Beyond Cartesian Representations for Local Descriptors

Patrick Ebel[1] , Anastasiia Mishchuk[1] , Kwang Moo Yi[2], Pascal Fua[1], Eduard Trulls[3]
1. École Polytechnique Fédérale de Lausanne, {patrick.ebel,anastasiia.mishchuk,pascal.fua}@epfl.ch
2. Visual Computing Group, University of Victoria, kyi@uvic.ca
3. Google Switzerland, trulls@google.com

Summary

Key idea: build better descriptor representations by resampling the patch with a **log-polar sampling scheme**.

1. Allows us to match when SIFT scale detections fail.
2. Can leverage much larger support regions (e.g. 64x larger than SIFT).
3. State-of-the-art results on three datasets.
4. Training code, data and models are available (**https://github.com/cvlab-epfl/log-polar-descriptors**).

Log-Polar sampling

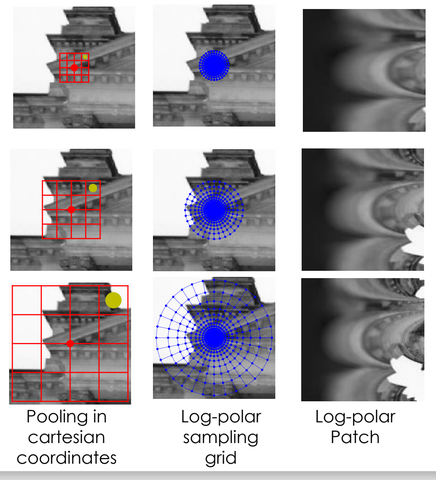
Given an image I of size H × W, a keypoint on I is fully described by its center coordinates, its scale and its orientation.

To extract a PS × PS patch around keypoint we use Polar Transformer Network that transforms coordinates accordingly:

$$x_i^s = x_i + e^{log(r_i)x_i^t/W} cos(\varphi_i)$$
$$y_i^s = y_i + e^{log(r_i)x_i^t/W} sin(\varphi_i)$$

(x_i^s, y_i^s) - source coordinates (x_i^t, y_i^t) - target coordinates

$\varphi_i = \frac{i+2\pi y_i^t}{H}$ - angle $\frac{\lambda}{2}\sigma_i$ - r (radius)

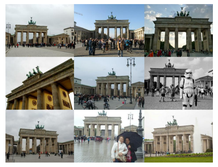


Rotations in cartesian space correspond to shifts on the polar axis in log-polar space (rotation equivariance)

Peripheral regions are undersampled, which enables us to leverage much larger support regions. Paired patches look similar even under drastic scale changes (scale equivariance).

Datasets

PhotoTourism dataset (train/evaluation)



Dataset contains 15 sequences for training (25.6k images) and validation and 11 for testing (4107 images).

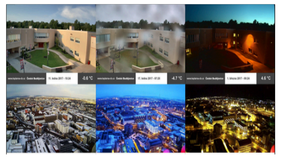
Matching correspondences are obtained by extracting SIFT keypoints and using the depth maps and applying multiple constraint checks .

HSequences dataset (evaluation)

We've used a subset of the HSequences dataset which contains of 57 sequences with significant geometric changes and 59 sequences with photometric changes.

We extracted SIFT keypoints and determine correspondences among them using provided homographies.

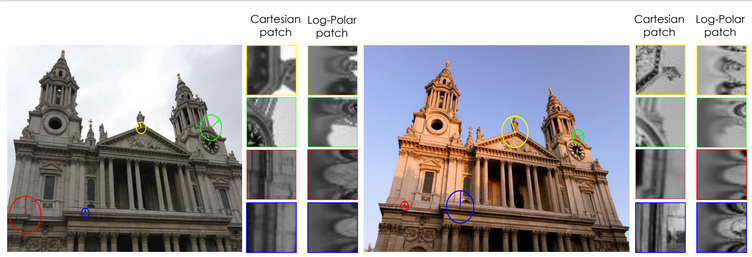
AMOS patches dataset (evaluation)



We've used a subset of the AMOS patches dataset, which was created from a large set of outdoor webcams to evaluate local feature descriptors on illumination and appearance changes.

