Feature Evaluation with High-Resolution Images

Kai Cordes, Lukas Grundmann, and Jörn Ostermann

Leibniz Universität Hannover, Germany,
{cordes,grundmann,ostermann}@tnt.uni-hannover.de

Abstract. The extraction of scale invariant image features is a fundamental task for many computer vision applications. Features are localized in the scale space of the image. A descriptor is build for each feature which is used to determine the correspondence to a second feature, usually extracted from a second image. For the evaluation of detectors and descriptors, benchmark image sets are used. The benchmarks consist of image sequences and homographies which determine the ground truth for the mapping between the images. The repeatability criterion evaluates the detection accuracy of the detectors while precision and recall measure the quality of the descriptors.
Current data sets provide images with resolutions of less than one megapixel. A recent data set provides challenging images and highly accurate homographies. It allows for the evaluation at different image resolutions with the same scene content. Thus, the scale invariant properties of the extracted features can be examined. This paper presents a comprehensive evaluation of state of the art detectors and descriptors on this data set. The results show significant differences compared to the standard benchmark. Furthermore, it is shown that some detectors perform differently on different resolutions. It follows that high resolution images should be considered for future feature evaluations.

1 Introduction

Scale invariant features play an important role in many computer vision applications, such as object recognition or scene reconstruction. These applications require discriminative and accurate features on images with large changes in illumination and perspective [9,18].
New approaches in feature detection [2,3,11,14,16,17] and description [1,2,3,7,11,15,16] usually use the reference test set and the evaluation protocols provided in [14,15]. It contains sequences of still images (800×640 pixel resolution) with changes in illumination, rotation, perspective, and scale. Only two out of eight sequences provide perspectively distorted images. The mapping from one image to the next is restricted to a homography. For the benchmark test, the ground truth homography matrices are provided. The most important criterion for the accuracy of the detectors is the repeatability criterion. The descriptors are evaluated with precision and recall curves.
Nowadays, high resolution images become more and more important. Resolutions of 4K (4000×3000 pixels) are required for the digital cinema. Even current smartphones provide large resolutions, such as the iPhone 6 with eight megapixels. However, feature evaluations are performed on images with resolutions of less than one megapixel. An evaluation of state-of-the-art feature detectors and

descriptors on high resolution images is still missing. Recently, a high resolution benchmark data set was published [1]. It provides image resolution of up to 8 megapixels [6] (a first step towards 4K) and focuses on the scenario of perspectively distorted images, which is demanded by scene reconstruction applications like in [9,18]. Our contribution is the evaluation of state-of-the-art feature detectors and descriptors on the benchmark [6] compared to [14]. We examine which of the detectors and descriptors are able to transfer their performance to large resolutions.

In the following Section 2, the feature detectors and descriptors are introduced. In Section 3, the experimental setup is presented. Section 4 shows the results and Section 5 gives the conclusions.

2 Overview

Several publications give informative overviews on scale invariant feature detectors and descriptors, e.g. [8,12]. Since we concentrate on the evaluation, we just give a short overview of the competitors. The evaluated detectors are Wave [17], A-KAZE [2], ORB [16], BRISK [11], SURF [3], and SIFT [13] (cf. Table 1). The evaluation criterion is the repeatability using the matlab script provided by the authors of [14]. The resulting best detector is used in the descriptor evaluation. The evaluated descriptors are A-KAZE [2], LIOP [19], MROGH [7], GLOH [15], and SIFT [13] (cf. Table 2). In our evaluation, we exclude the descriptors ORB, BRISK, and FREAK. These approaches concentrate on fast computation, and their performance in accuracy is to our experience equal to or lower than SIFT (cf. [4,8]).

Our evaluation aims at finding the most accurate detector together with the best possible descriptor. The implementations are taken as they are provided by the authors (cf. Table 1 and 2) using default parameters. For comparison, we added the computation times found in our experiments in milliseconds per feature, computed on i7 CPU, 3.50 GHz.

