A. Experimental Scenario

The algorithms presented in this paper have been tested in the domain of mobile robot topological localization on the IDOL2 (Image Database for rObot Localization 2 [10]) database. The database was introduced in [19] in order to test the robustness of an adaptive visual place recognition system in a real-world dynamic environment observed over a long period of time and under varying illumination conditions. The database comprises 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 perspective camera of resolution 320x240 pixels. In this paper we will use only the 12 data sequences acquired with the PowerBot.

The acquisition was performed in a five room subsection of a larger office environment, selected in such 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 Fig. 1. Example pictures showing interiors



(a) Variations introduced by illumination

(b) Variations observed over time

(c) Remaining rooms (at night)

Fig. 2. Examples of pictures taken from the IDOL2 database showing the interiors of the rooms, variations observed over time and caused by natural activity in the environment as well as introduced by changing illumination.

of the rooms are presented in Fig. 2. The appearance of the rooms was captured under three different illumination conditions: in cloudy weather, in sunny weather, and at night. The robots were manually driven through each of the five rooms while continuously acquiring images and laser range scans. Each image was then labelled as belonging to one of the rooms according to the position of the robot during acquisition. Since the database was originally designed to test the robustness of place recognition algorithms to variations that occur over a long period of time, the acquisition process 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 6 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.). It is important to note that, even for sequences acquired within a short time span under similar illumination conditions, variations still exist from everyday activities and viewpoint differences during acquisition. Example images illustrating the captured variability are shown in Fig. 2.

B. Single-cue Place Recognition

As a basis for the cue integration experiments, we used the place recognition systems presented in [20], [2] for visual cues and in [9] for laser range cues. The main principle behind both approaches is the same, as we can always find two main building blocks: a feature extractor and a classifier. For the work presented in this paper, we employed two discriminative classifiers to build models for the separate cues. The Support Vector Machines [4] were used both with visual and laser-based geometrical features, and the AdaBoost classifier [18] was used together with the geometrical features as described in [9]. Since we considered two different modalities, we also used different feature representations. In order to encode the visual information, we applied a rich global descriptor, Composed Receptive Field Histograms (CRFH) [8], and distinctive local features based on the SIFT descriptor [7]. Both has already been proved successful in the domain of vision-based localization [20], [2], [21]. To represent the information extracted from the laser, simple geometrical features were computed for each scan [9]. In the end, we constructed 4 different single-cue models: CRFH with SVM, SIFT with SVM, and laser range features with both SVM (L-SVM) and AdaBoost (L-AB).

We took a fully supervised approach and assumed that, during training of each of the models, the rooms are represented by collections of data capturing their visual and geometrical properties under various viewpoints, at fixed time and illumination setting. During testing, the algorithms were presented with data acquired in the same rooms, under roughly similar viewpoints but possibly under different illumination conditions and after some time. The goal was to recognize each single data sample provided to the system. 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 containing samples acquired at the rate of 5 fps. As a measure of performance we used the percentage of properly classified samples calculated separately for each of the rooms and then averaged with equal weights independently of the number of samples acquired in each room.

C. Feature Representation and Classification

In this work, we used two types of visual cues (global and local) extracted from the same image frame as well as simple geometrical features extracted from laser range scans.

As global image representation we used the Composed Receptive Field Histograms (CRFH) [8], a sparse multi-dimensional statistical representation of responses of several image filters. Following [20], we used histograms of 6 dimensions, with 28 bins per dimension, computed from second order normalized Gaussian derivative filters applied to the illumination channel at two scales. For the local feature extraction, we used the SIFT descriptor [7] which represents local image patches around interest points characterized by coordinates in the scalespace in the form of histograms of gradient directions. In order 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 [22].

In case of the laser sensor, we extracted a set of simple geometrical features from each scan [9]. We call them simple because they are represented by a single real value. The set of features used in this work was originally designed for laser scans covering 360^o field of view around the robot [9]. In this work, however, the scan covers only 180^o in front of the robot, therefore we set the rear values of the scan to zero.

