Jet-Based Local Image Descriptors

Anders Boesen Lindbo Larsen[1], Sune Darkner[1],
Anders Lindbjerg Dahl[2], and Kim Steenstrup Pedersen[1]

[1] Department of Computer Science, University of Copenhagen, DK
{abll,darkner,kimstp}@diku.dk
[2] Department of Informatics and Mathematical Modelling,
Technical University of Denmark, DK
abd@imm.dtu.dk

Abstract. We present a general novel image descriptor based on higher-order differential geometry and investigate the effect of common descriptor choices. Our investigation is twofold in that we develop a jet-based descriptor and perform a comparative evaluation with current state-of-the-art descriptors on the recently released DTU Robot dataset. We demonstrate how the use of higher-order image structures enables us to reduce the descriptor dimensionality while still achieving very good performance. The descriptors are tested in a variety of scenarios including large changes in scale, viewing angle and lighting. We show that the proposed jet-based descriptor is superior to state-of-the-art for DoG interest points and show competitive performance for the other tested interest points.

1 Introduction

Image characterizations based on local interest points and descriptors like SIFT [1], GLOH [2], PCA-SIFT [3], DAISY [4,5], and many others have had great impact on computer vision research. Such descriptors have been used in a range of applications for efficiently determining image similarities. Common for these approaches is their aim at a description of a local image patch, which is invariant with respect to photogrammetric variations like viewpoint, scale, and lighting change. Additionally, they all have low dimensionality compared to the original pixel representation, but the difference in dimensionality among the descriptors varies significantly.

These characterizations all describe the local geometry of the image, like the local distributions of first order derivatives in the SIFT descriptor [1], or the learned basis of PCA-SIFT [3] that encodes the parameters of this basis. The introduction of PCA-SIFT inevitable leads to the question of whether the complex structure and relatively high dimensionality of most descriptors are in fact over-representations and perhaps much simpler, theoretically sound, and compact formulations exist. This observation has inspired us to investigate the local k-jet [6], a higher order differential descriptor, as a natural basis for encoding the geometry of an image patch. Thus, the descriptor we propose can, similar to

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 640–652, 2012.

PCA-SIFT, be interpreted as a basis representation of the image patch. We will investigate the effect of a multi-scale representation in our descriptor as well as the effect of introducing multi-locality similar to the grid of histograms used in SIFT.

Local image descriptors based on higher order image differentials have previously been proposed [7,8] with little success for feature matching [2]. Compared to our descriptor, these descriptors do not rely on multi-locality and only use differential invariants (functions of the local jet) up to the third order. Our descriptor bears some resemblance to the jet-descriptor formulated by Laptev and Lindeberg in [9], which is based on the local 4-jet, but this does not include multi-locality and is only used for spatio-temporal recognition.

We investigate the matching performance of our method compared to current state-of-the-art descriptors on the DTU Robot dataset [10]. This dataset allows for superior performance studies for certain perturbation scenarios (view angle, scale and lighting) because of its large number of different scenes, and its systematic light variation and camera positions. The dataset used by Mikolajczyk and Schmid [11] contains the possibility for more perturbation scenarios including view angle, rotation, scale, focus, exposure level and compression artifacts, but consists of only eight scenes. In Winder et al. [4], a dataset with ground truth from 3D reconstructions of three different outdoor scenes from tourist photographs was used. This dataset represents the combination of many different perturbation factors (e.g. lighting, view angle, scale, noise, perspective). However, a bias towards the descriptor (SIFT) used for doing the 3D reconstruction is potentially present. Virtual scenes of photorealistic quality have recently been proposed [12,13]. These datasets offer better control over image perturbations and the ground truth is known, but are synthetic which may cause a bias on the descriptor performance.

The contributions of this paper include a new approach to the construction of local image descriptors based on the local k-jet as well as evaluation of local image descriptors on the DTU Robot dataset. We show that our descriptor is very competitive with state-of-the-art and that it outperforms these descriptors by a significant margin under certain conditions. Furthermore, our descriptor is not designed for a specific type of local image geometry as detected by corner or blob interest point detectors. As such, it is a general purpose descriptor applicable to any image patch. We will therefore investigate how performance of our descriptor varies with the choice of interest point detector.

