Message Propagation based Place Recognition with Novelty Detection

Lionel Ott and Fabio Ramos

1 Introduction

In long-term deployments the robot needs to cope with new and unexpected observations that cannot be explained by the robot’s current representation or model of the world. An example of such a representation could be the visual appearance of the environment observed by a camera. Such inexplicable observations can originate from different sources, such as sensor errors, adverse interaction of the sensor with the environment, or previously unobserved locations. One possible solution to deal with such measurements is to filter them using a method such as RANSAC [1], removing observations that do not fit a user defined model. However, in long-term autonomy defining such a model is challenging as the actual model can change over time and more importantly we expect to observe new locations that over time will become part of the model. From a clustering point of view, where clusters represent the model, novel observations can be considered outliers as they are far from any of the existing clusters. Adding the ability to detect and handle outliers in a clustering method allows building a model while retaining novel observations without the need for human supervision.

In this paper we formulate the problem of model building with novelty detection and its application to place recognition as an integer program and present an efficient message passing algorithm inspired by affinity propagation [2]. The resulting method builds the model by clustering the data without the user having to specify the number of clusters and automatically identifies the ℓ most novel observations. The method performs message passing on a factor graph consisting of two parts which are responsible for (i) building the model by finding a valid clustering solution and (ii) selecting the most novel observations as the points with the largest negative impact on the clustering. These two parts interact with each other in order to obtain a globally optimal solution.

The main contribution of this work is a novel algorithm for joint clustering with outlier selection applied to the task of visual place recognition. In experiments we demonstrate the ability of the method to build a valid model and selecting appropriate outliers when used to build a visual appearance based place recognition model.

2 Related Work

Image based place recognition has been an active area of research for some time now. FAB-Map [3] builds a generative model of the environment based on a bag-of-words representation of images which is improved by taking into account word co-occurrences modelled by a Chow-Liu tree. Schindler et al. [4] use vocabulary trees and demonstrate that using the most informative, as opposed to all available, features can considerably improve the place recognition performance. The system proposed by Galvez-Lopez and Tardos [5] is based on a binary bag of words approach using FAST+Brief features which are faster to compute than SIFT or SURF ones. This allows the algorithm to operate much higher rates then other methods. A different approach is taken by [6] which uses low resolution images instead of extracted features to describe a scene. To compensate for this the method matches image sequences against the traversed routes instead of image to image matches which results in good place recognition performance. Dayoub and Duckett [7] present an

*Lionel Ott and Fabio Ramos are with the School of Information Technologies, The University of Sydney, Australia.
adaptive method for topological place recognition which uses short and long-term memory concepts to handle changes in the appearance of the environment over time. The method adds and removes descriptors of locations by using a model of human memory. This allows the method to recognise places even in the event of changes in the environment with partial occlusions and noise. A supervised method for indoor place recognition is presented in [8] where each image is described by a receptive field histogram. A training set of these features is then used to train a support vector machine to classify future observations into the predefined locations. This method is expanded in [9] by the inclusion of an incremental SVM which allows retraining the classifier when environmental changes are detected which can improve recognition performance as well as runtime. Another supervised approach was proposed in [10] which trains a generative model based on features obtained from applying neighbourhood preserving dimensionality reduction techniques to the convolution of image patches with multiple Gabor filters.

Chawla and Gionis [11] have proposed k-means-- a practical and scalable algorithm for the k-means with outlier problem. The method is a two step approach where in each iteration of standard k-means, the ℓ points with the largest distance to their centroid are ignored in the next round of k-means. However, the algorithm inherits the weaknesses of classical k-means, i.e., (i) the requirement of setting the number of clusters k and (ii) initial selection of the k centroids.

Our proposed method approaches place recognition as a clustering task, as opposed to a classification task. Furthermore, the ability of our method to jointly cluster data while selecting outliers provides robustness as well as information about potentially interesting new places to explore.

3 Methodology

The goal of our method is to automatically build a model and select novel observations, which we achieve with a clustering method that selects the correct number of clusters from the data while at the same time selecting a fixed number of outliers which have the largest negative effect on the solution. For place recognition the clusters represent locations which are visually similar while the outliers represent the observations which are novel compared to the overall dataset. We formulate this problem as an integer program which we solve using a message passing algorithm inspired by affinity propagation.

