Multi-cue Discriminative Place Recognition*

Li Xing and Andrzej Pronobis

Centre for Autonomous Systems, The Royal Institute of Technology
SE100-44 Stockholm, Sweden
{lixing,pronobis}@kth.se

Abstract. In this paper we report on our successful participation in the RobotVision challenge in the ImageCLEF 2009 campaign. We present a place recognition system that employs four different discriminative models trained on different global and local visual cues. In order to provide robust recognition, the outputs generated by the models are combined using a discriminative accumulation method. Moreover, the system is able to provide an indication of the confidence of its decision. We analyse the properties and performance of the system on the training and validation data and report the final score obtained on the test run which ranked first in the obligatory track of the RobotVision task.

1 Introduction

This paper presents the place recognition algorithm based on multiple visual cues that was applied to the RobotVision task of the ImageCLEF 2009 campaign. The task addressed the problem of visual indoor place recognition applied to robot topological localization. Participants were given training, validation and test sequences capturing the appearance of an office environment under various conditions [1]. The task was to build a system able to answer the question "where are you?" (I am in the kitchen, in the corridor, etc) when presented with a test sequence imaging rooms seen during training, or additional rooms that were not imaged in the training sequence. The results could be submitted for two separate tracks: (a) obligatory, in case of which each single image had to be classified independently; (b) optional, where the temporal continuity of the sequences could be exploited to improve the robustness of the system. For more information about the task and the dataset used for the challenge, we refer the reader to the RobotVision@ImageCLEF'09 overview paper [2].

The visual place recognition system presented in this paper obtained the highest score in the obligatory track and constituted a basis for our approach used in the optional track. The system relies on four discriminative models trained on different visual cues that capture both global and local appearance of a scene. In order to increase the robustness of the system, the cues are integrated efficiently using a high-level accumulation scheme that operates on the separate models

* This work was supported by the EU FP7 integrated project ICT-215181-CogX. The supportis gratefully acknowledged.

C. Peters et al. (Eds.): CLEF 2009 Workshop, Part II, LNCS 6242, pp. 315–323, 2010.

adapted to the properties of each cue. Additionally, in the optional track, we used a simple temporal accumulation technique which exploits the continuity of the image sequences to refine the results. Since the misclassifications were penalized in the competition, we experimented with an ignorance detection technique relying on the estimated confidence of the decision.

Visual place recognition is a vastly researched topic in the robotics and computer vision communities and several different approaches have been proposed to the problem considered in the competition. The main differences between the approaches relate to the way the scene is perceived and thus the visual cues extracted from the input images. There are two main groups of approaches using either global or local image features. Typically, SIFT [3] and SURF [4] are applied as local features, either using a matching strategy [5,6] or the bag-of-words approach [7,8]. Global features are also commonly used for place recognition and such representations as gist of a scene [9], CRFH [10], or PACT [11] were proposed. Recently, several authors observed that robustness and efficiency of the recognition system can be improved by combining information provided by both types of cues (global and local) [5,12]. Our approach belongs to this group and four different types of features previously used in the domain of place recognition have been used in the presented system.

The rest of the paper gives a description of the structure and components of our place recognition system (Section 2). Then, we describe the initial experiments performed on the training and validation data (Section 3). We explain the procedure applied for parameter selection and study the properties of the cue integration and confidence estimation algorithms. Finally, we present the results obtained on the test sequence and our ranking in the competition (Section 4). The paper concludes with a summary and possible avenues for future research.

2 The Visual Place Recognition System

This section describes our approach to visual 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 properties under various viewpoints, at a fixed time and illumination setting. During testing, the algorithm is presented with data samples acquired under different conditions and after some time. The goal is to recognize correctly each single data sample provided to the system. The rest of the section describes the structure and components of the system.

2.1 System Overview

The architecture of the system is illustrated in Fig. 1. We use four different cues extracted independently from the visual input. We see that there is a separate path for each cue. Every path consists of two main building blocks: a feature extractor and a classifier. Thus separate decisions can be obtained for every cue. The outputs encoding the confidence of single-cue classifiers are combined using a discriminative accumulation scheme.

