Finding the Best Feature Detector-Descriptor Combination

Anders Lindbjerg Dahl
DTU Informatics, Technical University of Denmark
Lyngby, Denmark
abd@imm.dtu.dk

Henrik Aanæs
haa@imm.dtu.dk

Kim Steenstrup Pedersen
Department of Computer Science, DIKU, University of Copenhagen
Copenhagen, Denmark
kimstp@diku.dk

Abstract

Addressing the image correspondence problem by feature matching is a central part of computer vision and 3D inference from images. Consequently, there is a substantial amount of work on evaluating feature detection and feature description methodology. However, the performance of the feature matching is an interplay of both detector and descriptor methodology. Our main contribution is to evaluate the performance of some of the most popular descriptor and detector combinations on the DTU Robot dataset, which is a very large dataset with massive amounts of systematic data aimed at two view matching. The size of the dataset implies that we can also reasonably make deductions about the statistical significance of our results. We conclude, that the MSER and Difference of Gaussian (DoG) detectors with a SIFT or DAISY descriptor are the top performers. This performance is, however, not statistically significantly better than some other methods. As a byproduct of this investigation, we have also tested various DAISY type descriptors, and found that the difference among their performance is statistically insignificant using this dataset. Furthermore, we have not been able to produce results collaborating that using affine invariant feature detectors carries a statistical significant advantage on general scene types.

1. Introduction

The computational efficiency of a sparse image representation consisting of salient interest points, also referred to as features, is a major motivation for feature based methods for solving the image correspondence problem. Various detectors and descriptors have been proposed, but the question of how to optimally design an interest point characterization still remains open. The success of feature-based methods depends on the quality of the local characterization. In general it is not an easy task to judge the performance of such methods, because it is hard to validate if correspondence exist. However given knowledge about the geometry of the observed scene, it becomes easy to verify if two interest points corresponding in feature space also corresponds in the real scene. We therefore propose to use the DTU Robot dataset with known surface geometry presented in [1, 2] (see Sec. 2 for a brief description and Fig. 1). Based on this dataset we are able to systematically analyze the design of feature methods and due to the large variation in scene types we can judge the statistical significance of our findings.

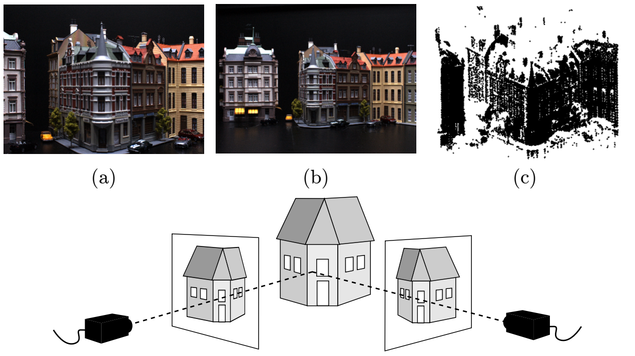

Figure 1. Example of data and setup. Two images of the same scene with one close up (a) and one distant from the side (b), and the reconstructed 3D points (c). Corresponding interest points can be found using the geometric information of the scene with known camera positions and 3D scene surface as schematically illustrated in (d). Illustration from [1].

Finding correspondence between image pairs using interest points is based on the assumption that common interest points will be detected in both images. For this to be useful, corresponding interest points have to be localized precisely on the same scene element, and the associated region around each interest point should cover the same part of the scene. Commonly, candidate points are detected using an interest point *detector* and a description of the local image structure – the so-called *descriptors* – surrounding the interest points are extracted. Following the extraction of descriptors, a comparison of these is made using a relevant similarity metric in order to determine correspondence between interest points. The rationale is that descriptors capture the essential visual appearance of the scene region covered by the interest point, and as a consequence the same scene point seen from different viewpoints and/or with different lighting should have similar descriptors. Therefore descriptors should preferably be invariant, or approximately, with respect to changes in viewpoint and lighting.

Early work on correspondence from local image features was based on rotation and scale invariant features [13, 19], and interest points from planer scenes was evaluated in [20]. Later the interest points have been adapted to affine transformation, to obtain robust characterization to larger viewpoint changes. These methods have been surveyed in [17], but the performance has been evaluated on quite limited datasets consisting of ten scenes each containing six images. The suggested evaluation criteria have since been used in numerous works together with this small dataset.

