IET Computer Vision

Research Article

IET Journals
The Institution of Engineering and Technology

RootsGLOH2: embedding RootSIFT 'square rooting' in sGLOH2

ISSN 1751-9632
Received on 6th September 2019
Revised 6th September 2019
Accepted on 30th January 2020
doi: 10.1049/iet-cvi.2019.0716
www.ietdl.org

Fabio Bellavia[1] ✉, *Carlo Colombo*[2]

[1]*Department of Mathematics and Computer Science, Università degli Studi di Palermo, Palermo, Italy*
[2]*Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy*
✉ *E-mail: bellavia.fabio@gmail.com*

Abstract: This study introduces an extension of the shifting gradient local orientation histogram doubled (sGLOH2) local image descriptor inspired by RootSIFT 'square rooting' as a way to indirectly alter the matching distance used to compare the descriptor vectors. The extended descriptor, named RootsGLOH2, achieved the best results in terms of matching accuracy and robustness among the latest state-of-the-art non-deep descriptors in recent evaluation contests dealing with both planar and non-planar scenes. RootsGLOH2 also achieves a matching accuracy very close to that obtained by the best deep descriptors to date. Beside confirming that 'square rooting' has beneficial effects on sGLOH2 as it happens on scale invariant feature transform, experimental evidence shows that classical norm-based distances, such as the Euclidean and Manhattan distances, only provide suboptimal solutions to the problem of local image descriptor matching. This suggests matching distance design as a topic to investigate further in the near future.

1 Introduction

Local image descriptors are fundamental for many computer vision applications such as image stitching [1], three-dimensional reconstruction [2], and visual odometry [3]. The relevant role played by local descriptors has granted an active interest in this research topic over the decades, still evolving together with the demand for the related applications.

The most common convention is to classify local image descriptors into handcrafted and data-driven [4] according to how descriptor vectors are extracted from the neighbourhood of local keypoints, carrying salient content in images [5].

Handcrafted descriptors mainly employ histograms to accumulate statistics reflecting some local patch property. The scale invariant feature transform (SIFT) descriptor [6], based on gradient orientation histograms, is the most popular local descriptor, due to its efficiency, robustness, and accuracy in general and common application scenarios. Other histogram-based descriptors use pixel ordering [7], Haar wavelets [8], kernel convolutions [9] or intensity value comparisons [10, 11]. However, most of the histogram-based descriptors are largely inspired by the authors of [12, 13] or even direct variants of SIFT [14–16]. RootSIFT [15] is a popular SIFT variant that replaces the Euclidean distance with the Hellinger's distance, which is more reliable for histogram comparisons. RootSIFT is nowadays considered as the true SIFT replacement due to the minimum amount of changes it requires in the descriptor matching process and its improved performances over the original SIFT.

Data-driven descriptors are those whose behaviour is tuned and refined according to data. The aim is to obtain low-dimensional binary descriptors [17, 18], to find an optimal parameter setup [19, 20], or both these objectives simultaneously [4, 21]. Quite recently, data-driven deep descriptors [22–24] have emerged, leveraging deep learning, modern hardware capability offered by graphics processing units (GPUs), and the availability of large datasets for training [25, 26]. Deep descriptors have shown in recent evaluations to outperform all other kinds of descriptors [27].

Despite the current research trend, strongly focused on deep descriptors, handcrafted descriptors still play a key role in descriptor design. Indeed, often handcrafted descriptors have been the source of inspiration for successful deep descriptors architectures. This is especially true for some recent state-of-the-art deep descriptors [28–30] that can be seen as efficient, parameter-optimised versions of the handcrafted descriptors under consideration. On the other hand, when computational efficiency on low-end or restricted hardware is demanded for non-deep descriptors that do not mandatorily require GPUs to run still provide the most efficient solutions.

Notwithstanding the recent advancements in the field, descriptor matching accuracy is still today far from perfect, especially when considering complex three-dimensional scenes. This justifies the continue efforts for improving descriptors, both the data-driven and the handcrafted. Among the latter, the recent shifting gradient local orientation histogram doubled (sGLOH2) descriptor [13] is currently one of the best in terms of matching accuracy. Inspired by the RootSIFT successful approach, in this study, sGLOH2 is further improved. The resulting RootsGLOH2 descriptor is shown to yield the best matching accuracy among state-of-the art non-deep descriptors, as witnessed by the results of recent evaluation contests on both planar and more challenging non-planar scenes. RootsGLOH2 performance is also very close to that of the best deep descriptors when the standard matching pipeline is employed.

The paper is organised as follows. Section 2 gives an overview of the current research on local image descriptors. In Section 3, RootsGLOH2 is defined after providing a brief description of its sources of inspiration, namely sGLOH2 and RootSIFT. Section 4 reports and discusses the results of RootsGLOH2 in recent evaluations, where it was compared with state-of-the-art descriptors. Finally, conclusions and future work are discussed in Section 5.