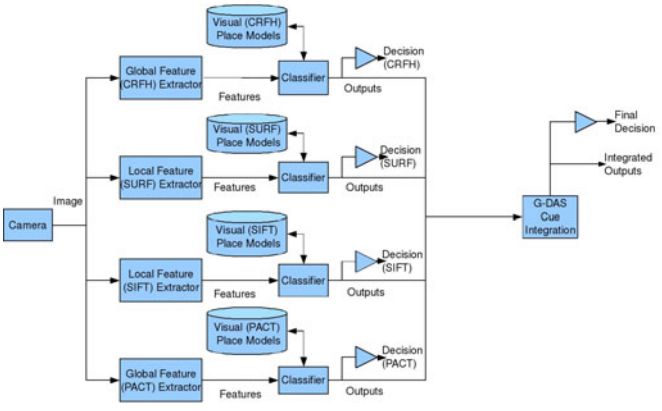


Fig. 1. Structure of the multi-cue visual place recognition system

2.2 Visual Features

The system relies on visual cues based on global and local image features. Global features are derived from the whole image and thus can capture general properties of the whole scene. In contrast, local features are computed locally, from distinct parts of an image. This makes them much more robust to occlusions and viewpoint variations. In order to capture different aspects of the environment, we combine cues produced by four different feature extractors.

Composed Receptive Field Histograms (CRFH). CRFH [13] is a multi-dimensional statistical representation (a histogram) of the occurrence of responses of several image descriptors applied to the whole image. Each dimension corresponds to one descriptor and the cells of the histogram count the pixels sharing similar responses of all descriptors. This approach allows to capture various properties of the image as well as relations that occur between them. On the basis of the evaluation in [10], we build the histograms from second order Gaussian derivative filters applied to the illumination channel at two scales.

PCA of Census Transform Histograms (PACT) Census Transform (CT) [11] is a non-parametric local transform designed for establishing correspondence between local patches. Census transform compares the intensity values of a pixel with its eight neighboring pixels, as illustrated in Figure 2. A histogram of the CT values encode both local and global information of the image. PACT [11] is a global representation that extracts the CT histograms for several image patches organized in a grid and applies Principal Component Analysis (PCA) to the resulting vector.

Scale Invariant Feature Transform (SIFT). As one of the local representations, we used a combination of the SIFT descriptor [3] and the scale, rotation and translation invariant Harris-Laplace corner detector [14]. The SIFT descriptor represents local image patches around interest points characterized by coordinates in the scale space in the form of histograms of gradient directions.

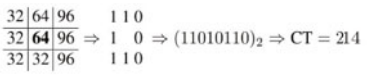


Fig. 2. Illustration of the Census Transform [11]

Speed-Up Robust Features (SURF). SURF [4] is a scale- and rotation-invariant local detector and descriptor which is designed to approximate the performance of previously proposed schemes while being much more computationally efficient. This is obtained by using integral images, a Hessian matrix-based measure for the detector and a distribution of Haar-wavelet responses for the descriptor.

2.3 Place Models

Based on its state-of-the-art performance in several visual recognition domains [15, 16], we used the Support Vector Machine classifier [17] to build the models of places for each cue. The choice of the kernel function is a key ingredient for the good performance of SVMs and we selected specialized kernels for each cue. Based on results reported in the literature, we chose in this paper the χ^2 kernel [18] for CRFH, the Gaussian (RBF) kernel [17] for PACT and the match kernel [19] for both local features. In order to extend the binary SVM to multiple classes, we used the one-against-all strategy for which one SVM is trained for each class separating the class from all other classes.

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 this work, we experimented with the distance-based methods proposed in [5], which define confidence as a measure of unambiguity of the final decision.