2 Jet Descriptor

To construct the jet descriptor we use the multi-scale k-jet as described by Florack et al. [6]. We consider the scale normalized derivatives of the linear scale-space of a 2D image $I(\boldsymbol{r}) : \mathbb{R}^2 \to \mathbb{R}$

$$L_{x^n y^m}(\boldsymbol{r}; \sigma) = \sigma^{n+m} \frac{\partial^{n+m}}{\partial x^n \partial y^m} (G * I)(\boldsymbol{r}; \sigma) \quad , \quad \boldsymbol{r} = (x, y) \quad , \tag{1}$$

where convolution is denoted by $*$ and G is the Gaussian aperture function

$$G(\boldsymbol{r};\sigma) = \frac{1}{(\sigma\sqrt{2\pi})^d} \exp\left(-\frac{\boldsymbol{r}\cdot\boldsymbol{r}}{2\sigma^2}\right) \quad , \quad \sigma \geq 0 \quad , \quad G(\boldsymbol{r};0) \equiv \delta \quad . \tag{2}$$

Here $\delta(0)$ denotes the Dirac delta function centered at zero. The local k-jet is then given by the vector $\mathcal{J}_k(\boldsymbol{r};\sigma) \in \mathbb{R}^K, K = \frac{(2+k)!}{2k!} - 1$

$$\mathcal{J}_k(\boldsymbol{r};\sigma) = (\{L_{x^n y^m}(\boldsymbol{r};\sigma) \,|\, 0 < n+m \leq k\})^T \quad . \tag{3}$$

By excluding the zeroth order term, the k-jet becomes invariant to additive changes to the intensities.

By construction, the scale space derivatives are correlated and under basic assumptions of the image statistics this correlation can be described analytically [14,15]. In order to create a scale normalized and decorrelated descriptor, we therefore perform a whitening of the local k-jet coefficients according to the covariance structure derived in [14,15]. This yields a vector where the elements are uncorrelated and of the same order of magnitude, allowing the use of the Euclidean distance as descriptor similarity measure. According to [14], the covariance between the jet coefficients $L_{x^i y^j}$ and $L_{x^k y^l}$, where both $n = i + j$ and $m = k + l$ are even, is given by

$$\mathrm{cov}(L_{x^i y^j}, L_{x^k y^l}) = (-1)^{\frac{n+m}{2}+k+l} \frac{\beta^2}{2\pi\sigma^{n+m}} \frac{n!m!}{2^{n+m}(n+m)\left(\frac{n}{2}\right)!\left(\frac{m}{2}\right)!} \quad . \tag{4}$$

σ is the scale parameter of the local k-jet, and β is a model parameter that is irrelevant in our context and will be set to $\beta = 1$. If either n or m is odd, the covariance is 0. Finally, we remark that this normalization method is related, but not identical, to the descriptor similarity measure proposed in [8]. After the whitening, the descriptor is L_2 normalized to achieve invariance to affine contrast changes.

Using the local k-jet above, we wish to investigate what descriptor configurations yield good discriminability. More specifically, we investigate the effect of sampling jets in a multi-scale or multi-local fashion. We will also explore to what extent we can increase the differential order k while getting performance improvements. This leads to the descriptor proposals listed in Table 1. We use the following naming convention. The suffix '-scale2' indicates that the jet is sampled at two different scales in a single point. The suffix '-grid$n$' indicates a multi-local jet sampling in a regular square grid of size $n \times n$ (similar to the spatial sampling grid used in SIFT [1]).

Following [2,16], image patches of size 64×64 pixels are extracted at three times the scale of the interest points generated by the interest point detector. Moreover, affine interest point regions are warped to isotropic regions. The jet descriptor parameters shown in Table 1 have been optimized manually using the dataset by the Oxford Visual Geometry Group[1]. The σ parameters denote the

[1] Data and code available at `http://www.robots.ox.ac.uk/~vgg/research/affine`.

jet scales in pixel units. The '-grid$n$' sampling location parameters p denote the x and y pixel offsets of the grid columns and rows respectively (the row and column offsets are the same since the grid is square, hence the shared p values).

The proposed jet descriptors are, by construction, invariant to changes in scale, translation, and contrast. The scale and translation invariance arise mainly from the interest point detectors which provide a localization in scale-space of the interest point and the resampling of the interest region around the point to 64×64 pixels. However, this localization includes some noise depending on the quality of the detector algorithm and its implementation (see e.g. [10]). We compensate for this by using fairly large scales relative to the 64×64 pixels patch making the jet descriptor robust towards small perturbations in the detected position and scale. We compensate for the reduction in derivative response level caused by the Gaussian blurring by using scale normalized derivatives (see (1)). The jet descriptors are not invariant to rotations around the optical axis as we do not orient the jets according to e.g. a dominant orientation. This could, however, easily be achieved by rotating the coordinate frame according to some fiducial orientation of the image patch and expressing the jets in this coordinate system.

