

University of Groningen

Data-efficient representation learning for visual place recognition

Leyva Vallina, María

DOI:
[10.33612/diss.736449452](https://doi.org/10.33612/diss.736449452)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2023

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Leyva Vallina, M. (2023). *Data-efficient representation learning for visual place recognition*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. <https://doi.org/10.33612/diss.736449452>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

*Data-Efficient Representation Learning
for Visual Place Recognition*

María Leyva Vallina

This research has been conducted at the Intelligent Systems group of Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence of the University of Groningen.

This work was partially funded by the European Horizon 2020 program, under the project TrimBot2020 (grant No. 688007).

Data-Efficient Representation Learning for Visual Place Recognition
María Leyva Vallina



Data-Efficient Representation Learning for Visual Place Recognition

PhD dissertation

to obtain the degree of PhD of the
University of Groningen
on the authority of the
Rector Magnificus Prof. J.M.A. Scherpen
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Tuesday 31 October 2023 at 09.00 hours

by

María Leyva Vallina

born on 23 February 1993
in Asturias, Spain

Supervisors

Prof. N. Petkov

Dr. M.H.F. Wilkinson

Co-supervisor

Dr. N. Strisciuglio

Assessment Committee

Prof. J. González Jiménez

Prof. E. Alegre

Prof. M. Biehl

*The Road goes ever on and on,
Down from the door where it began.
Now far ahead the Road has gone,
Let others follow, if they can!
Let them a journey new begin.
But I at last with weary feet
Will turn towards the lighted inn,
My evening-rest and sleep to meet.*

J.R.R. Tolkien

Contents

List of Figures	v
List of Tables	xi
1 Introduction	1
1.1 Thesis Organization	3
2 TB-places: A dataset for visual place recognition in garden environments	9
2.1 Introduction	10
2.2 Related work	11
2.3 Dataset and its acquisition	13
2.3.1 Hardware setup	15
2.3.2 TB-Places dataset	15
2.3.3 Ground truth	16
2.4 Evaluation	20
2.4.1 Image descriptors	20
2.4.2 Experiments	21
2.4.3 Performance measures	22
2.5 Results and discussion	24
2.5.1 Pose regression	28
2.6 Conclusions	29
3 Place recognition in gardens by learning visual representations	33
3.1 Introduction	34
3.2 Dataset	35
3.3 Evaluation	38
3.3.1 Performance measures	39
3.3.2 Baseline	39
3.3.3 Learning garden-specific representations	40
3.4 Results and discussion	42

3.5	Conclusions	45
4	Generalized Contrastive Optimization of Siamese Networks for Place Recognition	49
4.1	Introduction	50
4.2	Related works	53
4.2.1	Metric learning for place recognition	53
4.2.2	Data annotation	55
4.3	Methodology	56
4.3.1	Fully convolutional backbone and pooling	56
4.3.2	Generalized Contrastive Loss	57
4.3.3	Image search and place recognition	60
4.4	Training data and automatic labelling	61
4.4.1	2D Field-of-View overlap	62
4.4.2	3D Field-of-View overlap	64
4.4.3	Selection of training pairs	66
4.5	Experimental framework	67
4.5.1	Datasets	67
4.5.2	Implementation details	69
4.5.3	Evaluation metrics	70
4.6	Results and discussion	70
4.6.1	Large scale outdoor place recognition	70
4.6.2	Small scale outdoor place recognition	75
4.6.3	Indoor place recognition	76
4.6.4	Discussion	78
4.7	Additional results	83
4.7.1	Learned latent space	84
4.7.2	Network activation	85
4.7.3	Results with Global Average Pooling	88
4.7.4	Results on Pittsburgh250k and TokyoTM	88
4.7.5	Detailed results on RobotCar Seasons v2 and Extended CMU Seasons	88
4.7.6	Comparison to Thoma et al. (2020)	92
4.7.7	Performance for different ground truths	92
4.7.8	Extra results on 7Scenes	93
4.8	Field-of-View overlap	95
4.9	Gradient computations	95
4.9.1	Contrastive loss	96
4.9.2	Generalized Contrastive loss	97

CONTENTS

4.10	Conclusions	99
5	Regressing Transformers for Data-efficient Visual Place Recognition	103
5.1	Introduction	104
5.2	Related works	106
5.2.1	Visual place recognition.	106
5.2.2	Metric Learning for VPR.	107
5.2.3	Limitations of contrastive learning for VPR.	107
5.3	Regression for visual place recognition	108
5.3.1	Architecture and optimization	108
5.3.2	Training pairs and batch composition	109
5.3.3	Image search and retrieval	109
5.3.4	Training data	110
5.3.5	Evaluation data	110
5.3.6	Implementation details	111
5.4	Experiments and results	111
5.4.1	Results	111
5.4.2	Ablation experiments	116
5.5	Conclusions	119
6	Summary and Outlook	125
6.1	Summary	125
6.2	Outlook	126
	Bibliography	129
	Summary	137
	Samenvatting	139
	Acknowledgements	141
	Research Activities	143
	About the Author	147

