Interest point detectors stability evaluation on ApolloScape dataset

Jacek Komorowski[1][0000−0001−6906−4318], Konrad Czarnota[1], Tomasz Trzcinski[1,2][0000−0002−1486−8906], Lukasz Dabala[1], Simon Lynen[3]

[1] Warsaw University of Technology, Warsaw, Poland
[2] Tooploox
[3] Google

Abstract. In the recent years, a number of novel, deep-learning based, interest point detectors, such as LIFT, DELF, Superpoint or LF-Net was proposed. However there's a lack of a standard benchmark to evaluate suitability of these novel keypoint detectors for real-live applications such as autonomous driving. Traditional benchmarks (e.g. Oxford VGG) are rather limited, as they consist of relatively few images of mostly planar scenes taken in favourable conditions. In this paper we verify if the recent, deep-learning based interest point detectors have the advantage over the traditional, hand-crafted keypoint detectors. To this end, we evaluate stability of a number of hand crafted and recent, learning-based interest point detectors on the street-level view ApolloScape dataset.

Keywords: keypoint detectors · interest points stability

1 Introduction

Detection of local interest points in images is the fundamental part of many computer vision applications, including 3D reconstruction [1], panorama stitching [2] and monocular Simultaneous Localization and Mapping [3]. This topic has gained significant attention from the research community [4–13]. While traditional interest point detectors rely on hand-crafted features [7–9], more recent methods use machine learning techniques such as decision trees [10] or deep learning [4–6, 12, 13] to train a high-performing keypoint detector.

In the recent years, a number of novel, deep-learning based interest point detectors, such as LIFT [4], DELF [5], Superpoint [6] or LF-Net [13], was proposed. However there's a lack of a standard benchmark, to evaluate suitability of these novel keypoint detectors for real-live applications such as autonomous driving. Oxford VGG [14], a very popular benchmark for evaluation of local features, is rather limited, as it consists of only 8 sequences, each containing 6 images. Images are taken in a favourable environmental conditions and contain mostly planar scenes. The dataset does not capture variety of factors that impact the image of the visible scene, such as diverse weather conditions, different seasons or time of the day.

In the seminal work [15] on local feature detectors, the following properties of an ideal local interest points are listed: repeatability, distinctiveness, locality, quantity, accuracy and efficiency. Repeatability is considered as one of the most important properties of good features. High repeatability ensures that given two images of the same scene, taken under different viewing conditions, a high percentage of features detected on the scene part visible in both views can be found in both images. In this paper we evaluate repeatability of a wide range of hand-crafted and learning-based interest point detectors on the ApolloScape [16] street-level view dataset. Due to their size and diversity this datasets allows evaluation of interest point detectors in the real-life, variable and challenging conditions.

2   Related work

In this section we briefly review recently proposed, learning based interest point detectors.

TaSK (Task Specific Keypoint) [17] is one of the early attempts to use machine learning methods in order to improve interest point detector performance. It aims at improving repeatability of keypoints by learning a classifier to filter out detected keypoints to retain more stable features. However, the method is reliant on some other interest point detector, such as DoG [9], and only filters out points found by the base detector.

TILDE (Temporarily Invariant Learned DEtector) [18] is the learning-based method to detect repeatable keypoints under drastic changes of environmental conditions such as weather and lighting. It relies on other keypoint detector, DoG [9], only to generate training examples. Training set consists of multiple stacks of images showing the same scene and taken from the same viewpoint but at different season and time of the day. DoG features detected on training images are considered positive examples, if they are repeatable on majority images in the stack (set of images showing the same scene). The regressor is trained to compute a value (score map) for each patch of a given size of the input image. The loss function is constructed so the learned regressor produces a peaked shape on positive samples (patches centered near good keypoint location) and small value on negative samples.

LIFT (Learned Invariant Feature Transform) [4] is the first end-to-end, deep learning-based, method including feature detection, orientation estimation and robust descriptor extraction. The learning is based on SfM-verified DoG [9] keypoints detected in a large collection of images. The feature extraction pipeline is trained by presenting quadruples of image patches (base patch, positive example, negative example and a patch with no distinctive feature points) to the network having a Siamese-like architecture with four branches.