2 Related work

The approach used by SIFT is easily the most successful one among those employed for handcrafted descriptor design. Several SIFT extensions have appeared across the years aimed at improving different aspects of the descriptor, from robustness and matching accuracy to space and computational efficiency, as depicted in Fig. 1. Principal component analysis (PCA)-SIFT [14] applies PCA to the SIFT vector in order to simultaneously compress data and suppress noise. Affine SIFT [31] virtually generates new viewpoints of the local image patches in order to improve robustness. RootSIFT [15] efficiently replaces the Euclidean distance with the Hellinger's distance, more reliable for histogram comparisons, by simply 'square rooting' the SIFT

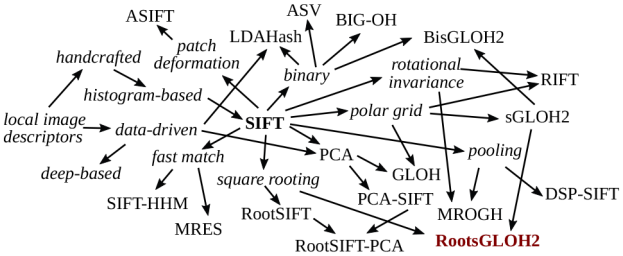

Fig. 1 *SIFT-centric taxonomy of local image descriptors. The proposed RootsGLOH2 is underlined in red (best viewed in colour)*

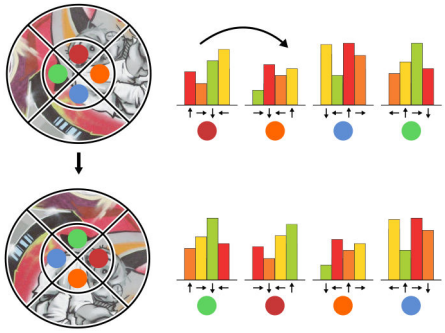

Fig. 2 *Rotation of an image patch by a factor* $(2\pi/m)$ *with the superimposed sGLOH grid (left), corresponding to a cyclic shift of the histogram bins inside each ring (right). In the example* $n = 2$ *and* $m = 4$*, colour labels on the patch grid identify the corresponding gradient orientation histograms on the descriptor vector (best viewed in colour)*

vector. RootSIFT-PCA [32] further extends RootSIFT by also applying PCA compression. Gradient local orientation histogram [33] replaces the original Cartesian grid of SIFT with a log-polar grid and applies PCA. In [34], an irregular grid with overlapping cells is employed to improve robustness. Multi-support region order-based gradient histogram [12] uses instead multiple support regions and a variable grid according to an intensity order pooling scheme. Intensity order pooling allows one to bypass the key point patch rotation according to the canonical orientation, whose estimation can be erroneous. Doing this before the actual descriptor computation yields a truly rotationally invariant descriptor. Domain size pooling SIFT [16] merges gradient data obtained at different scales to improve accuracy. In [35], the SIFT behaviour is analysed at different scales of the original patch using PCA so as to derive a new descriptor. Rotational invariant feature transform [36] replaces the grid with rings in order to achieve a rotationally invariant descriptor. Binarisation of gradient orientation histograms [37] gets a short-length binary descriptor by comparing consecutive SIFT vector elements. Accumulated stability voting [38] thresholds the differences between SIFT vectors for the same patch at different scales and sums up the results. Linear discriminant analysis hashing [18] defines thresholds on SIFT linear projections to achieve a binary descriptor. sGLOH2 [13] defines a rotating SIFT by arranging a circular grid organised so that discrete rotations of the local patch can be obtained by a circular shift of the descriptor vector. Binary sGLOH2 [13] further compares sGLOH2 vector elements to get a binary version of the original descriptor. SIFT handed hierarchical matching [39] achieves a fast match strategy by filtering on the most informative SIFT vector elements. Similarly, a multi-resolution exhaustive search [40] defines a fast hierarchical cascade matching at increasing resolution levels.

Among data-driven descriptors, deep descriptors have recently stepped into the limelight, thanks to the advent of effective convolutional neural network architectures, powerful GPUs and the availability of big data for training. Deep descriptors of the first generation mainly differed from each other by the loss function used, from triplet loss [22, 41] to hard negative mining [23] or ranking [42]. Following the recent trend in deep learning, the last generation of deep descriptors can rely on an even bigger amount of data for training with respect to the past [26, 43]. This has been exploited to constrain more the network architecture to follow specific behaviours [24, 30], either by taking inspiration from some handcrafted descriptors [28, 29, 43, 44] or by embedding more a priori geometric knowledge from the data [45]. As a matter of fact, modern deep descriptors achieved state-of-the-art results in matching accuracy and, according to the latest comparative evaluations [27, 46]; currently provide top notch performance in image matching with local image descriptors.

3 RootsGLOH2

RootsGLOH2 extends the state-of-the-art sGLOH2, a rotating SIFT providing robust matches, according to the 'square rooting' idea behind RootSIFT. The main features of both sGLOH2 and RootSIFT will be briefly described hereafter for the sake of completeness.

sGLOH2 is obtained by the concatenation of two sGLOH descriptors [47]. Fig. 2 illustrates the main sGLOH property. Following the general design of histogram-based descriptors, sGLOH is obtained by a concatenation of weighted oriented gradient histograms (such as SIFT), one for each grid region the local key point patch is divided into. Differently from other local descriptors, sGLOH uses a circular grid of n rings and m sectors and arranges histograms so that for each grid region, the first bin corresponds to the orientation pointing outside and the others follow in clockwise order. This implies that the minimal discrete rotation of $\alpha = (2\pi/m)$ of the patch corresponds to a permutation of the descriptor vector, specifically the one that cyclically shifts bins inside the histogram for each ring, thus without needing to recomputed the descriptor vector from scratch.