List of Figures

1.1	Example images within the TB-Places dataset. (a) shows a reference image and (b), (c), (d) and (e) are images with different annotated degrees of similarity. In (f) we show a 2D position map of the shown images in the reference system of the garden.	2
2.1	Four views of the TrimBot2020 garden at the Wageningen University and research campus.	13
2.2	The (a) camera rig used for the data recording session in 2016 has eight cameras arranged in an (b) octagonal shape, while the (c) rig used in spring 2017 has (d) five pairs of stereo cameras in a pentagonal frame.	14
2.3	Example of ambiguous cases where translation and orientation distances are the same: the cameras are oriented at (a) different places or at (b) the same place.	17
2.4	Examples of two field of view overlaps with (a) $d_t = 0.2$, $\Delta\phi = 40^\circ$, $FOVO = 0.71$ and (b) $d_t = 1.0$, $\Delta\phi = 60^\circ$, $FOVO = 0.75$	17
2.5	Relationship between the (a) camera viewpoint angle and FOV overlap for different translation distance values. (b) Approximation of function f by a polynomial regression, that computes the quaternion distance that maximizes the FOVO for a given translation distance.	18
2.6	Example images from the TB-Places dataset. For each reference image (left column), we show two positive matches (second and third column, surrounded by a green solid line) and one negative image (surrounded by a red dashed line).	20
2.7	Precision recall curves achieved on the (a) W17 and (b) R17 test sets.	26
2.8	F_1 -score as the camera viewpoint difference($\Delta\phi$) increases for (a) W17 and (b) R17 datasets.	27
3.1	(a) Robot platform in the Wageningen garden of TrimBot2020 project. (b) Camera rig employed for the recording sessions.	36

3.2	Top-view of the trajectories followed by the robot during the recording sessions in the Wageningen garden in (a) 2018 (W18) and (b) 2017 (W17).	37
3.3	Examples of TB-Places W18 dataset. The left column shows reference images, while the center shows positive matches (green squares), and the right column shows negative matches (dashed red squares).	38
3.4	Sketch of the employed architecture for learning garden-specific descriptors for place recognition. We feed a pair of images to a siamese CNN architecture and compute their representation with a Global Average Pool layer. The training is guided by optimizing a contrastive loss function $L(f_0, f_1)$.	41
3.5	Precision-recall curves achieved on train (W17) and test (W18) by the considered descriptors (a) before and (b) after training the models.	43
3.6	Train and test performance vs training epoch. The bullet indicates the best test AP achieved by each method, corresponding to epoch 6 for DenseNet, epoch 0 for NetVLAD VGG, and epoch 1 for ResNet.	44
4.1	Examples from MSLS dataset of a query image (a) and three matches with different degrees of similarity. A very close match is (b) with 86% of overlap, while (c) is a borderline case with 52% of commonality, and (d) is a negative match, with no features in common with the query.	51
4.2	Sketch of a siamese architecture where x_i and x_j are the input images, \hat{f} represents the convolutional backbone with a pooling layer and $\hat{f}(x_i)$ and $\hat{f}(x_j)$ are the representations of the input images. They are used as input for the Generalized Contrastive Loss (GCL) function $\mathcal{L}_{GCL}(\hat{f}(x_i), \hat{f}(x_j))$.	57
4.3	Example dataset consisting of three images. (A, B) and (B, C) are considered similar ($y = 1$) because they share part of their content, while (A, C) are dissimilar ($y = 0$).	58
4.4	(a) 2D Field-of-View representation with angle θ and radius r . The point (t_0, t_1) is the camera location in the environment, and α is the camera orientation in the form of a compass angle with respect to the north N . A (b) soft positive match: two cameras (with $\theta = 90^\circ$ and $r = 50m$) in the same position but with orientations 40° apart. A (c) soft negative example: two cameras (with $\theta = 90^\circ$ and $r = 50m$) located 25m apart but with the same orientation.	61