Convolutional feature detector proposed in [19] identifies regions of the input image that constitute good keypoints. The loss function consists of two terms: binary classification loss to classify patches as centered on the keypoint location and squared difference loss to penalize network responses on non-centered

patches. Training set is constructed by detecting interest points using a hand-crafted feature detector in a sequence of images. Tuples of points that survive SfM verification form positive training examples.

Quad-networks [12] is the first unsupervised interest point detection method, not relying on a hand crafted feature detector for the training set generation. Neural network is trained to map an image patch point to a single real-valued response and then to rank points according to the response. Image patch is mapped by the neural network to a 'heat map' with ranking of each pixel. The ranking is optimized to be repeatable under the desired transformation classes, such as rotation, translation or intensity change. If one point has the higher ranking than another one, it should still be higher after a transformation. Top and bottom quantiles of the response are repeatable and can be used as an interest points. Two training approaches are described. The first one uses an image dataset with ground truth data on camera poses and 3D scene structure from the LIDAR scan. Using the ground truth data, correspondence between image patches in two different images can be established. The other approach is fully unsupervised. The training set is constructed, by applying transformations, such as rotation and translation, to a set of base image patches.

Superpoint [6] is a self-supervised framework for training interest point detectors and descriptors. It operates on the full-sized images and jointly computes interest points locations and associated descriptors. VGG-style [20] convolutional network is used to reduce the dimensionality of the input image. Then the network branches into interest point decoder and descriptor decoder units. The interest point detector is pre-trained on the synthetic data consisting of a large set of computer generated images with pseudo-ground truth interest point locations. To boost the performance on the real data, the network is fine tuned using real images. Images are warped multiple times to help an interest point detector to see the scene from many different viewpoints and scales.

LF-Net (Local Feature Network) [13] is a novel deep architecture for sparse matching, which can be trained end-to-end and does not require using a hand-crafted feature detector to initialize the training process. The training requires image pairs with known relative pose and depth maps, so the correspondence between image patches in two images can be established. LF-Net runs the detector on the first image and finds the maxima on the produced score map. Then it optimizes the weights so that when the detector is run on the second image it produces a *clean response map* with sharp maxima at right locations.

3   Interest point detectors evaluation

3.1   Dataset

We performed evaluation of interest point detectors on the recently released ApolloScape [16] street-level view dataset. The dataset contains almost 150 thousand frames with high quality ground truth data including pixel-level semantic segmentation, pose information and depth maps for the static background. Image frames in the dataset are collected every one meter by the acquisition system

Fig. 1. Exemplary image (top) and corresponding depth map for the static background (bottom) from the ApolloScape [16] dataset.

with resolution 3384 x 2710. The dataset contains images taken in diverse locations under the varying weather conditions and with challenging lightning conditions (e.g. dark or very bright).

The data is acquired using Riegl VMX-1HA acquisition system consisting of two VUX-1HA laser scanners (360°FOV, range from 1.2m up to 420m), VMX-CS6 camera system (two front cameras with resolution 3384 x 2710), and the measuring head with IMU/GNSS (position accuracy $20 \sim 50$mm, roll and pitch accuracy $0.005°$, and heading accuracy $0.015°$).

3.2   Evaluation criteria

We evaluate stability of the interest point detectors using *repeatability* criteria introduced in [21]. *Repeatability* score for an interest point detector operating on a pair of images is defined as a ratio between the number of point-to-point correspondences that can be established between detected points and the number of points detected in both images.

Computing repeatability, in general case, is not a trivial task. First, correspondence between interest points on two images must be established. In standard benchmarks, such as Oxford VGG Affine Covariant Regions [14] or HPatches [22], this relation is defined by a homography, as they contain images of planar scenes or camera motion is purely rotational. In general case, where the visible scene is not planar and the camera movement is not restricted, establishing such correspondence is not easy. Second, an interest point detector in general will not detected an interest point at the exact position of the corresponding point, but at some close neighbourhood. Third, some interest points cannot be repeated as some parts of the scene can be observed by only one camera.