Different approaches have been taken when describing the local visual appearance of interest points. A majority of approaches extract some descriptive feature, such as histograms of differential geometric image properties in each pixel [13, 16, 19, 22], using integral images [5, 4], or the responses of steerable filters [9], differential invariants or local jets [3, 7, 12, 20]. The SIFT [13], GLOH [16], and DAISY [22, 23, 25, 26] descriptors also includes a spatial pooling step in order to agglomerate the descriptive feature in an arrangement around the interest point. A selection of descriptors have previously been evaluated in [16] on the same dataset as used in [17]. Again the limitations of the dataset restricts the ability to generalize the results from this survey to a wider class of scene types and more natural variation in illumination.

The ground truth in the data from [17] was obtained by an image homography. This limits the scene geometry to planar surfaces or scenes viewed from a large distance where a homography is a good approximation. Fraundorfer and Bishof [8] addressed this limitation by generating ground truth and requiring that a matched feature should be consistent with the camera geometry across three views. In Winder *et al.* [26, 11, 25, 6] results from Photo Tourism [21] were used as ground truth.

Moreels and Perona [18] evaluated feature descriptors similar to [8] based on pure geometry by requiring three view geometric consistency with the epipolar geometry. In addition they used a depth constraint based on knowledge about the position of their experimental setup. Hereby they obtained unique correspondence between 500-1000 detected points from each object. The limitation of their experiment is the use of relatively simple scenes with mostly single objects resulting in little self-occlusion. However, self-occlusions are very frequent in real world scenes and many interest points are typically found near occluding boundaries, limiting the applicability of their conclusions.

The aim of this work is to compare pairs of feature detectors and descriptors, to find the best combination. To keep the computational burden manageable the number of candidates have to be limited, and we thus only use candidates which have previously been reported to perform well. As for the detectors we choose Harris, Harris Affine, Harris Laplace, Hessian Laplace, Hessian Affine, MSER, and Difference of Gaussian (DoG), because they are popular and reported to work well in the literature [1, 24].

As for the feature descriptors, the state of the art is currently the SIFT [13] and DAISY descriptors [22, 23, 25, 26] which we choose to use and implement using the framework of Winder and Brown [26]. We also include conventional (normalized) cross correlation as a baseline, since much work has been done using this descriptor. The DAISY descriptors however cover a wide range of descriptors; as such we choose to divide our analysis into two, where we first identify the best DAISY descriptors on a subset of the detectors. This is the subject of Sec. 3, where 21 different variants of the DAISY descriptor are evaluated. Each combination is evaluated using ROC-curves (Receiver Operating Characteristics). Two representative descriptors are carried on to the last part of the analysis, reported in Sec. 4, where a matrix of the seven detectors and four descriptors are evaluated. A discussion of our results and recommendations is found in Sec. 5.

2. Data and Evaluation

In this investigation we use the DTU Robot dataset [1, 2][1] illustrated in Fig. 1. This dataset is constructed under controlled settings using an industrial robot. The set consists of 60 complex scenes, and Fig. 2 shows how each scene is viewed from 119 positions with known camera geometry. The dataset also incorporates light variation, but in this work we only focus on diffuse lighting. In addition the 60 scenes have been surface scanned using structured light. Together with the camera geometry this allows us to accurately determine the correct camera correspondences *without* matching visual features. In real outdoor scenes, as presented in [25], there is no alternative to have ground truth

[1] http://roboimagedata.imm.dtu.dk/

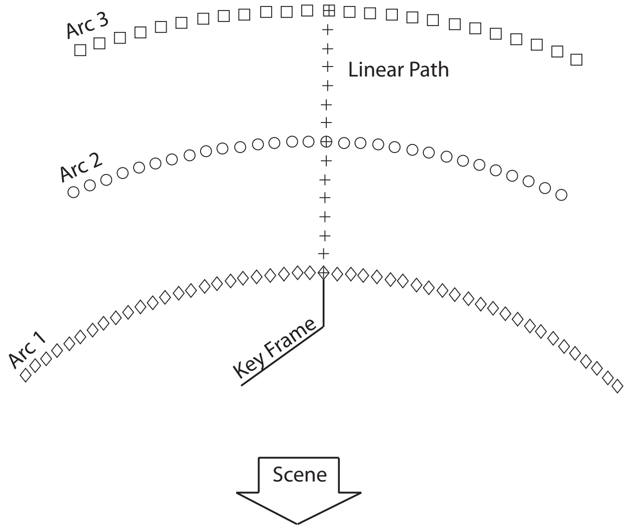

Figure 2. The central frame in the nearest arc is the key frame, and the surface reconstruction is attempted to cover most of this frame. The three arcs are located on circular paths with radii of 0.5 m, 0.65 m and 0.8 m, which also defines the range of the linear path. Furthermore, Arc1 spans $+/-40°$, Arc2 $+/-25°$ and Arc3 $+/-20°$. Illustration from [1].

based on feature matching, but this could likely bias the result.