Table 1: The detectors which are compared in the results section.

detector	implementation	year published	computation time [ms]
SIFT [13]	Hess code [10]	2004	4.38
SURF [3]	Author's binary	2006	0.54
BRISK [11]	OpenCV 2.4	2011	0.99
ORB [16]	OpenCV 2.4	2011	0.47
A-KAZE [2]	Author's code	2013	1.04
Wave [17]	Author's binary	2013	5.58

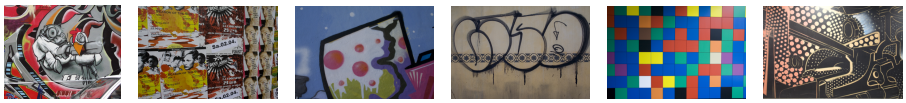
[1] http://www.tnt.uni-hannover.de/project/feature_evaluation/

Table 2: The descriptors with their default descriptor lengths d_l which are used for the comparisons in the results section.

descriptors	implementation	d_l	year published	computation time [ms]
SIFT [13]	Oxford binary	128	2004	1.74
GLOH [15]	Oxford binary	128	2005	1.87
MROGH [7]	Author's code	192	2011	2.35
LIOP [19]	Author's binary	144	2011	1.43
A-KAZE [2]	Author's code	61	2013	7.97

3 Experimental Setup

Most evaluations employ the benchmark provided in [14]. We mainly use the recently published benchmark data set [6] for two reasons: (1) it provides higher accuracy [5] and image resolution (even different resolutions for the same scenes), (2) it concentrates on the *perspective change* scenario which is in the focus of this evaluation. For comparison, we include the most popular perspective change sequence *Graffiti* of [14]. The first images of the sequences are shown in Figure 1. We use the repeatability criterion for the detectors evaluation while precision and recall determines the quality of the descriptors. The overlap error parameter is set to 0.4 [14].

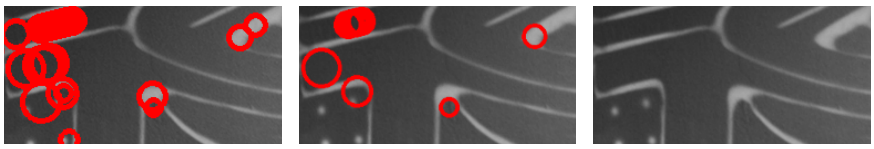


(a) Graffiti (b) Posters (c) There (d) Grace (e) Colors (f)Underground

Fig. 1: First images of the input image sequences. The resolution is 800×640 for (a) *Graffiti* and up to 3456×2304 for the sequences (b) - (f).

3.1 Feature Detection

The detectors provide a surprisingly large variation in extracted numbers of features. The numbers of features heavily dependent on texture, perspective, and resolution of the considered image. The detectors provided by OpenCV (ORB, BRISK) tend to extract many more features (sometimes more than 40000) than the others. Thus, we have to limit the number of detected points. For this purpose, the attribute *response* is used for each feature in OpenCV. For the evaluation, we sort the features by their response and choose the first n_f features. The number n_f is determined by the maximum of detected features by the others (A-KAZE, Wave, SURF, SIFT). In most cases, A-KAZE provides the largest number of features. The results for the repeatability are shown in Section 4.1.



(a) *Underground* 1365×1024 (b) *Underground* 2048×1365 (c) *Underground* 3456×2304

Fig. 2: Feature detection of the *Wave* detector on different resolutions.

3.2 Feature Descriptors

Since the A-KAZE detector provides the highest accuracy (cf. Section 4.1) and appropriate numbers of features for all of the sequences it is used for the detection task. Then, the descriptors are calculated by all methods as shown in Table 2. We use only original implementations from the authors (source code or binaries). For each detector, default parameters are used. Note, that for the descriptors different lengths d_l are provided by default (cf. Table 2). The results for precision and recall of the descriptors are shown in Chapter 4.2.

4 Experimental Results

The results for the detector evaluation is demonstrated in Section 4.1 while the results for descriptors is shown in Section 4.2. The approaches are subsumed in Table 1 and Table 2, respectively.

