Place Recognition in Challenging Conditions

Saravanabalagi Ramachandran* and John McDonald+

Department of Computer Science, Maynooth University, Ireland.

Abstract

Place recognition in a visual SLAM system helps build and maintain a map from multiple traversals of the same environment while closing loops to correct drift accumulated over time. Despite the marked success in visual place recognition research over the past decade, it remains a challenging problem in the context of variations caused due to different times of the day, weather, lighting and seasons. In this paper, we address this problem by progressively training convolutional neural networks in a siamese fashion to generate embeddings that encode semantic and visual features for sequence-aligned image pairs taken at different timescales and viewpoints. We present early results of the approach using Freiburg visual place recognition benchmark dataset consisting of aligned outdoor image sequences taken over extended time periods that include the variations mentioned above.

Keywords: Place Recognition, Deep Learning, Image Embeddings

1 Introduction

Estimating the pose of a robot and building a map of its environment at the same time, aka. Simultaneous Localization and Mapping (SLAM), is a central build building block of intelligent mobile robots. Place recognition is an important module in SLAM systems in that it plays a significant role in building more accurate maps by correcting the estimated trajectory of the robot when it revisits previously mapped regions. Although the traditional approaches [Sivic and Zisserman, 2003], [Cummins and Newman, 2011], [Gálvez-López and Tardós, 2012] using Bag of Visual Words (BoW) approach permitted a reliance on visual place recognition within SLAM systems, they lacked the repeatability and robustness required to a deal with the challenging variability that occurs in natural scenes caused due to different times of the day, weather, lighting and seasons (see Figure 1).

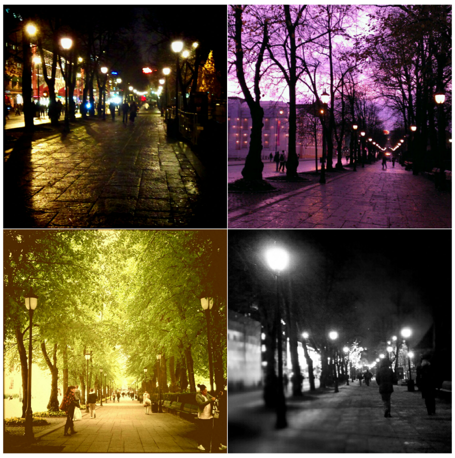

Figure 1: Images of the same place taken at different times, demonstrating the challenge of place recognition in the context of extreme changes in visual appearance. Image Credits: [Graver, 2013]

With the advent of Convolutional Neural Networks (CNN) and their compelling results over traditional methods in tasks such as semantic segmentation and feature learning, more recently researchers have sought to improve robustness by incorporating semantic, geometric, and topological information from the scene using CNNs. Segmap [Dubé et al., 2018] uses CNNs to eliminate moving objects using semantics and to generate compact embeddings for 3D objects. LoST [Garg et al., 2018] uses *conv5*

*Saravanabalagi.Ramachandran@mu.ie; +John.McDonald@mu.ie

This research was supported by Science Foundation Ireland grant 13/RC/2094 to Lero - the Irish Software Research Centre and grant 16/RI/3399.

layer of modified dense semantic segmentation neural network RefineNet [Lin et al., 2017] to generate embeddings. Embeddings are generated for each keyframe during the traversal, and each time an embedding is stored, it is compared to existing embeddings and the embeddings distances are computed. Images that have an embedding less than a defined threshold are further eliminated by matching maximally activated regions in the image to find the final loop closing candidate.

Validity of embeddings generated using pre-trained CNNs has a significant dependency on the training data. For example, RefineNet was pre-trained on CityScapes dataset. Cityscapes contains images from different cities and does not contain sequences that were shot at different times of the day, weather and scene types. Hence the network does not work as intended on images shot in conditions it has not previously seen. Retraining these networks with such images would be hard as it is a tedious and expensive task to obtain ground truth dense semantic segmentation labels.