When interest point correspondence is related by the homography H, repeatability definition introduced in [21] is commonly used. Let P be a 3D point and $p_1 = M_1 P$, $p_2 = M_2 P$ its projections of P onto the image I_1 and I_2 respectively. M_1 and M_2 are projections matrices related with images I_1 and I_2.

Two points $p1 \in I_1$ and $p2 \in I_2$ correspond, if they are related by homography H, that is if $dist(Hp_1, p_2) < \theta$, where θ is some fixed threshold. d_1 and d_2 denote points that could be potentially detected on both images, defined as as $d_1 = \{p_i \in I_1 | H_{21}p_i \in I_2\}$ and $d_2 = \{p_2 \in I_2 | H_{12}p_2 \in I_1\}$. The *repeatability* rate r is defined as:

$$r_{1,2} = \frac{|C(I_1, I_2)|}{\min(|d_1|, |d_2|)} \quad , \tag{1}$$

where $C(I_1, I_2) = \{(p_1 \in I_1, p_2 \in I_2) \, | dist(Hp_1, p_2) < \theta\}$ is the set of corresponding pairs of interest points.

In general case, when interest points are not related by homography, the above approach cannot be taken. If the dataset ground truth contains image poses and high quality depth maps, as is the case in ApolloScape [16] dataset, we can use it to establish the correspondence between interest points. The 3D point P corresponding to the interest point $p_1 \in I_1$ has the coordinates $P = (d\widetilde{x}, d\widetilde{y}, d)$, where d is the depth of the point p_1 and $(\widetilde{x}, \widetilde{y})$ are normalized coordinates of the point p_1 in the image I_1. Then, using the ground truth relative pose between images I_1 and I_2, we project P onto the second image I_2 to get its projection p_1^*. A pair of interest points $p_1 \in I_1$ and $p_2 \in I_2$ corresponds, if:

$$dist(p_1^*, p_2) < \theta. \tag{2}$$

3.3   Evaluation results

We evaluate stability of traditional interest point detectors: DoG [9], AKAZE [23], AGAST [24], Fast [10], ORB [11] and recently proposed detectors LIFT [4], Saddler [25], Superpoint [6], TILDE [18] and LF-Net [13].

For traditional detectors (DoG, AKAZE, AGAST, Fast and ORB) we use OpenCV [4] implementations. Evaluation of recent learning-based methods (LIFT, Superpoinit, TILDE, LF-Net) is performed using the code and pre-trained models released by authors.

We choose the following traversals from ApolloScape [16] dataset: 16, 32 from road id 1; 1, 5, 16 from road id 2; 1, 21 and 29 from road id 3. For each traversal we take every 20th frame (corresponding to 20 meter drive) as the base frame. We match keypoints in each base frame with keypoints in 19 subsequent frames. These 19 subsequent frames are at the distance of approximately $1, 2, \ldots, 19$ meters from the base frame. On each image we select $N = 10,000$ interest points with the strongest response. The threshold θ for interest point correspondence in Eq. 2 is set to 2.5 pixels.

Fig. 2, 3 and 4 show mean repeatability score as a function of a distance between camera centres in pairs of images for each evaluated traversal. The results are summarized in Tab. 1.

Recently proposed Saddler [25] detector has the best average repeatability for all evaluated traversals: 0.177. It has the highest repeatability in 6 out of 9 traversals and a relatively small gap to the best performing detectors in the

---

[4] https://opencv.org/

Table 1. Mean repeatability of interest point detectors evaluated on 9 traversals from ApolloScape dataset. Recently proposed Saddler detector has the highest average repeatability (0.177) for all evaluated sequences. Traditional FAST detector is a runner up with 0.164 average repeatability. The best deep-learning based detector Superpoint scores only 0.123.

