Comparative Evaluation of Hand-Crafted and Learned Local Features

Johannes L. Schönberger[1] Hans Hardmeier[1] Torsten Sattler[1] Marc Pollefeys[1,2]

CVPR July 21-26 2017 HONOLULU

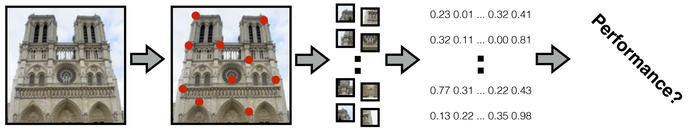
ETH zürich

http://cvg.ethz.ch/research/local-feature-evaluation

Results Table | Benchmark Protocol | Source Code

Overview

Matching **local image features** is a **key task** in **computer vision**. For more than a decade, **hand-crafted** features such as SIFT have been used for this task. Recently, new features **learned** from data have been proposed and shown to improve on SIFT in terms of discriminative power. This work is dedicated to an extensive **experimental evaluation** of local features in a **practical setting**.



Motivation

Most learned features evaluated on patch pair **classification task** measuring *false positive rate at 95% true positive rate (FPR95)* [3]

- Do **better FPR95** scores translate to **better matching** performance? What is the impact of typical filtering steps?
  (e.g., Lowe's ratio test and mutual nearest neighbor constraint to avoid ambiguous matches, geometric verification to prune outliers requires good precision for manageable runtimes)
- **More matches** between **similar images** do **not** necessarily imply a **better performance** under **extreme** illumination and viewpoint **changes**. How well do learned features perform under such conditions?

Benchmark

Single evaluation protocol to benchmark local image feature performance in a practical setting:

- Raw **image-to-image matching** performance
  (under, e.g., blur, exposure, day-night, scale, rotation, planar, internet, etc.)

- **Image-based reconstruction** performance
  (measuring impact of local feature matching performance on Bag-of-Words image retrieval, Structure-from-Motion, and Multi-View Stereo)

Insights

- **Patch classification** performance **does not translate** to more complex image-based **reconstruction task**
- **Previous** image-based reconstruction **datasets too easy** as a benchmark (Fountain, Herzjesu, South Building)
- **Learned features** better than RootSIFT but **not better** than **advanced hand-crafted features still better**
- **Learned features** exhibit **strong variation** in performance for **different datasets**
- Significant **room for improvement**, especially in the **hard cases** where all methods fail (e.g., day-night)

Methods

- **SIFT**: RootSIFT [1]
- **SIFT-PCA**: RootSIFT with PCA projection [4]
- **DSP-SIFT**: Domain-size pooled SIFT [5]
- **ConvOpt**: Learned descriptor using convex optim. [8]
- **DeepDesc**: Deep learned descriptor [7]
- **TFeat**: Shallow learned descriptor [2]
- **LIFT**: Learned keypoint detector and descriptor [6]

We used **pre-trained networks** provided by the authors.

	SIFT	SIFT-PCA	DSP-SIFT	ConvOpt	DeepDesc	TFeat	LIFT
Dimensionality	128	80	128	73	128	128	128
Size [bytes]	128	320	512	292	512	512	512
Platform	CPU	CPU	CPU	GPU	GPU	GPU	GPU
Extraction [s]	9.3	10.5	23.7	49.9	24.3	11.8	212.3
Matching [s]	0.14	0.11	0.14	0.10	0.14	0.14	0.14