In this paper, we present a deep learning based approach that aims to achieve this by leveraging the semantic and photometric information in an image within a single framework for place recognition. In particular, we train a deep convolutional neural network in a siamese configuration with a contrastive loss to map input images to an embedding space that aims to minimise the distance between embeddings of the same place, whilst maximise the distances between embeddings of different places. In this approach, we generate embeddings directly from images where the semantic and visual information are implicitly encoded by the network without the presence of ground truth labels. By training the network in a siamese fashion, we can remove the direct dependency on semantic labels or feature descriptions and instead train over pairs of images of the same place taken at different times. We provide the first results of the application of the approach to a publicly available place recognition benchmark dataset.

2 Methodology

We train our convolutional neural network in a siamese fashion to generate compact image embeddings, where the network implicitly encodes semantic and visual features. We force the network to output similar embeddings for images of the same place and embeddings that are farther apart in the embedding space for images that are not of the same place.

2.1 Loss Function

In a siamese type network, the network consumes two images (referred to as I_{left} and I_{right} from here) at once as input along with the ground truth label. The network uses two branches to get two embeddings (for I_{left} and I_{right}) but as weights are shared for left and right branches they do not take up twice the amount of space. We use Contrastive Loss [Hadsell et al., 2006] to achieve this behaviour which introduces two parameters to tune; margin m and threshold γ. These parameters can be adjusted empirically from the results of initial training. The contrastive loss is given by,

$$J(I_{\text{left}}, I_{\text{right}}, y) = yd^2 + (1-y)\max(m-d^2, 0) \tag{1}$$

where,

$$\begin{aligned}
J &= \text{the loss function}\\
I_{\text{left}}, I_{\text{right}} &= \text{image pair, one row of training data}\\
y &= \text{ground truth boolean, } \textit{true} \text{ if } I_{\text{left}} \text{ and } I_{\text{right}} \text{ are images of the same location, } \textit{false} \text{ otherwise}\\
d &= \text{the distance between the embeddings computed for } I_{\text{left}} \text{ and } I_{\text{right}}\\
m &= \text{margin}
\end{aligned}$$

During inference, match prediction $\hat{y}$ (if I_{left} and I_{right} are images of the same location) is computed as *True* when $d \leq \gamma$ and *False* otherwise. The matching threshold γ is initially set to half the margin during the training phase and then the Receiver Operating Characteristic (ROC) curve is be used to fine tune the threshold γ for inference by choosing the right trade-off between precision and recall.

2.2 Network Architecture

We use a siamese configuration with different backbones: VGG-16 [Deng et al., 2009], ResNetv2 [He et al., 2016] and Inceptionv3 [Szegedy et al., 2016], all pre-trained on Imagenet [Deng et al., 2009]. To get embeddings of same size, we include the final dense layers for VGG-16, and use *global average pool*ed outputs from *topless*[1] models for ResNetv2 and Inceptionv3.

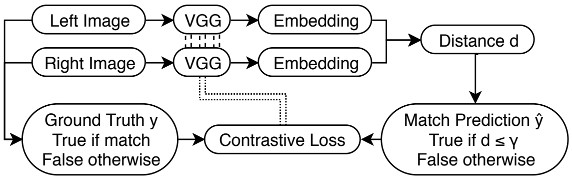

Figure 2: Network Architecture

2.3 Data Augmentation

Deep neural networks require large amounts of training data for better generalization. Due to the low number of training examples, we augmented our data with a series of random transformations (randomized for each image, every epoch) where the images were cropped, rotated, translated and zoomed. It should be noted that when performing these transformations we do not fill the empty pixels which may be filled using constant, same, reflect or wrap modes [Chollet et al., 2015], as this might lead to synthesizing images that may not imitate what we would encounter in the real world. Instead, we zoom the image and perform an aspect-ratio-preserving centre-crop to achieve a similar result.