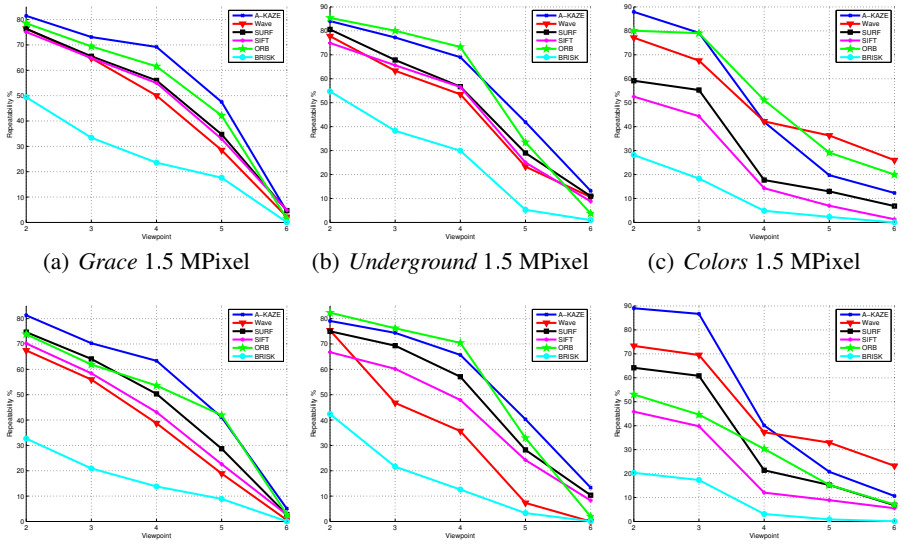
4.1 Detector Evaluation

The results for the repeatability are demonstrated in two sets:

1. A comparison between low-resolution and high-resolution for the same scenes in Figure 3 and Figure 4. (*Grace*, *Underground*, and *Colors*). Here, different performances are shown for some competitors.
2. The results for low-resolution input images (*Graffiti*, *Posters*, and *There*) in Figure 5. For these sequences, the results for higher resolution show no significant differences (*Graffiti* and *There*).

The first set shows that the performance decreases in general when using larger image resolutions. There are some examples, where the performance drops drastically. One example is the result of the *Wave* detector for the *Underground* sequence (cf. Figure 3). Here, the numbers of valid feature pairs for 8 megapixels are even smaller than the numbers for 1.5 megapixels. In Figure 2, the detection result of *Wave* is demonstrated on a part of the first image of *Underground*. On the full image, 7735 points are detected on resolution 1365×1024, 6821 on 2048×1365, and only 3282 on 3456×2304. On the contrary, *Wave* shows good performance on the *Colors* sequence. The *Colors* sequence provides a second example for a differing performance of a detector. The repeatability of *ORB* is significantly lower for 8 megapixels than for 1.5 megapixels. The *BRISK* detector gives poor results for the large resolution compared to the low resolution versions. The best results are provided by the *A-KAZE* detector.

The second set demonstrates results using smaller resolutions (cf. Figure 5). For *Graffiti* and *Posters*, *ORB* performs best, followed by *A-KAZE*. The challenging *There* sequence (strong viewpoint change) shows very low detection performance of *Wave*. It detects only 23 features in the first image of the sequence. Again, *A-KAZE* provides very good results for each of the sequences.
The overall results are subsumed in Table 3. The best results are achieved with the *A-KAZE* detector. *ORB* provides surprisingly good results for most sequences.



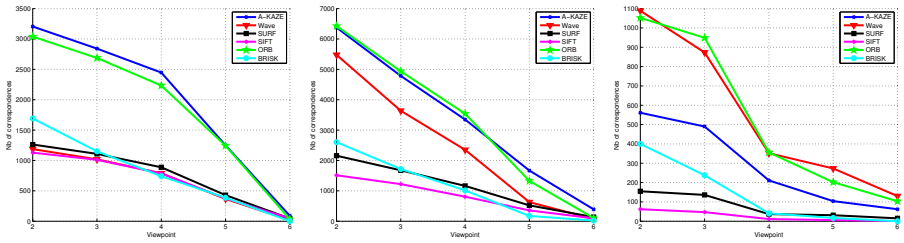
(a) *Grace* 1.5 MPixel (b) *Underground* 1.5 MPixel (c) *Colors* 1.5 MPixel

(d) *Grace* 8 MPixel (e) *Underground* 8 MPixel (f) *Colors* 8 MPixel

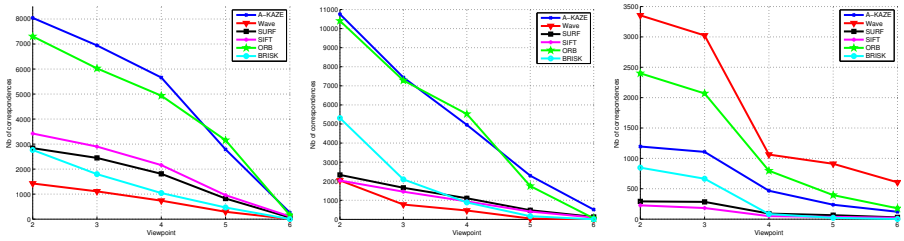
Fig. 3: Repeatability results for 1.5 megapixels (top row, 1365×1024) and 8 megapixels (bottom row, 3456×2304) for the sequences *Grace*, *Underground*, and *Colors*.