Traversal id / Keypoint	1-16-6	1-32-5	1-36-6	2-01-5	2-05-5	2-16-5	3-01-5	3-21-5	3-29-5	**Avg.**
					Repeatability					
Saddler [25]	0.103	0.089	0.153	**0.186**	**0.090**	**0.231**	**0.247**	**0.255**	**0.238**	**0.177**
FAST [10]	**0.113**	**0.093**	**0.200**	0.176	0.077	0.217	0.206	0.185	0.208	0.164
AGAST [24]	0.091	0.086	0.176	0.146	0.070	0.190	0.179	0.163	0.182	0.143
ORB [11]	0.078	0.056	0.138	0.140	0.038	0.192	0.236	0.196	0.207	0.142
AKAZE [23]	0.081	0.070	0.169	0.142	0.072	0.176	0.184	0.160	0.177	0.137
Superpoint [6]	0.075	0.070	0.119	0.136	0.080	0.152	0.168	0.156	0.151	0.123
DoG [9]	0.081	0.066	0.122	0.137	0.062	0.156	0.161	0.143	0.168	0.122
TILDE [18]	0.068	0.068	0.088	0.092	0.074	0.107	0.108	0.103	0.116	0.091
LF-Net [13]	0.060	0.065	0.095	0.082	0.034	0.104	0.114	0.101	0.103	0.084
LIFT [4]	0.050	0.052	0.054	0.058	0.049	0.056	0.068	0.061	0.072	0.058

remaining 3 traversals. Traditional FAST [10] (avg. repeatability 0.164) performs slightly worse.

Non of the recently proposed deep learning-based keypoint detectors shows an advantage other the traditional hand-crafted detectors. The best performing deep-learning based detector, Superpoint [6], has an average repeatability equal to 0.123. Other perform even worse: TILDE [18] scores 0.084, LF-Net [13] 0.04 and LIFT [4] 0.058 average repeatability.

Using semantic labels ground truth provided with the ApolloScape dataset, we analyzed dependency of the interest point repeatability on the semantic content of the visible scene. Tab. 2 shows semantic classes with top-5 and bottom-5 repeatability for the best performing detector (Saddler [25]), best traditional detector (FAST [10]) and best deep-learning based detector (Superpoint [6]). As expected, repeatability of keypoints detected on stable objects such as billboards, traffic lights and buildings is consistently higher than on movable or volatile objects such as car, motorcycles, or sky. Dependency of the interest point detector performance on the semantic class of the scene region is visualized on Fig. 5. These results proves, that in order to build high-performing interest point detector, high-level information on the semantic content of the visible scene must be taken into the account. Keypoints should not be detected on the movable or volatile regions of the scene, such as vehicles or sky.