Table 1. The jet descriptor proposals, their parameters and dimensionalities. Bottom rows: State-of-the-art descriptors that we compare the jet descriptors to in our experiments.

Variant	**Dim.**	**Parameters**
$\mathcal{J}_4$	14	$\sigma = 10.6$
$\mathcal{J}_5$	20	$\sigma = 10.6$
$\mathcal{J}_6$	27	$\sigma = 10.6$
$\mathcal{J}_7$	35	$\sigma = 10.6$
$\mathcal{J}_4$-scale2	28	$\sigma_1 = 7.5$, $\sigma_2 = 16$
$\mathcal{J}_5$-scale2	40	$\sigma_1 = 7.5$, $\sigma_2 = 16$
$\mathcal{J}_3$-grid2	36	$\sigma = 6.8$, $p_1 = 21$, $p_2 = 44$
$\mathcal{J}_4$-grid2	56	$\sigma = 6.8$, $p_1 = 21$, $p_2 = 44$
$\mathcal{J}_5$-grid2	80	$\sigma = 6.8$, $p_1 = 21$, $p_2 = 44$
$\mathcal{J}_3$-grid4	144	$\sigma = 5.2$, $p_1 = 15$, $p_2 = 26$, $p_3 = 38$, $p_4 = 50$
SIFT [1]	128	-
GLOH [2]	128	-
SURF [17]	64	-
PCA-SIFT [3]	36	-

3 Dataset and Evaluation Criteria

The DTU Robot dataset[2] [10,16] consists of 60 different scenes containing object categories such as miniature buildings, fabrics, books and groceries (see Fig. 1a

[2] Dataset and code available at `http://roboimagedata.imm.dtu.dk`.

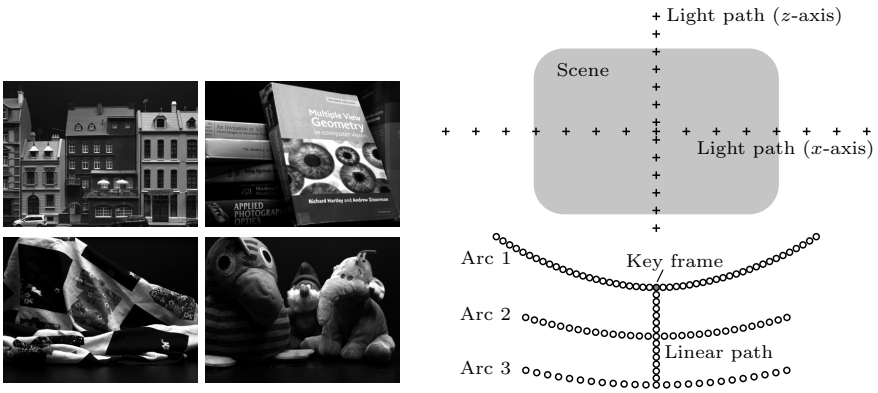

Fig. 1. The DTU Robot dataset. (a): Examples of different scenes. (b): A top-down view of all camera positions (circles) and light source positions (pluses). The key frame in the front center is used as reference image when evaluating descriptor performance. Arc 1 is at a distance of 0.5 m from the scene and spans $\pm 40°$, Arc 2 has distance 0.65 m and spans $\pm 25°$ and Arc 3 has distance 0.8 m and spans $\pm 20°$. The linear path spans the range $[0.5\,\text{m}; 0.8\,\text{m}]$.

for examples). Each scene is photographed systematically with different configurations of camera position and light source. The camera positions are placed along four paths. Three of these paths are horizontal arcs at different distances to the scene. The fourth camera path is linear along the depth axis (z-axis). Note that there are no vertical variations in the placement of the camera, nor are there any rotations along the optical axis. The illumination possibilities cover light sources placed on two linear paths along the horizontal axis (the x-axis) and the depth axis (z-axis) respectively. Note that the light sources are created artificially from the original dataset as described in [16]. An overview of the camera and light configurations is shown in Fig. 1b.

Descriptor performance is calculated following the evaluation criteria described in [16]. That is, for each feature descriptor in the *key frame* (see Fig. 1b), we perform a *nearest neighbor distance ratio matching* with the feature descriptors of another image. We then calculate the *receiver operating characteristic* (ROC) curve from the number of correct and incorrect feature matchings. In order to get a single value to quantify the descriptor performance on an image pair, we compute the *area under the curve* (AUC). For each configuration of camera and lighting position, we compute the mean AUC over the 60 scenes and use it to quantify the descriptor performance.