4.2 Descriptor Evaluation

Since *A-KAZE* provides the best results for feature detection, this approach is used. For the descriptor evaluation, the sequence test set is extended with the sequences *Wall*, *Boat*, and *Bikes* [14]. The results are shown in Figure 6 (*Graffiti* and *Wall*) and in Figure 7 (*Boat* and *Bikes*) for the lower resolution images (0.5 megapixels). The comparisons with different resolution (1.5 megapixels and 8 megapixels) of the same scene are demonstrated Figure 8 (*Grace*), in Figure 9 (*Underground*), in Figure 10 (*Colors*), and in Figure 11 (*There*). For the *Posters* sequence, too many features are extracted for the large resolution version (> 45000) to evaluate with the matlab script. We show the results of the 1.5 megapixels sequence in Figure 12. The overall results are subsumed in Table 4.
Like in the detectors evaluation, there are several examples with varying performances of descriptors on different resolutions but the same scene. For the

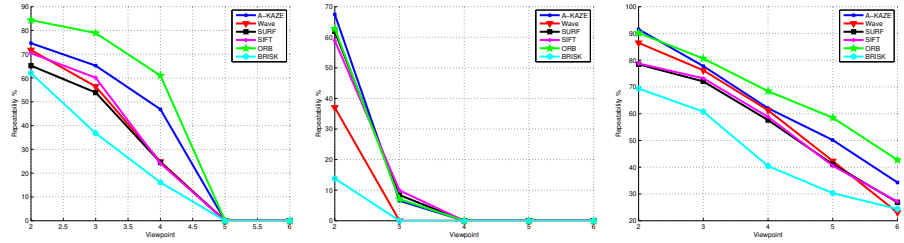


(a) *Grace* 1.5 MPixel (b) *Underground* 1.5 MPixel (c) *Colors* 1.5 MPixel

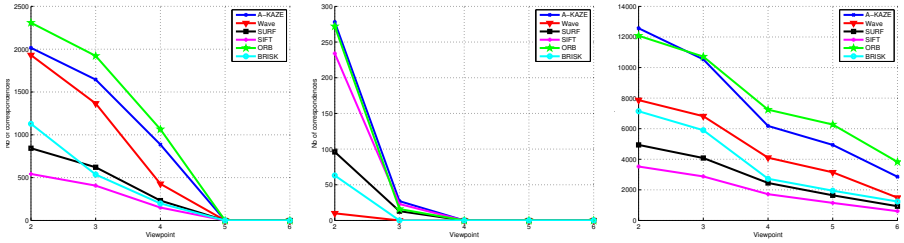


(d) *Grace* 8 MPixel (e) *Underground* 8 MPixel (f) *Colors* 8 MPixel

Fig. 4: Absolute numbers or valid feature pairs for 1.5 megapixels (top, 1365×1024) and 8 megapixels (bottom, 3456×2304) for *Grace*, *Underground*, and *Colors*.



(a) *Graffiti*: Repeatability (b) *There*: Repeatability (c) *Posters*: Repeatability



(d) *Graffiti*: Correspondences (e) *There*: Correspondences (f) *Posters*: Correspondences

Fig. 5: Repeatability (top) and absolute numbers of features for the sequences *Graffiti* (0.5 megapixels), *There*, and *Posters* (1.5 megapixels).

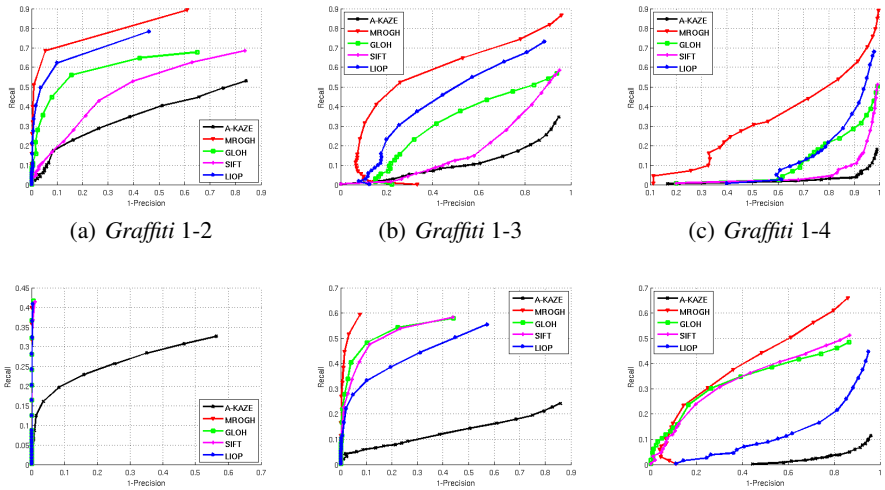
Table 3: The results for the detectors test field.