2.1. Evaluation criteria

The evaluation framework used is similar to the one reported in [1], which only includes an evaluation of the matching performance of different detector methods on the DTU Robot dataset. We want to determine if a pair of corresponding features are correct or not, where correspondence is found by the Euclidean distance between feature descriptors. Fig. 2 illustrates how the features are matched between one key frame and all other images. Fig. 3 shows the two criteria that we use for determining correct correspondence. Correct matches have to be within 2.5 pixels of the epipolar line *and* the corresponding 3D point must be within a 5 pixel error margin corresponding to approximately 3 mm.

Given an image pair, where one image is the key frame, a detector-descriptor pair is evaluated by

1. For each feature in the key frame find the distance to the best δ_b and the *second* best δ_s matching feature in the other image.

2. For each feature correspondence compute the ratio, $r = \frac{\delta_b}{\delta_s}$, between the match score of the second best and the best correspondence. It is also determined if the best match is correct or not.

3. Using this ratio, r, as a predictor for correct matches, c.f. [13], the ROC (Receiver Operating Characteristic) curve, as a function of r, is constructed based on all features in an image pair. We compare the area under the ROC curve (AUC). The area is between zero and one, where one indicates perfect performance of the detector-descriptor pair.

4. The AUC is used as the performance measure of a detector-descriptor combination on a pair of images.

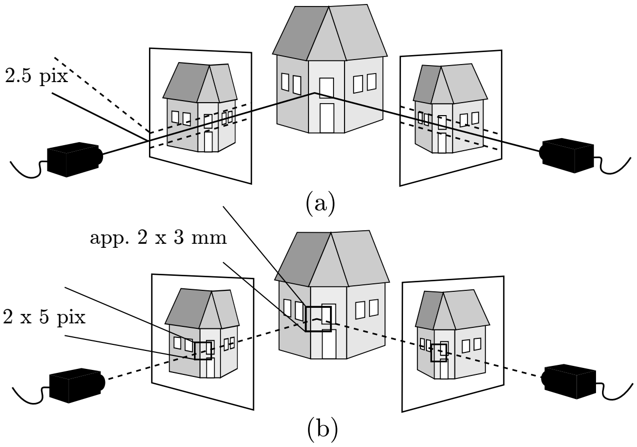

Figure 3. Matching criteria for interest points. This figure gives a schematic illustration of a scene of a house and two images of the scene from two viewpoints. (a) The consistency with epipolar geometry, where corresponding descriptors should be within 2.5 pixels from the epipolar line. (b) Window of interest with a radius of 5 pixels and corresponding descriptors should be within this window, which is approximately 3 mm on the scene surface. Ground truth is obtained from the surface geometry. Illustration from [1].

These AUCs are the basis for our statistical analysis. The AUC is chosen as a performance measure, in line with [25], because it elegantly removes the need to balance between many false positive or many false negatives. As a result it strongly relates to the underlying discriminative power of the method.

We compare different detector-descriptor methods by computing the mean performance, i.e. the mean AUC over the 60 sets for each position, c.f. Fig. 6, 7 and Tab. 2. Based on the central limit theorem, we assume these means to be normal distributed. We compare the means using students t-test

$$\frac{\mu_1 - \mu_2}{\hat{\sigma}} \ , \tag{1}$$

where μ_1 and μ_2 are the two means to be compared and $\hat{\sigma}$ is an estimate of the standard deviation. When computing an estimate of the variances, $\hat{\sigma}^2$, we perform an analysis of variance, assuming that for a given method and a given problem, performance is given by two factors

$$\text{Performance} = \text{Problem Difficulty} + \text{Method} + \text{Noise} \ .$$

