

Selecting local region descriptors with a genetic algorithm for real-world place recognition

Leonardo Trujillo ^a, Gustavo Olague ^a,
Francisco Fernández de Vega ^b and Evelyne Lutton ^c

^a EvoVisión Project, CICESE Research Center, Ensenada, B.C. México.

^b Grupo de Evolución Artificial, Universidad de Extremadura, Mérida, Spain.

^c APIS Team, INRIA-Futurs, Parc Orsay Universit 4, ORSAY Cedex, France.

trujillo@cicese.mx, olague@cicese.mx, fcofdez@unex.es, evelyne.lutton@inria.fr

Abstract. The basic problem for a mobile vision system is determining where it is located within the world. In this paper, a recognition system is presented that is capable of identifying known places such as rooms and corridors. The system relies on a bag of features approach using locally prominent image regions. Real-world locations are modeled using a mixture of Gaussians representation, thus allowing for a multimodal scene characterization. Local regions are represented by a set of 108 statistical descriptors computed from different modes of information. Therefore, the system needs to determine which subset of descriptors captures regularities between image regions of the same location, and also discriminates between regions of different places. A genetic algorithm is used to solve this selection task, using a fitness measure that promotes: 1) a high classification accuracy; 2) the selection of a minimal subset of descriptors; and 3) a high separation among place models. The approach is tested on two real world examples: a) using a sequence of still images with 4 different locations; and b) a sequence that contains 8 different locations. Results confirm the ability of the system to identify previously seen places in a real-world setting.

1 Introduction

Building an artificial system that is capable of answering the question “*Where am I?*” is one of the central problems studied in computer vision research. This task has only been partially solved in constrained real-world situations. To solve an instance of this problem, and many vision problems in general, three design issues must be accounted for [1]: 1) What information should be extracted from the output of visual sensors? 2) How is the information extracted? 3) How should the information be represented? 4) How will it be used to solve higher-level tasks?

This contribution introduces a system that performs place recognition using only local image information and probabilistic models for each location. The design questions stated above are addressed in the following manner. Questions 1 and 4 are answered using common computer vision techniques that are applicable to different types of problems. The extracted information are local image regions, what corresponds to a *bags of features* approach [2–4]. In this way, the system can be robust to occlusions and avoids the need for prior segmentation. The information gathered from the images

is used to create a probabilistic mixture model for each of the known locations. On the other hand, questions 2 and 3 were answered using design techniques based on evolutionary computation. More precisely, information is extracted using GP-evolved region detectors [5–7]. In addition, local image regions are represented by a subset of statistical descriptors that are selected by a genetic algorithm (GA) [4]. The proposal is related to vision based recognition, where a model correspondence is sought. This contrasts with image indexing approaches where specific image instances are retrieved or used for comparison.

This paper is organized as follows. Section 2 reviews related research and gives a working definition for the problem of place recognition. Section 3 presents a detailed description of the proposed method. Section 4 describes the experimental setup, the test sets used and the obtained results. Finally, concluding remarks and future perspectives are given in Section 5.

2 Related work and problem definition

This section presents some examples of place recognition systems. Then, a more precise problem statement is given that illustrates the difficulty of the problem.

Related work. Due to paper size considerations the review of past works is not exhaustive, however it does give a general overview of the type of approaches that are currently being used to solve the problem of place recognition. For instance, Torralba *et al.* [8] present a combined place and object recognition system that uses context dependent information. The system employs global features for place recognition, and object detection is contingent on the location that has been identified. The system also relies on the spatial relations between different locations, basically employing a topological map of the environment represented by a hidden Markov model. On the other hand, the present work is concerned with place recognition that only employs visual information without a Markovian assumption regarding the temporal and spatial relationships between frames of an image sequence. Therefore, the proposed method is closely related with the problem of object class recognition. Furthermore, instead of using a holistic image description, the current work relies on a sparse and local description of each image [2–4]. Another relevant example is the work by Wang *et al.* [9], where vision-based localization is performed for a mobile robotic system. In that work, the authors utilize scale invariant features detected using the Harris-Laplace detector and characterize each local region using the popular SIFT descriptor. Their approach is to utilize an image indexing technique, which contrasts with the recognition based approach presented here. Moreover, the use of SIFT limits the type of information that the system can employ for recognition purposes, more notably it excludes any type of color information. Both aforementioned examples show how different types of modelling are possible. Therefore, a more concrete definition for the problem of place recognition is necessary to clearly express the goal behind the current work.