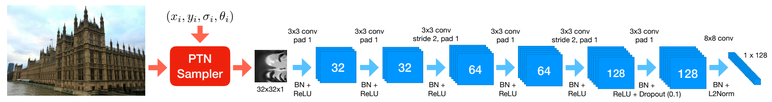
We used given keypoints and correspondences to create subset for the retrieval task.

Cartesian sampling vs. Log-polar sampling

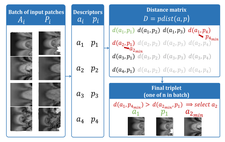


Two images taken from different viewpoints with four pairs of corresponding keypoints.

Architecture



We extract patches size 32×32 in-line with a sampler (pictured: PTN) on the desired keypoints following a HardNet architecture, for the descriptor



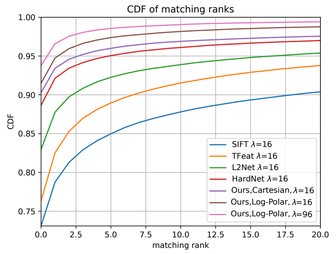
Loss: We use a triplet loss with in-batch hard negative mining

$$\mathcal{L} = \frac{1}{n}\sum_{i=1,n} \max\left(0, 1 + d(a_i, p_i) - \min\left(d(a_i, p_{j_{min}}), d(a_{k_{min}}, p_i)\right)\right)$$

PhotoTourism dataset evaluation

Sequence	SIFT	TFeat	L2-Net	Geodesc	HardNet	Ours ($\lambda = 16$) Cart	Ours ($\lambda = 16$) LogPol	Ours ($\lambda = 96$) LogPol
'british_museum'	7.52	4.05	3.94	4.38	2.86	2.78	2.29	**1.13**
'florence_cathedral_side'	9.30	4.44	1.65	9.42	0.77	0.50	0.30	**0.27**
'lincoln_memorial_statue'	6.47	9.61	5.48	6.69	3.59	2.77	1.86	**1.65**
'milan_cathedral'	19.01	6.13	4.60	7.97	1.23	0.76	0.46	**0.22**
'mount_rushmore'	40.08	20.12	9.63	12.11	2.57	1.86	0.78	**0.71**
'reichstag'	5.90	1.43	5.91	2.12	0.36	0.21	0.17	**0.13**
'sagrada_familia'	24.79	9.10	5.50	9.98	1.67	1.08	0.44	**0.19**
'st_pauls_cathedral'	15.40	4.53	3.05	10.76	1.26	0.72	0.55	**0.23**
'united_states_capitol'	16.68	7.32	5.41	15.60	3.45	1.81	1.02	**0.59**
Average	16.13	7.42	5.02	8.78	1.97	1.39	0.87	**0.57**

False positive rate at true positive rate equal to 95% (FPR95), is reported (lower is better)
We benchmark our models against the baselines with patches extracted at the SIFT scale, λ = 12.