LIST OF FIGURES

4.5	Example image pairs from the MSLS dataset. The first row shows the query images, the second row shows the corresponding matches from the map set, and the third row shows the estimated 2D FoV overlap. The query image is associated with the red camera, while the map image with the blue camera.	62
4.6	Example image pairs from the TB-Places dataset. The first row shows a positive pair with FoV overlap of 74%. The second row depicts a soft negative pair with FoV overlap equal to 41%. The red camera corresponds to the query image, while the blue one to the map image.	64
4.7	3D Field-of-View overlap examples from the 7Scenes dataset. The first column shows a positive pair with 3D FoV overlap of 75%. The second column depicts a borderline pair with 50% 3D FoV overlap. The last column shows a soft negative image pair with 25% 3D FoV overlap. The red pointcloud corresponds to the query 3D FoV, the blue one is the map, and the magenta represents the overlap between them.	65
4.8	Configurations of the experiments on the TB-Places dataset. (a) W17 subset is the map set, and W18 is the query. (b) We divide W18 into map and query. For visualization purposes, the trajectories have been downsampled.	69
4.9	Comparison of the results achieved by our methods with those of state-of-the-art methods and models trained with a binary contrastive loss for (a) MSLS validation and (b) MSLS test. TL stands for models trained with the Triplet Loss, CL for Contrastive Loss and GCL for Generalized Contrastive Loss. The results obtained by our models are displayed in red, with a solid line for those trained with a GCL function, and a dashed one for those trained with a CL function.	74
4.10	Comparison of the results for the TB-Places dataset. In (a) we report the results when using W17 as map and W18 as query set. In (b), we show the top-k recall achieved when dividing W18 into map and query. The results achieved by our method are shown in red, the results of the models trained with the binary Contrastive Loss are in blue, and the results of NetVLAD are displayed in black.	76
4.11	Recall@K results achieved on the 7 Scenes dataset. The results of the models trained with our Generalized Contrastive Loss are shown in red, while those of the models trained with the binary Contrastive Loss are shown in blue. The Recall@K achieved by NetVLAD off-the-shelf is plotted in black.	77

4.12	Similarity ground truth distribution for 10000 randomly selected pairs in the MSLS train set when using different batch composition strategies. The vertical axis is in log scale.	80
4.13	Results obtained on the MSLS validation, MSLS test, Pittsburgh30k and Tokyo 24/7 datasets by our models with and without PCA whitening. For all the datasets, reducing the dimensionality of the latent space vectors and applying the whitening transform contribute to an increase of the retrieval performance.	82
4.14	Visualization of the learned embedded space. We selected 1000 random positive pairs and 1000 random negative pairs from the MSLS Copenhagen set, computed the differences between their representations and projected them into a 2D space using T-SNE.	84
4.15	CNN activations for the ResNet50-GeM and the VGG16-GeM models with CL and GCL for several input image pairs. The first two pairs, corresponding to the first four columns, are part of the MSLS test set. The third and fourth belong to the Pittsburgh30k and Tokyo 24/7 test set, respectively. We show the activations for the last layer of the backbone overlapped with the input images.	86
4.16	Comparison of the results achieved by methods trained with GCL and CL for (a) different distance thresholds and (b) different ψ threshold values. GCL-based methods are plotted in red, while CL-based methods are shown in blue.	93
4.17	Precision-recall curves achieved on the 7Scenes dataset by ResNet18 and ResNet34 models trained using the Contrastive Loss function (CL, blue lines) and the proposed Generalized Contrastive loss (GCL, red lines) function.	94
4.18	Relation of 2D FoV overlap with translation and orientation distance. This data is computed from a subset of the MSLS training set.	96
5.1	A reference image (leftmost) and four match images, taken at different distances. The larger the distance w.r.t the reference image, the lower the annotated similarity ground truth ψ , and the smaller the amount of shared visual features.	105
5.2	Sketch of a siamese architecture where x_i and x_j are the input images, $\theta(\cdot)$ represents the encoder and $\theta(x_i)$ and $\theta(x_j)$ are the representations of the input images. They are used as input for the loss function.	109

LIST OF FIGURES

5.3	Retrieval performance every 10k training iterations on the (a) MSLS-val, (b) Pitts30k, and (c) Tokyo24/7 for the same ViT-R50 encoder trained with CL, GCL, and MSE loss functions. We ran each experiment three times and report the average, minimum and maximum R@5.	113
5.4	Attention maps extracted from the last layer of our ViT-R50-GCL and ViT-R50-MSE models for pairs of similar images from the MSLS validation dataset (rows 1-2), the Pitts30k test dataset, (rows 3-4) and the Tokyo24/7 dataset (rows 5-6).	117
5.5	Results obtained on the MSLS validation, MSLS test, Pittsburgh 30k and Tokyo 24/7 datasets by MSE-trained models with and without PCA whitening. Reducing the dimensionality of the descriptors and applying the whitening transform contribute to an increase of the retrieval performance.	119

