A Discriminative Approach to Robust Visual Place Recognition

A. Pronobis, B. Caputo, P. Jensfelt, and H.I. Christensen
Centre for Autonomous Systems
Royal Institute of Technology,
SE-100 44 Stockholm, Sweden
[pronobis,caputo,patric,hic]@nada.kth.se

***Abstract*— An important competence for a mobile robot system is the ability to localize and perform context interpretation. This is required to perform basic navigation and to facilitate local specific services. Usually localization is performed based on a purely geometric model. Through use of vision and place recognition a number of opportunities open up in terms of flexibility and association of semantics to the model. To achieve this the present paper presents an appearance based method for place recognition. The method is based on a large margin classifier in combination with a rich global image descriptor. The method is robust to variations in illumination and minor scene changes. The method is evaluated across several different cameras, changes in time-of-day and weather conditions. The results clearly demonstrate the value of the approach.**

I. Introduction

A fundamental competence in mobile robotics is the ability to localize, i.e., to determine its position in the world. The methods used are either based on metric geometric models or discrete topological. Semantics has rarely been associated with such models. However as robot break-down the fences and start to interact with people there is a need to include semantics in the models and to enable use of *place* information as a way to introduce contextual information into the system.

Place recognition allow localization in topological mapping and provide a method for loop closing or recovery from the kidnaped robot problem. In particular, the research on topological mapping has pushed methods for place recognition. Scalability issues have been at the forefront of the issues to be addressed.

Early work on place recognition was based on sonar and/or laser data, as robust sensory modalities [1]. Recently advances in vision has made this a viable modality opening up for a richer variety of places and more robust detection.

This paper presents a vision-based algorithm able to recognize places on the basis of their visual appearances, under different illumination conditions and across a significant span of time. We apply an appearance-based recognition technique, from the object classification domain, composed by: (a) a rich visual descriptor consisting of a high dimensional receptive field histogram. This descriptor has shown remarkable performances coupled with computational efficiency on challenging object recognition scenarios [2]; (b) a support vector machine, a discriminative classifier which has become the algorithm of choice for several visual recognition domains [2], [3]. The method was assessed on a thorough set of experiments, using 3 different camera devices and image data gathered under varying conditions and times. Results show that the method is able to recognize places with high precision and robustness, even when training on images from one camera device and testing on another.

The rest of the paper is organized as follows: after a review of previous literature in the field (Section II), we describe our visual recognition algorithm (Section III). Section IV describes the experimental setup and Section V presents experiments showing the effectiveness of the proposed approach. Conclusions are drawn and potential avenues for future research outlined in Section VI.

II. Related Works

The research on *place recognition* has been mostly conducted in the mobile robotics community. In [4] a system using a sequential AdaBoost classifier with simple geometric features is presented. The features are extracted from two laser scanners mounted back to back on a robot and correspond to for example the average laser beam length and the freespace area. The different classes to distinguish between are room, door, corridor and hallway.

Several approaches to the vision-based place recognition have been proposed. These methods employ either regular cameras ([5], [6]) or omni-directional sensors ([7], [8], [9], [10], [11]) in order to acquire images. 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 localization techniques make use of either artificial or natural landmarks in order to extract information about position. An interesting approach to the problem was presented by Mata *et al.* [12]. The system uses information signs as landmarks, and interprets them through its ability to read text and recognize icons. Local image features may also be regarded as natural landmarks. The SIFT descriptor [13] was successfully used by Se *et al.* [14] and Andreasson *et al.* [11] (with modifications), while Tamimi and Zell [6] employed Kernel PCA to extract features from local patches. Global features are also commonly used for place recognition. Torralba [15] suggested to use a representation called the "gist" of the scene, which is a vector of principal components of outputs of a filter bank applied to the image. Several

other approaches use color histograms [8], [9], eigenspace representation of images [7] or Fourier coefficients of low frequency image components [10].

We are not aware of any other evaluation of visual place recognition algorithms conducted under conditions realistic for applications, i.e. with varying illumination conditions and over time; thus, this is one of the major contribution of this paper.

III. Discriminative Place Recognition

This section describes our approach to visual place recognition, and the algorithm we propose to this purpose. Following [5], we assume that the encoding of the global configuration of a real-world scene is informative enough to represent and recognize it. We apply an appearance-based classification method, successfully used for object recognition in realistic settings [2]. The method is fully supervised, and assumes that each room is represented, during training, by a collection of images which capture its visual appearance under different viewpoints, at a fixed time and illumination setting. During testing, the algorithm is presented with images of the same rooms, acquired under roughly similar viewpoints but possibly under different illumination conditions, and after some time (where the time range goes from some minutes to several months). The goal is to recognize correctly each single image seen by the system.

