Rotation Equivariant Siamese Networks for Tracking

Deepak K. Gupta*, Devanshu Arya†, Efstratios Gavves*
*QUVA Lab, University of Amsterdam, The Netherlands
†Informatics Institute, University of Amsterdam, The Netherlands
{d.k.gupta,d.arya,e.gavves}@uva.nl

A. Datasets

We provide here the details related to the three datasets that have been used to benchmark the performance of RE-SiamNets in this paper. The three datasets are Rot-MNIST, Rot-OTB100 and ROB sequences. Details follow below.

A.1. Rot-MNIST

Rot-MNIST, as the name implies, comprises rotating MNIST digits on backgrounds of natural images extracted from the original GOT-10k [2] training set. Each video comprises MNIST digits floating around and the target digit rotates. Both the motions, translation as well as rotation, are governed by Brownian equation. The training set comprises 2500 sequences exhibiting only translational motion, while the test set contains 100 sequences comprising translation as well as rotation. Each sequence of the training set as well as the test set contains 100 frames. Randomly sampled frames from 3 sequences of the test set are shown in Figure 1.

A.2. Rot-OTB100

Rot-OTB100 is a dataset built through rotating the image frames of the original OTB100 dataset. For this purpose, each sequence is taken and starting from the first frame, every frame is rotated 0.5 degrees counter-clockwise with respect to its previous frame. For keeping the dataset structure similar to that of OTB, we do not use rotated bounding boxes, rather the regular ones. The bounding box is chosen in a way such that the rotated version of the original bounding box (obtained after rotating) fits tightly within it. Thus, Rot-OTB100 is a testset with exactly same number of sequences as OTB100, except that the frames are rotated.

A.3. ROB

Details of Rotating Object Benchmark (RTB) as well as example sequences have already been shown in the main paper. All the 35 videos have been acquired at 10 fps and comprise 300-500 frames each.

B. Implementation details

B.1. Models

The discussion on the model details are provided below with respect to the three benchmarking datasets as outlined earlier.

Rot-MNIST. For Rot-MNIST, we develop a reduced rotation equivariant variant of SiamFC [1], comprising 999K paramters. RE-mSiamFC differs in terms of the size of the kernels used. Comparisons are made with the non-rotated equivariant of the same model comprising equal number of parameters. For 4 rotational groups (R4), the resultant model comprises 5 convolutional layers, comprising 62, 75, 157 and 160, and kernels of sizes 3×3 in all the layers. Note that the number of channels for any choice of rotational groups is made such that the number of parameters is approximately 999K. Further, all except the last layer are followed by Batchnorm and ReLU activation layers. Finally, pooling is used across the rotational groups to obtain a single set of feature maps from the last layer.

The baseline model for comparison is the non-rotational equivariant version comprising similar number of parameters. This model is referred as mSiamFC. The model is similar to that of RE-mSiamFC, except two differences. First, all rotation equivariant modules are replaced with non-rotation equivariant counterparts. Further, the 4 layers comprise 96, 128, 256 and 256 in the four layers, respectively.

Rot-OTB100 and ROB. Compared to Rot-MNIST, Rot-OTB100 and ROB datasets are relatively more complex and larger models are needed. In this regard, we build two rotation equivariants of SiamFC, referred as *RE-SiamFC* and RE-SiamFCv2. RE-SiamFC has an architecture similar to that of the original SiamFC [1], and comprises 2.33M parameters approximately. The 5 convolutional layers comprise 72, 160, 240, 240 and 160 channels, respectively. The respective kernel sizes are 11×11 and 5×5 in the first two layers, and 3×3 in the last three layers. The choice of padding, stride and pooling is similar to that of SiamFC [1], but rotation equivariant. The number of channels stated

Figure 1: Sampled frames from 3 sequences of the test set of Rot-MNIST dataset. The backgrounds are taken from sequences of GOT-10k dataset [2]. Further, to avoid clutter around the target, we have avoided labelling the bounding boxes in the examples above.

above are for R4. For R8 and R16, these are scaled down keeping the number of parameters same.

RE-SiamFCv2 is similar to RE-SiamFC, except that it uses larger kernel sizes, thus reduced number of channels, thereby keeping the number of parameters equal to 2.33M. It uses 4 convolutional layers, and for R4, these layers are composed of 64, 96, 128 and 163 channels, respectively. The corresponding kernel sizes are 9×9, 7×7, 7×7 and 6×6, respectively. Accordingly, the number of channels for R8 and R16 are 49, 71, 85, 118 and 36, 48, 60, 80, respectively. All except the last convolutional batchnorm and ReLU activation layers, and pooling is performed across the different groups after the last convolutional layer.

B.2. Extension: Training details.