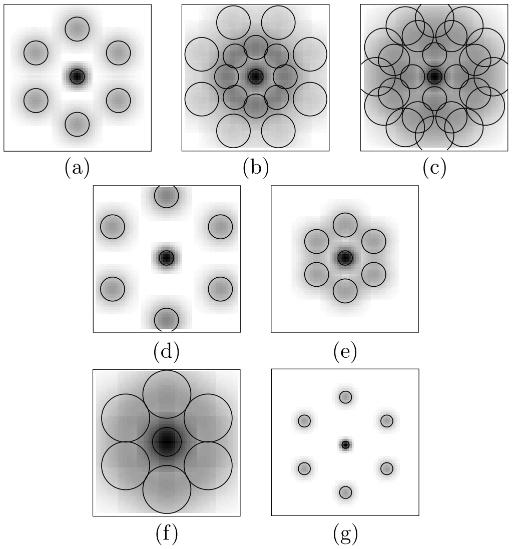

Figure 4. Layout of the descriptors for spatial summation. The circles mark the size of the sample points and the dark color shows the Gaussian weighing. *First row* one ring with six samples – (1-6) (a), two rings with eight samples in each – (1-8-8) (b), three rings with four, eight and twelve samples – (1-4-8-12) (c). *Second row* one ring with six samples – large footprint – (1-6 lf) (d), small footprint – (1-6 sf) (e). One ring with six samples – large sample area – (1-6 lg) (f), small sample area – (1-6 sg) (g).

Since we are interested in comparing the methods the variance due to the Problem Difficulty is factored out, which reduces the overall variance, $\hat{\sigma}^2$, making it easier for a difference in means to be significant.

2.2. Implementation

All feature detectors are computed by implementations provided by the authors of [13, 14, 15][2], whereas we implemented our own interest point descriptors. They are estimated on an affine warped image patch sampled according to the parameters obtained from the interest point detection and rotated to one dominant gradient direction. The image patch is sampled with a radius of three times the scale of the feature point and we discard points that exceed the image borders. We found this to be a good tradeoff between performance and number of discarded sample points. In the experiments described in Sec. 3 we use a patch size of 66 × 66 pixels whereas the patches in the experiments in Sec. 4 are 30 × 30. This is especially a consequence of the pixel similarity estimates where we have feature vectors of 900 dimensions. Using the 66 × 66 pixel patches this would be 4356 dimensions, which approximately slows the calculation down with a factor four. We only observed a minor loss in precision, which is shown in Tab. 1 "spatial layout – 1-8-8" should be compared to "HesAff" and "HarAff" – "DAISY-I" and "DAISY-II" in Tab. 2. It shows a performance loss of 0.013 caused by reduction in patch size.

Our implementation of the DAISY descriptor closely follows the description of Winder *et al.* [25][3]. To ensure that the only difference between the DAISY and SIFT descriptors were the sampling, we chose to implement our own SIFT descriptor. To validate the performance we did a small experiment to compare to the original implementation of Lowe [13][4], and we obtained similar performance with patches of 66 × 66 pixels and about 5% fewer matching descriptors with the 30 × 30 patches.

3. Comparing DAISY descriptors

Brown *et al.* [6] presents a framework for optimizing feature descriptors. They have chosen the DAISY-type descriptor presented in Winder and Brown [26], because it is easily reconfigurable. The optimization is based on three outdoor scenes where ground truth is obtained from the bundler software [21], which is based on the SIFT framework [13]. In this experiment we have performed a similar investigation to Brown *et al.*, but based on the extended DTU Robot dataset, where ground truth geometry is based on precise calibration and structured light scanning. In order to keep the computational burden manageable, we only did this experiment on the Harris affine and Harris Laplace features.

The descriptors proposed by Brown *et al.* [6] are varied in the spatial layout and differential-geometric response. The spatial layout that we have tested are illustrated in Fig. 4. We have varied the number of sample points, the size of sample points and their relative distance. We employ three differential-geometric responses – the directional binned gradients in four and eight directions (Type 1), average positive and negative gradients (Type 2) and steerable filters (Type 3). The experimental result is summarized in Tab. 1. This approach closely follows Winder *et al.* [25].

The results show that the effect of changing the differential-geometric response is limited, so it is clearly an advantage to select either Type 1 or 2, because the computational cost of these descriptors is much lower. There is a small advantage in selecting a spatial layout where two rings are sampled, but three ring sampling does not give an improvement. Fig. 5 shows that this advantage is seen for all positions. The dimensionality difference arise from number of sample directions in Type 1, combinations of positive and negative gradients in Type 2, and number of directions of the steerable filters in Type 3. But there is almost no difference in selecting the large dimensionality over the small. There is a clear difference in Scene type, where the AUC is significantly higher for less specular objects like fabric

