SeqMatchNet: Contrastive Learning with Sequence Matching for Place Recognition & Relocalization

Sourav Garg, Madhu Vankadari and Michael Milford

Centre for Robotics — QUT — University of Oxford

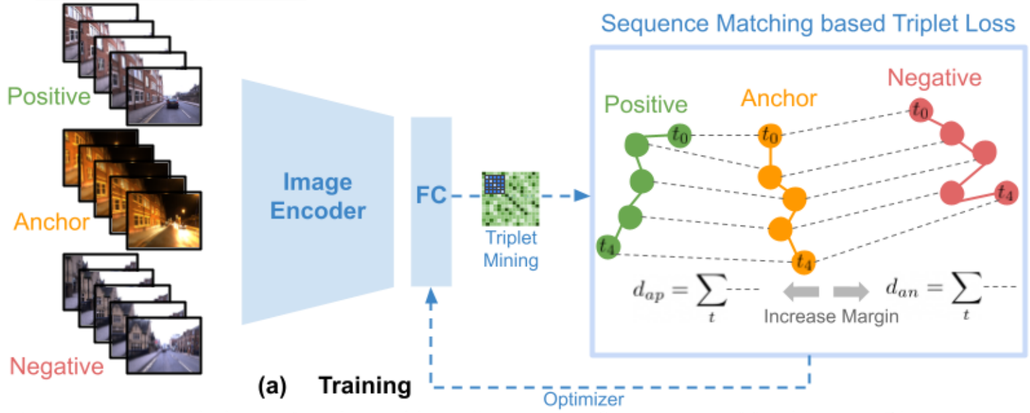
CoRL 2021

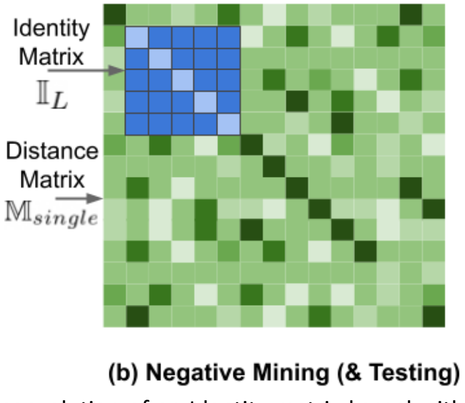
Research Problem:

A mobile robot can be localized by recognizing a previously seen image (or a sequence of images) of a revisited place. This can be achieved through sequence-based Visual Place Recognition (VPR), which often requires robust single image representations along with sequence matching to deal with challenging appearance variations such as those caused by day-night or seasonal cycles.

In this work, for the first time, we bridge the gap between single image representation learning and sequence matching through *SeqMatchNet,* which transforms the single image descriptors such that they become more responsive to the sequence matching metric.

Proposed Approach:





Images are encoded using existing global descriptor method (e.g. NetVLAD).

A learnable linear transform (FC) is applied *independently* to single image descriptors.

Sequence Matching: Distance between two sequences is computed as the sum of temporal-order corresponding pairwise Euclidean distances.

2D convolution of an Identity matrix kernel with the single image-based distance matrix is introduced for sequence matching to efficiently mine negatives online during training and for testing.

Source Code:
github.com/oravus/SeqMatchNet

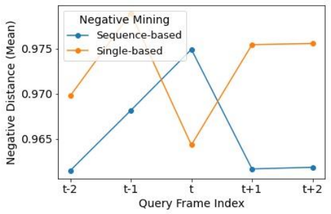
Results & Analyses:

Recall@1/5/20 when using single/sequence based loss and negative mining for representation learning. It can be observed that using sequence matching metric consistently improves performance.

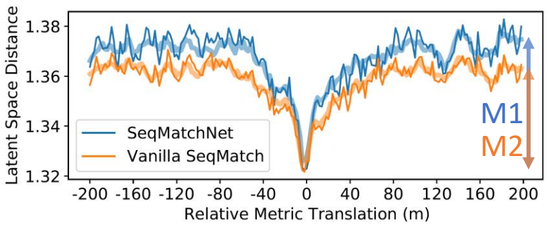
	Loss Type	Mining Type	Oxford	Nordland
Existing:	Single	Single	0.76/0.89/0.97	0.59/0.75/0.86
Proposed:	SeqMatch	Single	**0.79/0.90**/0.97	0.61/0.76/0.87
	Single	SeqMatch	0.76/0.89/0.97	**0.66/0.81/0.91**
	SeqMatch	SeqMatch	*0.78*/0.89/0.97	**0.66/0.81/0.91**

Qualitative Matches:

	Methods	Oxford	Brisbane
No Sequence:	NetVLAD [4]	0.47/0.70/0.85	0.20/0.28/0.41
Sequential Descriptor:	Smoothing [94]	0.59/0.72/0.85	0.20/0.25/0.32
	Delta [94]	0.37/0.55/0.74	0.20/0.33/0.50
	SeqNet (S_5) [95]	**0.62/0.76/0.88**	***0.32/0.40/0.55***
Sequence Matching:	NetVLAD [4]	0.67/0.79/0.90	0.21/0.27/0.37
	S_1 [95]	***0.71***/0.83/0.93	0.28/0.36/0.48
	S_1 [95] + GISM [58]	0.65/-/-	0.26/-/-
	S_1 [95] + GRH [5]	0.34/-/-	0.18/-/-
	Ours: SeqMatchNet	0.70/***0.84/0.94***	**0.29/0.38/0.50**
Hierarchical:	HVPR (S_5, S_1) [95]	***0.71***/0.82/0.88	0.29/***0.40/0.55***
	Ours: HVPR (S_5, SeqMatchNet)	***0.71***/0.82/0.88	**0.30/*0.40/0.55***



While single image based negative mining (low distance = harder) only focuses on standalone hardest negatives (orange dip at t), sequence matching mines negatives considering the neighbourhood of the images as well, even when the central element (t) of the sequence is not necessarily the hardest negative (blue curve).



Distance margins between the correct match and false positives are higher (M1 > M2) for learnt SeqMatchNet (blue) than the vanilla NetVLAD based sequence matching (orange). Day-Night VPR with train/test as Brisbane/Oxford.

State-of-the-art VPR Benchmarking using Recall@1/5/10:

We compare different modes of VPR using sequential information including *Sequential Descriptors* where a sequence of single descriptors is aggregated into a summary vector, *Sequence Matching* where a sequence of single match scores is aggregated using various techniques, and a *Hierarchical* combination of both.