Input		Ranking					
Sequence	Resolution	1ST	2ND	3RD	4TH	5TH	6TH
Grace	1.5 MP	A-KAZE	ORB	SURF/SIFT		Wave	BRISK
Grace	8.0 MP	A-KAZE	ORB	SURF/SIFT		Wave	BRISK
Underground	1.5 MP	A-KAZE/ORB		Wave/SURF/SIFT			BRISK
Underground	8.0 MP	A-KAZE/ORB		SURF	SIFT	Wave	BRISK
Colors	1.5 MP	A-KAZE/ORB/Wave			SURF	SIFT	BRISK
Colors	8.0 MP	A-KAZE/Wave		SURF	ORB	SIFT	BRISK
Graffiti	0.5 MP	ORB	A-KAZE	Wave/SURF/SIFT			BRISK
There	1.5 MP	ORB	A-KAZE/SURF		SIFT	BRISK	Wave
Posters	1.5 MP	A-KAZE/ORB		Wave/SURF/SIFT			BRISK

Grace sequence, the *A-KAZE* descriptor provides better results than SIFT for 1.5 megapixels while being worse than SIFT for 8 megapixels (cf. Figure 8). A second example is the LIOP descriptor on the *Underground* sequence(cf. Figure 9). For 1.5 megapixels, it performs very good (ranking 2ND in the test field) while the performance drops significantly for 8 megapixels. The LIOP descriptor provides the most varying result for different sequences. For *Wall*, it ranks 4TH (cf. Figure 6) while providing the best results for *Posters* (cf. Figure 12).
As shown in Table 4, the overall best descriptor results are provided by MROGH, followed by LIOP. The MROGH descriptor is theoretically rotational invariant [7]. Thus, the estimation of a dominant orientation is not required. The *Boat* (cf. Figure 7) sequence illustrates this strength. Interestingly, MROGH provides the best results for nearly every sequence tested in this evaluation. Although the A-KAZE detector provides the most accurate features (cf. Section 4.1), its descriptor is only ranked 6TH in this test field. As expected, GLOH is slightly better than SIFT, ranking 3RD and 4TH.

5 Conclusions

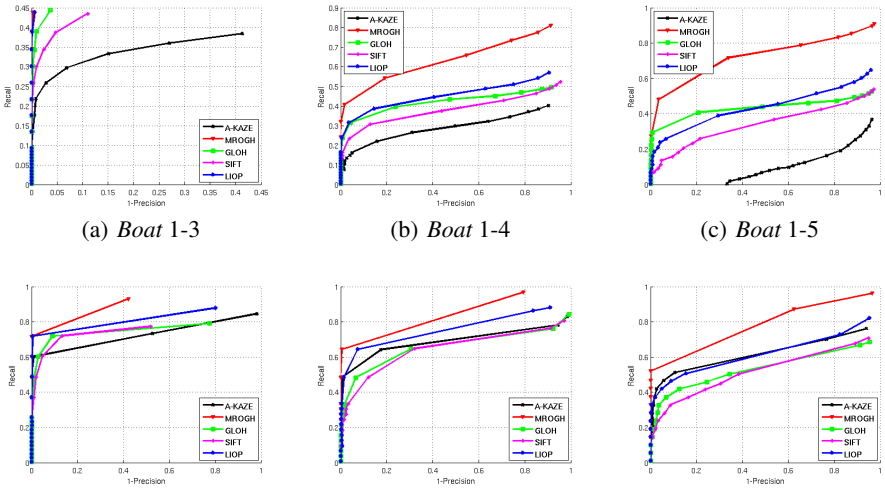
A recent benchmark data set [6] allows for the evaluation of scale invariant feature detectors and descriptors on high resolution images. The benchmark enables the comparison of the approaches on different resolutions of the same scene. The evaluation presented in this paper concentrates on the accuracy of detectors and descriptors at different image resolutions. It is shown that different resolutions can lead to significantly different results for detectors (Wave, ORB) and descriptors (LIOP, A-KAZE). The most accurate detector is the A-KAZE detector. The A-KAZE regions are used for the evaluation of state of the art descriptors. Here, the MROGH descriptor leads to the best results.
The evaluation shows that the benchmark [6] offers new and interesting results regarding accuracy and high image resolutions. Furthermore, it offers the unique possibility to examine the behavior of the detectors and descriptors on different resolutions of the same scene.