The rest of this section describes the feature descriptor (Section III-A) and the classifier we used (Section III-B). A comprehensive description of the experimental setup is given in Section IV.

A. High Dimensional Composed Receptive Field Histograms

Recent work has shown that receptive field responses summarized into histograms are highly effective for recognition of objects [16], [17] and spatio-temporal events [18]. Here we used histogram features of very high dimensionality (6-16 dimensions), which should be able to capture the rich visual appearance of indoor places. When using histograms of such high dimensionality, computational problems can easily occur. Thus, we used the method proposed by [2], which makes use of a sparse and ordered representation allowing to define efficient operations on them (for instance, a 16-dimensional histogram of a 256×256 image can be computed in about 0.1 s on a 1GHz Sun Fire). High dimensional composed receptive field histograms can be computed from several types of image descriptors (and various combinations of these):

- Normalized Gaussian derivatives, obtained by computing partial derivatives $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ from the scale-space representation $L(\cdot,\cdot;\ t) = g(\cdot,\cdot;\ t) * f$ obtained by smoothing the original image f with a Gaussian kernel $g(\cdot,\cdot;\ t)$, and multiplying the regular partial derivatives by the standard deviation $\sigma = \sqrt{t}$ raised to the order of differentiation [19].
- Differential invariants, invariant to rotations in the image plane, mainly the normalized gradient magnitude $|\nabla_{\text{norm}} L| = \sqrt{t(L_x^2 + L_y^2)}$, the normalized Laplacian $\nabla^2_{\text{norm}} L = t(L_{xx} + L_{yy})$, the normalized determinant of the Hessian $\det(\mathcal{H}_{\text{norm}} L) = t^2(L_{xx}L_{yy} - L_{xy}^2)$.
- Chromatic cues obtained from RGB-images according to $C_1 = (R - G)/2$ and $C_2 = (R + G)/2 - B$.

We tested a wide variety of combinations of image descriptors, with several scale levels σ and numbers of histogram bins per dimension (for a comprehensive report on these experiments see [20]). On the basis of these results, here we used composed receptive field histograms of six dimensions, with 28 bins per dimension, computed from second order normalized Gaussian derivative filters applied to the illumination channel.

B. Support Vector Machines

Support Vector Machines (SVMs, [21], [22]) belong to the class of large margin classifiers. Consider the problem of separating the set of training data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots (\boldsymbol{x}_m, y_m)$ into two classes, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label (for the multi-class extensions, we refer the reader to [21], [22]). If we assume that the two classes can be separated by a hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, and that we have no prior knowledge about the data distribution, then the optimal hyperplane (the one with the lowest bound on the expected generalization error) is the one which has maximum distance to the closest points in the training set. The optimal values for $\boldsymbol{w}$ and b can be found by solving the following constrained minimization problem:

$$\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{subject to} & y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, \forall i = 1, \ldots m \end{array} \tag{1}$$

Solving it using Lagrange multipliers α_i $(i = 1, \ldots m)$ results in a classification function

$$f(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \cdot \boldsymbol{x} + b\right), \tag{2}$$

where α_i and b are found by using an SVC learning algorithm [21], [22]. Most of the α_i's take the value of zero; $\boldsymbol{x}_i$ with nonzero α_i are the "support vectors". In cases where the two classes are non-separable, the solution is identical to the separable case except for a modification of the Lagrange multipliers into $0 \leq \alpha_i \leq C, i = 1, \ldots m$, where C determines the trade-off between margin maximization and error minimization. To obtain a nonlinear classifier, one maps the data from the input space $\Re^N$ to a high dimensional feature space $\mathcal{H}$ by $\boldsymbol{x} \rightarrow \Phi(\boldsymbol{x}) \in \mathcal{H}$, such that the mapped data points of the two classes are linearly separable in the feature space. Assuming there exists a kernel function K such that $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$, then a nonlinear SVM can be constructed by replacing the inner product $\boldsymbol{x} \cdot \boldsymbol{y}$ in the linear SVM by the kernel function $K(\boldsymbol{x}, \boldsymbol{y})$

$$f(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right). \tag{3}$$

This corresponds to constructing an optimal separating hyperplane in the feature space. Kernels commonly used include

polynomials $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y})^d$, which can be shown to map into a feature space spanned by all order d products of input features, and the Gaussian RBF kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp\{-\gamma||\boldsymbol{x} - \boldsymbol{y}||^2\}$. In this paper we use the χ^2 kernel ([23]):

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\{-\gamma\chi(\boldsymbol{x} - \boldsymbol{y})^2\}, \tag{4}$$

which has shown to give good performances for histogram-like features [24], [3] in vision applications.

