Visual place recognition using bayesian filtering with markov chains

Mathieu Dubois, Hervé Guillaume, Frenoux Emmanuelle, Philippe Tarroux

Visual place recognition using Bayesian Filtering with Markov Chains*

Mathieu Dubois[1,2], Hervé Guillaume[1,2], Emmanuelle Frenoux[2,1] and Philippe Tarroux[2,3]

1- Univ Paris-Sud - Dept of Computer Science
Orsay, F-91405 - France

2- LIMSI - CNRS
B.P. 133, F-91403 - France

3- Ecole Normale Supérieure
45 rue d'Ulm, Paris, F-75230 - France

Abstract. We present a novel idea to use Bayesian filtering in the case of place recognition. More precisely, our system combines global image characterization, Learned Vector Quantization, Markov chains and Bayesian filtering. The goal is to integrate several images seen by a robot during exploration of the environment and the dependency between them. We present our system and the new Bayesian filtering algorithm. Our system has been evaluated on a standard database and shows promising results.

1 Introduction

Traditional approaches to the problem of robot localization have focused on metric localization (*i.e.*the ability to position objects in a common coordinate frame) or topological localization (*i.e.*the ability to build a graph of interesting places). However in those approaches, a place does not necessarily coincide with the human concept of rooms or regions. Place recognition or semantic localization is the ability for a mobile robot to recognize the place it is currently in. Place categorization is the ability to determine the nature of the environment (kitchen, room, corridor, etc.). The semantic category of a place can be used as a contextual information. Such context is a rich source of information to foster object detection and recognition (giving priors on objects identity, location and scale). This research focuses on place recognition without object recognition. Another interesting application could be to solve the kidnapped robot problem.

Several solutions have been proposed for place recognition and place categorization using vision alone or combined with several types of telemeters (laser, sonar). Vision provides richer information than telemeters which is an essential advantage for fine discrimination. That's why we are interested in *visual* place recognition and categorization. This differs from visual scene classification because a single frame may be ambiguous, non-informative or even misleading for place recognition. For instance, think about a robot in an office but facing a window: the scene will contain misleading outside objects (trees, cars, etc). Scene classification considers what the robot is seeing while place recognition allows it

*This version contains some corrections.

to deduce the place it is in from the image. The task is difficult because, due to perceptual aliasing, there is a huge semantic gap between the human notion of a place and the data that can be extracted from an image.

2 Related work

The vast majority of research on place recognition use techniques developed for visual scene classification. We can distinguish methods using global images features (see [1, 2]) and methods using local descriptors computed around interest points (see [3, 4, 5]). Colors and orientations are the most used features. As we said earlier, an image or a feature may be misleading. It is therefore necessary to use methods to disambiguate perception.

In [6] several cues from the same image are combined iteratively until a sufficient confidence rate is reached (or no more cues are available). Note that recognition is carried out using only one image.

While an image may be ambiguous, a collection of images is generally not. Therefore it makes sense to use several images to mutually disambiguate perception. In [7], the authors use a simple spatio-temporal accumulation process to filter the decision of a discriminative confidence-based place recognition system (which uses only one image to recognize the place). One problem with this method is that the system needs to wait some time before giving a response. Also, special care must be taken to detect places boundaries and to adjust the size of the bin. In [1] the authors use a Hidden Markov Model (HMM) where each place is a hidden state of the HMM and the feature vector stands for the observation. Recognition is conducted with standard Bayesian techniques. The drawback is that the input space is continuous and high-dimensional. The learning procedure is then computationally expensive.

Bag-of-Words (BoW) has been developed to reduce the complexity of learning in scene classification (see [8]). An image is represented by a histogram of the visual words found in it. The words are extracted from a dictionary of features learned by mean of a Learned Vector Quantization (LVQ) algorithm. The major advantage is that the learning space is discretized. BoW has been used in place recognition (see [4]).