sGLOH packs m different descriptors of the same patch at different orientations so that two descriptor vectors $\boldsymbol{H}$ and $\boldsymbol{H}'$ are compared using the distance

$$\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}') = \min_{k = 0, \ldots, m-1} \bar{\mathscr{D}}(\boldsymbol{H}, \boldsymbol{H}'_{\alpha k}) \tag{1}$$

induced by a generic distance $\bar{\mathscr{D}}$, such as the Euclidean or Manhattan distance, where $\boldsymbol{H}'_{\alpha k}$ corresponds to the permuted descriptor vector $\boldsymbol{H}'$ according to rotation αk. Matching strategies for sGLOH can be designed so as to exploit the additional orientation information provided by limiting the rotations to check. This can reduce the number of wrong matches since some of these are dropped and cannot be selected by chance. In particular, the shifting global orientation matching strategy uses information provided by the scene context to get a global reference orientation, under the reasonable assumption that all keypoints of the scene roughly undergo the same rotation αg, not known a priori. The range of discrete orientations in (1) is modified to $k = (g-1) \bmod m, g, (g+1) \bmod m$, where $g \in \{0, 1, \ldots, m-1\}$, can be robustly estimated as the orientation maximising the number of best matches [47]. In [13], it was observed that sGLOH matching can suffer from performance degradations when the relative rotation between corresponding patches approaches the value in-between two discrete rotations, i.e. it is of the form $k(2\pi/m) + (\pi/m)$ for $k = 0, \ldots, m-1$. The sGLOH2 descriptor was designed to solve this issue: it concatenates the standard sGLOH descriptor of the original patch with the sGLOH descriptor obtained after applying a rotation of π/m to the patch. sGLOH2, can handle up to $2m$ discrete rotations of π/m degrees. Fig. 3 shows SIFT, sGLOH, and sGLOH2 matching accuracy in terms of mean average precision (mAP) for a simple test considering some images matched against their corresponding rotated versions. mAP is computed as the average on a set of different test images (see [13] for more details). sGLOH2 solves sGLOH issues when rotation approaches the one in-between two consecutive discrete rotations. The gap in terms of matching accuracy between SIFT and sGLOH2, not so evident for this simple case, increases when more complex image transformations than bare rotations are considered (see the experimental section). The upright SIFT, i.e. when key point patch is not rotated according to the canonical orientation [6] before SIFT computation is reported too for completeness. In this sense, sGLOH2 is a 'rotating SIFT'. sGLOH2 matching can also take advantage of the global reference orientation as for sGLOH,

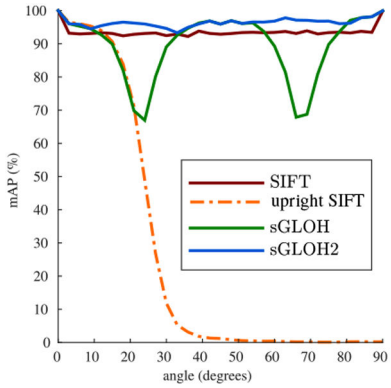

Fig. 3 *Descriptor matching accuracy for SIFT, sGLOH, and sGLOH2 (see text for details) (best viewed zoomed in and in colour)*

Fig. 4 *Sample image pairs for the WISW contest in the case of planar (top row) and non-planar (middle row) scenes, and for the IMW challenge (bottom row) (best viewed in colour)*
(a) Graffiti, ***(b)*** Spidey, ***(c)*** Castle, ***(d)*** Horse, ***(e)*** Florence, ***(f)*** London bridge

and at the meantime speed up by an efficient adaptive run-time cascade filtering matching. The resulting matching strategy is named sGOr2a⋆ [13].

RootSIFT [15] manipulates SIFT descriptors so that the Euclidean distance between two RootSIFT descriptor vectors $\boldsymbol{h}'$ Q1and $\boldsymbol{w}'$ becomes equivalent to the Hellinger's distance between the corresponding original SIFT descriptor vectors $\boldsymbol{h}$ and $\boldsymbol{w}$, defined as

$$\mathscr{D}_{\mathrm{H}}(\boldsymbol{h},\boldsymbol{w}) = \sum_i \sqrt{\frac{\boldsymbol{h}_i}{\sum_j h_j}\frac{\boldsymbol{w}_i}{\sum_j w_j}} \tag{2}$$

The vector element h'_i of the RootSIFT descriptor $\boldsymbol{h}'$ is

$$h'_i = \sqrt{\frac{\boldsymbol{h}_i}{\sum_j h_j}} \tag{3}$$

and similarly for $\boldsymbol{w}'$. Hence, the squared Euclidean distance between two RootSIFT descriptors is

$$\begin{aligned} \| \boldsymbol{h}' - \boldsymbol{w}' \|^2 &= \boldsymbol{h}'\boldsymbol{h}'^{\mathrm{T}} + \boldsymbol{w}'\boldsymbol{w}'^{\mathrm{T}} - 2\boldsymbol{h}'\boldsymbol{w}'^{\mathrm{T}} \\ &= 2 - 2\sum_i \sqrt{\frac{\boldsymbol{h}_i}{\sum_j h_j}\frac{\boldsymbol{w}_i}{\sum_j w_j}} = 2 - 2\mathscr{D}_{\mathrm{H}}(\boldsymbol{h},\boldsymbol{w}) \end{aligned} \tag{4}$$

i.e. it is equal to the Hellinger's distance up to a constant factor. The Hellinger's distance is generally preferable to the Euclidean distance at comparing histograms [15], of which SIFT descriptors are a particular instance. The reason for this superiority lies in the observation that when matching two histograms, the Euclidean distance tends to emphasise large errors occurring on a few bins with respect to small errors on the remaining majority of bins, while the Hellinger's distance does the opposite. As suggested in [13], the lower order Manhattan distance can also be usefully employed in the place of the Euclidean distance for mitigating this issue.