As classifiers, we used AdaBoost [18] for the laser range features and the Support Vector Machines [4] described in Section III-A for all cues. The key idea behind AdaBoost is to create an accurate strong classifier by combining a set of weak classifiers. The requirement for each weak classifier is that its accuracy is better than a random guessing. The input to the algorithm is a set of labeled training examples which have assigned a weight distribution. In a series of rounds, the algorithm selects a new weak classifier based on the weight distribution, which is then modified. The final strong classifier is a weighted majority vote of the selected weak classifiers. The original algorithm was designed for binary classifications and outputs. However, in this work we used a modified version which permits us to classify several classes and to obtain a confidence value for each class as shown in [9].

In case of SVMs, special care must be used in choosing an appropriate kernel function. In this work, we used the χ^2 kernel [23] for the global CRFH descriptors, and the match kernel proposed in [24] for the local SIFT descriptors. Both have been used in our previous work on SVM-based place recognition, obtaining good performances [20], [2]. For the laser range features, we used a Radial Basis Function (RBF) kernel [4], which we selected through a set of reference experiments.

V. EXPERIMENTAL EVALUATION

We conducted several series of experiments on the IDOL2 database in order to analyse the properties of each of the four types of single-cue models (the SVM models trained on CRFH and SIFT as well as the SVM and AdaBoost models trained on the laser range cues) and evaluate the performance of both cue integration schemes (G-DAS and SVM-DAS). We present the results in successive subsections and give a brief summary and discussion on the efficiency of different solutions in Section V-C. First, we considered each of the cues separately and we benchmarked them on experiments of increasing difficulty (Section V-A). Then, we tested the accumulation schemes on several scenarios (Section V-B).

