An Evaluation of Recent Local Image Descriptors for Real-World Applications of Image Matching

Fabio Bellavia, Carlo Colombo, University of Florence, Italy
{fabio.bellavia, carlo.colombo}@unifi.it

Summary

An experimental comparison of the most recent local descriptors is carried out on increasingly complex image matching tasks. The evaluation includes both planar and more challenging non-planar scenes.

Evaluation protocol

- Keypoint extraction by the HarrisZ detector
- Patch alignment by SIFT dominant orientation
- Descriptor computation
- Decriptor matching with the Nearest-Neighbor Ratio (NNR)
- Accuracy performance is computed using automatically generated ground-truth correspondences

Planar scenes

The planar dataset consists of 65 homography-related image pairs (13 sequences of 6 images each) from the Oxford and Viewpoint datasets, mainly including perspective transformations.

Ground-truth correspondences are computed according to the overlap error

$$\mathrm{OE}(E_1, E_{2\to1}) = 1 - \frac{E_1 \cap E_{2\to1}}{E_1 \cup E_{2\to1}}$$

where E_1 is the elliptical patch on the reference image and $E_{2\to1} \sim H^{-\mathrm{T}} E_2 H^{-1}$ is the reprojection onto the reference image of the elliptical patch E_2 on the other image.

Non-planar scenes

The non-planar dataset contains 42 fundamental matrix-related image pairs (each sequence is composed of 2 or 3 images).

Fundamental matrices are estimated using manually selected correspondences and used for automatic ground-truth computation.

The ground truth is not based on SfM, as usually done. It is built using the approximated overlap error

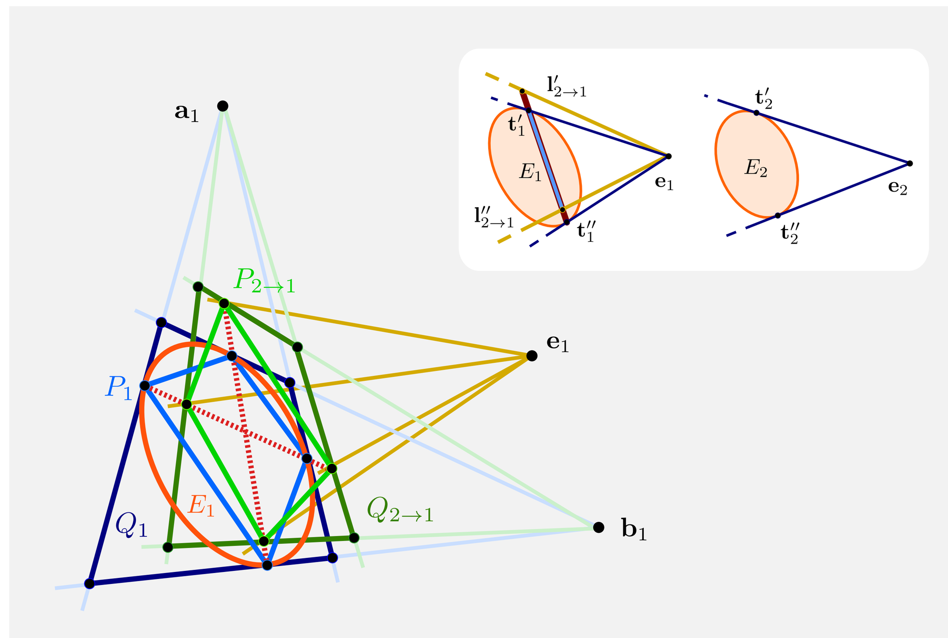
$$\mathrm{AOE} = \frac{\mathrm{OE}(P_1, P_{2\to1}) + \mathrm{OE}(Q_1, Q_{2\to1})}{2}$$

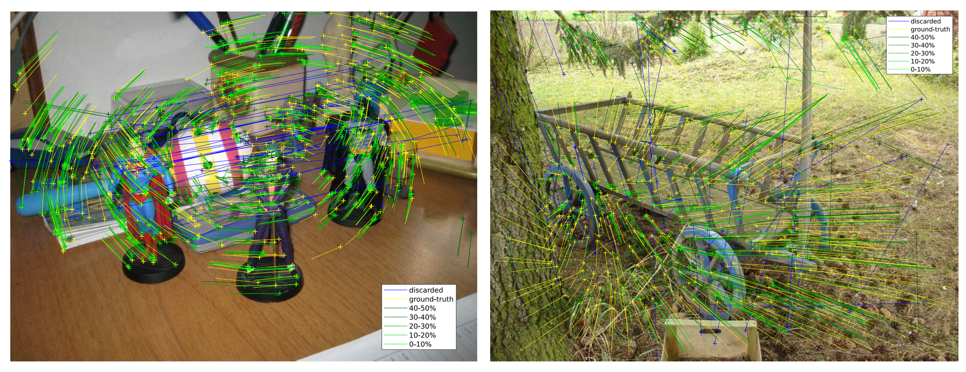
which is an extension of the overlap error for scenes with parallax. Inspired by the linear overlap error, AOE approximates each elliptical patch by a pair of quadrilaterals obtained by tangency and epipolar constraints.

False positives of the ground-truth are filtered out according to local flow length:

$$\| \mathbf{c}_1 - \mathbf{c}_2 \| > \mu + 2.5\sigma$$

where μ and σ are the median and MAD flow values around the putative matches $\mathbf{c}_1$, $\mathbf{c}_2$.





Results

◇ Descriptors are ranked by mean Average Precision (mAP) on non-planar scenes (non-planar).

◇ Each descriptor employs the matching distance that gives the best results. Odd as it may seem, SIFT works better with L_1 than with L_2.