According to these observations, the RootsGLOH2 descriptor vector is defined and computed by 'square rooting' the corresponding sGLOH2 vector. Differently from the case of SIFT, sGLOH2 vectors are normalised, to sum up to 1 by design, so normalisation is not needed. As with sGLOH2, RootsGLOH2 matching is performed by the sGOr2a⋆ strategy [13] using the Manhattan distance as matching distance, which was found to perform better than the Euclidean distance with sGLOH-like descriptors. Obviously, in this case, it is not possible to relate the resulting metric to the Hellinger's distance, as it was done before in the Euclidean case. Nevertheless, the key idea to avoid emphasising large errors on a few histogram bins at the expense of small errors on most of the bins is still valid.

4 Evaluation

Two recent contests for local image descriptor matching will be considered for the evaluation of RootsGLOH2, namely the 'which is which contest' (WISW) [27] and the 'image matching workshop challenge' (IMW) [46], held, respectively, at the '18th International Conference on Computer Analysis of Images and Patterns (CAIP 2019)' and the '2019 IEEE Conference of Computer Vision and Pattern Recognition (CVPR 2019)'.

4.1 WISW benchmark setup

WISW relies on the well-consolidated evaluation of correct matches defined according to the overlap error between putative corresponding patches that, in the case of planar scenes, is the standard evaluation approach. The Oxford benchmark [33] and its evolution HPatches [48] are the most representative benchmarks of this kind. In addition to HPatches, WISW allows for custom patch orientations to maximise the rotational invariance of the descriptors, and considers viewpoint transformation combined with illumination changes, blur and noise effects, instead of analysing these kinds of transformations one at a time. WISW results are expressed in terms of mAP of correct matches. A match is considered correct if the patch reprojection overlap error does not exceed 50%. WISW uses 15 different scenes of six images each, of which only one is used as reference inside each scene, yielding a total of $15 \times (6-1) = 75$ image pairs. The scenes include 'bar', 'boat', 'graffiti' and 'wall' from the Oxford dataset, the whole viewpoint dataset [49] and six new scenes, each including more than one image transformation. Some image pair examples are shown in Fig. 4 (top row). Evaluation on planar scenes is not enough to gain an effective insight into descriptor behaviour on non-planar scenes, which represent nowadays the true field of application for image descriptors. For instance, it would be quite hard to derive how descriptors work in the presence of self-occlusions on the basis of planar scene analysis only. In order to overcome this limitation, WISW incudes a further evaluation on non-planar scenes according to a piecewise approximation [50] of the overlap error. In this case, the dataset employed in the evaluation contains images from 35 different scenes used in previous works (19 having three images, the remaining 16 with two images only), for a total of $19 \times 3 + 16 = 73$ image pairs, some of which are shown in Fig. 4 (middle row).

4.2 IMW benchmark setup

Another possible approach to deal with non-planar scenes is the one exploited by IMW, i.e. somewhat complementary to the one used in WISW. Specifically, the IMW approach relies on an indirect evaluation of the descriptors according to the reconstruction quality they achieve when employed in a structure-from-motion (SfM) pipeline, similar to the approach proposed in [51, 52]. In detail, the state-of-the-art SfM COLMAP [53] is

Table 1 Results for WISW (see Sections 4.1 and 4.3)

			mAP (%)	
			Planar	Non-planar↓
	SOSNet	[24]	∘	76.30	53.40
	AffNet + HardNet2	[56]	∘	74.11	52.34
	HardNet2	[26]	∘	74.29	50.09
	L2Net	[22]	∘	69.49	48.79
	RootsGLOH2	—	•	70.68	48.20
	HardNet	[23]	∘	71.49	47.80
	GeoDesc	[45]	∘	75.60	47.56
	sGLOH2	[13]	•	67.25	44.86
Q2	DOAP	[42]	∘	69.80	40.66
Q2	MKD	[9]	•	59.52	39.05
	RootSIFT	[15]	•	58.46	37.73
Q2	MIOP	[7]	•	56.83	33.38
Q2	LIOP	[7]	•	54.51	32.05

↓, sorting column; ∘, deep descriptor; •, non-deep descriptor.

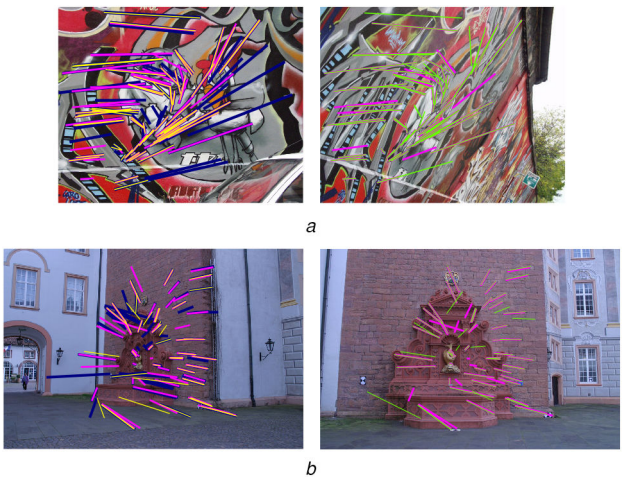

Fig. 5 *Examples of planar (top row) and non-planar (bottom row) image pairs of the WISW contest, with superimposed the optical flow of correct matches found by RootSIFT (yellow), sGLOH2 (green), RootsGLOH2 (magenta) and SOSNet (blue) according to the WISW evaluation (best viewed zoomed in and in colour)*
(a) Graffiti, ***(b)*** Fountain