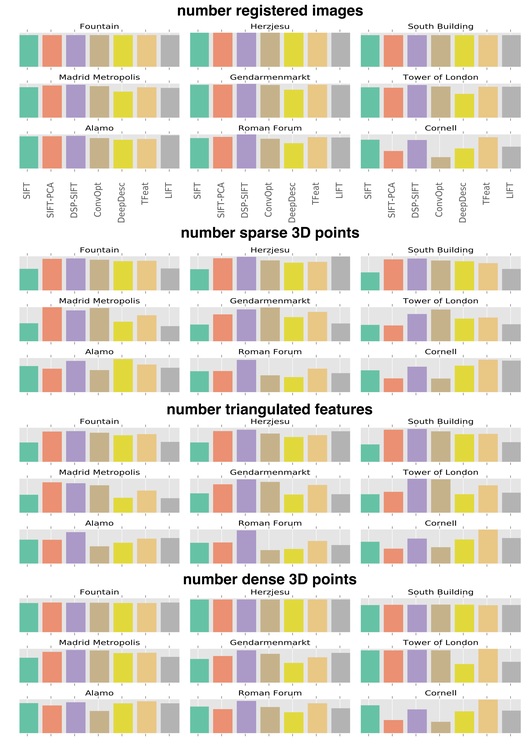
Raw Matching Performance

	SIFT	SIFT-PCA	DSP-SIFT	ConvOpt	DeepDesc	TFeat	LIFT
Putative Match Ratio in %							
Blur	3.7	5.7	7.0	5.2	4.6	4.2	6.5
JPEG	20.9	29.3	34.0	26.8	24.4	22.9	27.5
Exposure	33.0	34.1	35.3	32.8	10.4	31.2	34.9
Day-Night	5.5	6.8	6.2	7.2	3.6	5.5	5.4
Scale	12.1	25.2	23.4	23.8	23.0	21.5	19.6
Rotation	12.8	17.6	17.3	10.0	11.9	8.7	1.3
Scale-rotation	2.4	6.0	5.8	4.7	4.5	3.7	2.0
Planar	5.9	10.0	10.1	9.4	7.7	8.0	8.0
Non-planar	7.8	8.8	8.7	8.4	7.4	7.2	8.3
Internet	3.2	4.6	4.4	4.3	2.7	3.4	4.8
Precision in %							
Blur	43.8	46.5	48.4	45.2	41.9	46.3	44.5
JPEG	98.5	96.5	98.3	94.1	91.6	95.8	95.9
Exposure	99.3	98.0	98.6	96.6	68.0	97.3	97.5
Day-Night	93.8	80.4	77.8	73.9	37.8	76.5	71.2
Scale	43.0	95.5	95.5	92.2	89.1	94.3	89.1
Rotation	33.2	33.1	33.1	32.2	32.3	32.3	7.9
Scale-rotation	32.8	46.7	46.8	42.3	39.1	43.9	18.7
Planar	33.9	37.3	39.9	34.3	32.5	33.6	33.2
Non-planar	43.3	42.2	43.1	38.4	34.5	39.3	40.4
Internet	39.8	40.3	39.7	35.6	27.2	36.6	37.1
Matching Score in %							
Blur	3.7	5.5	6.8	4.9	4.1	4.0	6.2
JPEG	20.8	28.8	33.7	26.1	23.5	22.6	27.1
Exposure	32.8	33.5	34.9	31.8	9.1	30.5	34.2
Day-Night	5.3	5.9	5.5	5.8	1.8	4.7	4.3
Scale	11.7	24.4	22.8	22.6	21.3	20.7	18.2
Rotation	12.8	17.5	17.2	9.7	11.6	8.5	0.9
Scale-rotation	2.4	5.8	5.6	4.3	3.9	3.5	1.6
Planar	5.7	9.6	9.9	8.7	6.9	7.5	7.4
Non-planar	7.7	8.4	8.4	7.8	6.5	6.9	7.7
Internet	3.1	4.1	4.0	3.5	1.8	2.8	4.1
Recall in %.							
Blur	17.0	22.4	27.2	20.0	16.9	17.0	17.9
JPEG	37.9	51.6	62.8	46.6	41.0	39.2	51.5
Exposure	79.0	81.0	84.1	76.5	18.2	73.1	64.0
Day-Night	25.6	29.2	26.2	28.9	8.4	22.9	19.3
Scale	22.4	84.0	73.9	76.1	71.9	68.9	98.4
Rotation	20.8	28.5	28.1	16.1	19.1	14.1	2.3
Scale-rotation	6.4	16.4	15.2	12.0	10.9	9.6	5.3
Planar	11.4	18.0	18.6	16.4	13.3	14.2	17.9

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. CVPR, 2012.
[2] V. Balntas, E. Riba, D. Ponsa, K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. BMVC, 2016.
[3] M. Brown, G. Hua, S. Winder. Discriminative Learning of Local Image Descriptors. PAMI, 2011.

Reconstruction Performance

- Evaluation using Structure-from-Motion and Multi-View Stereo
- Exhaustive image matching for Fountain (11 images), Herzjesu (8 images), South Building (128 images), Madrid Metropolis (1344 images), Gendarmenmarkt (1463 images), Tower of London (1576 images)
- Image retrieval with matching against top-100 retrieved images for Alamo (2915 images), Roman Forum (2364 images), Cornell (6514 images)



[4] A. Bursuc, G. Tolias, and H. Jégou. Kernel local descriptors with implicit rotation matching. ACM Multimedia, 2015.
[5] J. Dong and S. Soatto. Domain-size pooling in local descriptors: DSP-SIFT. CVPR, 2015.
[6] M. Kwang, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned Invariant Feature Transform. ECCV, 2016.
[7] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. ICCV, 2015.
[8] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. PAMI, 2014.

[1]Eidgenössische Technische Hochschule Zürich [2]Microsoft Corp.

This project received funding from the European Union's Horizon 2020 research and innovation program under grant No. 688007 (TrimBot2020).