[2] http://www.robots.ox.ac.uk/~vgg/research/affine/

[3] http://cvlab.epfl.ch/~brown/patchdata/patchdata.html

[4] http://www.cs.ubc.ca/~lowe/keypoints/

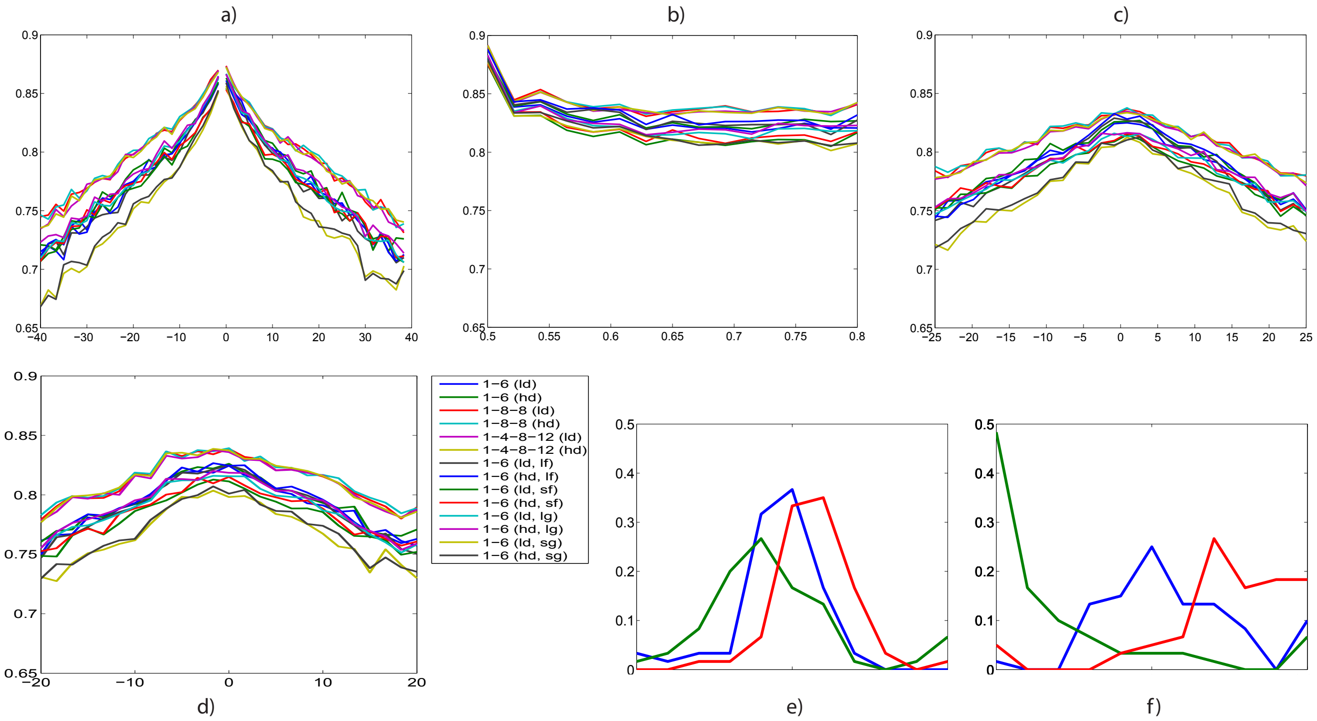

Figure 5. Performance evaluation of the DAISY descriptor. Average AUCs for Type 2 descriptors are shown in (a - d). The vertical axis in the graphs show the AUC, and the horizontal is the angle (a,c,d) and distance (b) relative to the key-frame. Each graph corresponds to the sample path shown in Fig. 2, with Arc 1 (a), Linear Path (b), Arc 2 (c) and Arc 3 (d). The labels relate to the descriptor design shown in Fig. 4. In (e - f) probability density functions for different descriptor designs are shown for a 30° angle where (e) is affine interest points and (f) is non-affine. This shows that with a sparse sampling the performance goes down for the non-affine, but the affine invariance can be compensated by a dense sampling.

than for specular objects like beer cans. The difference between affine and non-affine feature detectors is surprisingly small, which might be a result of complexity of the evaluated scenes with many occluding boundaries. The findings regarding affine detectors are confirmed by the experiments presented in Sec. 4.

From this study, we choose the two-ring DAISY descriptor with small (DAISY-I) and large dimensionality (DAISY-II) for further analysis. This is done together with SIFT and a vector of simple pixel intensities (normalized cross correlation). These four descriptors are analyzed in combination with seven feature detectors.