employed to get the dense 3D ground-truth reconstructions from 11 scenes of popular landmarks—see Fig. 4 (bottom row) for some sample images. Each scene is reconstructed from a high number of input images in order to get high-quality reconstructions. Descriptors are then evaluated according to the pose estimation error resulting from SfM reconstruction (using only a very small subset of images for each sequence), or stereo matching. In particular, robust matching by RANSAC is applied to each possible Q2 image pair of the subsets, and the surviving inliers are used to retrieve the relative pose between the two cameras. Matching accuracy is then measured by using the angular difference between the estimated and ground truth vectors for both rotation and translation. To reduce these to one value, a variable threshold (set to the same value for rotation and translation) is used in order to determine each pose as correct or not, and the area under the curve up to a defined angular threshold is finally computed. According to the IMW organisers, the empirical threshold of 15° is an adequate proxy for wide-baseline stereo matching performance. Other reconstruction statistics, such as the number of 3D points, the average key point track length, the 3D to 2D key point reprojection error, and the ratio of successfully registered images within the model are also presented but are less relevant for descriptor evaluation. As pointed out in [54], the obtained stereo matching results are not so reliable due to the RANSAC parameter setup of the evaluation. Hence, only SfM results will be considered hereafter in the analysis of IMW results.

4.3 WISW benchmark results

Table 1 reports WISW results in terms of mAP according to the evaluation protocol described in Section 4.1. Descriptors are ranked according to their performances on non-planar scenes that are more relevant for practical applications. For a clearer evaluation, keypoints are all extracted with the same HarrisZ detector [55]. Examples of correct matches found by RootSIFT, sGLOH2, RootsGLOH2, and SOSNet according to the WISW ground-truth in complex scenes for both the planar and non-planar cases are reported in Fig. 5. As shown by the results, RootsGLOH2 clearly improves upon sGLOH2 by increasing the mAP by about 4% in both the planar and non-planar cases. This suggests that avoiding to emphasise large errors on a few histogram bins at the expense of a few errors on the vast majority of bins is effective at improving descriptor distance. Moreover, the mAP gap between RootsGLOH2 and the best deep-descriptors is quite limited: except for the very recent SOSNet [24], the gap is no more than 4% for any kind of scene, while this gap is about 8% for the current second best non-deep descriptor sGLOH2. RootsGLOH2 turns out to be well-aligned in terms of matching accuracy with HardNet and L2Net, neglecting the minimal differences from the planar to the non-planar cases. Additionally, when it comes to non-planar scene matching, RootsGLOH2 works slightly better than the recent GeoDesc descriptor, which is comparable to the top best SOSNet and HardNet2 (reported as HardNetA in the WISW evaluation) on planar scenes. In this sense, GeoDesc appears somewhat overfitted on planar transformations only at the expense of non-planar scene transformations that are more general, complex, and relevant for actual applications. Except for GeoDesc, rank is roughly preserved between the planar and non-planar evaluation. Notice also that 5% of mAP discrepancy on a base of 70% mAP, which happens in the planar case is less problematic for any application than the same difference on a base of 50% mAP baseline, obtained for the non-planar case.

4.4 IMW benchmark results

Table 2 reports IMW results, both in terms of pose mAP up to a tolerance of 15° and in terms of the number of images that were correctly registered to the model, according to the setup protocol described in Section 4.2. Besides the average mAP results over all the subsets considered in the SfM, by which descriptors are ranked in the table, results considering only subsets of five and 25 images, respectively, the second smallest and the largest subset sizes are reported (the minimum subset size of three images is very close to a stereo matching evaluation, and as previously stated it is not reliable). As for the WISW evaluation (see Table 1), ranking is roughly preserved among columns. In the case, descriptor information and details have not been yet released, the not available (na) mark denotes the missing reference. Notice that, differently from WISW, IMW does not limit the key point extraction method to use. Since IMW allows more submissions of the same detector + descriptor pairs with different parameters, in order to better focus on the evaluation, Table 2 only reports results for the best setups, excluding matching strategies relying on key point localisation. This choice is motivated by the fact that in the latter case the evaluation would be somewhat unfair, since matches for submissions that apply geometric matching strategies similar to [57, 58] would be clearly better filtered for the successive RANSAC step, independently from the descriptor employed. Specifically, the default matching strategy for the results reported in the table is the nearest neighbour (NN), except for HarrisZ + RsGLOH2 using sGOr2a⋆, and the descriptors superscripted with ‘ + ‘, that use the mutual NN. The maximum number of keypoints allowable per image is 8000, except for any version of SuperPoint and DELF (2048), and ELF-SIFT (512). Inspecting the results, except for the SIFT + ContextDesc+ pair, which employs the very recent ContextDesc [28] descriptor with mutual NN, the Harrisz + RootsGLOH2 pair mAP gap with respect to state-of-the-art detector/descriptor pair using deep learning, is quite limited (<5%) as in WISW. For the second best non-deep descriptor considered, i.e. AKAZE, this is more relevant (up to about 12%). Notice also that mAP increases and performance gap decreases as the image