A. Experiments with Separate Cues

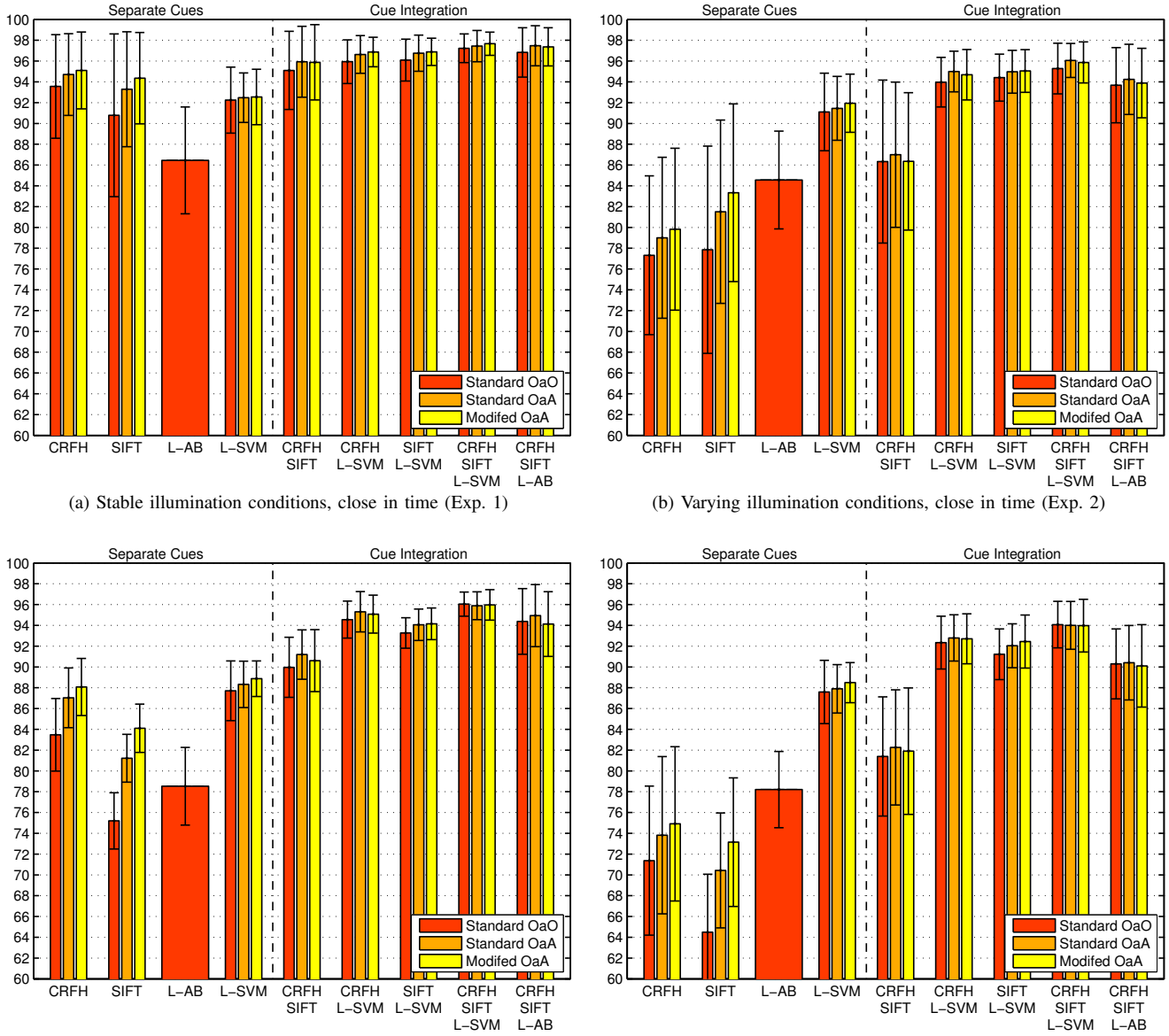
We conducted four sets of experiments for each cue. The first set consisted of 12 experiments, performed on different combinations of training and test data acquired closely in time and under similar illumination conditions. For the second set of experiments, we used 24 pairs of sequences captured still at relatively close times, but under different illumination conditions. In this way we increased the complexity of the problem [20], [2]. In the third set of experiments, we tested the robustness of each of the cues to long-time variations introduced by natural activity in the environment (objects/furniture being moved and reorganized). Therefore, we conducted 12 experiments, where we used for testing data acquired 6 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 6 months from each other and under different illumination settings. For all experiments, model parameters were determined via cross validation.

We evaluated the performance of all four types of 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 III-A. The results of all four sets of experiments for these models are presented in Fig. 3a-d (the first four bar groups). As a first remark, we see that, according to expectations, the recognition systems based on visual cues (CRFH and SIFT) suffer from changes in illumination, while the geometrical laser-based features don't. Moreover, variations that occurred over a long period of time pose a challenge for both modalities. We can observe differences in performance also between the two visual cues. The models based on global features (CRFH) suffer more from the illumination variations, while the SIFT features are less robust to variations introduced by natural activities in the environment. 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). This illustrates the potential of visual cues, but also stresses the need for more robust solutions.

One of the contributions of the paper is the place recognition algorithm based on simple-valued geometrical features [9] and SVMs. Fig. 3a-d presents a comparison of performance of our new method and the previous solution using the AdaBoost classifier [9]. We can see that the difference in performance is very significant in favour of the SVM-based method (from 6.1% for Exp. 1 to 10.3% for Exp. 4 in average) which allows to conclude that the robustness of the system was greatly improved by implementing a more complex classifier.

As already mentioned, all the presented 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 [2] - in case of single cue and G-DAS experiments, the modified version gives the best performance independently of the modality on which the classifier was trained.