List of Tables

2.1	Details about the TB-Places dataset, with information on each subset. We also specify the percentage of similar image pairs on the total number of image pairs.	16
2.2	Details on the whole-image descriptors used for the baseline performance comparison analysis.	21
2.3	Dimension reduction performed by computing Principal Component Analysis (PCA) on the training data of the TB-Places dataset. We evaluated different feature normalization, namely no normalization, L1 and L2 norms. In the last column, we report the percentage of dimensions that are discarded by the PCA.	23
2.4	Average Precision that we achieved using the considered descriptors, with and without normalization and PCA, using Cosine (C), and Euclidean (E) distances, on the Wageningen 2017 and Renningen 2017 test sets.	25
2.5	Pose regression results (errors) on the W17 test set.	29
3.1	Details on the composition of the extended TB-Places dataset with the new W18 set of images. We report, for each subset, the number of image pairs labelled as similar and their percentage among all the possible image pairs.	36
3.2	Details on the descriptors that we considered in the benchmark analysis.	40
3.3	Details on the trained models. The second column displays the number of features of the learned holistic descriptors. The selected α values correspond to the threshold that produces the best F_1 -score in the training set.	42
3.4	Achieved performance (Average Precision) in W17 and W18 datasets before (Baseline) and after training the models.	42

4.1	Ablation study on the considered datasets: all the models are trained on the MSLS train set. CL stands for Contrastive Loss and while GCL for our Generalized Contrastive Loss. For the cases in which PCA whitening is applied we report the dimensionality that achieves the best results on the MSLS validation set.	72
4.2	Comparison of our results with those of state-of-the-art approaches on the considered datasets. For each model, we report if PCA whitening is used and the dimensionality of the learned image latent vector. The best result by our method is underlined, and the overall best is presented in bold font.	73
4.3	Average Precision results obtained by the networks trained with the proposed Generalized Contrastive loss function on the 7Scenes dataset, compared with those achieved by the same network architectures trained using the binary Contrastive loss function and by the NetVLAD off-the-shelf model.	78
4.4	Recall@5 in the MSLS validation and test sets for GCL models when using different batch composition strategies. Strategy A is 50% $\psi \in [0.5, 1]$, 25% $\psi \in (0, 0.5)$ and 25% $\psi = 0$. B is 25% $\psi \in [0.75, 1]$, 25% $\psi \in [0.5, 0.75]$, 25% $\psi \in (0, 0.5)$ and 25% $\psi = 0$. For C, we have 33.3% $\psi \in [0.5, 1]$, 33.3% $\psi \in (0, 0.5)$ and 33.3% $\psi = 0$. Finally, for strategy D we have 50% $\psi \in [0.5, 1]$ and 50% $\psi \in [0, 0.5)$	79
4.5	Results achieved on the MSLS validation set by our models trained by backpropagating the gradient only partially through the network. The <i>last trained layer</i> column indicates the block of the backbone until which the gradient is backpropagated.	81
4.6	Training time on MSLS using one NVIDIA V100 GPU. In the second column, the number in parenthesis is the amount of time a model took to converge. The total training time we report is until convergence.	83
4.7	Ablation study on the considered datasets: all the models are trained on the MSLS train set and deploy a global average pooling layer. CL stands for Contrastive Loss and while GCL for our Generalized Contrastive Loss. For the cases in which PCA whitening is applied we report the dimensionality that achieves the best results on the MSLS validation set.	87
4.8	Generalization results of the models trained on the MSLS dataset for the Pittsburgh250k and TokyoTM benchmarks.	89
4.9	Detailed results on the Extended CMU dataset. The symbol * denotes the models for which PCA whitening has been applied.	90

LIST OF TABLES

4.10	Detailed results on the RobotCar Seasons v2 dataset, divided by wheather and ilumination conditions. The symbol * denotes the models for which PCA whitening has been applied.	91
4.11	Comparison with Thoma et al. (2020), that outperformed several triplet-based loss functions on the CMU Seasons dataset.	92
4.12	Median translation and rotation errors (in cm and degrees, respectively) on the 7Scenes dataset using the descriptors computed using the models trained with the binary Contrastive Loss and the Generalized Contrastive Loss for retrieval and InLoc for localization.	95
5.1	Comparison to state-of-the-art methods trained on the MSLS dataset. The mark * denotes methods that perform re-ranking. We underline the best results by our methods and show the best results overall in bold.	112
5.2	Comparison on recall, Mean Reciprocal Ranking and Kullback-Leibler divergence between the same model trained with a GCL (Leyva-Vallina et al. (2023)) and an MSE (ours).	115
5.3	Ablation study in encoder (NetVLAD, VGG16 GeM, ResNeXt GeM, ViT and ViT-R50) and PCA. All considered models are trained on the MSLS dataset by optimizing an MSE loss.	116
5.4	Comparison between equivalent models trained with a GCL (Leyva-Vallina et al. (2023)) and an MSE (ours), using the Cosine and the Euclidean distance.	120