In [9] the authors propose to use a BoW model in conjunction with Bayesian filtering. It allows to efficiently integrate several images but relies on the assumption that the images are independent. [10] propose the temporal Bag-of-Words (tBoW) model, a simple alternative which allows to use BoW with a global signature. A place is represented by a histogram of prototypes over time (using several images). One of the main drawback of this method is the necessity to use a mechanism to inform the system when it moves from one place to another. In this paper we push the idea of using several images a step further.

We first present our system which can be seen as an extension of the tBoW method (section 3). We subsequently address the reset problem by adapting the Bayesian filtering algorithm (section 4).

3 Image signature and vector quantization

At each time step t, the input image I_t is presented to the system and its signature is computed. Thanks to LVQ, the signature can be mapped to a prototype. One image is then represented by an integer $z_t \in \{1, 2, \ldots, S\}$ which identifies the prototype (S is the number of prototypes). The set of prototypes is the vocabulary that will be used to learn places. Training is supervised so each image is labeled with the name of the place, $x_t \in \mathcal{C}$, the robot is currently in. $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ denotes the set of places in the environment.

To characterize the images we use two recent descriptors that have been developed in the context of place recognition or categorization. Those descriptors are global *i.e.*they use all the pixels to compute the signature (no interest points are extracted). Thus we say that they capture the visual context of the image. GIST (see [1]) is based on the Principal Component Analysis (PCA) of the output of a Gabor-like filter bank and thus captures the most significant spatial structure in the image (we use 4 scales and 6 orientations and project the 1152-dimensional output on the first 80 principal components which explains more than 99% of the variance). CENsus TRansform hISTogram (CENTRIST) (see [9]) is based on the Census Transform (CT) of the edges. The CT captures the local intensity pattern and is robust to illumination and gamma changes. Note that unlike [9] we do not divide the image in sub-windows. Instead, we use only one histogram for all the pixels in the image.

The LVQ algorithm chosen in this paper is the Self-Organizing Map (SOM). In the current set-up the training of the SOM is performed off-line. It has been shown that this process gives clusters of similar images [11]. The set of all the prototypes forms the dictionary of visual scenes. In this paper we use SOM of size 20×20 giving $S = 400$. For more details see [10].

4 Bayesian filtering with Markov chains

The classical Bayesian filtering model assumes that the hidden state x_t is complete, meaning that the observation z_t is conditionally independent on previous measurements $z_{1:t-1}$: $\mathrm{P}(z_t|z_{1:t-1}, x_t) = \mathrm{P}(z_t|x_t)$. In our case the hidden state is a place. As said earlier there is a huge semantic gap between the human notion of a place and the data that can be extracted from an image. Therefore it makes sense to suppose that the hidden state is not complete and then introduce dependence between observations.

As we use quantized global image characterization, the flow of images seen by the robot during exploration of a place will be coded by a sequence of words. As the sensory-motor capacities of a robot are limited, the order of exploration of a place is constrained. Therefore the words won't be independent. We propose to use the transition between words (and not only the frequency of each word) as a place model. Such transitions depend on the sensory-motor capacities of the robot and then implicitly encode the spatial configuration of a place.

4.1 Modeling transition between words

For a place c_i, the transition between prototypes is modeled as a Markov Chain *i.e.*the probability of a prototype at time t given the entire observation history $z_{1:t-1}$, depends only on the previous prototype: $\mathrm{P}\left(z_t|z_{1:t-1}, x_t = c_i\right) = \mathrm{P}\left(z_t|z_{t-1}, x_t = c_i\right)$. As the vocabulary is discrete and finite we can use a non-parametric approach. A place is represented by a transition matrix $\mathbf{A}$ where a_{ij} is the probability of switching from prototype i to prototype j, $\mathrm{P}\left(z_t = j|z_{t-1} = i\right)$. To avoid null values when counting transitions we use the Laplace estimator.

By contrast, the classical Bayesian filtering framework with a BoW model assumes that the observations are independent and train a Naive Bayes Classifier. While the BoW uses an unordered set of words our model can be said to model phrases where each word is the prototype describing the current image. In this sense we model paths in the place rather than the place itself. Note that this Markov chain can be used as a classifier.