Note that the ROC analysis differs slightly from the evaluation criteria used in the popular study by Mikolajczyk and Schmid [2], where the recall vs. $1 -$ precision curve is used instead of the ROC curve. We have experimented with both evaluation measures and have found that they reveal the same descriptor performance as the ordering of the descriptors does not change.

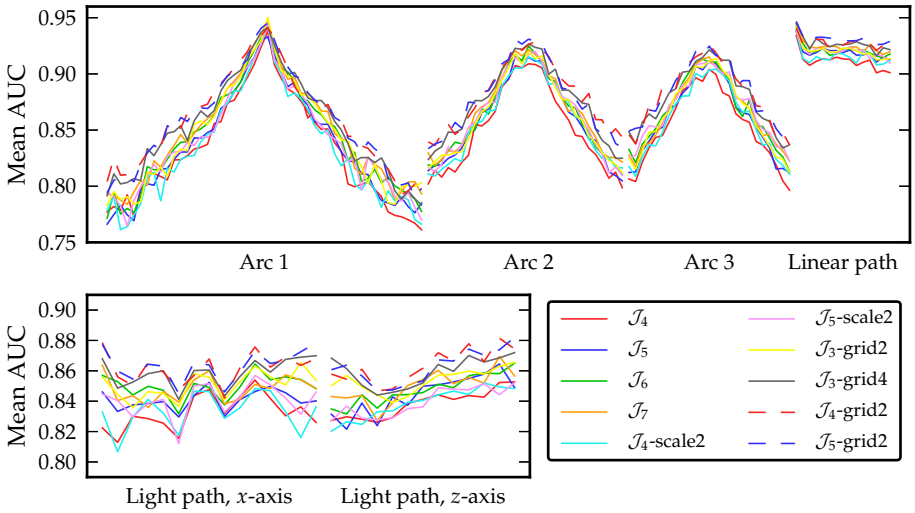

Fig. 2. Performance of different jet descriptors using MSER interest points

4 Experiments

Using the DTU Robot dataset, we investigate the performance of our different jet descriptor configurations from Table 1. For this experiment, we use interest points generated by the MSER detector [18] as it has been shown to perform well [16]. The experimental results are shown in Fig. 2. We see that for the single jets, the performance improvements stagnate around $k = 6$. The multi-scale approach does not seem to increase the discriminability significantly as $\mathcal{J}_4$-scale2 and $\mathcal{J}_5$-scale2 only have a slight edge over $\mathcal{J}_4$ and $\mathcal{J}_5$ respectively. The multi-scale descriptors are slightly more discriminative than the single-scale descriptors for changes in view point. Under illumination variations, the performance of the multi-scale descriptors is no better than the single-scale descriptors. Conversely, the multi-local approach yields a visible improvement as it consistently achieves better scores than the single- and multi-scale jets. We see, to some extent, that multi-locality can be replaced with higher-order image structure as $\mathcal{J}_4$-grid2 perform slightly better than $\mathcal{J}_3$-grid4. Among the multi-local approaches, $\mathcal{J}_4$-grid2 offers a good trade-off between descriptor dimensionality and performance.

We compare our jet descriptors with the state-of-the-art descriptors listed in Table 1. We choose these descriptors because they are among the top performers in the comparative study by Dahl et al. [16]. These descriptors are generated using code made available by the Oxford Visual Geometry Group and from the SURF website[3]. Since our dataset does not support rotations around the optical axis, we have disabled rotational invariance for all descriptors to better reveal their discriminative ability. As representative for our multi-local jet descriptors,