◇ The best performing descriptors are those that capture both the local image context and the global scene structure (see GeoDesc, sGLOH2, SOSNet, HardNet$_A$).

◇ Most descriptors exhibit a gradual performance degradation in the transition from planar, through viewpoint, to non-planar scenes.

◇ Deep descriptors have the best overall performance on all datasets.

◇ Deep descriptors strongly depend on training data. For example, HardNetPS (trained with SfM) and HardNet++ (trained on Brown and HPatches) switch ranking when passing from planar to non-planar scenes.

◇ Some binary descriptors exhibit a good balance between length (i.e., memory storage and computational efficiency) and accuracy.

◇ In MVA 2019 (data from Nov. 2018), the best descriptors of all are the deep GeoDesc (with and without quantization) and the hand-crafted sGLOH2 (binary and non-binary).

◇ The "Which is Which?" (WISW, Apr. 2019) contest included still unpublished descriptors (SOSNet, RsGLOH2 and RalNet Shuffle) and the brand new HardNet$_A$ descriptor.

MVA 2019

		family	rot. inv.	mAP (%) planar	mAP (%) viewpoint	mAP (%) non-planar	dim	type
L_2	GeoDesc$_Q$	dp		78.78	65.03	51.53	128	uchar
	GeoDesc	dp		78.75	65.10	51.51	128	float
	L2-Net$_{CS}$	dp		67.00	54.64	48.12	256	float
	HardNet++	dp		70.73	58.37	47.54	128	float
	HardNetPS	dp		73.94	59.86	45.77	128	float
	L2-Net	dp		59.91	48.62	43.00	128	float
	MIOP	dd	✓	76.36	57.02	40.54	128	float
	DeepDesc	dp		55.38	47.84	38.35	128	float
L_1	sGLOH2	hc	✓	75.64	63.51	50.68	256	uchar
	LIOP	hc	✓	74.11	55.22	39.52	144	uchar
	RootSIFT	hc		63.71	49.09	38.88	128	float
	SIFT	hc		63.93	47.48	37.58	128	uchar
Hamming	BisGLOH2	hc	✓	74.26	61.49	49.31	1152	bit
	BiL2-Net$_{CS}$	dp		61.42	49.35	43.31	256	bit
	RFD$_G$	dd		68.77	55.63	40.25	406	bit
	RFD$_R$	dd		68.26	54.13	38.48	293	bit
	BiL2-Net	dp		48.70	36.58	34.33	128	bit
$*$	MKD	hc		62.65	48.89	40.67	238	float
	MKD$_W$	dd		62.84	48.64	40.10	128	float

family [hc: hand-crafted | dd: data-driven | dp: deep-based]
rotationally invariant $*$ dot product
planar viewpoint only non-planar

WISW 2019 contest

		family	rot. inv.	mAP (%) viewpoint	mAP (%) viewpoint$_+$	mAP (%) non-planar	mAP (%) non-planar$_+$	dim	type
L_2	SOSNet	dp		74.01	76.30	60.76	53.40	128	float
	AffNet+HardNet$_A$	dp		71.71	74.11	59.98	53.34	128	uchar
	HardNet$_A$	dp		72.14	74.29	57.47	50.08	128	uchar
	OriNet+HardNet$_A$	dp		71.14	73.50	57.09	49.92	128	uchar
	L2Net$_{CS}$	dp		66.97	69.49	56.46	48.79	256	float
	GeoDesc$_Q$	dp		71.83	75.60	55.47	47.56	128	uchar
	HardNet++	dp		68.86	71.49	55.37	47.80	128	uchar
	RalNet Shuffle	dp		62.76	65.51	49.75	41.53	128	uchar
	DOAP	dp		67.19	69.80	44.99	41.77	128	float
	DeepDesc	dp		56.32	53.24	44.93	37.03	128	float
	MIOP	dd	✓	52.13	56.83	39.33	33.38	128	float
L_1	RsGLOH2	hc	✓	67.84	70.68	56.11	48.19	256	float
	sGLOH2	hc	✓	63.50	67.25	52.49	44.86	256	uchar
	RootSIFT	hc		56.74	58.46	44.77	37.73	128	float
	LIOP	hc	✓	49.50	54.51	37.93	32.05	144	uchar
Hamming	BisGLOH2	hc	✓	62.27	66.04	51.64	44.08	1152	bit
	BiL2-Net$_{CS}$	dp		61.06	63.11	50.86	43.33	256	bit
	BiDOAP	dp		52.74	54.24	41.41	34.57	128	bit
	RFD$_G$	dd		50.75	53.58	40.40	34.17	406	bit
	RFD$_R$	dd		50.28	52.62	39.31	32.96	293	bit
$*$	MKD$_W$	dd		56.40	59.52	45.70	39.05	128	float

viewpoint$_+$ = viewpoint plus 30 new viewpoint pairs non-planar$_+$ = non-planar plus 30 new non-planar pairs

◇ In WISW 2019, most of the latest deep descriptors outperform the best hand-crafted descriptors.

◇ WISW uses a slightly different evaluation protocol. Specifically, SIFT-based patch alignment is replaced by a deep-based one. Moreover, a symmetric version of NNR is employed for descriptor matching. These changes remarkably improve the matching accuracy.

◇ The WISW dataset extends the MVA dataset with 30 more viewpoint and 30 more non-planar image pairs. The new viewpoint pairs do not add information on descriptor behavior, as the results remain almost unchanged. The new non-planar pairs induce instead a performance loss. Hence, current descriptors cannot successfully deal yet with many patch deformations occurring in non-planar scenes.