(a) *Graffiti* 1-2 (b) *Graffiti* 1-3 (c) *Graffiti* 1-4

(d) *Wall* 1-3 (e) *Wall* 1-4 (f) *Wall* 1-5

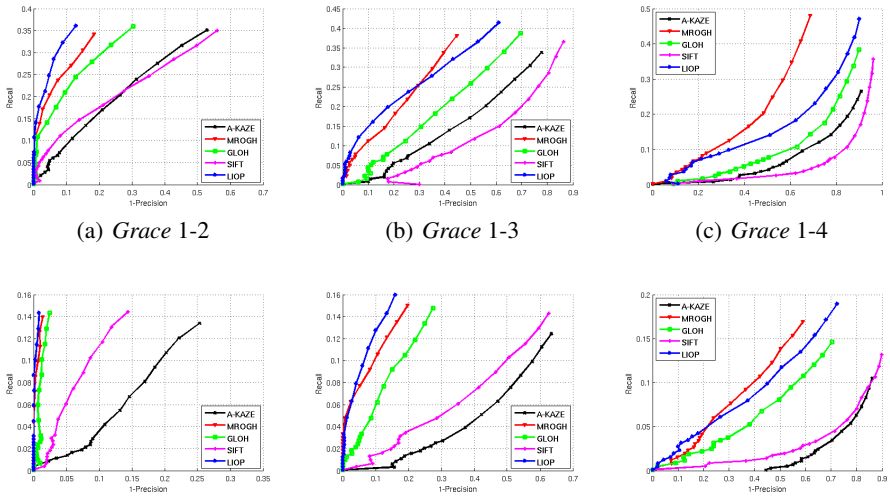
Fig. 6: Precision-recall diagrams for *Graffiti* (top row) for the image pairs 1-2, 1-3, and 1-4 and *Wall* (bottom row) for the image pairs 1-3, 1-4, 1-5.



(a) *Boat* 1-3 (b) *Boat* 1-4 (c) *Boat* 1-5

(d) *Bikes* 1-2 (e) *Bikes* 1-4 (f) *Bikes* 1-6

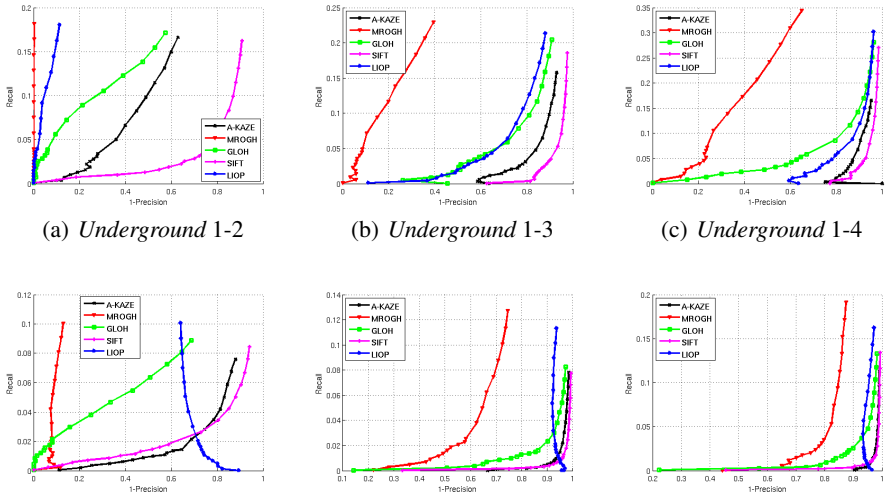
Fig. 7: Precision-recall diagrams for *Boat* (top row, image pairs 1-3, 1-4, 1-5) and *Bikes* (bottom row, image pairs 1-2, 1-4, and 1-6). The *Boat* sequence shows scale and rotation change. The *Bikes* shows differences in image blur.