Table 2 Results for IMW (see Sections 4.2 and 4.4)

			mAP at 15°, %			Successfully registered images, %		
			All↓	5	25	All	5	25
SIFT + ContextDesc+	[28]	∘	60.17	53.13	84.15	98.10	96.60	98.30
SIFT + HardNet (Larger Patches)	[23]	∘	54.81	45.11	82.63	97.90	96.10	97.90
Superpoint+ (new version)	[na]	∘	54.40	46.16	81.14	95.60	91.00	97.60
Scale-invariant Desc	[59]	∘	54.27	43.91	83.63	97.60	94.90	98.10
SIFT + ContextDesc	[28]	∘	53.99	43.92	82.95	97.90	95.90	98.10
SIFT + GeoDesc	[45]	∘	53.17	43.05	83.91	97.30	94.50	98.20
HesAffNet + HardNet2	[56]	∘	52.84	44.00	81.54	96.80	93.70	97.70
SIFT + L2Net	[22]	∘	50.87	40.57	81.20	97.30	94.10	98.00
Hessian + HardNet2	[26]	∘	50.55	41.40	75.83	95.10	91.00	95.50
HarrisZ + RootsGLOH2	–	•	50.40	40.84	80.13	96.60	92.90	98.40
AKAZE + SphereDesc	[na]	∘	48.65	37.51	78.00	96.60	92.70	97.80
Superpoint (new version)	[na]	∘	47.78	36.49	78.21	95.00	89.30	97.40
SIFT + TFeat	[60]	∘	46.43	33.47	78.18	96.50	92.30	97.30
AKAZE (OpenCV)	[61]	•	42.85	26.77	78.45	94.30	88.00	96.90
SuperPoint	[62]	∘	42.67	31.38	71.69	91.20	85.10	92.30
SIFT (OpenCV)	[6]	•	41.46	26.74	76.79	93.60	87.00	96.50
D2Net	[43]	∘	41.02	32.44	65.14	93.90	89.90	93.50
Q2 SURF (OpenCV)	[8]	•	30.07	14.51	65.47	90.10	80.40	94.80
Q2 Brisk + SSS	[na]	∘	26.94	14.37	55.52	92.90	85.80	95.50
Q2 SIFT-AID	[63]	∘	26.88	12.20	59.40	88.70	79.80	92.90
Q2 ORB (OpenCV)	[21]	•	23.07	11.45	51.74	87.60	79.40	91.30
ELF-SIFT	[na]	∘	16.78	6.65	40.87	72.50	70.50	69.50
DELF	[30]	∘	16.29	9.20	33.00	87.90	79.90	90.30

↓, sorting column; ∘, deep descriptor; •, non-deep descriptor (numerical columns refer to evaluation with different image subset sizes).

initial subset for estimating the SfM model is enlarged. Similar considerations hold for the number of correctly registered images, where it can also be noted that for a subset size of 25 images Harrisz + RootsGLOH2 achieve the topmost rate of registered images, witnessing again the robustness of RootsGLOH2.

4.5 Running times

Concerning running times, RootsGLOH2 descriptor computation is practically the same of sGLOH2 (square rooting can be neglected compared to the other operations needed), which is less than half of the time needed to compute a RootSIFT descriptor, since patch rotation in the canonical orientation (and its estimation) needs not to be computed [13]. For instance, on a Intel i5-2500 CPU @ 3.30 GHz with 16 Gb of RAM, the extraction of 2048 descriptors takes >2 s for SIFT and <1 s for sGLOH2, while for deep descriptors based on L2Net (including SOSNet, HardNet2, and Geodesc), excluding patch normalisation and canonical orientation estimation that take more than 1 s, 4.5 s are needed on a CPU. For completeness, SIFT GPU implementation is almost 150% faster than GeoDesc on an NVidia GeForce GTX1080 [45]. On the other hand, amortised computational time for matching two descriptors, using single-threaded SSE 4.1 optimised code, changes from about 400 to 650 ns as one moves from sGLOH2 to RootsGLOH2. This is due to the fact that RootsGLOH2 requires float operations instead of integer operations due to the presence of the square rooting operation. For reference, the corresponding amortised matching time for RootSIFT and any other real-value descriptor, including deep descriptors, is 50 ns since there is no need to check distances at several patch orientations. Finally note that, differently from deep descriptors, handcrafted descriptors such as RootsGLOH2 do not need high-capability GPU hardware to run efficiently.

5 Conclusion and future work

This study proposed to embed the RootSIFT 'square rooting' idea into the sGLOH2 handcrafted descriptor. The resulting RootsGLOH2 descriptor provides clear improvements upon sGLOH2, as RootSIFT does for SIFT. The results obtained give a further evidence of the fact that both the classical Euclidean and Manhattan distances (used by SIFT and sGLOH2, respectively) are suboptimal solutions for the associated histogram-based descriptors, as they tend to emphasise the importance of large errors on a few histogram bins instead than that of small errors on the majority of bins. Future work will be devoted to extending the square rooting concept, by investigating which transformations can be said to be truly optimal for each given kind of handcrafted descriptor.

The evaluation of RootsGLOH2, taken out according to very recent benchmark comparing the best and latest descriptors, also shows that RootsGLOH2 matching accuracy on both the planar and non-planar cases is very close to that of the top notch deep descriptors, currently the unquestionable rulers of this research area, and clearly better than the matching accuracy achieved by other non-deep descriptors. This suggests that investigating on handcrafted descriptors, spending time, and resources, even nowadays is not a waste of time as it can provide fresh and novel ideas, capable of making deep descriptors less black-boxed, also considering that the implicit design of current state-of-the-art deep descriptors is often inspired by handcrafted descriptor approaches.

6 Acknowledgment

This work was partially supported by 'PON Ricerca e Innovazione 2014-2020', issued by the Italian Ministry of Education and Research (MIUR), cofunded by the European Social Fund (ESF), CUP B74I18000220006, id. proposta AIM 1875400 – linea di attività 2, Area Cultural Heritage.