[3] `http://www.vision.ee.ethz.ch/~surf`

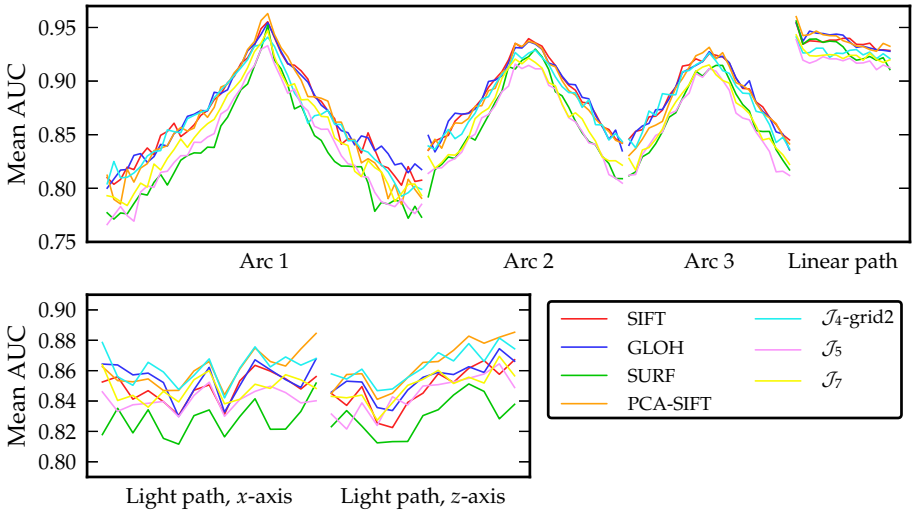

Fig. 3. Performance comparison with other state-of-the-art descriptors using MSER interest points

we select $\mathcal{J}_4$-grid2. We also include $\mathcal{J}_5$, $\mathcal{J}_7$ as we want to investigate the performance when relying on a single, low-dimensional jet. We do not compare with previously published differential invariant based descriptors, such as [7,8,9], because they have in our experiments (not part of the paper) been shown to be significantly under par.

In Fig. 3, we have plotted the performance of the descriptors on MSER interest points. We see that for changes in view angle (Arc 1), $\mathcal{J}_4$-grid2 is able to follow the top performers SIFT, GLOH and PCA-SIFT decently (though a bit under par). Under variations in scale (Linear path), the jet descriptors perform significantly under par. For changes in lighting, however, $\mathcal{J}_4$-grid2 is a top performer together with PCA-SIFT.

In order to investigate the effect of different interest point detectors, and thereby different types of local image geometry, we show the descriptor performance on Difference of Gaussians (DoG) interest points [1] illustrated in Fig. 4. The picture is quite different from MSER interest points as $\mathcal{J}_4$-grid2 is the top performer by a visible margin in all perturbation scenarios. Even our single scale jets perform similar to SIFT at a much lower dimensionality. Interestingly, $\mathcal{J}_5$ performs better than $\mathcal{J}_7$ under changes in viewpoint and vice versa under illumination changes. It is also interesting to notice that PCA-SIFT and $\mathcal{J}_7$ have similar performance with $\mathcal{J}_7$ slightly under par in some situations.

We also show the performance of our descriptors on Harris-Laplace (HarLap) and Harris-Affine (HarAff) interest points [19] in Fig. 5 and Fig. 6 respectively. At this point, we recognize a trend as $\mathcal{J}_4$-grid2 consequently shows superior robustness towards changes in lighting while being very competitive under viewpoint variations. The single jets are very competitive under illumination

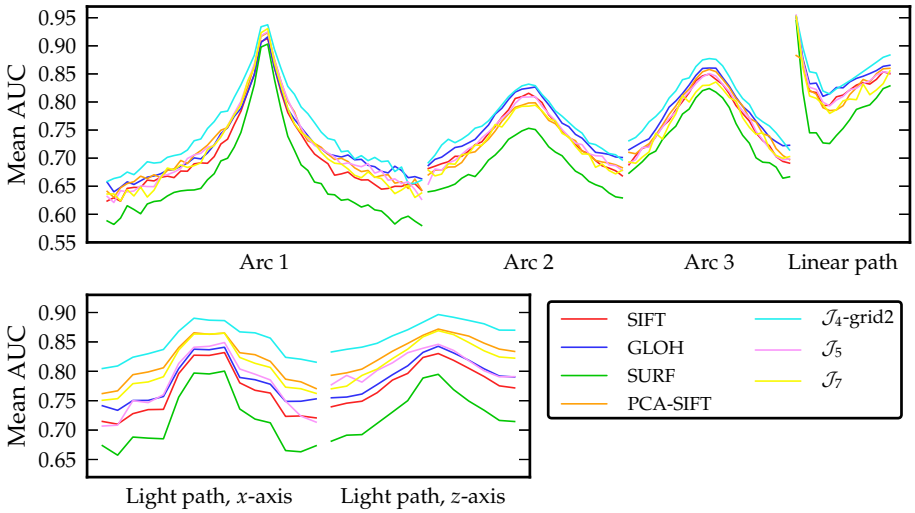

Fig. 4. Descriptor performance on DoG interest points

variations, however, for changes in viewpoint, their performance is mediocre in most cases. The performance of SIFT and GLOH is very correlated with an advantage to GLOH. Lastly, the SURF descriptor performs poorly in the majority of the evaluation scenarios.