2.4 Cue Integration and Temporal Accumulation

As indicated in [5], different properties of viual cues result in different performance and error patterns on the place classification task. The role of the cue integration scheme is to exploit this fact in order to increase the overall performance. Our place recognition system uses the Discriminative Accumulation Scheme (DAS) [16] that was proposed for the place classification problem in [5]. It accumulates multiple cues, by turning classifiers into experts. The basic idea is to consider real-valued outputs of a multi-class discriminative classifier 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. In the presented system, this can be expressed by the equation $\boldsymbol{O}_\Sigma = a \cdot \boldsymbol{O}_{CRFH} + b \cdot \boldsymbol{O}_{PACT} + c \cdot \boldsymbol{O}_{SIFT} + d \cdot \boldsymbol{O}_{SURF}$, where a, b, c, d are the weights assigned to each cue and $a + b + c + d = 1$. The vectors $\boldsymbol{O}$ represent the outputs of the multi-class classifiers for each cue.

We used a very similar scheme to improve the robustness of the system operating on image sequences. For this, we exploited the continuity of the sequences

and accumulated the outputs (of a single cue or integrated cues) for the current sample and N previously classified samples. The result of accumulation was then used as the final decision of the system.

3 Experiments on the Training and Validation Data

We conducted several series of experiments on the training and validation data in order to analyze the behavior of our system and select parameters. We present the analysis and results in successive subsections.

3.1 Selection of the Model Parameters

The first set of experiments was aimed at finding the values of parameters of the place models, i.e. the SVM error penalty C and the kernel parameters. The experiments were performed separately for each visual cue (CRFH, PACT, SIFT and SURF). To find the parameters, we performed cross validation on the training and validation data. For every training set, we selected parameters that resulted in highest classification rate on all available test sets acquired under different conditions. The classification rate was calculated in a similar way as the final score used in the competition i.e. as the percentage of correctly classified images in the whole testing sequence.

Figure 3 presents the results obtained for the experiments with the *dum-night3* training set which was selected for the final run of the competition. It is apparent that the model based on the SIFT features provides the highest recognition rate on average. However, we can also see that different cues have different characteristics as their performance changes according to different patterns. This suggests that the overall performance of the system could be increased by integrating the outputs of the models.

3.2 Cue Integration and Temporal Accumulation

The next step was to integrate the outputs of the models and choose the proper values of the DAS weights for each model. We performed an exhaustive search for

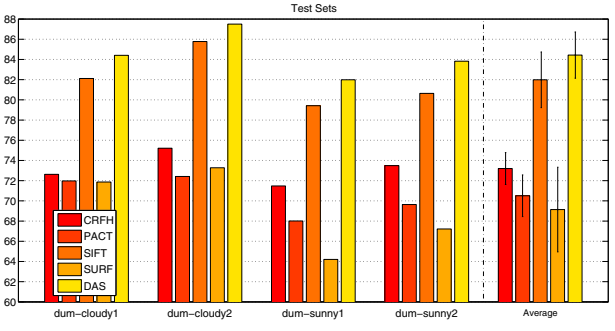


Fig. 3. Classification rates for the best model parameters and the *dum-night3* training set. Results are given separately for each test set as well as averaged over all sets.