Problem Statement. The place recognition problem can be defined as follows. Given a set of l physical real-world locations L , and a set of n representative images for each location, train a system capable of recognizing which location is being viewed in each frame of a sequence of test images that are different from the images used for



Fig. 1. The problem of place recognition. 1) First row: different views of the same location (Research Lab); 2) Second row: Images from four different locations (Students Lab, Computer Lab, Research Lab, and Office).

training. The only constraint is that the testing sequence only contains images from the l locations learned during training. This constraint could be easily relaxed by adding an *unknown* class to the list of locations.

In order to grasp the difficulty of the problem some visual examples are shown in Figure 1. The first row contains images from different views of the same location. A high degree of variability exists within this location, what represents a single class. The second row, contains images from four different locations; nevertheless, it is possible to observe that many features are shared amongst them. Indeed, finding the set of features that can best discriminate between classes and at the same time capture properties that are special to each class, represents a complex search problem. This work addresses these issues using an evolutionary search and a probabilistic modelling.

3 Outline of the recognition system

With a clear understanding of the place recognition task, it is now possible to introduce each aspect of the current proposal. In accordance with the introductory discussion, this Section starts by giving a description of each of the main design choices. Then, the GA learning loop is described, and finally the proposed recognition criteria are given.

Extracting local image information. As stated above, the system employs sparsely distributed local image regions, also know as a “bags of features” approach. Salient image regions are extracted based on their distinctive properties that make them unique when compared with neighboring regions. Employing this approach has two principal advantages. First, relevant image regions are extracted without the need for prior segmentation thereby eliminating what is considered to be a very difficult task. Second, the extraction of these type of regions is robust to partial occlusions or scene variations.

In order to extract locally prominent regions a scale adapted interest region detector is employed [7]. Selecting a characteristic scale for local image features is a process in which local extrema of an operator’s response, embedded into a linear scale-space, are found over different scales [10]. The interest operator employed in the current work was synthesized with Genetic Programming and is optimized for high repeatability and global region separability [5, 6]. The operator is named K_{IPGP1^*} and is given by

$$K_{IPGP1^*}(\mathbf{x}; t_j) = G_{t_j} * |G_{t_j} * I(\mathbf{x}) - I(\mathbf{x})|, \quad (1)$$



Fig. 2. Scale invariant regions detected on three test images. Top Row: original images; Bottom Row: detected regions. Columns contain a test image from one of the locations in Experiment 2 of this paper, from left to right: Students lab; Computer Lab; Research Lab.

Features	Description
Gradient information	<i>Gradient</i> , <i>Gradient magnitude</i> and <i>Gradient Orientation</i> (∇ , $\ \nabla\ $, ∇_ϕ).
Gabor filter response	The sum of <i>Gabor filters</i> with 8 different orientations (<i>gab</i>).
Interest operators †	The response to 3 stable interest operators: <i>Harris</i> , <i>IPGP1</i> and <i>IPGP2</i> (K_{Harris} , K_{IPGP1} , K_{IPGP2}).
Color information	All the channels of 4 color spaces: <i>RGB</i> , <i>YIQ</i> , <i>Cie Lab</i> , and <i>rg chromaticity</i> (R , G , B , Y , I , Q , L , a , b , r , g).

† K_{IPGP1} is proportional to a DoG filter, and K_{IPGP2} is based on the determinant of the Hessian [5, 7].