4. Comparing Detector-Descriptor Combinations

In this section we present the evaluation of detector-descriptor combinations with the aim of finding the best performers. We compare a combination of the four feature descriptors (SIFT, DAISY-I, DAISY-II and cross correlation) with seven feature detectors. These detectors are Harris corner detector [10], Harris Laplace, Harris affine, Hessian Laplace, Hessian affine [17], MSER [14], and Difference of Gaussians (DoG) [13]. The combined result is summarized in Tab. 2 and Fig. 6

To evaluate the significance of the performance difference we have estimated the average standard deviation $\hat{\sigma}$ of (1). Overall we obtain $\hat{\sigma} = 0.08$, but if we exclude cross correlation, which has a higher variance than all others, then we obtain $\hat{\sigma} = 0.05$. To give an idea of significance based on Student's t-test from (1) we consider a difference larger than 0.05 as significantly different on a 84% confidence level and 0.1 as significant on a 98% level.

The performance is computed for all 28 combinations on all 119 camera positions, where the distribution of the performance was evaluated over all 60 scenes. Our central evaluation criterion is the mean over these 60 scenes for a given position and detector-descriptor combination. Due to space limitations we are only able to present a summarized evaluation as shown in Fig. 6 and Tab. 2 outlining our conclusions.

Fig. 6 shows a combination with the same detector but different descriptors. Cross correlation is clearly outperformed by the other descriptors. SIFT and DAISY has almost identical performance, and Tab. 2 shows that their average difference is less than 0.015, which is statistically in-a)

b)

c)

d)

Figure 6. Mean AUC for the MSER detector displayed for all four descriptors and for all positions. The vertical axis in the graphs show the AUC, and the horizontal is the angle (a,c,d) and distance (b) relative to the key-frame. Each graph corresponds to the sample path shown in Fig. 2, with Arc 1 (a), Linear Path (b), Arc 2 (c) and Arc 3 (d). It is seen that the SIFT and the two DAISY descriptors have very similar performance, compared to a $\hat{\sigma} = 0.05$, but outperform the correlation.

Comparison	**Type**	**Performance**
Descriptor type	Type 1	0.781
	Type 2	0.785
	Type 3	0.791
Spatial layout	1-6	0.786
	1-8-8	0.804
	1-4-8-12	0.802
	1-6 lf	0.784
	1-6 sf	0.778
	1-6 lg	0.784
	1-6 sg	0.763
Descriptor dimensionality	Small	0.783
	Large	0.788
Scene types	Houses	0.751
	Books	0.791
	Fabric	0.831
	Greens	0.799
	Beer cans	0.696
Affine vs. Laplace	Laplacian	0.783
	Affine	0.788

Table 1. Mean AUC for different groupings of the descriptor types. The table shows mean value of all positions. In Fig. 4 the spatial layout is shown.

	Corr	**SIFT**	**DAISY-I**	**DAISY-II**	**Avg.**
Har	0.615	0.767	0.729	0.741	0.713
HarAff	0.629	0.818	0.791	0.798	0.759
HarLap	0.635	0.814	0.784	0.790	0.756
HesAff	0.636	0.795	0.773	0.779	0.746
HesLap	0.630	0.757	0.740	0.742	0.717
MSER	0.648	**0.846**	0.826	0.832	0.788
DOG	0.646	**0.849**	0.837	**0.843**	0.794
Avg.	0.634	0.807	0.783	0.789	0.753

Table 2. Mean AUC over all positions for the feature detector and descriptor combinations. Top 3 performers highlighted with boldface (**Har** is Harris corners, **HarAff** is Harris affine, **HarLap** is Harris Laplace, **HesAff** is Hessian affine and **HesLap** is Hessian Laplace feature detectors respectively).

significant.

In Fig. 7 the SIFT descriptor is shown in combination with the seven detectors. We chose to show SIFT, but very similar results were obtained for the DAISY descriptors. Here there is a difference in performance where MSER and DOG detectors perform about one standard deviation better than the Harris affine and Harris Laplace detectors, and about two standard deviations better than the Harris based detectors, which is statistically significant. Harris corner detector with no scale adaption performs well when the