(a) *Grace* 1-2 (b) *Grace* 1-3 (c) *Grace* 1-4

(d) *Grace* 1-2 (e) *Grace* 1-3 (f) *Grace* 1-4

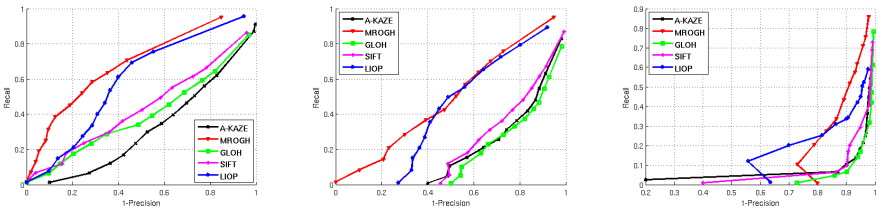
Fig. 8: Precision-recall diagrams for *Grace* with the resolutions 1536×1024 (top) and 3456×2304 (bottom).



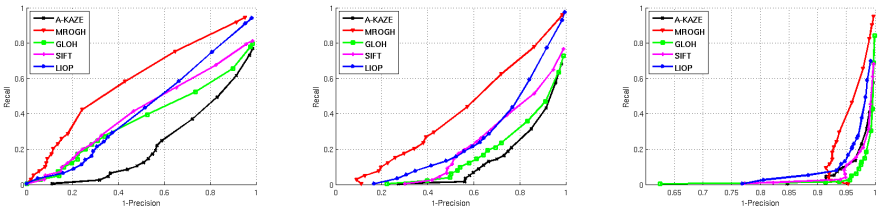
(a) *Underground* 1-2 (b) *Underground* 1-3 (c) *Underground* 1-4

(d) *Underground* 1-2 (e) *Underground* 1-3 (f) *Underground* 1-4

Fig. 9: Precision-recall diagrams for *Grace* with the resolutions 1536×1024 (top) and 3456×2304 (bottom).

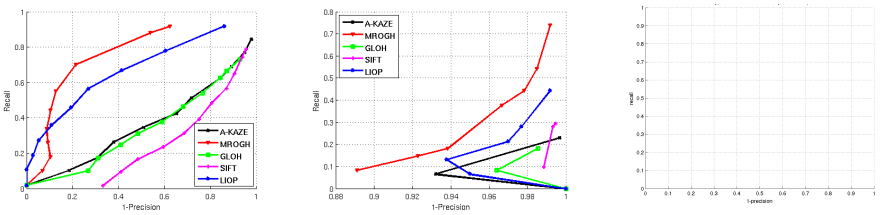


(a) *Colors* 1-2 (b) *Colors* 1-3 (c) *Colors* 1-4

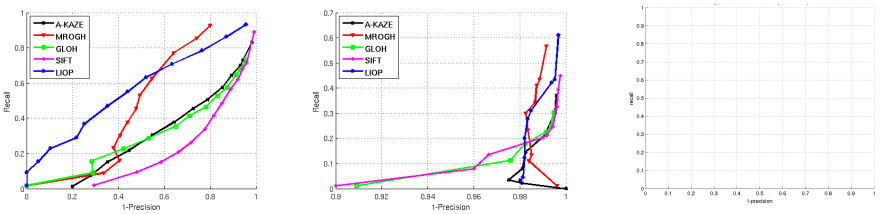


(d) *Colors* 1-2 (e) *Colors* 1-3 (f) *Colors* 1-4

Fig. 10: Precision-recall diagrams for *Colors* with the resolutions 1536×1024 (top) and 3456×2304 (bottom).

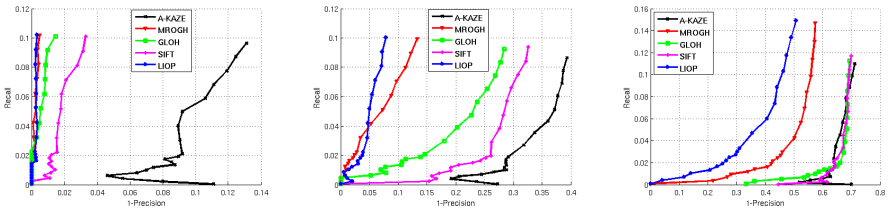


(a) *There* 1-2 (b) *There* 1-3 (c) *There* 1-4



(d) *There* 1-2 (e) *There* 1-3 (f) *There* 1-4

Fig. 11: Precision-recall diagrams for *There* with the resolutions 1536×1024 (top) and 3456×2304 (bottom). For this challenging sequence, none of the descriptors provide a result for the pair 1-4.



(a) *Posters* 1-2 (b) *Posters* 1-3 (c) *Posters* 1-4

Fig. 12: Precision-recall diagrams for *Posters* with the resolutions 1536×1024.