A further analysis of the results can be performed, that serves as a motivation for integrating different visual cues and modalities. Fig. 4 shows the distribution of errors for each actual class (room) made by the four models. It is



(a) Stable illumination conditions, close in time (Exp. 1)

(b) Varying illumination conditions, close in time (Exp. 2)

(c) Stable illumination conditions, distant in time (Exp. 3)

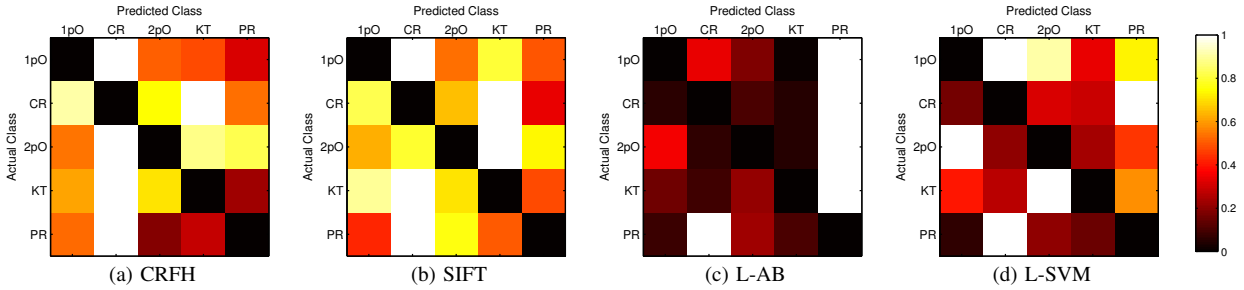
(d) Varying illumination conditions, distant in time (Exp. 4)

Fig. 3. Results of a separate evaluation of each of the cues and performance of the SVM-DAS cue integration scheme on four types of problems.

apparent that each of the cues makes errors according to a different pattern. At the same time, similarities occur between the same modalities. We can see that visual models are biased towards the corridor, while the geometrical models tend to misclassify places as the printer area. A straightforward explanation can be offered for that phenomenon. The vision-based models were trained on images acquired with perspective camera with constrained viewing angle. As a result, similar visual stimuli coming from the corridor is present in the images captured by the robot leaving each of the rooms. The same area close to doorway, from the geometrical point of view, is similar to the narrow passage in the printer area. Ideally, the cue integration scheme should learn to trust more different cues with respect to different classes.

B. Experiments with Cue Integration

The accumulation schemes presented in this paper perform high level cue integration. As a result, separate models should be trained for each of the combined cues. In our evaluation, we used the models obtained during single-cue experiments. In order to be used for real applications, an integration scheme should perform and generalize well in presence of any type of variability it might encounter. For that reason, 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.



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

Fig. 4. Distribution of errors made by the four models for each actual class (bright colors indicate errors). The diagonal elements were removed.