3 Experiments

We use sequence-aligned images from the Freiburg visual place recognition dataset [Naseer et al., 2014], which contains 674 image pairs. For each matching pair of images in the dataset, we generated a non-matching pair picking a reference image at random[2]. We obtained a total of 1348 image pairs, exactly half of which are matching pairs, while the rest are non-matching. We sliced the data twice with different regions as test data. Split 1: Here the first one-third (450 image pairs) is used as test data for evaluation and the remaining two-thirds of the data (898 image pairs) is used for training. Split 2: Here the last one-third is used as test data and the first two-thirds is used as training data. In both cases, the last 33% of the training data (300 of 898 image pairs) was used as validation data to monitor and improve the performance of the network. All images were centre-cropped preserving aspect ratio to $224 \times 224 \times 3$ for VGG (fixed resolution due to dense layers) and $270 \times 480 \times 30$ for ResNet and Inception. Each network was trained with a margin m of 1.0. We used a value of 0.5 for γ, the separation threshold for matches and non-matches. Experimental results are shown in Table 1.

It should be noted that in each of the experiments above, the dataset is limited in scale and so the results only provide an indicative measure of each model's potential performance. We further noticed that two models do *not* perform well in Split 2, due to repeated frames i.e. the models tend to overfit due to lack of variability in a limited data setting. As such, in order to comprehensively evaluate the proposed approach a considerably larger dataset is required. For example, here both the training and test samples, although separated, are drawn from the same database. In order to truly evaluate a system's real-world performance, it is important to utilise both a large image dataset and to incorporate images from different data sources. This is the focus of our current work.

Backbone	VGG-16	ResNet v2	Inception v3
		Split 1	
AuC	0.9533	0.9681	0.9198
AP	0.9534	0.9607	0.9354
		Split 2	
AuC	0.6533	0.9668	0.5281
AP	0.5768	0.9743	0.5814

Table 1: Experiment results showing Area under Curve (AuC) in the ROC graph and Average Precision (AP) for each model in siamese settings.

[1] with the final fully connected and classification layer removed

[2] random but not within the previous or proceeding 20 frames of the actual reference image

4 Conclusion

We have presented an approach to robust place recognition for challenging conditions, invariant to viewpoints, changes caused by seasons and different times of the day. To do this, we use convolutional neural networks in a siamese fashion to learn an image embedding space. Results were presented where we evaluated the trained models on images from the Freiburg visual place recognition dataset demonstrating promising performance.

We are currently carrying out further training on different and more extensive datasets with sequences aligned, both with and without the help of local feature descriptors. Our aim here is to ensure that the model will benefit from using local image features, but also, operate on images with challenging conditions where local feature descriptors would struggle to find matches. A comparison with state of the art over image pairs chosen from a variety of datasets is necessary to give a closer measure of the true performance of our model and helps to estimate the real world performance. Further, evaluation needs to be performed by integrating the place recognition system within a SLAM pipeline, to better understand its performance in a complete system.

References

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras. https://keras.io/preprocessing/image/#imagedatagenerator-class.

[Cummins and Newman, 2011] Cummins, M. and Newman, P. (2011). Appearance-only slam at large scale with fab-map 2.0. *Int. J. Rob. Res.*, 30(9):1100–1123.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[Dubé et al., 2018] Dubé, R., Cramariuc, A., Dugas, D., Nieto, J. I., Siegwart, R., and Cadena, C. (2018). Segmap: 3d segment mapping using data-driven descriptors. *ArXiv*, abs/1804.09557.

[Gálvez-López and Tardós, 2012] Gálvez-López, D. and Tardós, J. D. (2012). Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197.

[Garg et al., 2018] Garg, S., Suenderhauf, N., and Milford, M. (2018). Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics. *Proceedings of Robotics: Science and Systems XIV*.

[Graver, 2013] Graver, J. (2013). From the accidental to the deliberate. http://mortalmuses.com/2013/01/15/accidental-to-deliberate/.

[Hadsell et al., 2006] Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE CVPR*, volume 2, pages 1735–1742.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Lin et al., 2017] Lin, G., Milan, A., Shen, C., and Reid, I. (2017). RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*.

[Naseer et al., 2014] Naseer, T., Spinello, L., Burgard, W., and Stachniss, C. (2014). Robust visual robot localization across seasons using network flows. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 2564–2570. AAAI Press.

[Sivic and Zisserman, 2003] Sivic and Zisserman (2003). Video google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE ICCV*, pages 1470–1477 vol.2.

[Szegedy et al., 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE CVPR*, pages 2818–2826.