Table 1. The complete feature space Φ .

with $j = 0, 1, \dots, k$, and k is the number of scales to be analyzed, here it is set to $k = 5$. The size of a detected region is proportional to the scale at which it obtained its extrema value. For the sake of uniformity, all regions are scaled to a size of 41×41 pixels using bi-cubic interpolation before region descriptors are computed, as in [11]. Figure 2 shows sample interest regions extracted with the K_{IPGP1} * operator. It is important to note, however, that the recognition system does not depend on any particular region detector. To exhibit this, the first experimental setup employs the Kadir and Brady detector that relies on a local entropy measure of intensity values [12].

Feature Space. Local image regions are more discriminantly characterized using different types of numerical descriptors. In the current work, the space of all possible descriptors Φ includes 18 different modes of color and texture related information, see Table 1. To characterize the information contained along different information channels, six statistical descriptors are computed: *mean* μ , *standard deviation* σ , *skewness* γ_1 , *kurtosis* γ_2 , *entropy* H and *log energy* E . This yields a total of 108 possible descriptor values that can be used to model regions from each location.

Place models. It is necessary to model how regions are mapped to descriptor space. It is expected that regions extracted from images of the same location, each offering a different view, will create distinctive clusters in Φ . Hence, when a test image is ob-

tained and local regions are extracted, to determine which location those regions correspond with it is only necessary to map those regions to descriptor space and compute their class membership. This is a classification problem, and a Gaussian mixture model (GMM) is chosen to solve this task. GMMs are able to represent multimodal data, a property that can be expected from different image regions taken from a real-world location, see Figure 1. Formally, a GMM pdf is defined as,

$$p(\mathbf{x}; \Theta) = \sum_{c=1}^C \alpha_c \mathcal{N}(\mathbf{x}; \mu_c, \Sigma_c), \quad (2)$$

where $\mathcal{N}(\mathbf{x}; \mu_c, \Sigma_c)$ is the c th multivariate Gaussian component with mean μ_c , covariance matrix Σ_c , and an associated weight α_c . Estimation of the mixture model parameters is done using the EM algorithm when a fixed number of components is assumed. Alternatively, if a variable number of component is desired, with a maximum bound, it is possible to use Figueiredo-Jain (FJ) algorithm [13]. Classification with GMMs can be easily done employing Bayes rule. Additionally, it is possible to estimate the amount of separation between two class models using a closed form solution.

Fisher’s Linear Discriminant. Fisher defined the separation between two distributions \mathcal{N}_i and \mathcal{N}_j as the following ratio

$$S_{i,j} = \frac{(\mathbf{w}(\mu_i - \mu_j))^2}{(\mathbf{w}^T(\Sigma_i + \Sigma_j)(\mathbf{w}))}, \quad (3)$$

where $\mathbf{w} = (\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j)$ [14]. Note that S is defined for unimodal pdfs, hence a weighted version \hat{S} that accounts for the weight α_i and α_j of the associated Gaussian components in a GMM is proposed, such that

$$\hat{S}_{i,j} = \frac{S_{i,j}}{1 + \alpha_i + \alpha_j}. \quad (4)$$

Hence, the separation between components with a small combined weight (they have less influence over their associated models) will *appear* to be larger with respect to the separation between components with larger weights. Therefore, let C_a and C_b represent the number of components of $p_a(\mathbf{x}; \Theta_a)$ and $p_b(\mathbf{x}; \Theta_b)$ respectively, then $S^{a,b}$ represents the *apparent* separation matrix of size $C_a \times C_b$ that contains the weighted separation $\hat{S}_{i,j}$ of every component of p_a with respect to every component of p_b . The final *apparent* separation measure \mathcal{S} between p_a and p_b is defined as

$$\mathcal{S}^{a,b} = \inf(S^{a,b}). \quad (5)$$

3.1 The GA training algorithm

The system recognizes a total of l different locations. For each location L , n different images are taken as *representative* views, these images are used for training. For every representative image, scale invariant regions are extracted, and for all such regions the complete 108 descriptor values are computed off-line. Then, the GA performs feature

selection and learns appropriate GMMs, one for each location, in a single run. The GA employs fitness proportional selection, mask crossover, single bit mutation and an elitist survival strategy. Solution representation and fitness evaluation are described next.