IV. Experimental Setup

In this section we describe the experimental scenario and the data acquisition devices employed for the evaluation of our visual place recognition system. We tested it on two mobile robot platforms, "Minnie" and "Dumbo", as well as on images captured with a standard camera. The robot platforms are shown in Fig. 1. For the purpose of the experiments with the camera, we acquired a new database, INDECS (INDoor Environment under Changing conditionS), comprising pictures of places. The database represents one of the contributions of this paper, and together with all the other visual data used during the experiments, will be made publicly available upon acceptance of the paper.

The rest of the section is organized as follows: Section IV-A presents the working scenario, as to say the environment where we conducted the experiments and the image acquisition procedure. Then, Section IV-B gives detailed information on the robot platforms. Finally, Section IV-C provides a brief description of the INDECS database.

A. Experimental scenario

The experiments were conducted within a five room subsection of a larger office environment. Each of the five rooms represents a different type of functional area: a one-person office, a two-persons office, a kitchen, a corridor, and a printer area (in fact a continuation of the corridor). The rooms are physically separated by sliding glass doors, with the exception of the printer area which was treated as a separate room only due to its different functionality. Example pictures showing the interior of each room are presented in Fig. 2. Fig. 5 provides top views of the environment.

(a) Minnie (b) Dumbo

Fig. 1. Robot platforms employed in the experiments.

As already mentioned, the visual data were acquired with three different devices. In each case, the appearance of the rooms was captured under three different illumination and weather conditions: in cloudy weather (natural and artificial light), in sunny weather (direct natural light dominates), and at night (only artificial light). The image acquisition was spread over a period of time of three months, for the INDECS database, and over two weeks for the robot platforms. Additionally, the INDECS database was acquired ten months before the experiments with the robots. In this way we captured the visual variability that occurs in the real-world environments due to varying illumination and natural activities in the rooms (presence/absence of people, furniture relocated, changed, added or, removed). Fig. 3 presents a comparison of images taken under different illumination conditions and using various devices.

B. Robot platforms

Both robots, the PeopleBot Minnie and the PowerBot Dumbo, are equipped with the pan-tilt-zoom Cannon VC-C4 camera. However, as can be seen from Fig. 1, the cameras are mounted at different height. On Minnie the camera is 98cm above the floor, whereas on Dumbo it is 36cm. Furthermore, the camera on Dumbo was tilted up approximately 13° to reduce the amount of floor captured in the images. All images were acquired with a resolution of 320x240 pixels, with the zoom fixed to wide-angle[1], the auto-exposure and the auto-focus modes enabled.

We followed the same procedure during image acquisition with both robot platforms. The robot was manually driven (average speed around 0.3-0.35m/s) through each of the five rooms while continuously acquiring images at the rate of five frames per second. For the different illumination conditions (sunny, cloudy, night), the acquisition procedure was performed twice, resulting in two image sequences acquired one after another giving a total of six sequences across a span of over two weeks. Example images can be seen in Fig. 3. Due to the manual control, the path of the robot was slightly different for every sequence. Example paths are presented in Fig. 5. Each image sequence consists of 1000-1300 frames. To automate the process of labeling the images for the supervision, the robot pose was estimated during the acquisition process using a laser based localization method. Each image was then labeled as belonging to one of the five rooms based on the position from where it was taken. As a consequence of this, images taken, for example, from the corridor, but looking into a room are labeled as corridor.

[1]Roughly 45° horizontal and 35° vertical field of view.

Fig. 2. Example pictures taken from the INDECS database showing the interiors of the five rooms used during the experiments.