As the final experiment, we will examine to what extent descriptor performance is affected by surface structure of the scene content. More specifically, we want to see if planar surfaces vs. non-planar surfaces have influence on the discriminability of the descriptors. To represent planar surfaces, we use a total of 17 scenes containing miniature houses, books and building materials (wooden planks and bricks). To represent non-planar surfaces, we use a total of 22 scenes containing fabric, plush toys, vegetables and beer cans. We evaluate the descriptor performance using DoG and HarAff interest points (note that HarAff estimates an affine transformation of the interest points, which cannot be assumed for non-planar surfaces). The results are shown in Fig. 7. We have omitted the results for illumination changes to save space and because they show similar results. We see a notable difference in descriptor performance depending on the scene type. Feature matching on non-planar scenes is more difficult as we see the AUC is generally lower than for planar scenes in the extreme camera positions on the arcs and the linear path. It seems that the jet descriptors are more robust on non-planar surfaces as their performance decreases less than the other descriptors (especially for large view angle changes). Notice also the difference between the Harris-Affine and DoG detectors; in general the Harris-Affine detector produce more noisy AUC mean curves, which indicates the presence of more noise in the matching of these points. This is also the case for the planar scenes for which the underlying assumptions in the Harris-Affine detector should be fulfilled. Furthermore, DoG has a higher mean AUC close to the reference image

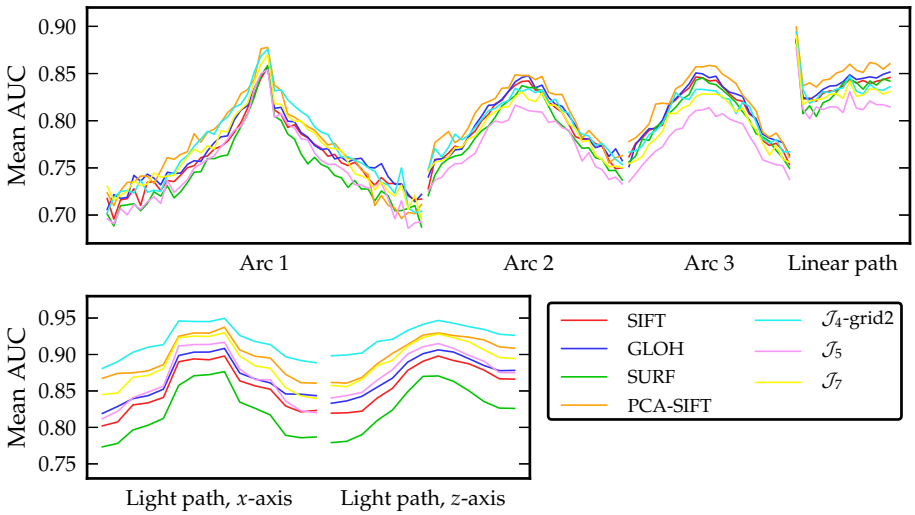

Fig. 5. Descriptor performance on HarLap interest points

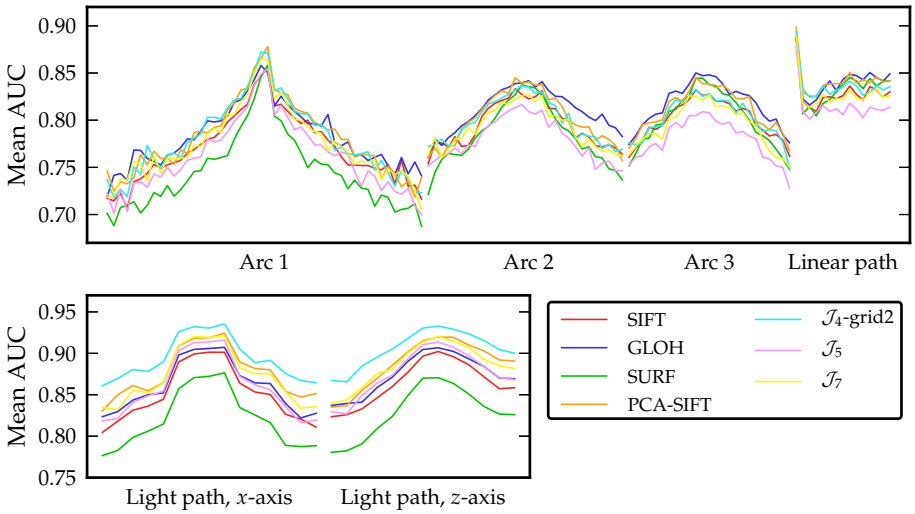

Fig. 6. Descriptor performance on HarAff interest points

in Arc 1. One may speculate whether this difference arises from the different geometries detected by DoG and Harris-Affine – blobs versus corners – or from the affine correction performed by the Harris-Affine detector.