To train RE-mSiamFC, RE-SiamFCv1 and RE-SiamFCv2, we follow a training procedure similar to the default SiamFC [1][1] Each model was trained for 50 epochs with batch size of 8 on a single NVIDIA GTX GPU. For R16 variants, we use batch sizes of 8 and 150 epochs. The initial learing rate is set to 1e-2 and it is decayed to 1e-5 during the course of training. The weight decay and momentum terms are set to 1e-4 and 0.9, respectively. For training the models, we use GOT-10k training set.

For RE-SiamRPN++, we follow training details similar to the baseline SiamRPN++ model. We separately trained a ResNet50 architecture using rotation equivariant modules. This backbone was trained for 50 epochs using batch sizes of 128. The details of model training are same as the standard training of SiamRPN++, as specified in `pysot`[2] pytorch library.

Type	Range	$SR_{0.1}$	$SR_{0.3}$	$SR_{0.5}$	$SR_{0.7}$	$SR_{0.9}$
R4	$\pm\frac{\pi}{4}$	0.52	0.56	0.61	0.73	0.95
R8	$\pm\frac{\pi}{8}$	0.48	0.52	0.60	0.73	0.95
	$\pm\frac{\pi}{4}$	0.67	0.72	0.79	0.87	0.98
R16	$\pm\frac{\pi}{16}$	0.12	0.14	0.16	0.32	0.87
	$\pm\frac{\pi}{8}$	0.17	0.20	0.22	0.38	0.88
	$\pm\frac{\pi}{4}$	0.30	0.34	0.38	0.51	0.92

Table 1: Performance scores measured in terms of success rate at different overlap thresholds for orientation estimation using SiamFCv2 for Rot-OTB100.

C. Extension: Results

We show here a few additional results related to our experiments. Figure 2 shows a few examples of predictions made by SiamFCv2 as well as the equivalent RE-SiamFCv2 variant. Further, we provide precision and success plots of OPE for SiamFCv2-R8 on Rot-OTB100 in Figure 3. Further, we show success rates at different overlap thresholds on different orientation estimates in Table 1.

[1]For SiamFC, we use the Pytorch code available at `https://github.com/huanglianghua/siamfc-pytorch`.

[2]For SiamRPN++ and its rotation equivariant modifications, we use `pysot` library available at `https://github.com/STVIR/pysot`.

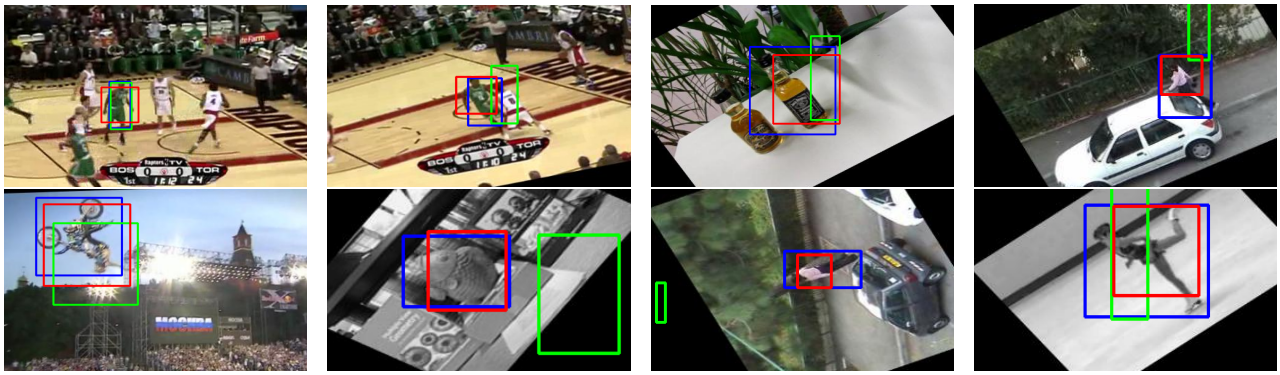

Figure 2: Results on Rot-OTB100 obtained with SiamFC-Netv2 (green) and RE-SiamFCv2 with R8 (red).

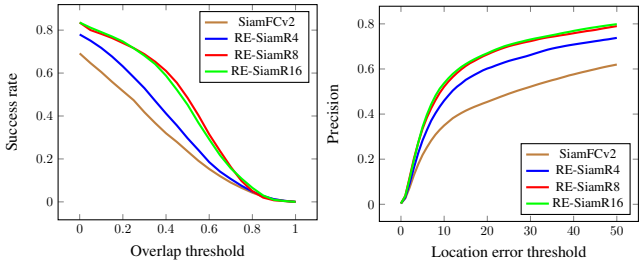

Figure 3: Performance curves for Rot-OTB100 dataset obtained using SiamFCv2 and RE-SiamNet with different choices of equivariant rotation groups. All networks chosen here used 233K optimization parameters.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision – ECCV 2016 Workshops*, pages 850–865, 2016. 1, 2

[2] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2