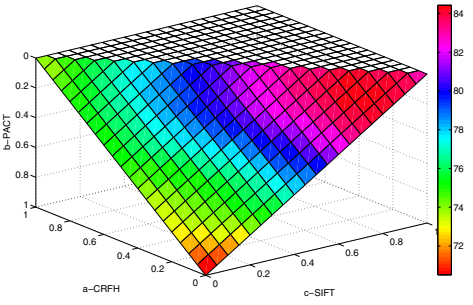


Fig. 4. Classification rates obtained for various values of the DAS weights

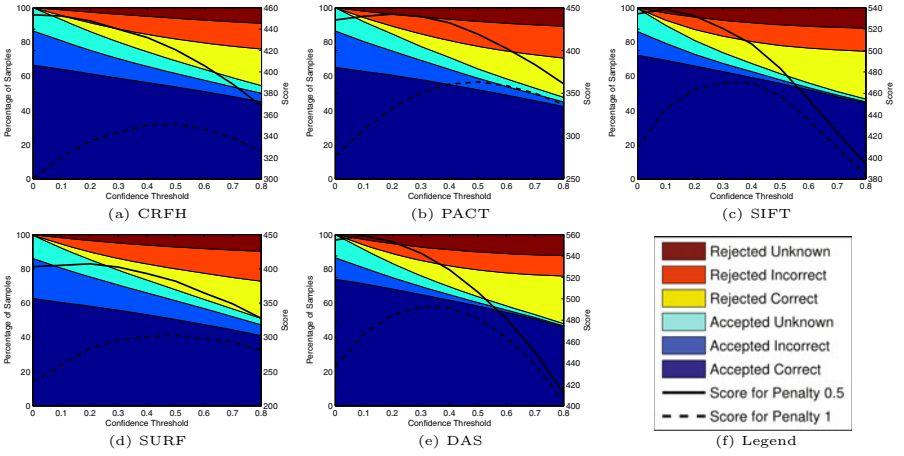
the weights on the training and validation data independently for each training set. Then, we selected the values that provided the highest average classification rate over all test sets. The results are presented in Figure 3.

This weight selection procedure revealed that once SIFT is used as one of the cues, there is no benefit of adding SURF (the weight for SURF was selected to be 0). This is not surprising since SURF captures similar information as SIFT, while employing some heuristics in order to make the feature extraction process more efficient. According to the results presented in the previous section, those heuristics decrease the overall performance of the system, while not introducing any additional knowledge. Figure 4 illustrates how the the average classification rates for the *dum-night3* training set and all test sets changed for various values of the weights used for CRFH, PACT and SIFT (the weight used for SURF is assumed to be 0). The following weights were selected and used for further experiments: $a = 0.1$, $b = 0.15$, $c = 0.75$, $d = 0$.

We performed similar experiments to find the number of past samples we should accumulate over in order to refine the results in case of the optional track. The results revealed that we obtain the highest score when 4 past test samples are accumulated with equal weights with the currently classified sample.

3.3 Confidence Estimation

According to the performance measure used in the competition, the classification errors were penalized. Therefore, we experimented with an ignorance detection mechanism based on the confidence of the decision produced by the system. In order to simulate the case of unknown rooms in the test set, we always removed one room from the training set. Figure 5a-e presents the obtained average results. We gradually increased the value of confidence that is required in order to accept the decision of the system and measured the statistics of the accepted and rejected decisions. In both cases, we measured the percentage of test samples that were classified correctly, misclassified or unknown during training. We can see from the plots that the confidence thresholding procedure rejects mostly samples from unknown rooms and samples that would be incorrectly classified. This increases the classification rate for the accepted samples. At the same time, the



(a) CRFH (b) PACT (c) SIFT (d) SURF (e) DAS (f) Legend

Fig. 5. Average results of the experiments with confidence-based ignorance detection for separate cues and cues integrated

Table 1. Results and scores obtained on the final test set

	Obligatory Track	Optional Track
Score	793.0	853.0
Rank	**1**	4

(a) Scores and ranks

Predicted → True↓	1-person Office	Corridor	2-person Office	Kitchen	Printer Area	Unkn. Room
1-person Office	119 (129)	25 (23)	12 (8)	4 (0)	0 (0)	0 (0)
Corridor	4 (2)	570 (580)	6 (3)	10 (6)	1 (0)	0 (0)
2-person Office	1 (0)	4 (0)	131 (134)	25 (27)	0 (0)	0 (0)
Kitchen	1 (0)	5 (0)	2 (0)	152 (161)	1 (0)	0 (0)
Printer Area	5 (0)	138 (139)	10 (7)	3 (2)	120 (128)	0 (0)
Unkn. Room	13 (14)	206 (206)	22 (24)	11 (6)	89 (91)	0 (0)

(b) Confusion matrix. Values in brackets are for the optional track.

plots show the score used for the competition calculated for the accepted samples only. If we use the penalty equal to 0.5 points for each misclassified sample (as used in the competition), the number of rejected errors must be twice as large as the number of rejected samples that would be classified correctly. As a result, the ignorance detection scheme provided only a slight improvement of the final score and we decided not to use confidence thresholding for the final run. However, as shown in Figure 5, if the penalty was increased to 1 point, the improvement would be significant.