4   Conclusions and future work

The best performing interest point detector on ApolloScape dataset is recently proposed Saddler [25]). Average Saddler repeatability across all evaluated se-
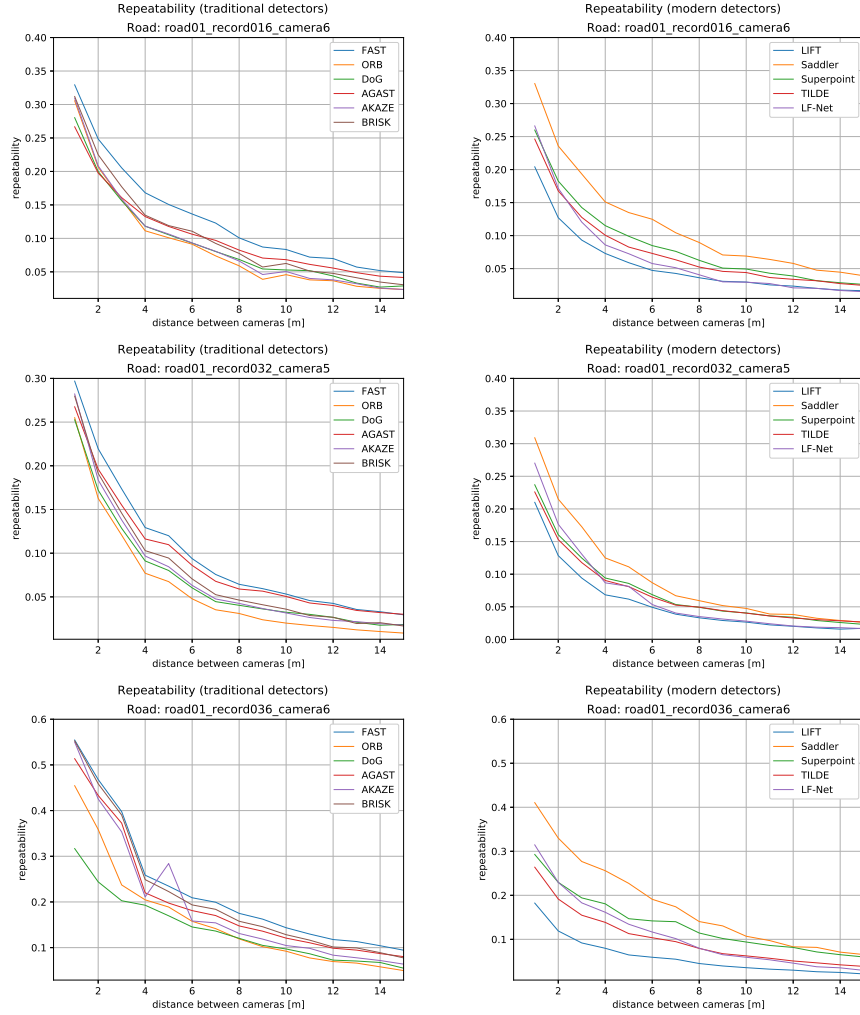
Fig. 2. Repeatability of keypoint detectors evaluated on record 016 (top), record032 (middle) and record 036 (bottom) from road 01 as the function of distance between cameras. (Left column) Traditional keypoint detectors, (right column) modern keypoint detectors.
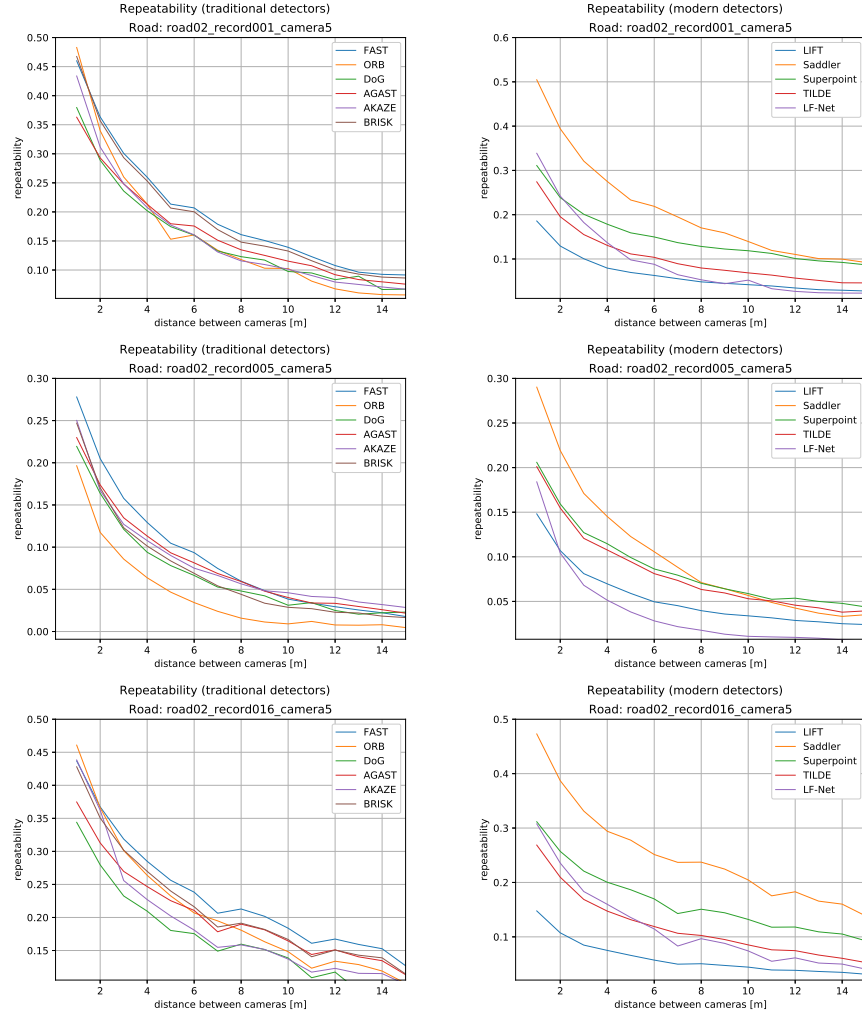