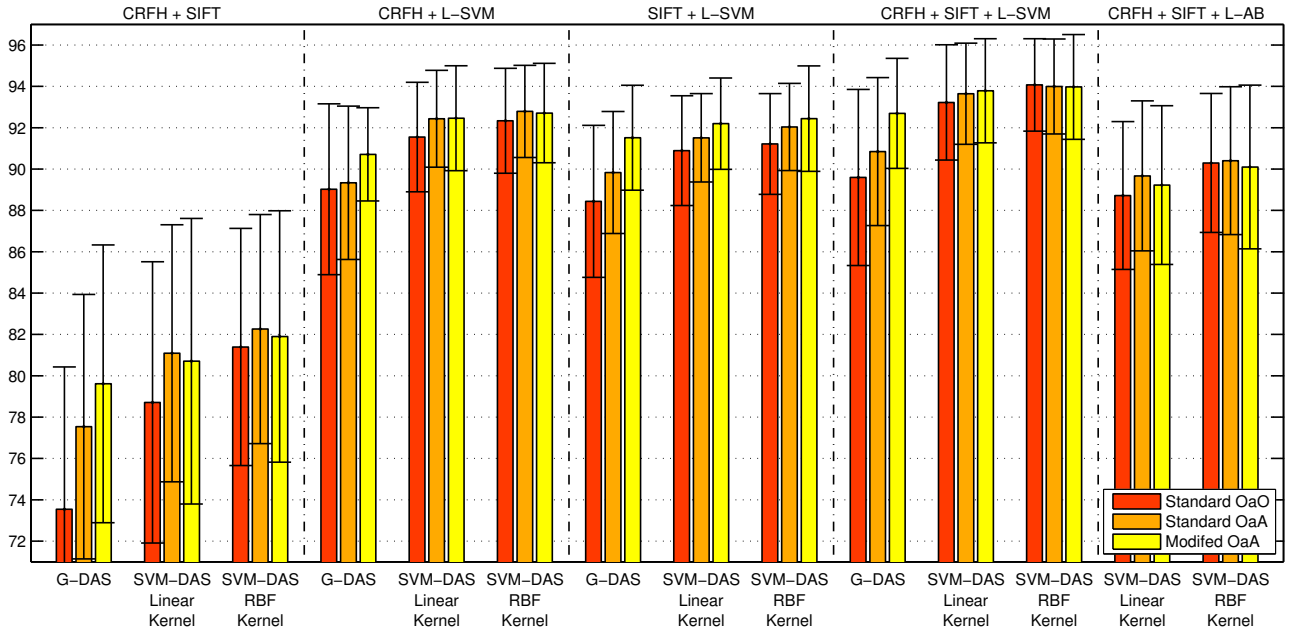


Fig. 5. Comparison of performance of all integration methods for the most complex problem (Exp. 4).

For the final experiments, we selected three different cue accumulation methods: Generalized DAS (G-DAS) and SVM-DAS with two kernel types (linear and Radial Basis Function). In all experiments, we found that the SVM-DAS with RBF kernel outperforms the other methods (the difference in performance with respect to G-DAS was statistically significant at the confidence level of 95%). As a result, for space reasons, we report results of each of the experiments only using that method (Fig. 3a-d, last 5 bar groups). Detailed comparison of all integration methods for the most complex problem (Exp. 4) is given in Fig. 5.

We tested the methods for several combinations of different cues and modalities. First, we combined the two visual cues, obtaining similar results as in [2]. We see that the generalization of purely visual recognition system can be greatly improved by integrating different types of cues, in this case local and global. This can be observed especially for Exp. 4, where the algorithms had to tackle largest variability. Despite that, according to the error distributions in Fig. 4, we should expect largest gain when different modalities are combined. As we can see from Fig. 3 this is the case indeed. By combining one visual cue and laser range cue (e.g. CRFH + L-SVM), we exploit the descriptive power of vision in 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. The gain in this case is statistically significant at the confidence level 95%.

As it was mentioned in Section III-C, SVM-DAS can be applied for problems where outputs of different classifiers need to be integrated. To test this ability in practice, we combined the SVM models trained on visual cues with AdaBoost model based on geometrical features (L-AB). We present the results in Fig. 3a-d (last bar group). It can be observed that the method obtained large improvement in comparison to each of the individual cues. For instance for Exp. 4, the recognition rate increased by 12.2% in average. This proves the versatility of our approach.

C. Discussion

The results of the extensive experimental evaluation presented in this section clearly show that SVM-DAS performs significantly better than G-DAS and can be used to integrate outputs of classifiers of different characteristics employing different multi-class algorithms. We also showed that by using more sophisticated kernel types, it is possible to

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

TABLE I

AVERAGE PERCENTAGES (WITH STANDARD DEVIATIONS) OF TEST SAMPLES FOR WHICH ALL CUES HAD TO BE USED IN ORDER TO OBTAIN THE MAXIMAL RECOGNITION RATE.

perform non-linear cue accumulation. The experiments (see Fig. 5) show that although there is no drastic improvement, we can expect better results with the RBF kernel (especially for the OaO multi-class extension). As a result, we suggest that the kernel was selected according to the constraints put 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.

At this point, a comment should be made on the computational cost of using multiple cues in general. Although, it is clear that generalization performance can be significantly improved by using multiple cues or modalities, each of the cues introduces additional cost. Therefore, there is always a trade-off between the complexity of the solution and the overall performance. For example, a solution based on global visual features, laser range cues and SVM-DAS runs in real-time at a rate of approximately 5fps, which would not be possible if additional visual cue such as SIFT was used. It should be noted, however, that due to the high level integration architecture, not all of the cues have to be always extracted and used, especially that, in most cases, decision based on one cue only is correct. The computational cost can be significantly reduced by taking the approach presented in [2]. 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 [2]. Table I 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 I) and we can retain the maximal performance by using additional cues only in approximately 30% of cases. Additional cues will be used more often when the variability is large, and rarely for less difficult cases.