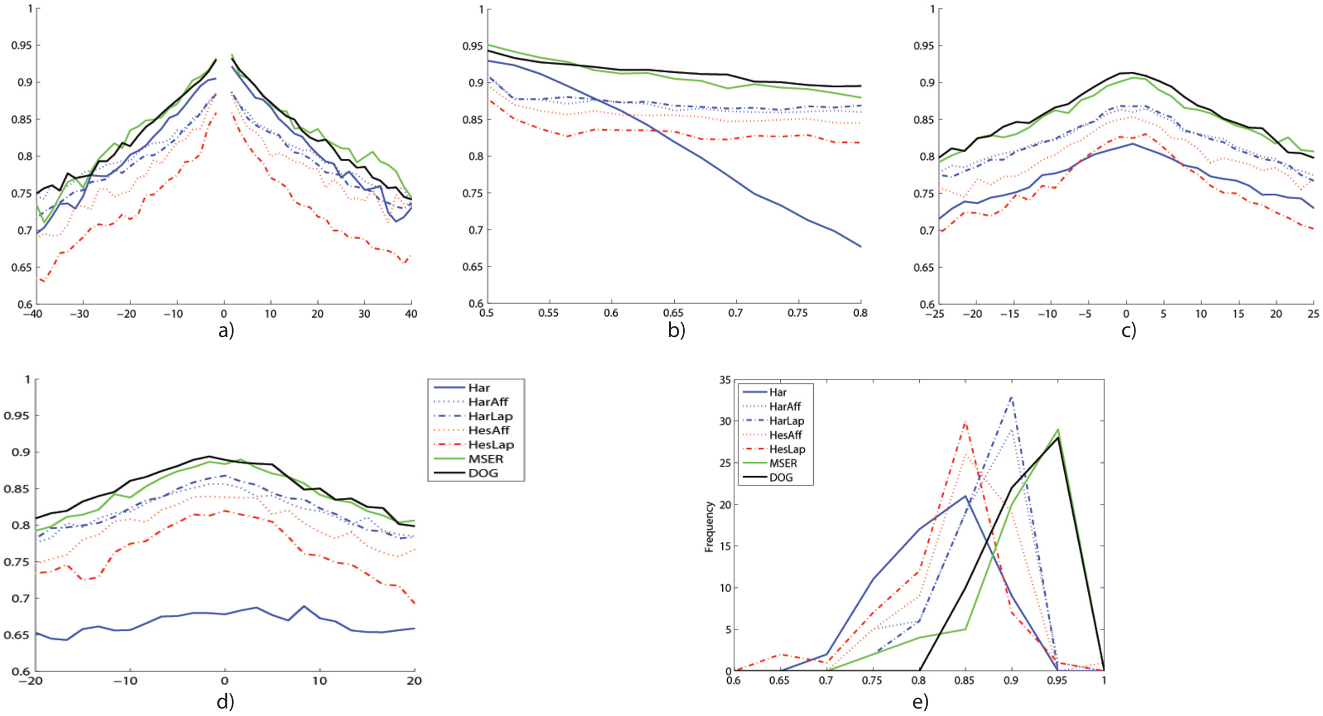

Figure 7. Mean AUC for the SIFT descriptor displayed for all seven detectors and all positions. The vertical axis in the graphs show the AUC, and the horizontal is the angle (a,c,d) and distance (b) relative to the key-frame. Each graph corresponds to the sample path shown in Fig. 2, with Arc 1 (a), Linear Path (b), Arc 2 (c) and Arc 3 (d). Here it is seen that the MSER and DOG detectors are the top performers, outperforming the Harris based detectors on a statistically borderline level, and significantly outperforming the hessian based descriptors. The performance of the 'pure' Harris corner detector is very scale dependent. Similar results are obtained for the two DAISY descriptors, as indicated in Fig. 6. The validity of our findings is further cooperated by considered the probability distribution functions for each position, in (e) the pdf is shown for 0.86° of Arc 2.

scale change is not to large.

So, our experiments suggest that the best choice is a DOG or MSER detector with a SIFT or DAISY descriptor, or a perhaps a Harris corner detector if the scale change is low. The dataset used also has different categories of scene types like 'fabric', 'books', 'model houses', etc. and running the experiments on a specific scene type did not change the overall picture. Compared to the results in [1], where the recall rate of detectors was evaluated on the same dataset, it is interesting to see that the best performers in a full feature tracking frame work are not identical to the ones with the best recall rates. Again this implies that the discriminative power of the extracted features vary for different feature detectors. This last point is especially noteworthy for the MSER detectors. Both the descriptor experiment presented in Sec. 3 and this combined experiment show that an affine detector has an advantage, but this advantage is small compared to variance making it statistically insignificant, see Fig. 8.

5. Discussion

Based on the experiments reported in this paper the general conclusion is that the best detector-descriptor combination is either the DOG or MSER detectors and SIFT or DAISY descriptors. If the scale change is low a Harris corner detector would be superior and also faster and simpler to run and implement. The experiments also show, that many other performance differences exist, which confirm other studies, but these differences are not statistically significant. This demonstrates a need for considering statistical significance when performing these type of comparisons, necessitating the use of large datasets to make meaningful estimates of significance and variance, such as the dataset used here [1].