Fig. 3. Repeatability of keypoint detectors evaluated on record 001 (top), record005 (middle) and record 016 (bottom) from road 02 as the function of distance between cameras. (Left column) Traditional keypoint detectors, (right column) modern keypoint detectors.
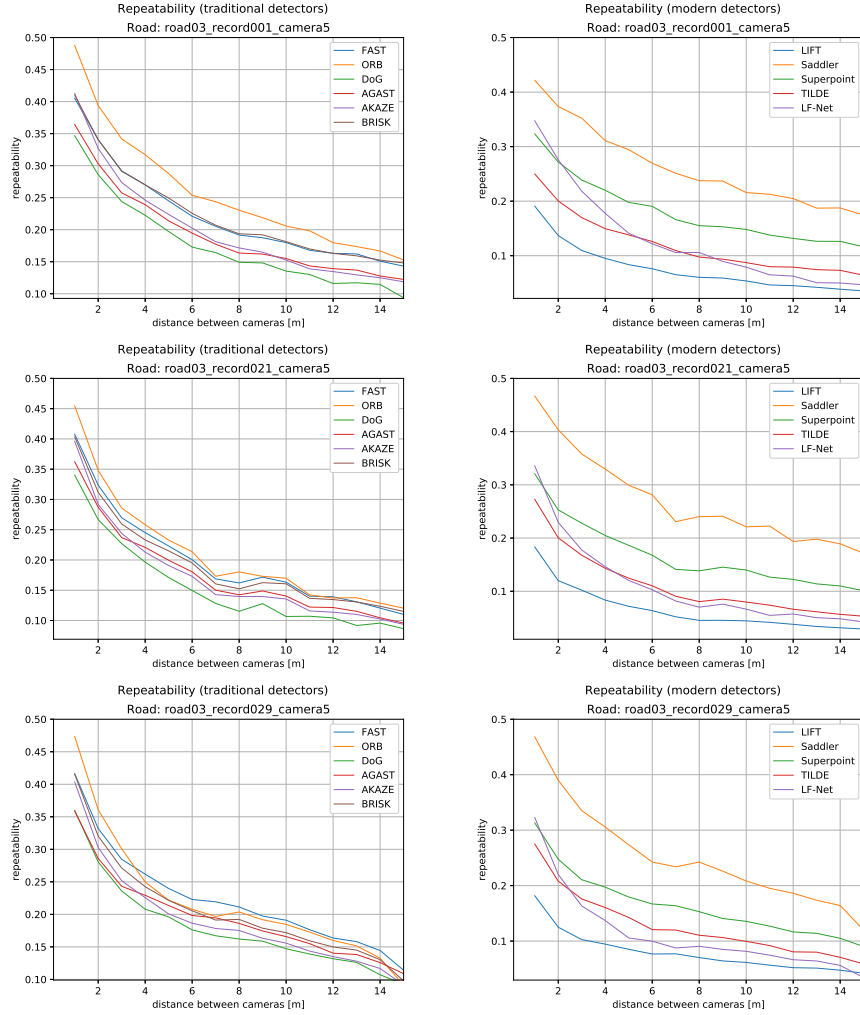
Fig. 4. Repeatability of keypoint detectors evaluated on record 001 (top), record021 (middle) and record 029 (bottom) from road 3 as the function of distance between cameras. (Left column) Traditional keypoint detectors. (right column) modern keypoint detectors.