5 Discussion

From the experiments, we have seen that jet-based descriptors offer a viable alternative to state-of-the-art local image descriptors. More specifically, the multi-local jet descriptor $\mathcal{J}_4$-grid2 has shown superior results for illumination changes, competitive performance for changes in viewing angle and a performance slightly under par for scale variations, except for DoG interest points on which we obtain superior performance for all perturbations. Thus, we have shown that the popular histogram approaches, as applied in e.g. SIFT [1] and its extensions [2,4,5], is not crucial for achieving good performance. Additionally, we have shown that the use of higher-order differential image structure allows us to reduce the complexity of the multi-locality (recall that e.g. SIFT is constructed from 16 histogram sampling points). In fact, a single jet can be competitive with both SIFT and GLOH in some situations. The reduction of multi-local sampling points leads to a much lower descriptor dimensionality, which has typically been obtained by adding a PCA step to the description algorithm.

The jet descriptors have, similar to PCA-SIFT, an interpretation as a basis representation of the image patch. However, instead of learning a basis for the patch, our descriptor relies on the monomial basis of the Taylor expansion of the patch. This interpretation give us a low-dimensional descriptor which allows us to reconstruct the patch directly from the descriptor itself [20].

Descriptor performance has been evaluated using interest points generated by different popular detectors. We consider this an important part of our descriptor evaluation, as we have seen that the relative descriptor performance to a large extent is dependent on the detector. Thus, one should be very careful when drawing conclusions from a single interest point detector as it clearly will favor some descriptors over others. From our results, we have seen that the jet descriptors offer superior performance on DoG interest points in all perturbation scenarios. For MSER, HarAff and HarLap interest points, the jet approach offers competitive performance.

The promising results of our jet descriptors are at first glance peculiar considering previous non-favorable evaluations of descriptors based on image differentials [2,8]. Our approach differs in the following ways: We use the local k-jet directly (instead of differential invariants), we extract differentials to a higher order and we employ multi-local sampling of the jets. Including spatial sampling of jets and increasing differential order is essential to the success of our descriptor. We speculate that part of the explanation for the good performance of our jets is that the DTU Robot dataset contains non-planar surfaces (as we have seen, jet descriptors are more robust in these sitations). Previous evaluations of differential descriptors have been on datasets containing to a large extent planar or near-planar surfaces.