Solution Representation. Each individual in the population is coded as a binary string $B = (b_1, b_2, \dots, b_{108})$ of 108 bits. Each bit is associated with one of the statistical descriptors in Φ . Therefore, if bit b_i is set to 1 its associated descriptor will be selected, with the opposite being true if $b_i = 0$. The feature vector \mathbf{x}_λ for each region λ is given by the concatenation of all the selected descriptors.

Fitness Evaluation: Here is where object models are learned and fitness is assigned. For every physical location L a corresponding GMM $p_L(\mathbf{x}; \Theta_L)$ is trained using an *all vs. all*¹ strategy. Only 70% of the regions extracted from the representative views are used for learning the GMMs, employing the descriptor values selected by B . This generates a set $\mathcal{P} = \{p_L(\mathbf{x}; \Theta_L)\}$ of GMMs, one for each location, with $|\mathcal{P}| = l$. Afterwards, the remaining 30% of image regions are used for validation to compute an accuracy score \mathcal{A} using Bayes rule. Fitness is given for minimization by,

$$f(B) = \begin{cases} \frac{B_{ones}^\alpha + 1}{\mathcal{A}^2 \cdot \inf(\mathcal{S}^{p_i, p_j})} \quad \forall p_i, p_j \in \mathcal{P}, i \neq j, \text{ when } \mathcal{A} > 0, \\ \frac{B_{ones} + 1}{\varepsilon} \quad \text{otherwise.} \end{cases} \quad (6)$$

In the above equation, B_{ones} is the number of ones in string B , $\varepsilon = 0.01$, and α is a weight parameter. The first case in Eq. 6 is applied when the GMM training algorithm converges; fitter individuals will minimize the number of selected descriptors and maximize the average validation accuracy \mathcal{A} . Furthermore, the term $\inf(\mathcal{S}^{p_i, p_j})$ promotes between-class model separation by selecting the infimum of all the apparent separation measures between all models in \mathcal{P} . The second case in Eq. 6 is applied when the training algorithm fails to converge.

Pruning the descriptor space with a two-stage GA. Previous results of a similar algorithm [4], used for object recognition, suggests that the described approach is capable of solving complex recognition problems. However, several limitations were noticeable. For instance, space Φ is quite large, thus all GA runs would converge towards models with only one Gaussian component in a large subspace of Φ , using between 27 and 43 dimensions. This made the use of GMMs completely unnecessary. Moreover, this makes the obtained solutions less desirable because of the large amount of numerical descriptors that they require. Therefore, in order to overcome these limitations a two-stage GA is proposed.

In the first stage, the GA runs using the process described above, with the parameter $\alpha = 2$ in Eq. 6. In this way, the term B_{ones} will have a greater influence on the fitness score. Thus, the GA will favor solutions that use a small subset of Φ . Additionally, due to the large dimensions of Φ , the EM algorithm is used for training with only one Gaussian component in each model. After a fixed number of generations, set to 50 in all experiments, the best solution identifies a subspace of Φ denoted by Φ^* .

The second stage works the same as the first with three modifications. First, the search space employed is defined by Φ^* instead of Φ , what is normally a substantially

¹ All vs all learning implies that all class models are learned in a single step or EM execution.



Fig. 3. Sample images for each of the four locations used in Experiment 1.

more compact search space. Second, with the smaller search space the dimensions of the Gaussian components are expected to be smaller thereby encouraging a more multimodal characterization. Hence, the FJ algorithm is used for training the GMMs with maximum of 10 components per GMM. Third, the weight parameter in Eq. 6 is set to $\alpha = 1$ thereby reducing the influence that B_{ones} has over fitness and focusing fitness on the accuracy term \mathcal{A} . The output of this two-stage GA is a set $F \subset \Phi^*$ of descriptors that best characterizes the regions from each location L , and a set of trained GMMs \mathcal{P}^o used to classify unseen test images.