4.2 Adapting the Bayesian filtering algorithm

The purpose of Bayesian filtering is to estimate the posterior $\mathrm{bel}(x_t) = \mathrm{P}\left(x_t|z_{1:t}\right)$. We assume a Markovian property between the places: $\mathrm{P}\left(x_t|x_{1:t-1}\right) = \mathrm{P}\left(x_t|x_{t-1}\right)$ (this shall not be confused with the Markovian property on the observable state). Therefore we have:

$$\begin{aligned} \mathrm{bel}\left(x_t\right) &\propto \mathrm{P}\left(z_t|z_{1:t-1}, x_t\right)\mathrm{P}\left(x_t|z_{1:t-1}\right) && (1)\\ &\propto \mathrm{P}\left(z_t|z_{t-1}, x_t\right)\sum_{c_i}\mathrm{P}\left(x_t|x_{t-1} = c_i\right)\mathrm{bel}\left(x_{t-1} = c_i\right) && (2)\end{aligned}$$

where the observation probability $\mathrm{P}\left(z_t|z_{t-1}, x_t\right)$ is given by the Markov chain. Note that unlike the classical approach, the observation at time t, z_t, does not depend solely on the hidden state, x_t, but also on the previous observation z_{t-1}. Our method is called Bayesian Filtering with Markov Chains (BFMC). As in the classical approach the belief can be recursively computed so the computational cost is very low.

To use BFMC we have to specify:

1. The prior place distribution $\mathrm{P}\left(x_0\right)$. In this paper we always used a uniform distribution (a robotic system can start exploring the environment from any place).
2. The place transition distribution $\mathrm{P}\left(x_t|x_{t-1}\right)$. Following [9] we define the transition matrix as $\mathrm{P}\left(x_t|x_{t-1}\right) = p_e$ if $x_t = x_{t-1}$; the rest of the probability mass is shared uniformly among all other transitions.

5 Experimental results

5.1 Database and protocol

It is hard to compare previous place recognition methods because they were tested on different databases. The COsy Localization Database [3] has been pro-

posed as a standard database to evaluate vision-based place recognition systems. It consists of sequences acquired by a human-driven robot in different laboratories across Europe under different illumination conditions (night, cloudy, sunny) and several times. In each laboratory part, two paths were explored (standard and extended). In our experiments we discarded sequences known to contain errors (missing places). All the experiments were carried out with the perspective images. For fair comparison we reproduce the place recognition experiment done in [3]. The experiment addresses the problem of recognizing a place seen during training but under different imaging conditions: later in the day and/or under different illumination. Training is performed on a sequence taken from one laboratory, part, path and illumination condition. Testing is performed against a sequence taken from the same laboratory, part and path. Results are averaged on all permutations of training and testing sequences (ensuring that they are different).

5.2 Results

In a first series of experiments we systematically varied the parameter p_e in the interval $[0.90, 0.99]$ and assessed the system with the average classification rate. Results show that performance increases almost linearly with p_e for CENTRIST (going from 57.67% to 66.53%) and to a lower extent for GIST (going from 72.76% to 74.98%). Other experiments are performed with $p_e = 0.99$ (as in [9]).

The results show that GIST outperforms CENTRIST especially in the Ljubljana laboratory. Both descriptors show good robustness to illumination with an advantage for CENTRIST. Those results are coherent with the ones reported in [10].