(a) Corridor (b) Kitchen

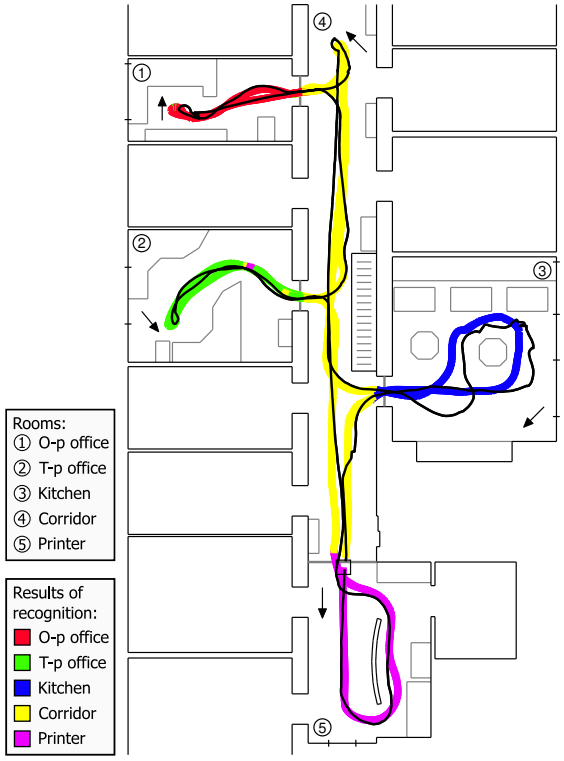
Fig. 3. Example pictures acquired with the camera and the two robot platforms under various illumination conditions. Pictures on the left show the influence of the the illumination, while the examples on the right illustrate the differences between pictures acquired in a cluttered environment using different devices. Additional variability caused by natural activities in the rooms is also apparent (presence of people, relocated furniture).

C. The INDECS Database

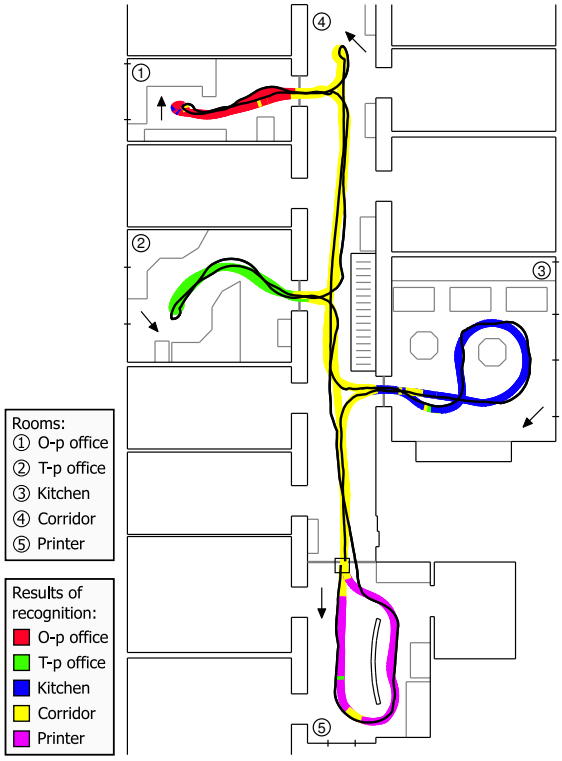
The INDECS database consists of pictures of the environment described above gathered under different viewpoints and locations. We marked several points in each room (approximately one meter apart) where we positioned the camera for each acquisition. The number of points changed with the dimension of the room, from a minimum of 9 for the one-person office to a maximum of 32 for the corridor. At each location we acquired 12 pictures, one every 30°, even when the tripod was located very close to a wall or furniture. Images were acquired using an Olympus C-3030ZOOM digital camera mounted on a tripod. The height of the tripod was constant and equal to 76 cm; all images in the INDECS database were acquired with a resolution of 1024x768 pixels, the auto-exposure mode enabled, flash disabled, the zoom set to wide-angle mode, and the auto-focus enabled. In this paper the INDECS images were subsampled to 512x386 before being used in the experiments. Again, the images were labeled according to the position of the point at which the acquisition was made. The images were taken across a span of three months and, as in the previous case, under various illumination conditions (sunny, cloudy and night). Fig. 3 illustrates types of variability captured for some rooms. In total there are 3264 images (324 for the one-person office, 492 for the two-persons office, 648 each for the kitchen and the printer area, and 1152 for the corridor) in the INDECS database.

