Discriminative Learning of Deep Convolutional Feature Point Descriptors

Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, Francesc Moreno-Noguer

1 Waseda University, Tokyo, Japan
2 CVLab, École Polytechnique Fédérale de Lausanne, Switzerland
3 Catchoom Technologies, Barcelona, Spain
4 CentraleSupélec and INRIA-Saclay, Chatenay-Malabry, France
5 Institut de Robótica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain

1 First two authors contributed equally

Objective
- Learn compact, discriminative representations of image patches with Convolutional Neural Networks.
- Optimize for comparisons with the L2 norm, i.e., no metric learning. Our descriptors work within existing pipelines.

Main features
- Drop-in replacement for SIFT: 128f, compared with the L2 norm.
- Consistent improvements over the state of the art.
- Trained in one dataset, but generalizes very well to scaling, rotation, deformation and illumination changes.
- Computational efficiency (on GPU: 0.76 ms; dense SIFT: 0.14 ms). Code is available: [GitHub.com/etrulls/deepdesc-release](https://github.com/etrulls/deepdesc-release)

Key observation
1. We train a Siamese architecture with pairs of patches. We want to bring matching pairs together and otherwise pull them apart.
2. Problem: Randomly sampled pairs are already easy to separate.
3. Solution: To train discriminative networks we use hard negative and positive mining. This proves essential for performance.

Model & Training
Our model is a 3-Layer Convolutional Neural Network. For training we use a siamese architecture with weight sharing and SGD.

- Input size: 64 × 64
- Filter size: 7 × 7, 6 × 6, 5 × 5
- Output channels: 32, 64, 128
- Pooling & Normalization: 2 × 2, 3 × 3, 4 × 4
- Non-linearity: Tanh, Tanh, Tanh
- Stride: 2, 3, 4

We minimize the hinge embedding loss with 3D point indices $p_1, p_2$:

$$l(x_1, x_2) = \max(0, C - \|D(x_1) - D(x_2)\|_2)$$

This penalizes corresponding pairs that are placed far apart, and non-corresponding pairs that are less than $C$ units apart.

Methodology: Train over two sets and test over third (leave-one-out), with cross-validation. Metric: precision-recall (PR). 'Needle in a haystack' setting: pick 10k unique pairs and generate one positive pair and 1k negative pairs for each, i.e. 10k positives vs. 10M negatives. Results summarized by 'Area Under the Curve' (AUC).

Effect of mining
(a) Forward-propagate positives $n_p \geq 128$ and negatives $n_n \geq 128$.
(b) Pick the 128 with the largest loss (for each) and back-propagate.

We perform random SIFT mining, but this allows us to train discriminative models with a small number of parameters (−45k), which also alleviates overfitting concerns.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Training</th>
<th>Test</th>
<th>LY</th>
<th>ND</th>
<th>YO</th>
<th>LY+ND</th>
<th>LY+YO</th>
<th>LY+YO+ND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>LY+ND</td>
<td>0.993</td>
<td>0.630</td>
<td>0.456</td>
<td>0.218</td>
<td>0.088</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LY+YO</td>
<td>0.959</td>
<td>0.677</td>
<td>0.424</td>
<td>0.195</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>YO+ND</td>
<td>0.922</td>
<td>0.685</td>
<td>0.439</td>
<td>0.228</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG [4]</td>
<td>LY</td>
<td>0.894</td>
<td>0.632</td>
<td>0.400</td>
<td>0.174</td>
<td>0.067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG [4]</td>
<td>ND</td>
<td>0.880</td>
<td>0.596</td>
<td>0.372</td>
<td>0.182</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG [4]</td>
<td>YO</td>
<td>0.879</td>
<td>0.582</td>
<td>0.365</td>
<td>0.166</td>
<td>0.064</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daisy [6]</td>
<td>-</td>
<td>0.835</td>
<td>0.594</td>
<td>0.363</td>
<td>0.172</td>
<td>0.032</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIFT [5]</td>
<td>-</td>
<td>0.722</td>
<td>0.532</td>
<td>0.308</td>
<td>0.138</td>
<td>0.053</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 1: PR curve, training LY+YO, test ND.*

Generalization: Wide-Baseline Matching
Data from [8]. We match a set of points from view 3 against 4 to 8 (increasing baseline) and build PR curves, as before. No re-training.

Generalization: Deformation and Illumination
Our models outperform the state-of-the-art on illumination changes and non-rigid deformations [8] without re-training or fine-tuning.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Training</th>
<th>Deformation</th>
<th>Illumination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>LY+YO</td>
<td>LY+YO</td>
<td>LY+YO+ND</td>
</tr>
<tr>
<td>Ours</td>
<td>LY+YO</td>
<td>LY+YO</td>
<td>LY+YO+ND</td>
</tr>
<tr>
<td>Ours</td>
<td>LY+YO</td>
<td>LY+YO</td>
<td>LY+YO+ND</td>
</tr>
<tr>
<td>Ours</td>
<td>LY+YO</td>
<td>LY+YO</td>
<td>LY+YO+ND</td>
</tr>
</tbody>
</table>

References