Place Recognition The final place recognition process proceeds as follows. Given an *unseen* image I from one of the known places, interest regions are detected and their corresponding descriptors, specified in F , are computed. The extracted regions are classified using Bayes rule with the models in \mathcal{P}^o . Therefore, if a majority of the regions are classified to a model $p_L \in \mathcal{P}^o$ then it is said that location L is viewed in the imaged scene I .

4 Experimental setup and results

This section is divided in two parts, one for each experimental configuration.

Experiment 1. The first test for the proposed recognition system contains four locations: 1) Room, 2) WC, 3) Diner, and 4) Lounge. The training and test images were chosen from the same image sequence of 1 Mb color photos, representative images of each location are shown in Figure 3. The size of the images is larger than what is normal for this type of system, however the larger image allows the region detector to extract more image patches for training and testing. Regions were extracted using the entropy-based Kadir & Brady detector. In order to simplify the learning of GMM parameters, a max number of training regions was set to 3,500, which were randomly selected from all the regions extracted from the training images. From this subset, 30% are used in validation and the rest with the learning algorithm (EM or FJ) for the GMMs. Table 2 gives further details regarding the number of photos per location, the cardinality of each GMM learned, the total of test images, the number of misclassified images. Additionally, the first row of Table 3 presents the characteristics of the best individual found by the two-stage GA, describing: the fitness value, the size of Φ^* , the cardinality of F , the descriptors in F , and validation accuracy \mathcal{A} .

Experiment 2. The second setup contains eight locations from our research center, these are: 1) Students Lab, 2) Computer Lab, 3) Research Lab, 4) Lockers, 5) 1st Floor, 6) 2nd Floor, 7) Office, and 8) Mail. Two sequences of images were taken from each

<i>Location</i>	Training Im. GMM components		Test Im. Error	
Room	7	3	2	0
WC	5	6	2	0
Diner	9	4	2	0
Lounge	9	8	2	0

Table 2. Description of Experiment 1, setup and results.

<i>Experiment</i>	Fitness	$ \Phi^* $	$ F $	Features	\mathcal{A}
1)	0.0061	30	7	$\nabla_{\phi(\sigma)}, K_{IPGP1(\gamma_1)}, R_{(\mu,\sigma)}, Q_{(\gamma_2)}, a_{(\sigma)}, b_{(\mu)}$	70.75%
2)	0.0053	36	14	$gab_{(\mu,\gamma_2)}, K_{IPGP2(H)}, G_{(\sigma)}, B_{(\sigma)}, Y_{(\gamma_1)}, I_{(\mu,\sigma)}, Q_{(\sigma)}, L_{(\sigma)}, a_{(H)}, b_{(\mu,H)}, r_{(\sigma)}$	74.73%

Table 3. Performance and selected features; see text for further details.

location, one during the morning and the other in the afternoon thus providing different lighting conditions. Sample images of each location are shown in Figure 4. All images are color jpeg photos of size 320×340 . In this experimental setup the K_{IPGP1^*} scale invariant detector is used, without any restrictions in the amount of regions that are used for training; Table 4 gives further details. Some locations have more training regions than others, this is a result of the fact that some regions have many textured objects, such as the Office and all the Labs, while other locations are much simpler, such as the corridors on each floor, the Mail area, and the Locker area. The second row of Table 3 describes the best individual found by the GA. Additionally, Table 5 presents the confusion matrix for this experiment, here it is possible to see that most of the recognition errors occur with the simpler less textured places. This suggests that the the learning algorithm builds more discriminant models for those regions with more training regions, something that can be expected beforehand.

5 Conclusions and future work

This paper described a learning approach to place recognition which is an essential task for any mobile vision system, such as those used by autonomous robots. This proposal relies on local scale invariant regions and builds probabilistic models using mixtures of Gaussians. The regions are described using statistical values related to texture and color information. The numerical descriptors used are chosen by a two-stage GA from a maximum of 108 different values. The evolutionary learning process searches for the smallest possible subset of descriptors, while also attempting to maximize classification accuracy and the distinctiveness of the GMMs that represent each physical location. Experimental results confirm the validity of the approach by solving two real-world problems of place recognition, and doing so using only visual information. However, results from Experiment 2 indicate that building a recognition system that only relies on visual cues represents a very difficult problem because of self-similarities between locations within most office buildings. Future extensions of this work will center on integrating the system with an autonomous robot in order to facilitate localization dur-



Fig. 4. Sample images for each of the eight locations used for Experiment 2.