V. Results

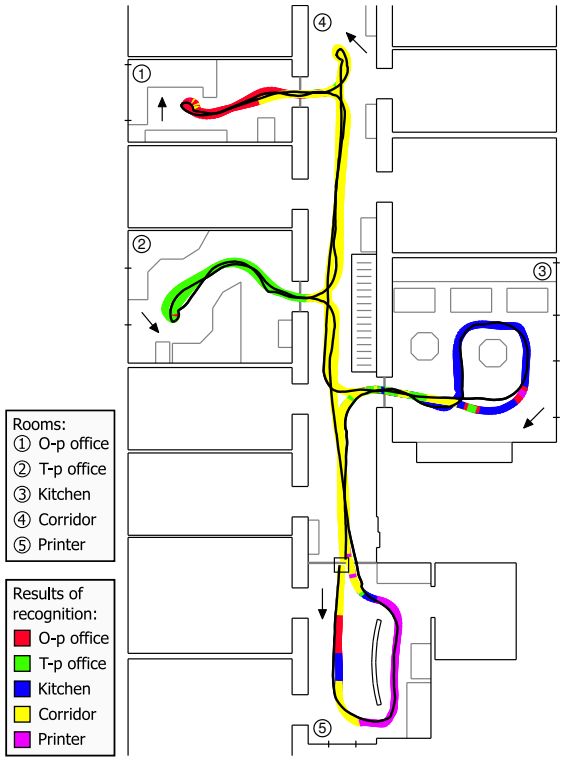
We conducted three sets of experiments in order to evaluate the performance of our system and test its robustness to different types of variations. We present the results in successive subsections and give a brief summary in Section V-D. We started with a set of reference experiments evaluating our method under stable illumination conditions (Section V-A). Next, we increased the difficulty of the problem and tested the robustness of the system to changing illumination conditions as well as to other variations that may occur in real-world environments (Section V-B). Finally, we conducted a series of experiments aiming to reveal whether a model trained on images acquired with one device can be useful for solving localization problems with a different device (Section V-C). In every case, the system performed the recognition on the basis of only one input image. In future work we intend to extend this by fusing information over time, but the aim of the current work is to investigate the performance of the underlying recognition system. In view of the fact that the number of acquired images varied across the rooms, each