VI. SUMMARY AND CONCLUSION

This paper presented a new cue integration method, able to combine multiple cues derived by a single modality, as well as cues obtained by multiple sensors. For each cue, it trains a large margin classifier and computes as a set of outputs, related to the confidence of the decision. The outputs are then used as input to a Support Vector Machine, that combines optimally the different cue contributions. The method was tested in the domain of robot localization. A thorough experimental evaluation using multiple visual cues alone, and combined with laser range features, clearly show the value of our approach.

In the future, we plan to use this method for attacking the scalability issue with geometrical localization methods. Also, we plan to combine this approach with incremental extensions of the SVM algorithm ([19], [25]), so to obtain a system able to learn continuously from multiple sensors.

REFERENCES

[1] B. Kuipers and P. Beeson, "Bootstrap learning for place recognition," in *Proc. AAAI'02*.

[2] A. Pronobis and B. Caputo, "Confidence-based cue integration for visual place recognition," in *Proc. IROS'07*.

[3] M. E. Nilsback and B. Caputo, "Cue integration through discriminative accumulation," in *Proc. CVPR'04*.

[4] N. Cristianini and J. S. Taylor, *An Introduction to SVMs and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[5] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, 1985.

[6] B. Caputo and G. Dorko, "How to combine color and shape information for 3D object recognition: Kernels do the trick," in *NIPS'02*.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, 2004.

[8] O. Linde and T. Lindeberg, "Object recognition using composed receptive field histograms of higher dimensionality," in *Proc. ICPR'04*.

[9] O. M. Mozos, C. Stachniss, and W. Burgard, "Supervised learning of places from range data using adaboost," in *Proc. ICRA*, 2005.

[10] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "The IDOL2 database," KTH, CAS/CVAP, Tech. Rep., 2006, Available at `http://cogvis.nada.kth.se/IDOL/`.

[11] A. Tapus and R. Siegwart, "Incremental robot mapping with fingerprints of places," in *Proc. IROS'05*.

[12] J. Triesch and C. Eckes, "Object recognition with multiple feature types," in *Proc. ICANN'98*.

[13] J. Matas, R. Marik, and J. Kittler, "On representation and matching of multi-coloured objects," in *Proc. ICCV'95*.

[14] J. Clark and A. Yuille, *Data fusion for sensory information processing systems*. Kluwer Academic Publisher, 1990.

[15] A. Rottmann, O. M. Mozos, C. Stachniss, and W. Burgard, "Semantic place classification of indoor environments with mobile robots using boosting." in *Proc. AAAI*, Pittsburgh, PA, USA, 2005.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley, 2001.

[17] J. Aloimonos and D. Shulman, *Integration of Visual Modules: an Extension of the Marr Paradigm*. Academic Press, 1989.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *EuroCOLT'95*.

[19] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "Incremental learning for place recognition in dynamic environments," in *Proc. IROS'07*.

[20] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen, "A discriminative approach to robust visual place recognition," in *Proc. IROS'06*.

[21] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Proc. ICRA'01*.

[22] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *IJCV*, vol. 66, no. 3, 2006.

[23] O. Chapelle, P. Haffner, and V. Vapnik, "SVMs for histogram-based image classification," *IEEE Trans. Neur. Netw.*, vol. 10, no. 5, 1999.

[24] C. Wallraven, B. Caputo, and A. Graf, "Recognition with local features: the kernel recipe," in *Proc. ICCV'03*.

[25] F. Orabona, C. Castellini, B. Caputo, J. Luo, and G. Sandini, "Indoor place recognition using online independent support vector machines," in *Proc. BMVC'07*.