4 The Final Test Run

The test sequence and the ID of the training sequence (*dum-cloudy3*) were released in the final round of the competition. For the final run, we used the parameters identified on the training and validation data. In order to obtain the results for the obligatory track, we applied the models independently to each image in the test sequence and integrated the results using the selected weights. We did not perform ignorance detection. In order to obtain the results for the optional task, we applied the temporal averaging to the results submitted to

the obligatory track. Table 1a presents our scores and ranks in both tracks. Table 1b shows the confusion matrix for the test set. We can see that the temporal averaging filtered out many single misclassifications in the test sequence.

5 Conclusions

In this paper we presented our place recognition system applied to the RobotVision task of the ImageCLEF'09 campaign. Through the use of multiple visual cues integrated using a high-level discriminative accumulation scheme, we obtained a system that provided robust recognition despite different types of variations introduced by changing illumination and long-term human activity.

The most difficult aspect of the task turned out to be the novel class detection. We showed that the confidence of the classifier can be used to reject unknonwn or misclassified samples. However, we did not provide any principled way to detect the cases when the classifier dealt with a novel room. Our future work will concentrate on that issue.

References

1. Luo, J., Pronobis, A., Caputo, B., Jensfelt, P.: Incremental learning for place recognition in dynamic environments. In: Proc. of IROS 2007 (2007)
2. Caputo, B., Pronobis, A., Jensfelt, P.: Overview of the CLEF 2009 Robot Vision task (2009)
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2) (2004)
4. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
5. Pronobis, A., Caputo, B.: Confidence-based cue integration for visual place recognition. In: Proc. of IROS 2007 (2007)
6. Valgren, C., Lilienthal, A.J.: Incremental spectral clustering and seasons: Appearance-based localization in outdoor env. In: Proc. of ICRA 2008 (2008)
7. Filliat, D.: A visual bag of words method for interactive qualitative localization and mapping. In: Proc. of ICRA 2007 (2007)
8. Cummins, M., Newman, P.: FAB-MAP: Probabilistic localization and mapping in the space of appearance. International Journal of Robotics Research 27(6) (2008)
9. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: Proc. of ICCV 2003 (2003)
10. Pronobis, A., Caputo, B., Jensfelt, P., Christensen, H.I.: A discriminative approach to robust visual place recognition. In: Proc. of IROS 2006 (2006)
11. Wu, J., Rehg, J.M.: Where am I: Place instance and category recognition using spatial PACT. In: Proc. of CVPR 2008 (2008)
12. Weiss, C., Tamimi, H., Masselli, A., Zell, A.: A hybrid approach for vision-based outdoor robot localization using global and local image features. In: Proc. of IROS 2007 (2007)
13. Linde, O., Lindeberg, T.: Object recognition using composed receptive field histograms of higher dimensionality. In: Proc. of ICPR 2004 (2004)

14. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proc. of ICCV 2001 (2001)
15. Pronobis, A., Martínez Mozos, O., Caputo, B.: SVM-based discriminative accumulation scheme for place recognition. In: Proc. of ICRA 2008 (2008)
16. Nilsback, M.E., Caputo, B.: Cue integration through discriminative accumulation. In: Proc. of CVPR 2004 (2004)
17. Cristianini, N., Taylor, J.S.: An Introduction to SVMs and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)
18. Chapelle, O., Haffner, P., Vapnik, V.: Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks 10(5) (1999)
19. Wallraven, C., Caputo, B., Graf, A.: Recognition with local features: the kernel recipe. In: Proc. of ICCV 2003 (2003)