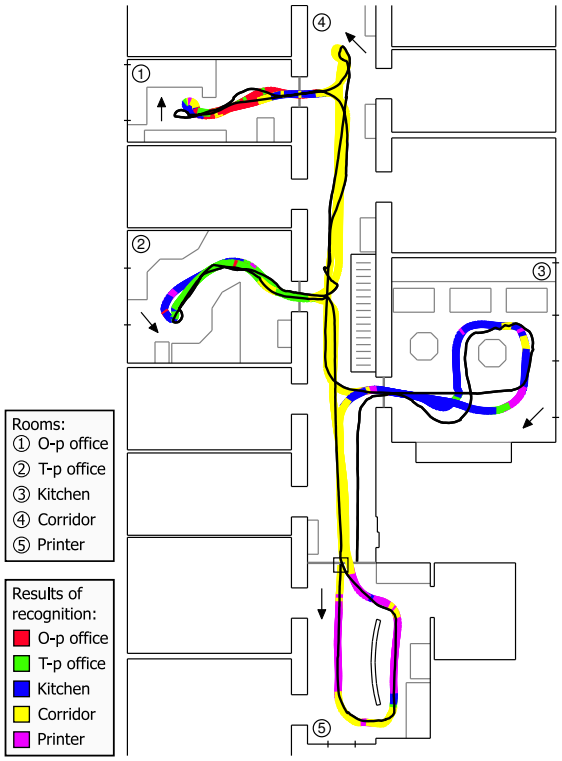
(a) Minnie Cloudy2 ⇒ Minnie Cloudy1



(b) Dumbo Cloudy2 ⇒ Dumbo Sunny2

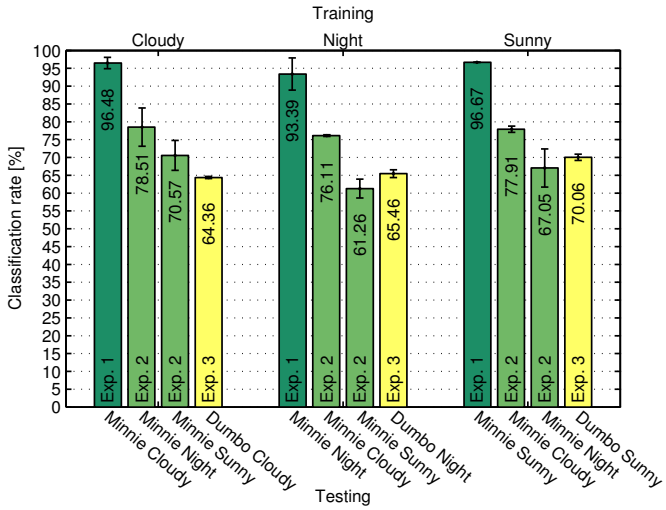


(c) Dumbo Cloudy1 ⇒ Dumbo Night1

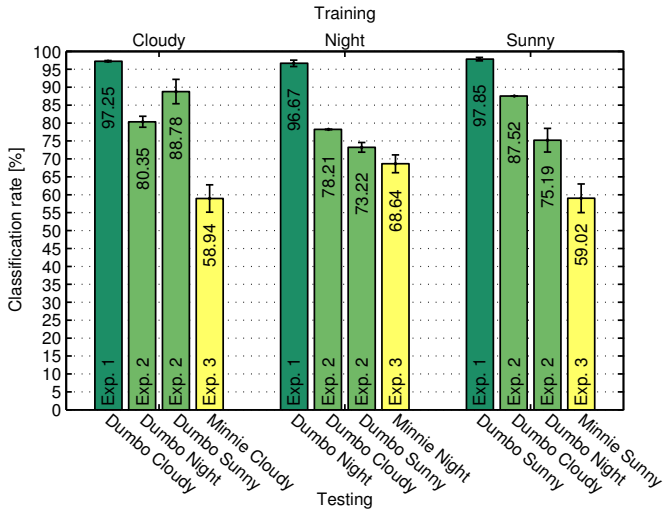


(d) Minnie Night2 ⇒ Dumbo Night1

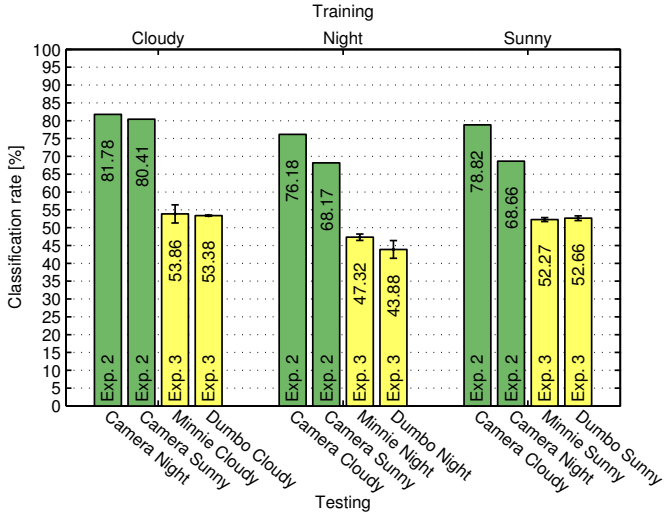
Fig. 5. Maps of the environment with plotted paths of the robot during acquisition of the training and test sequences. The training path is plotted with the thin black line, while the thick line shows the test path. The color of each point indicates the result of recognition, and the arrows show the direction of driving. Each experiment started at the point marked with square. The position of the furniture (plotted with gray line) is approximate and could vary between the experiments.



(a) Training on images acquired with Minnie



(b) Training on images acquired with Dumbo



(c) Training on the INDECS database

Fig. 4. Average results of the experiments with both robot platforms and the standard camera. The results are grouped according to the type of illumination conditions under which the training images were acquired. The bottom axes indicate the platform and illumination conditions used for testing. The uncertainties are given as one standard deviation.

room was considered separately during the experiments. The final classification rate was then computed as an average to which the results for each room contributed equally. For all the experiments we used our extended version of the *libSVM* software, and we set $C = 100$. After a preliminary set of experiments, we decided to use the χ^2 kernel and constant parameters of the feature extractor. The parameters were, however, different for the experiments with the robot platforms (scale $\sigma = 1$ and 4) and for images acquired with the camera ($\sigma = 2$ and 8). Such approach was motivated by the fact that the cameras mounted on the robots offered lower image quality, and the movement introduced additional distortions. Kernel parameters were determined via cross-validation.

A. Stable illumination conditions

In order to evaluate our method under stable illumination conditions, we trained and tested the system on pairs of image sequences acquired one after the other using the same robot. We did not use the INDECS database for these experiments since only one set of data for each illumination condition was available. Although the illumination conditions for both training and test images were in this case very similar, the algorithm had to tackle other kinds of variability such as viewpoint changes caused mainly by the manual control of the robot and presence/absence of people. The results of the performed experiments are presented in Fig. 4a,b. For each platform and type of illumination conditions used for training, the first bar presents an average classification rate over the two possible permutations of the image sequences in the training and test sets[2]. On average, the system classified properly 95.5% of the images acquired with Minnie and 97.2% of images acquired with Dumbo. Detailed results for one of the experiments are shown in Fig. 5a. It can be observed that the errors are usually not a result of viewpoint variations (compare the training and test paths in the kitchen) and mostly occur near the borders of the rooms. This can be explained by the relatively narrow filed of view of the cameras as well as the fact that the images were not labeled according to their content but to the position of the robot at the time of acquisition. Since these experiments were conducted with the sequences captured under similar conditions, we treat them as a reference for other results.