Table 4: The results for the descriptors test field.

Input		Ranking				
Sequence	Resolution	1ST	2ND	3RD	4TH	5TH
Graffiti	0.5 MP	MROGH	LIOP	GLOH	SIFT	A-KAZE
Wall	0.5 MP	MROGH	GLOH/SIFT		LIOP	A-KAZE
Boat	0.5 MP	MROGH	LIOP/GLOH		SIFT	A-KAZE
Bikes	0.5 MP	MROGH	LIOP	SIFT/GLOH/A-KAZE		
Grace	1.5 MP	MROGH/LIOP		GLOH	A-KAZE	SIFT
Grace	8.0 MP	MROGH/LIOP		GLOH	SIFT	A-KAZE
Underground	1.5 MP	MROGH	LIOP/GLOH		A-KAZE	SIFT
Underground	8.0 MP	MROGH	GLOH	LIOP/A-KAZE/SIFT		
Colors	1.5 MP	MROGH/LIOP		SIFT	A-KAZE/GLOH	
Colors	8.0 MP	MROGH	LIOP/SIFT		GLOH	A-KAZE
There	1.5 MP	MROGH/LIOP		A-KAZE/GLOH		SIFT
There	8.0 MP	MROGH/LIOP		A-KAZE/GLOH		SIFT
Posters	1.5 MP	LIOP	MROGH	GLOH	SIFT	A-KAZE

References

1. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: Fast retina keypoint. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 510–517 (2012)
2. Alcantarilla, P., Nuevo, J., Bartoli, A.: Fast explicit diffusion for accelerated features in nonlinear scale spaces. In: British Machine Vision Conference (BMVC) (2013)
3. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 3951, pp. 404–417. Springer (2006)
4. BloodAxe: OpenCV Features Comparison. `https://github.com/BloodAxe/OpenCV-Features-Comparison` (2014), [Online; accessed 20-February-2015]
5. Cordes, K., Rosenhahn, B., Ostermann, J.: Increasing the accuracy of feature evaluation benchmarks using differential evolution. In: IEEE Sympo-

sium Series on Computational Intelligence (SSCI) - IEEE Symposium on Differential Evolution (SDE) (2011)
6. Cordes, K., Rosenhahn, B., Ostermann, J.: High-resolution feature evaluation benchmark. In: Wilson, R. (ed.) 15th International Conference on Computer Analysis of Images and Patterns (CAIP). Lecture Notes in Computer Science, vol. 8047, pp. 327–334. Springer (2013)
7. Fan, B., Wu, F., Hu, Z.: Aggregating gradient distributions into intensity orders: A novel local image descriptor. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2377–2384 (2011)
8. Figat, J., Kornuta, T., Kasprzak, W.: Performance evaluation of binary descriptors of local features. In: Chmielewski, L., Kozera, R., Shin, B.S., Wojciechowski, K. (eds.) Computer Vision and Graphics, Lecture Notes in Computer Science, vol. 8671, pp. 187–194. Springer International Publishing (2014)
9. Frahm, J.M., Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building rome on a cloudless day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) European Conference on Computer Vision. Lecture Notes in Computer Science (LNCS), vol. 6314, pp. 368–381. Springer (2010)
10. Hess, R.: An open-source siftlibrary. In: Proceedings of the International Conference on Multimedia. pp. 1493–1496. MM '10, ACM, New York, NY, USA (2010)
11. Leutenegger, S., Chli, M., Siegwart, R.: BRISK: Binary robust invariant scalable keypoints. In: IEEE International Conference on Computer Vision (ICCV). pp. 2548–2555 (2011)
12. Li, Y., Wang, S., Tian, Q., Ding, X.: A survey of recent advances in visual feature detection. Neurocomputing 149, Part B, 736 – 751 (2015)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) 60(2), 91–110 (2004)
14. 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 (IJCV) 65(1-2), 43–72 (2005)
15. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27(10), 1615–1630 (2005)
16. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: IEEE International Conference on Computer Vision (ICCV). pp. 2564–2571 (2011)
17. Salti, S., Lanza, A., Di Stefano, L.: Keypoints from symmetries by wave propagation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2898–2905 (2013)
18. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. International Journal of Computer Vision (IJCV) 80, 189–210 (2008)
19. Wang, Z., Fan, B., Wu, F.: Local intensity order pattern for feature description. In: IEEE International Conference on Computer Vision (ICCV). pp. 603–610 (2011)