7 References

[1] Brown, M., Lowe, D.G.: 'Automatic panoramic image stitching using invariant features', *Int. J. Comput. Vis.*, 2007, **74**, (1), pp. 59–73

[2] Snavely, N., Seitz, S.M., Szeliski, R.: 'Modeling the world from internet photo collections', *Int. J. Comput. Vis.*, 2008, **80**, (2), pp. 189–210

[3] Fanfani, M., Bellavia, F., Colombo, C.: 'Accurate keyframe selection and keypoint tracking for robust visual odometry', *Mach. Vis. Appl.*, 2016, **27**, (6), pp. 833–844

[4] Fan, B., Kong, Q., Trzcinski, T., *et al.*: 'Receptive fields selection for binary feature description', *IEEE Trans. Image Process.*, 2014, **26**, (6), pp. 2583–2595

[5] Bertini, M., Colombo, C., Bimbo, A.D.: 'Automatic caption localization in videos using salient points'. IEEE Int. Conf. on Multimedia and Expo (ICME 2001), 2001, pp. 68–71

Q3 [6] Lowe, D.G.: 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vis.*, 2004, **60**, (2), pp. 91–110

[7] Wang, Z., Fan, B., Wang, G., *et al.*: 'Exploring local and overall ordinal information for robust feature description', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (11), pp. 2198–2211

[8] Bay, H., Ess, A., Tuytelaars, T., *et al.*: 'Speeded-up robust features (SURF)', *Comput. Vis. Image Underst.*, 2008, **110**, (3), pp. 346–359

[9] Mukundan, A., Tolias, G., Chum, O.: 'Multiple-kernel local-patch descriptor'. British Machine Vision Conf. (BMVC), 2017

[10] Leutenegger, S., Chli, M., Siegwart, R.: 'BRISK: binary robust invariant scalable keypoints'. Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2011

[11] Alahi, A., Ortiz, R., Vandergheynst, P.: 'Freak: fast retina keypoint'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 510–517

[12] Fan, B., Wu, F., Hu, Z.: 'Rotationally invariant descriptors using intensity order pooling', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (10), pp. 2031–2045

[13] Bellavia, F., Colombo, C.: 'Rethinking the sGLOH descriptor', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, **40**, (4), pp. 931–944

[14] Ke, Y., Sukthankar, R.: 'PCA-SIFT: a more distinctive representation for local image descriptors'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2004, pp. 506–513

[15] Arandjelovíc, R., Zisserman, A.: 'Three things everyone should know to improve object retrieval'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2911–2918

[16] Dong, J., Soatto, S.: 'Domain-size pooling in local descriptors: DSP-SIFT'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015

[17] Trzcinski, T., Christoudias, M., Lepetit, V., *et al.*: 'Learning image descriptors with the boosting-trick'. Advances in Neural Information Processing Systems, 2012, pp. 269–277

Q4 [18] Strecha, C., Bronstein, A.M., Bronstein, M.M., *et al.*: 'LDAHash: improved matching with smaller descriptors', *IEEE Trans Pattern Anal Mach Intell*, 2012, **34**, (1), pp. 66–78

[19] Brown, M.A., Hua, G., Winder, S.A.J.: 'Discriminative learning of local image descriptors', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (1), pp. 43–57

[20] Balntas, V., Tang, L., Mikolajczyk, K.: 'BOLD – binary online learned descriptor for efficient image matching'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2367–2375

[21] Rublee, E., Rabaud, V., Konolige, K., *et al.*: 'ORB: an efficient alternative to SIFT or SURF'. Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2011, pp. 2564–2571

[22] Tian, Y., Fan, B., Wu, F.: 'L2-net: deep learning of discriminative patch descriptor in Euclidean space'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6128–6136

[23] Mishchuk, A., Mishkin, D., Radenovic, F., *et al.*: 'Working hard to know your neighbor's margins: local descriptor learning loss'. Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems (NIPS), 2017, pp. 4829–4840

[24] Tian, Y., Yu, X., Fan, B., *et al.*: 'SOSNet: second order similarity regularization for local descriptor learning'. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2019

[25] Li, Z., Snavely, N.: 'Megadepth: learning single-view depth prediction from internet photos'. Computer Vision and Pattern Recognition (CVPR), 2018

[26] Pultar, M., Mishkin, D., Matas, J.: 'Leveraging outdoor webcams for local descriptor learning', 2019

Q6 [27] Bellavia, F., Colombo, C.: "Which is which?' evaluation of local descriptors for image matching in real-world scenarios'. Int. Conf. on Computer Analysis of Images and Patterns (CAIP), 2019. Available from http://cvg.dsi.unifi.it/wisw.caip2019

[28] Luo, Z., Shen, T., Zhou, L., *et al.*: 'Contextdesc: local descriptor augmentation with cross-modality context', Computer Vision and Pattern Recognition (CVPR), 2019

[29] Mukundan, A., Tolias, G., Bursuc, A., *et al.*: 'Understanding and improving kernel local descriptors', *Int. J. Comput. Vis.*, 2018

[30] Noh, H., Araujo, A., Sim, J., *et al.*: 'Large-scale image retrieval with attentive deep local features'. IEEE Int. Conf. on Computer Vision (ICCV), 2017

[31] Morel, J.M., Yu, G.: 'ASIFT: a new framework for fully affine invariant image comparison', *SIAM J. Imaging Sci.*, 2009, **2**, (2), pp. 438–469

[32] Bursuc, A., Tolias, G., Jégou, H.: 'Kernel local descriptors with implicit rotation matching'. ACM Int. Conf. on Multimedia Retrieval (ICMR), 2015

[33] Mikolajczyk, K., Schmid, C.: 'A performance evaluation of local descriptors', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (10), pp. 1615–1630