<i>Location</i>	Training Im.	Training Regions	GMM components	Test Im.	Error
Students Lab	17	5469	3	17	4
Computer Lab	14	5477	3	14	0
Research Lab	27	4941	2	26	12
Lockers	14	507	3	14	14
1st Floor	13	4534	3	12	11
2nd Floor	20	2263	2	19	12
Office	17	6756	2	16	1
Mail	10	1139	3	10	10

Table 4. Description of Experiment 2, setup and results.

ing *kidnapping* events in real time. Moreover, restrictions should be included, such as spatial relationships between different locations thereby making the recognition process more robust. Additionally, it would be of interest to expand the amount and type of descriptors available to the GA search.

Acknowledgements Research funded by UC MEXUS-CONACyT, the Ministerio de Educación y Ciencia (project Oplink - TIN2005-08818-C04), the LAFMI project, and the Junta de Extremadura Spain. First author supported by scholarship 174785 from CONACyT México.

References

1. Olivier Faugeras. *Three-Dimensional Computer Vision (Artificial Intelligence)*. The MIT Press, November 1993.
2. Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering objects and their localization in images. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV 2005), 17-20 October 2005, Beijing, China*, volume 1, pages 370–377. IEEE Computer Society, 2005.
3. Jutta Willamowski, Damian Arregui, Gabriela Csurka, Chris Dance, and Lixin Fan. Categorizing nine visual classes using local appearance descriptors. In *Proceedings of ICPR*

	Students	Computer	Research	Lockers	1st Floor	2nd Floor	Office	Mail
Students Lab	13	1	0	0	0	0	3	0
Computer Lab	0	14	0	0	0	0	0	0
Research Lab	0	2	14	0	0	0	10	0
Lockers	0	4	0	0	0	0	10	0
1st Floor	1	5	0	0	1	0	5	0
2nd Floor	0	0	2	0	0	7	10	0
Office	0	1	0	0	0	0	15	0
Mail	0	1	2	0	0	0	7	0

Table 5. Confusion matrix for Experiment 2.

2004, *Workshop on Learning for Adaptable Visual Systems, 23-26 August 2004, Cambridge, United Kingdom*. IEEE Computer Society, 2004.

4. Leonardo Trujillo, Gustavo Olague, Francisco Fernández de Vega, and Evelyne Lutton. Evolutionary feature selection for probabilistic object recognition, novel object detection and object saliency estimation using gmms. In *BMVC '03: Proceedings of the 18th British Machine Vision Conference*, volume 2, pages 630–639. British Machine Vision Association, 2007.
5. Leonardo Trujillo and Gustavo Olague. Synthesis of interest point detectors through genetic programming. In Mike Cattolico, editor, *Proceedings of GECCO 2006*, volume 1, pages 887–894. ACM, 2006.
6. Leonardo Trujillo and Gustavo Olague. Using evolution to learn how to perform interest point detection. In *Proceedings of ICPR 2006, 20-24 August 2006, Hong Kong, China*, volume 1, pages 211–214. IEEE Computer Society, 2006.
7. Leonardo Trujillo and Gustavo Olague. Scale invariance for evolved interest operators. In Mario Giacobini et al., editor, *Proceedings of EvoWorkshops 2007*, volume 4448 of *Lecture Notes in Computer Science*, pages 423–430. Springer, 2007.
8. Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 273, Washington, DC, USA, 2003. IEEE Computer Society.
9. Junqiu Wang, Hongbin Zha, and Roberto Cipolla. Coarse-to-fine vision-based localization by indexing scale-invariant features. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 36(2):413–422, 2006.
10. Tony Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.
11. Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.
12. Timor Kadir and Michael Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.
13. Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):381–396, 2002.
14. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.