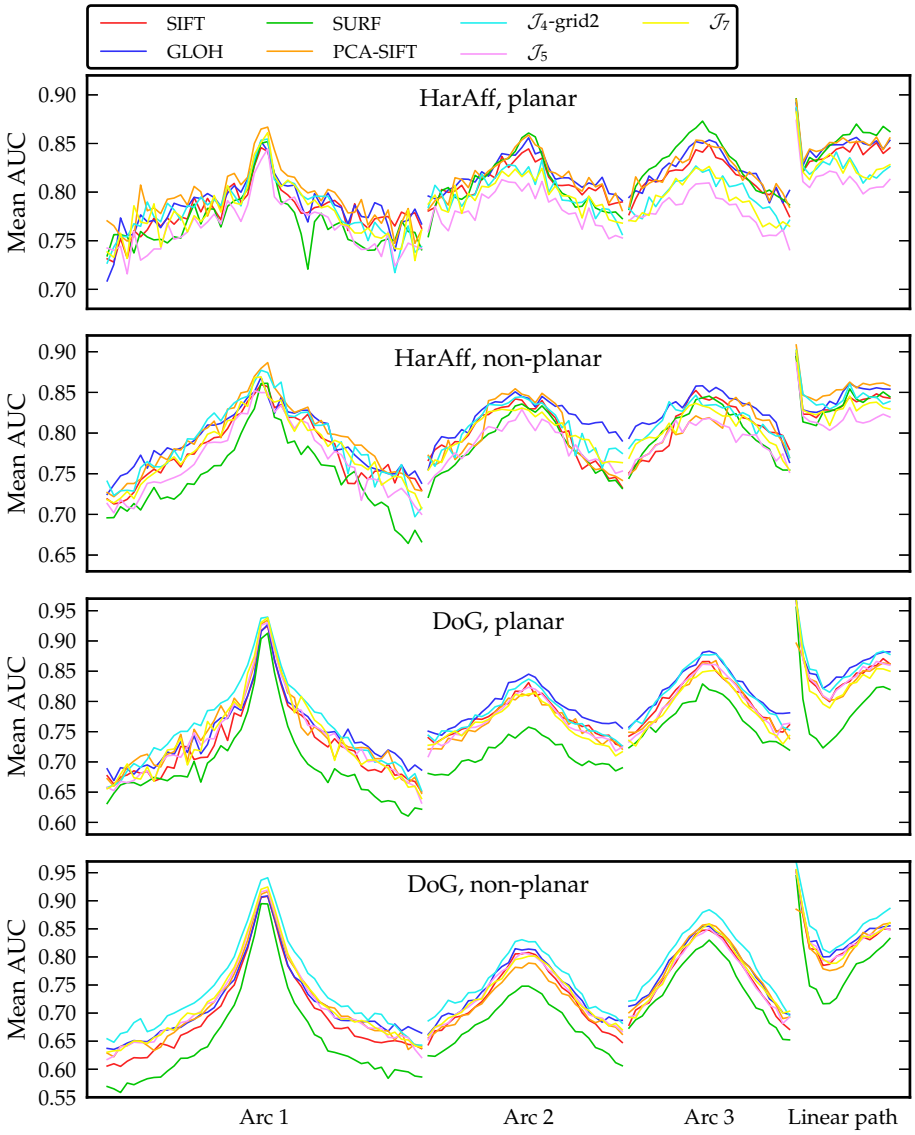

Fig. 7. Descriptor performance for planar and non-planar image surface structures

Our jet descriptors are very simple to implement and rely on very few parameters (compared to histogram-based descriptors) making them easy to configure. We remark that $\mathcal{J}_4$-grid2 requires 14 convolutions of the input patch which allows for a fast implementation in hardware. Single-scale jets can be implemented even more efficiently from a series of point-wise products.

Taking the positive performance results and the simplicity of the descriptor into account, we believe that the multi-local jet descriptor can prove to be valuable for many computer vision applications.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
2. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1615–1630 (2005)
3. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), vol. 2, pp. 506–513 (2004)
4. Winder, S., Hua, G., Brown, M.: Picking the best daisy. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 178–185 (2009)
5. Tola, E., Lepetit, V., Fua, P.: Daisy: An efficient dense descriptor applied to wide-baseline stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 815–830 (2010)
6. Florack, L., ter Haar Romeny, B.M., Viergever, M., Koenderink, J.: The gaussian scale-space paradigm and the multiscale local jet. International Journal of Computer Vision 18, 61–75 (1996)
7. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 530–535 (1997)
8. Balmashnova, E., Florack, L.: Novel similarity measures for differential invariant descriptors for generic object retrieval. Journal of Mathematical Imaging and Vision 31, 121–132 (2008)
9. Laptev, I., Lindeberg, T.: Local Descriptors for Spatio-temporal Recognition. In: MacLean, W.J. (ed.) SCVMA 2004. LNCS, vol. 3667, pp. 91–103. Springer, Heidelberg (2006)
10. Aanæs, H., Dahl, A., Steenstrup Pedersen, K.: Interesting interest points: A comparative study of interest point performance on a unique data set. International Journal of Computer Vision 97, 18–35 (2012)
11. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision 65, 43–72 (2005)
12. Kaneva, B., Torralba, A., Freeman, W.T.: Evaluating image feaures using a photorealistic virtual world. In: IEEE International Conference on Computer Vision (2011)
13. Vedaldi, A., Ling, H., Soatto, S.: Knowing a Good Feature When You See It: Ground Truth and Methodology to Evaluate Local Features for Recognition. In: Cipolla, R., Battiato, S., Farinella, G. (eds.) Computer Vision. SCI, vol. 285, pp. 27–49. Springer, Heidelberg (2010)
14. Pedersen, K.S.: Statistics of natural image geometry. PhD thesis, University of Copenhagen, Department of Computer Science, Denmark (2003)

15. Markussen, B., Pedersen, K.S., Loog, M.: Second order structure of scale-space measurements. Journal of Mathematical Imaging and Vision 31, 207–220 (2008)
16. Dahl, A.L., Aanæs, H., Pedersen, K.S.: Finding the best feature detector-descriptor combination. In: The First Joint Conference of 3D Imaging, Modeling, Processing, Visualization and Transmission (2011)
17. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110, 346–359 (2008)
18. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22, 761–767 (2004)
19. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision 60, 63–86 (2004)
20. Lillholm, M., Nielsen, M., Griffin, L.D.: Feature-based image analysis. International Journal of Computer Vision 52, 73–95 (2003)