B. Varying illumination conditions

We also conducted a series of experiments aiming to test the robustness of our method to changing illumination conditions as well as to other variations caused by normal activities in the rooms. The experiments were conducted on the INDECS database and the visual data captured using both robot platforms. As with the previous experiments, the same device was used for both training and testing. This time, however, the training and test sets consisted of images acquired under different illumination conditions and usually on different days. Fig. 4a,b show average results of the experiments with the

[2]Training on the first sequence, testing on the second sequence, and vice versa.

robots for each permutation of the illumination conditions used for training and testing (the two middle bars for each type of training conditions). Fig. 4c gives corresponding results obtained on the INDECS database.

We see that in general the system performs best when trained on the images acquired in cloudy weather. The explanation for this is straightforward: the illumination conditions on a cloudy day can be seen as intermediate between those at night (only artificial light) and on a sunny day (direct natural light dominates). In such case, the average classification rate computed over two testing illumination conditions (sunny and night) was equal to 84.6% for Dumbo, 74.5% for Minnie, and 81.0% for the INDECS database. Fig. 5b,c present detailed results for two example runs. The errors occur mainly for the same reasons as in the previous experiments and additionally in places heavily affected by the natural light e.g. when the camera is directed towards a bright window. In such cases, the automatic exposure system with which all the cameras are equipped causes the pictures to darken. Minnie was more susceptible to that phenomenon due to the higher position of the camera.

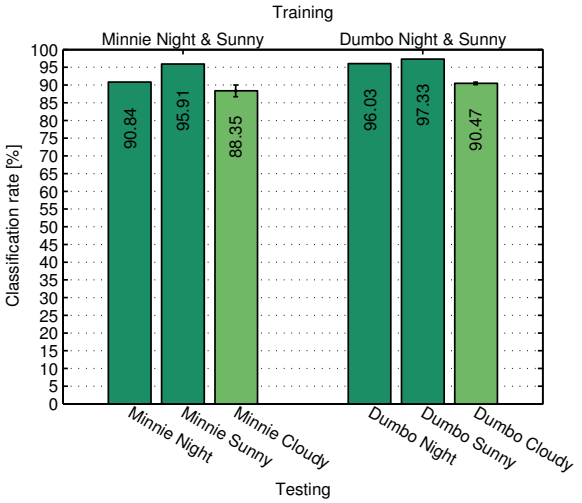


Fig. 6. Performance of the system trained on two image sequences acquired under different illumination conditions (sunny and night) for both mobile platforms. The classification rates in case of experiments using different illumination conditions for training and testing were averaged over two test image sequences. The uncertainties are given as one standard deviation.

C. Recognition across platforms

The final set of experiments was designed to test the portability of the acquired model across different platforms. For that purpose we trained and tested the system on images acquired under similar illumination conditions using different devices. We started with the experiments with both robot platforms. We trained the system on the images acquired using either Minnie or Dumbo and tested with the images captured with the other robot. We conducted the experiments for all illumination conditions. The main difference between the platforms from the point of view of our experiments lies in the height at which the cameras are mounted. The results presented in Fig. 4a,b indicate that our method was still able to classify up to about 70% of images correctly. The system performed better when trained on the images captured with Minnie. This can be explained by the fact that the lower mounted camera on Dumbo provided less diagnostic information. It can also be observed from Fig. 5d that in general the additional errors occurred when the robot was positioned close to the walls or furniture. In such cases the height at which the camera was mounted influenced the content of the images the most.

We followed a similar procedure using the INDECS database as a source of training data and different image sequences captured with the robot platforms for testing. It is important to note that the database was not intended to be used for this purpose, and was acquired ten months before the experiments with the robots. Additionally, the points at which the pictures were taken were positioned approximately 1m from each other and, in case of the kitchen, covered different area of the room due to reorganization of the furniture. Consequently, the problem required that the algorithm was invariant not only to various acquisition techniques but also offered great robustness to large changes in viewpoint and the appearance of the rooms. The experimental results are presented in Fig. 4c. We see that the algorithm obtains a recognition performance of about 50%. While this result is surely disappointing if compared to the 70% reported above, obtained by using the two robot platforms, it is still quite remarkable considering the very high degree of variability between training and test data, and that results are significantly above chance (which in this case would be 20%).

D. Discussion

The results of the extensive experimental evaluation presented in this section indicate that our method is able to perform place recognition using standard visual sensors with high precision. It offers good robustness to changes in the illumination conditions as well as to additional variations introduced by the natural variability that occurs in real-world environments. As the system is to be used on a robot platform, it must not only be accurate but also effective. For this reason we tried to provide the highest possible robustness using relatively small amount of training data acquired during only one run. We managed to achieve a recognition time of about 350ms per frame on a Pentium IV 2.6 GHz where the bulk of the time (300ms) is spent in a piece of code that has not yet been ported from MATLAB to C/C++.