Furthermore, it is interesting to note that the DOG detectors perform much better than the Hessian type detectors, although they are very similar, i.e. the DOG is basically a well-engineered approximation of the Laplace filter, which is equal to the trace of the Hessian. This indicates that perhaps a better-engineered version of the Harris Laplace corner detector, inspired by the DOG detector, could be made. This is especially interesting in the light that the Harris corners performed better than the Hessians.

A last point of curiosity is that we have not been able to produce results collaborating that using affine invariant feature detectors carries a statistical significant advantage on general scene types. However this type of invariance may have merit in e.g. 3D reconstruction of urban type scenes or other near-planar scenes.

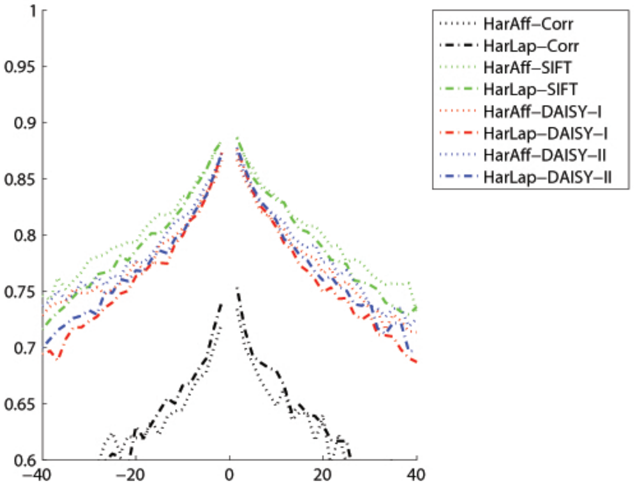

Figure 8. Affine vs. non-affine (Lap). Affine performs slightly better with large angles, but the improvement is not significant.

References

[1] H. Aanæs, A. L. Dahl, and K. S. Pedersen. On recall rate of interest point detectors. In *3DPVT*, 2010. 1, 2, 3, 7, 8

[2] H. Aanæs, A. L. Dahl, and V. Perfernov. Technical report on two view ground truth image data. Technical report, DTU Informatics, Technical University of Denmark, 2009. 1, 2

[3] E. Balmashnova and L. Florack. Novel similarity measures for differential invariant descriptors for generic object retrieval. *JMIV*, 31(2-3):121–132, 2008. 2

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 2

[5] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *ECCV*, pages 404–417, 2006. 2

[6] M. Brown, G. Hua, and S. Winder. Discriminative Learning of Local Image Descriptors. *IEEE T-PAMI*, 2010. 2, 4

[7] L. Florack, B. ter Haar Romeny, J. Koenderink, and M. Viergever. Cartesian differential invariants in scale-space. *JMIV*, 3(4):327–348, 1993. 2

[8] F. Fraundorfer and H. Bischof. Evaluation of local detectors on non-planar scenes. In *Proc. 28th workshop of AAPR*, pages 125–132, 2004. 2

[9] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE T-PAMI*, 13(9):891–906, 1991. 2

[10] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conf.*, pages 147–151, 1988. 5

[11] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. *ICCV*, pages 1–8, 2007. 2

[12] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987. 2

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2, 3, 4, 5

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. 4, 5

[15] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004. 4

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE T-PAMI*, 27(10):1615–1630, 2005. 2

[17] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool. A comparison of affine region detectors. *IJCV*, 65(1-2):43–72, 2005. 2, 5

[18] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. *IJCV*, 73(3):263–284, 2007. 2

[19] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE T-PAMI*, 19(5):530–535, 1997. 2

[20] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *IJCV*, 37(4):151–172, 2000. 2

[21] N. Snavely, S. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, 2008. 2, 4

[22] E. Tola, V. Lepetit, and P. Fua. A Fast Local Descriptor for Dense Matching. In *CVPR*, 2008. 2

[23] E. Tola, V. Lepetit, and P. Fua. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *IEEE T-PAMI*, 32(5):815–830, May 2009. 2

[24] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280, 2008. 2

[25] S. Winder, G. Hua, and M. Brown. Picking the best daisy. In *CVPR*, 2009. 2, 3, 4

[26] S. A. J. Winder and M. Brown. Learning local image descriptors. In *CVPR*, pages 1–8, 2007. 2, 4