3.1 Optimisation Problem

Given an assignment cost matrix $d_{ij}$, which represents the distance or similarity between pairs of data points, cluster creation costs $c_j$, which is the cost assigned with selecting a point as an exemplar, and the number of outliers $\ell$. An exemplar is a representative point of a cluster, similar to the centroid in k-means only that the exemplar is an actual data point. In appearance based place recognition the value $d_{ij}$ indicates the dissimilarity between two images $i$ and $j$. The task of clustering and outlier selection is defined as the problem of finding the assignments to the binary exemplar indicators $y_i$, outlier indicators $o_i$ and point assignments $x_{ij}$ that minimises the following cost function:

\[
\begin{align*}
\text{minimise} & \quad \sum_j c_j y_j + \sum_i \sum_j d_{ij} x_{ij} \\
\text{subject to} & \quad x_{ij} \leq y_j \quad (1a) \\quad x_{ij} = 1 \quad (1b) \\quad \sum_i o_i = \ell \quad (1c) \\quad x_{ij}, y_j, o_i \in \{0, 1\} \quad (1d)
\end{align*}
\]

The above constraints enforce valid solutions, i.e.:  
- points can only be assigned to valid exemplars (Eq. (1b));  
- every point must be assigned to exactly one other point or be declared an outlier (Eq. (1c));  
- exactly $\ell$ outliers have to be selected (Eq. (1d));  
- only integer solutions are allowed (Eq. (1e)).
3.2 Message Passing Algorithm

Our algorithm solves the above integer program by representing it with the factor graph shown in Figure 1a. The factor graph consists of two parts. The left side is identical to affinity propagation, with the difference that it is connected to the right side responsible for the selection of outliers. The constraints of the integer program Eq. (1a) are encoded by the following energy function:

$$\max \sum_{ij} S_{ij} + \sum_j E_j(x_{ij}) + \sum_i I_i(x_i; o_i) + \sum_k P_k(o_k),$$  \hspace{1cm} (2)$$

where

$$S_{ij} = \begin{cases} -c_i & \text{if } i = j \\ -d_{ij} & \text{otherwise} \end{cases}$$  \hspace{1cm} (3)$$

$$I_i(x_i; o_i) = \begin{cases} 0 & \text{if } \sum_j x_{ij} + \sum_k o_{ik} = 1 \\ -\infty & \text{otherwise} \end{cases}$$  \hspace{1cm} (4)$$

$$E_j(x_{ij}) = \begin{cases} 0 & \text{if } x_{jj} = \max_i x_{ij} \\ -\infty & \text{otherwise} \end{cases}$$  \hspace{1cm} (5)$$

$$P_k(o_k) = \begin{cases} 0 & \text{if } \sum_i o_{ik} = 1 \\ -\infty & \text{otherwise} \end{cases}$$  \hspace{1cm} (6)$$

with $x_i = x_{i1}, \ldots, x_{iN}$ and $o_k = o_{1k}, \ldots, o_{Nk}$. Due to the nature of the constraints we maximise the energy function and use negative distances. The three constraints can be interpreted as follows:

1. 1-of-$N$ Constraint ($I_i$). Each data point has to choose exactly one exemplar or be declared as an outlier (Eq. (1c)).
2. Exemplar Consistency Constraint ($E_j$). For point $i$ to select point $j$ as its exemplar, point $j$ must declare itself an exemplar (Eq. (1b)).
3. Select $\ell$ Outliers Constraint ($P_k$). For every outlier selection exactly one point is assigned (Eq. (1d)).

These constraints are enforced by associating an infinite cost with invalid configurations, resulting in obviously suboptimal solutions. As with affinity propagation we find the maximum a posteriori setting (MAP) using the max-sum algorithm [12]. The message update rules required for this are obtained by using the above constraints to simplify the general update messages:

$$\mu_{v \rightarrow f}(x_v) = \sum_{f^* \in \text{ne}(v) \setminus f} \mu_{f^* \rightarrow v}(x_v),$$  \hspace{1cm} (7)$$

$$\mu_{f \rightarrow v}(x_v) = \max_{x_1, \ldots, x_M} \left[ f(x_v, x_1, \ldots, x_M) + \sum_{v^* \in \text{ne}(f) \setminus v} \mu_{v^* \rightarrow f}(x_v) \right],$$  \hspace{1cm} (8)$$

where $f$ is a factor, or a function over a subset of variables, $\mu_{v \rightarrow f}(x)$ is the message sent from node $v$ to factor $f$, $\mu_{f \rightarrow v}(x_v)$ is the message from factor $f$ sent to node $v$, $\text{ne}(\cdot)$ is the set of neighbours of the given factor or node, and $x_v$ is the value of node $v$.

The messages exchanged by our method, message passing for outlier clustering (MPOC), are shown in Figure 1b. Since only binary variables are involved, it is sufficient to compute the difference between the two variable settings. Combining and simplifying these messages by exploiting knowledge about assignments resulting in invalid solutions we obtain the final set of update equations as:

$$\rho_{ij} = s_{ij} + \min_{t \neq j} \left[ -\max_{t \neq j}(\alpha_{it} + s_{it}), -\max_{t}(\omega_{it}) \right],$$  \hspace{1cm} (9)$$

$$\alpha_{ij} = \begin{cases} \sum_{t \neq j} \max(0, \rho_{ij}) & i = j \\ \min \left[ 0, \rho_{jj} + \sum_{t \neq \{i,j\}} \max(0, \rho_{ij}) \right] & i \neq j \end{cases},$$  \hspace{1cm} (10)$$
The method starts by initialising the messages $\alpha_{ij}$, $\rho_{ij}$ and $\lambda_{ik}$ to 0 and $\omega_{ij}$ to the median of $S$. Once the messages are initialised they are updated in turn with damping until convergence is achieved. Typically this is the case when the energy of the solution is stable over a few iterations. Outliers are determined as the $\ell$ points with the largest values of $\max_k(\lambda_{ik} + \omega_{ik})$. From the remaining points, the exemplars are then selected as the points for which $(\alpha_{ii} + \rho_{ii}) > 0$ is true. All other points $i$ are assigned to the exemplar $e$ satisfying $\arg\max_e(\alpha_{ie} + \rho_{ie})$.