	CENTRIST	GIST	CENTRIST+BFMC	GIST+BFMC
Saar. - Std	$36.39\%^{\pm 1.90\%}$	$67.50\%^{\pm 4.97\%}$	$85.38\%^{\pm 3.51\%}$	$86.73\%^{\pm 3.78\%}$
Saar. - Ext	$34.36\%^{\pm 3.43\%}$	$55.17\%^{\pm 5.17\%}$	$75.27\%^{\pm 9.90\%}$	$75.28\%^{\pm 5.52\%}$
Frei. - Std	$36.27\%^{\pm 2.23\%}$	$66.82\%^{\pm 4.38\%}$	$76.56\%^{\pm 5.43\%}$	$83.87\%^{\pm 3.61\%}$
Frei. - Ext	$33.51\%^{\pm 1.52\%}$	$50.69\%^{\pm 3.81\%}$	$68.55\%^{\pm 2.78\%}$	$74.65\%^{\pm 3.19\%}$
Ljub. - Std	$17.31\%^{\pm 2.34\%}$	$71.76\%^{\pm 5.76\%}$	$60.75\%^{\pm 8.25\%}$	$89.68\%^{\pm 1.78\%}$
Ljub. - Ext	$13.78\%^{\pm 2.04\%}$	$67.08\%^{\pm 4.30\%}$	$54.10\%^{\pm 4.26\%}$	$82.15\%^{\pm 3.48\%}$
Average	$28.60\%^{\pm 2.24\%}$	$64.60\%^{\pm 4.73\%}$	$70.10\%^{\pm 4.69\%}$	$82.06\%^{\pm 3.56\%}$

Table 1: Average recognition rate in congruent conditions without (left) and with (right) Bayesian filtering. The error is given as one standard deviation.

Table 1 shows the results with and without BFMC when training and testing illumination are the same. When using BFMC, performance increases by 41.50% for CENTRIST and by 17.46% for GIST. [9] reports performance increase of 7% with classical Bayesian filtering but on a harder categorization task (and a different database). The overall performance is slightly lower than in [3]. However our system is more suited to a robot because it is simple, incremental and can be learned on-line (once the training of the SOM is done).

6 Conclusion

This article presents the idea of using transitions between words and global signature for place recognition together with Bayesian filtering. This new model uses an explicit dependence between observations to incorporate image history. This model is closely related to the n-grams models used in natural language processing. Experiments show encouraging results. Future work include categorization and the use of n-grams for modeling transitions between words. We think that this idea could be promising if combined with an attention-based navigation system (as opposed to a human-driven robot) because attention should induce correlations across observations.

References

[1] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the Nineth IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*, pages 273–280. IEEE Computer Society, October 2003.

[2] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen. A discriminative approach to robust visual place recognition. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, pages 3829–3836, Beijing, China, 2006.

[3] M. M. Ullah, A. Pronobis, B. Caputo, J. Luo, P. Jensfelt, and H. I. Christensen. Towards robust place recognition for robot localization. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2008)*, Pasadena, USA, 2008.

[4] D. Filliat. Interactive learning of visual topological navigation. In *Proceedings of the 2008 IEEE International Conference on Intelligent Robots and Systems (IROS 2008)*, 2008.

[5] K. Ni, A. Kannan, A. Criminisi, and J. Winn. Epitomic location recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2158–2167, 2009.

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

[7] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensefllt. Multi-modal semantic place classification. *The International Journal of Robotics Research*, 29(2-3):298–320, 2010.

[8] D. Gokalp and S. Aksoy. Scene classification using bag-of-regions representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, pages 1–8, Minneapolis, USA, 2007.

[9] J. Wu, H.I. Christensen, and J.M. Rehg. Visual place categorization: Problem, dataset, and algorithm. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009 (IROS 2009)*, pages 4763–4770, St. Louis, USA, 2009. IEEE.

[10] H. Guillaume, M. Dubois, P. Tarroux, and E. Frenoux. Temporal Bag-of-Words: A Generative Model for Visual Place Recognition using Temporal Integration. In L. Mestetskiy and J. Braz, editors, *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP 2011)*. INSTICC, SciTePress, 2011.

[11] H. Guillaume, N. Denquive, and P. Tarroux. Contextual priming for artificial visual perception. In *European Symposium on Artificial Neural Networks (ESANN 2005)*, pages 545–550, Bruges, Belgium, 2005.