[34] Cui, Y., Hasler, N., Thormählen, T., *et al.*: 'Scale invariant feature transform with irregular orientation histogram binning'. Proc. Int. Conf. on Image Analysis and Recognition (ICIAR), 2009, pp. 258–267

[35] Hassner, T., Mayzels, V., Zelnik.Manor, L.: 'On SIFTs and their scales'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1522–1528

[36] Lazebnik, S., Schmid, C., Ponce, J.: 'A sparse texture representation using local affine regions', *IEEE Trans. Pattern. Anal. Mach. Intell*, 2005, **27**, (8), pp. 1265–1278

[37] Baber, J., Dailey, M.N., Satoh, S., *et al.*: 'BIG-OH: binarization of gradient orientation histograms', *Image Vis. Comput.*, 2014, **32**, (11), pp. 940–953

[38] Yang, T., Lin, Y., Chuang, Y.: 'Accumulated stability voting: a robust descriptor from descriptors of multiple scales'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 327–335

[39] Treen, G., Whitehead, A.: 'Efficient SIFT matching from keypoint descriptor properties'. Proc. Workshop on Applications of Computer Vision (WACV), 2009, pp. 1–7

[40] Tsai, C.Y., Tsao, A.H., Wang, C.W.: 'Real-time feature descriptor matching via a multi-resolution exhaustive search method', *J. Softw.*, 2013, **8**, (9), pp. 2197–2201

[41] Simo.Serra, E., Trulls, E., Ferraz, L., *et al.*: 'Discriminative learning of deep convolutional feature point descriptors'. Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2015

[42] He, K., Lu, Y., Sclaroff, S.: 'Local descriptors optimized for average precision'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2018

[43] Dusmanu, M., Rocco, I., Pajdla, T., *et al.*: 'D2-Net: a trainable CNN for joint detection and description of local features'. Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019

[44] Ono, Y., Trulls, E., Fua, P., *et al.*: 'LF-Net: learning local features from images'. NIPS'18: Proc. 32nd Int. Conf. on Neural Information Processing Systems, 2018

Q7 [45] Luo, Z., Shen, T., Zhou, L., *et al.*: 'Geodesc: learning local descriptors by integrating geometry constraints'. Proc. European Conf. on Computer Vision (ECCV), 2018

[46] 'Image Matching Workshop (IMW) challenge at (CVPR2019)', 2019, Available from https://image-matching-workshop.github.io

[47] Bellavia, F., Tegolo, D., Valenti, C.: 'Keypoint descriptor matching with context-based orientation estimation', *Image Vis. Comput.*, 2014, **32**, (9), pp. 559–567

[48] Balntas, V., Lenc, K., Vedaldi, A., *et al.*: 'HPatches: a benchmark and evaluation of handcrafted and learned local descriptors'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3852–3861

[49] Yi, K.M., Verdie, Y., Fua, P., *et al.*: 'Learning to assign orientations to feature points'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1–8

[50] Bellavia, F., Valenti, C., Lupascu, C.A., *et al.*: 'Approximated overlap error for the evaluation of feature descriptors on 3D scenes'. Proc. Int. Conf. on Image Analysis and Processing (ICIAP), 2013, pp. 270–279

[51] Fan, B., Kong, Q., Wang, X., *et al.*: 'A performance evaluation of local features for image based 3d reconstruction'. arXiv, 2018

Q8 [52] Schönberger, J.L., Hardmeier, H., Sattler, T., *et al.*: 'Comparative evaluation of hand-crafted and learned local features'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017

[53] Schönberger, J.L., Frahm, J.M.: 'Structure-from-motion revisited'. Conf. on Computer Vision and Pattern Recognition (CVPR), 2016

[54] 'On the art of establishing correspondence', 2019. Available from https://image-matching-workshop.github.io/slides/slides-matas.pdf

[55] Bellavia, F., Tegolo, D., Valenti, C.: 'Improving Harris corner selection strategy', *IET Comput. Vis.*, 2011, **5**, (2), pp. 86–96

[56] Mishkin, D., Radenovic, F., Matas, J.: 'Repeatability is not enough: learning affine regions via discriminability'. Proc. European Conf. on Computer Vision (ECCV), 2018

[57] Bian, J., Lin, W.Y., Matsushita, Y., *et al.*: 'GMS: grid-based motion statistics for fast, ultra-robust feature correspondence'. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2828–2837

[58] Ma, J., Zhao, J., Jiang, J., *et al.*: 'Locality preserving matching', *Int. J. Comput. Vis.*, 2019, **127**, (5), pp. 512–531

[59] Ebel, P., Mishchuk, A., Yi, K.M., *et al.*: 'Beyond Cartesian representations for local descriptors'. IEEE Int. Conf. on Computer Vision (ICCV), 2019

[60] Balntas, V., Riba, E., Ponsa, D., *et al.*: 'Learning local feature descriptors with triplets and shallow convolutional neural networks'. Proc. British Machine Vision Conf. (BMVC), 2016, pp. 119.1–119.11

[61] Alcantarilla, P.F., Nuevo, J., Bartoli, A.: 'Fast explicit diffusion for accelerated features in nonlinear scale space'. Proc. British Machine Vision Conf. (BMVC), 2013

[62] DeTone, D., Malisiewicz, T., Rabinovich, A.: 'Superpoint: self-supervised interest point detection and description'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018

[63] Rodriguez, M., Facciolo, G., von Gioi, R.G., *et al.*: 'SIFT-AID: boosting SIFT with and affine invariant descriptor based on convolutional neural networks'. Proc. Int. Conf. of Image Processing (ICIP), 2019