## 4 Experiments

The Malaga dataset [13] consists of image, laser, and GPS data collected in the city of Malaga spanning 37 km. We use the image data to build an appearance model for place recognition and use the GPS data to validate the place predictions. The dataset contains images recorded at 20 Hz, however, we build the appearance model using data sampled at 0.5 Hz. The model is thus built from 2840 images, or an average of one image every 13 m. Each image is represented by HSV colour histograms and local binary pattern [14] histograms capturing texture. The entries $d_{ij}$ of the cost matrix are computed using the Bhattacharyya distance:

$$d_B(a, b) = \sqrt{1 - \frac{1}{\sqrt{\sum_i a_i \sum_i b_i} N^2} \sum_i \sqrt{a_i b_i}},$$

where $a$ and $b$ are two histograms and $N$ is the number of bins in the histograms. For this experiment the number of outliers was set to $\ell = 100$, while the cost of creating a cluster $c_i$ was set to $\theta$ median($d_{ij}$) with $\theta \in [2, 10]$.

The images in Figure 2a show some of the exemplars obtained from clustering the images. We can see that these images represent visually distinct scenes. Looking at the outliers selected during the clustering, shown in Figure 2b, we can see that several images have bad image quality due to sun glare, contrast issues or underexposure. Some of the other images show scenes that occur infrequently such as underpasses. All of these novel or surprising images were selected automatically, without us informing the algorithm of what is considered an uncommon image. The algorithm selected these images because they were the most dissimilar ones in the dataset.

In order to perform place recognition we use the visual appearance model as a search structure. For each query image we find the most similar cluster and within that cluster the image most similar to the query. The location of this most similar image is then used as the location of the query image. Since the number of clusters is much smaller then the total number of images in the model this saves considerable time when searching for the best match compared to a linear search. In Figure 3 we
Figure 2: (a) Some of the exemplars selected when clustering the Malaga dataset, representing visually distinct locations within the dataset. (b) Outliers selected from the dataset representing images that have bad image quality or represent uncommon scenes.

Figure 3: Visualisation of the place recognition output for a section of the dataset. Each point represents the predicted location of a query image with the colour indicating the cluster used in the search.

show the location prediction as well as cluster used in the search for a small area of the dataset. Each point indicates the predicted location of an observation while the colour corresponds to the cluster used in the search. Since the cameras are forward facing we can see that driving along a road in opposing directions produces a different cluster for each direction. We can also see how the majority of the straight sections are clustered into their own distinct cluster.

Table 1 summarises the training time, time needed to query 11,355 images, and percentage of place predictions that were incorrect for MPOC, $k$-means--, and linear search. A prediction is considered incorrect if it is more than 100 m from its true location. In comparison to $k$-means-- the query time of MPOC is slightly better but with much better place recognition accuracy. The training time of $k$-means-- is much lower, however, this is offset by the issue that the number of clusters needs to be specified, which in practice is impracticable. Comparing MPOC to the linear search we see that MPOC has much better query time, however, since the linear search considers every image its accuracy is slightly better. This could potentially be addressed by considering the first two or three most similar clusters as opposed to only the most similar one. Overall our method provides good place recognition accuracy with excellent query performance while being extremely easy to build as only an image similarity metric to compute the $d_{ij}$ matrix is needed.
### Table 1: Information about the different methods including the time required for training, time needed to answer 11355 queries and the percentage of failed place predictions.

<table>
<thead>
<tr>
<th>Method</th>
<th>MPOC</th>
<th>k-means--</th>
<th>Linear search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training time(s)</td>
<td>93.4</td>
<td>1.9</td>
<td>0.0</td>
</tr>
<tr>
<td>Query time (s)</td>
<td>2.9</td>
<td>3.7</td>
<td>122.7</td>
</tr>
<tr>
<td>Failed predictions</td>
<td>9.2%</td>
<td>20.3%</td>
<td>6.7%</td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper we proposed a novel method to automatically, and without supervision build a place model based on images while at the same time selecting the most novel images. The problem of combined model building and novelty detection is formulated as an integer program to which we propose a message passing based algorithm. In experiments we demonstrate the ability to learn a model suitable for place recognition which additionally provides the set of images which are the most surprising in the dataset.

References