Additional experiments indicate, that it is possible to improve the robustness by incorporating images acquired during two runs under different illumination conditions into one training set. In such case, however, the user pays the price of the recognition time and the memory requirements. For example, if the system was trained using the images captured during sunny weather and at night, the average classification rate for testing image sequence acquired with cloudy weather was equal to 90.5% for Dumbo and 88.4% for Minnie (see Figure 6). Consequently, the classification rate improved by 10% in case of Dumbo and 18% in case of Minnie for testing

conditions not known during training, while keeping the same rates for testing conditions used also for training. Since the number of support vectors in such case usually doubles, the recognition time increased by about 50ms.

VI. Summary and Conclusion

This paper presented a vision-only recognition algorithm for place classification under varying illumination conditions, across a significant span of time and with training and test performed on different acquisition devices. The method used rich global descriptors and support vector machines as discriminative classifier; this algorithm has proved successful in the object recognition domain. We tested our approach with a very extensive set of experiments, which showed that our method is able to perform place recognition with high precision, remarkable robustness and a recognition time per frame of 350 ms.

This work can be extended in many ways: firstly, we plan to incorporate invariance to illumination changes in the feature descriptors, to achieve a higher robustness. Secondly, we want to move from recognition of single images by fusing information over time. Finally, we want to extend the system to be able to perform room categorization. Future work will address these issues.

Acknowledgment

This work was partially sponsored by the SSF through its Centre for Autonomous Systems (CAS), the EU as part of the project CoSy IST-2004-004450. and the Swedish Research Council contract 2005-3600 - Complex. The support is gratefully acknowledged.

References

[1] I. Nourbakhsh, R. Powers, and S. Birchfield, "Dervish: An office navigation robot," *AI Magazine*, vol. 16, no. 2, pp. 53–60, 1995.

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

[3] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh, "On the significance of real-world conditions for material classification," in *Proc Eccv04*.

[4] O. Martínez Mozos, C. Stachniss, and W. Burgard, "Supervised learning of places from range data using adaboost," in *Proc. ICRA'05*.

[5] A. Torralba and P. Sinha, "Recognizing indoor scenes," AI Memo, Tech. Rep. 2001-015, 2001.

[6] H. Tamimi and A. Zell, "Vision based localization of mobile robots using kernel approaches," in *Proc IROS'04*.

[7] J. Gaspar, N. Winters, and J. Santos-Victor, "Vision-based navigation and environmental representations with an omni-directional camera," *IEEE Trans RA*, vol. 16, no. 6, 2000.

[8] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proc. ICRA'00*.

[9] P. Blaer and P. Allen, "Topological mobile robot localization using fast vision techniques," in *Proc ICRA'02*.

[10] E. Menegatti, M. Zoccarato, E. Pagello, and H. Ishiguro, "Image-based monte-carlo localisation with omnidirectional images," *Robotics and Autonomous Systems*, vol. 48, no. 1, 2004.

[11] H. Andreasson, A. Treptow, and T. Duckett, "Localization for mobile robots using panoramic vision, local features and particle filter," in *Proc ICRA'05*.

[12] M. Mata, J. M. Armingol, A. de la Escalera, and S. M. A., "Using learned visual landmarks for intelligent topological navigation of mobile robots," in *Proc ICRA'03*.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc ICCV'99*.

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

[15] A. Torralba, "Contextual priming for object detection," *IJCV*, vol. 53, no. 2, 2003.

[16] B. Schiele and J. Crowley, "Recognition without correspondence using multidimensional receptive field histograms," *IJCV*, vol. 36, no. 1, pp. 31–50, January 2000.

[17] M. Swain and D. Ballard, "Color indexing," *IJCV*, vol. 7, no. 1, pp. 11-32, 1991.

[18] L. ZelnikManor and M. Irani, "Event-based analysis of video." in *Proc. CVPR'01*.

[19] T. Lindeberg, *Scale-space theory in computer vision*. Kluwer, 1994.

[20] A. Pronobis, "Indoor place recognition using support vector machines," Master's thesis, NADA/CVAP, KTH, 2005, available at `http://www.nada.kth.se/~pronobis/`.

[21] N. Cristianini and J. S. Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[22] V. Vapnik, *Statistical learning theory*. New York: Wiley and Son, 1998.

[23] S. Belongie, C. Fowlkes, F. Chung, and J. Malik, "Spectral partitioning with indefinite kernels using the nyström extension," in *Proc Eccv02*.

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