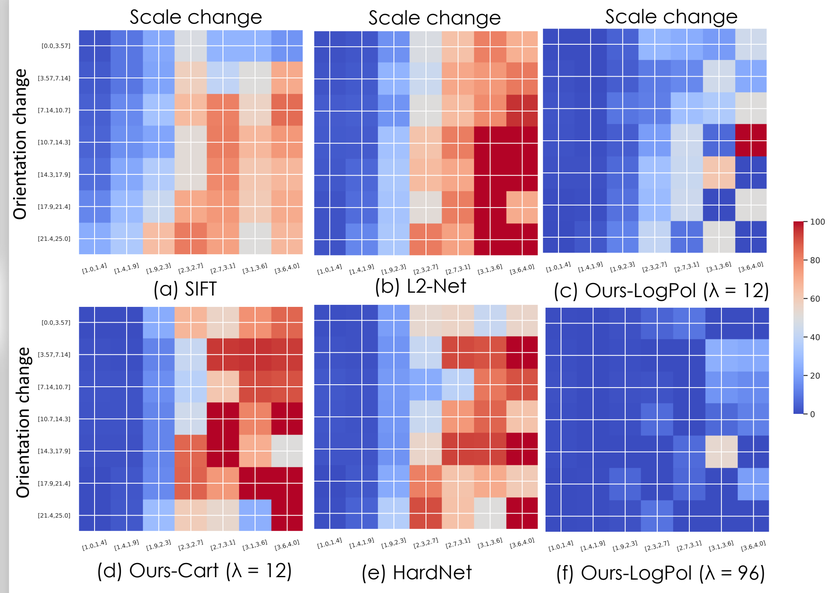


Patch retrieval. The cumulative distribution function of the rank in a patch retrieval scenario with a large number of distractors (higher is better).

Type	Method	Stereo task mAP[15°]	Stereo task Rank†	Multi-view task mAP[15°]	Multi-view task Rank†
DoG	SIFT (IJCV'04)	0.0277	9	0.4146	8
	TFeat (BMVC'16)	0.0357	8	0.4643	7
	L2-Net (CVPR'17)	0.0400	6	0.5087	5
	HardNet (NIPS'17)	0.0425	4	**0.5481**	1
	GeoDesc (ECCV'18)	0.0368	7	0.5298	4
	ContextDesc (CVPR'19)	0.0439	3	0.5399	3
e2e	SuperPoint (CVPR'18)	0.0415	5	0.4778	6
	D2-Net (CVPR'19)	**0.0490**	1	0.3967	9
Ours (DoG)	Ours-Cartesian, $\lambda = 16$	0.0405	—	0.5208	—
	Ours-LogPol, $\lambda = 32$	0.0420	—	0.5389	—
	Ours-LogPol, $\lambda = 64$	0.0432	—	0.5396	—
	Ours-LogPol, $\lambda = 96$	**0.0448**	2	**0.5427**	2

PhotoTourism challenge. Metric is mean average precision in pose estimation (higher is better).
We rank 2nd on both tracks, and 1st on average.

Influence of scale and orientation changes



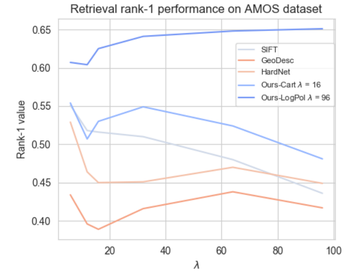
PhotoTourism dataset FPR95 vs Scale and orientation changes.

- All baselines degrade quickly under scale changes.
- Training deep networks with cartesian patches with scale changes is not sufficient.
- By contrast, our log-polar representation enables to learn scale invariance.

HSequences, AMOS patches: retrieval task

Method	Viewpoint split	Illumination split
SIFT, $\lambda = 12$	0.740	0.607
HardNet, $\lambda = 12$	0.813	0.707
GeoDesc, $\lambda = 12$	**0.879**	0.727
Ours-Cart, $\lambda = 12$	0.828	0.722
Ours-Cart, $\lambda = 16$	0.831	0.732
Ours-Cart, $\lambda = 32$	0.825	0.736
Ours-Cart, $\lambda = 64$	0.752	0.666
Ours-Cart, $\lambda = 96$	0.681	0.616
Ours, LogPol, $\lambda = 12$	0.833	0.729
Ours, LogPol, $\lambda = 16$	0.838	0.743
Ours, LogPol, $\lambda = 32$	0.849	0.764
Ours, LogPol, $\lambda = 64$	0.849	0.774
Ours, LogPol, $\lambda = 96$	0.847	**0.774**

Results on HPatches.
Rank-1 performance (% of keypoints that are found in the first positions while ranking across all the keypoints 1.0 – perfect matching score) on the viewpoint and illumination splits of the HSequences dataset (higher is better).



Results on AMOS patches.
Rank-1 performance on the AMOS patches dataset (higher is better).
For this dataset, extracting descriptors with smaller patches produces better results for most baselines, so we consider λ = 6.

References

[Polar Transformer Networks] Esteves et.al. Polar Transformer Networks , ICLR, 2018

[HardNet] Mishchuk et.al. Working hard to know your neighbor's margins: Local descriptor learning loss, NIPS, 2017