Table 2. Semantic classes with top-5 and bottom-5 repeatability for FAST, Saddler and Superpoint keypoint detectors. Repeatability of keypoints detected on the stable structures such as billboards, traffic light or traffic signs) is significantly higher than repeatability of the keypoints detected on movable or volatile objects, such as cars, motorcycles or sky.

Detector	Top 5		Bottom 5	
	Semantic class	Avg. rep.	Semantic class	Avg. rep.
FAST [10]	billboard	0.552	rover	0.021
	traffic light	0.523	sky	0.132
	traffic sign	0.502	unlabelled	0.138
	road pile	0.491	car	0.151
	dustbin	0.418	motorcycle	0.160
Saddler [25]	traffic light	0.640	rover	0.012
	billboard	0.538	sky	0.075
	pole	0.526	motorcycle	0.115
	traffic sign	0.487	person	0.130
	building	0.482	car	0.133
Superpoint [6]	billboard	0.538	rover	0.030
	traffic sign	0.486	sky	0.075
	road pile	0.395	unlabelled	0.086
	overpass	0.381	car	0.097
	building	0.373	motorcycle	0.107

quences is 0.177. It has a considerable advantage over the second best detector, FAST (0.164 mean repeatability).

Interestingly, all recently proposed deep learning-based interest point detectors perform noticeable worse compared to best hand-crafted detector (Saddler). The best deep learning-based detector, Superpoint [6], scores average repeatability equal to 0.123. Other evaluated deep learning-based descriptors (TILDE [18], LF-Net [13], LIFT [4]) perform even worse.

Analysis of the stability of interest points detected in objects with different semantic class labels proves that keypoints detected on the stable objects, such as road signs or buildings, have the highest repeatability. This suggest, that in order to build high-performing interest point detector, high-level information on the semantic content of the visible scene must be taken into the account. Keypoints should not be detected on the movable or volatile objects, such as vehicles or sky.

As the future work we plan to verify if the results generalize to other large-scale, stree-level view datasets, such as Oxford RobotCar [26] and Berkeley Deep-Drive [27].

A very promising direction of the research is development of the repeatable and discriminant learning-based interest point detector and descriptor. Recently released large-scale street-view datasets contain millions of images taken in mul-
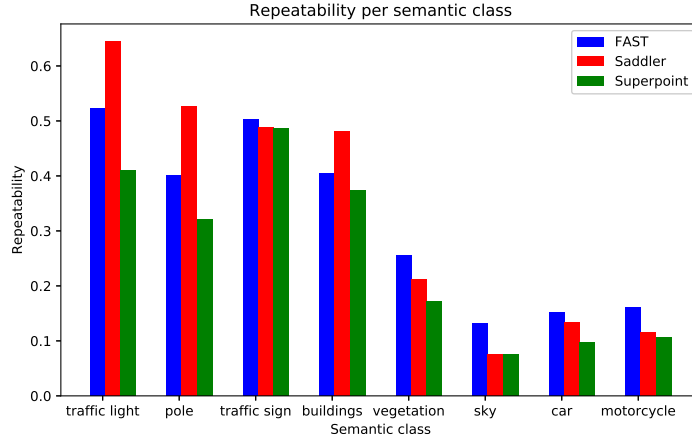
Fig. 5. Repeatability of FAST, Saddler and Superpoint keypoints on objects with different semantic labels. Repeatability for stable objects such as traffic signs or buildings (left side) is consistently the highest for all detectors. Repeatability for movable objects (right side) is the lowest.

tiple locations in diverse weather conditions, and at the different times of the day and year. Ground truth, in the form of high quality dense depth maps or point clouds, allows establishing correspondences between interest point in different images, This should allow training high quality interest point detectors and descriptors, invariant to viewpoint change and variable environmental conditions.

Acknowledgement

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Commun. ACM **54**(10) (October 2011) 105–112
2. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. IJCV **74**(1) (2007) 59–73
3. Lynen, S., Sattler, T., Bosse, M., Hesch, J.A., Pollefeys, M., Siegwart, R.: Get out of my lab: Large-scale, real-time visual-inertial localization. In: Robotics: Science and Systems. (2015)
4. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European Conference on Computer Vision, Springer (2016) 467–483
5. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Largescale image retrieval with attentive deep local features. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3456–3465
6. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. arXiv preprint arXiv:1712.07629 (2017)
7. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey vision conference. Volume 15., Citeseer (1988) 10–5244
8. Lindeberg, T.: Feature detection with automatic scale selection. International journal of computer vision **30**(2) (1998) 79–116
9. Lowe, D.G.: Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Volume 2., Ieee (1999) 1150–1157
10. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European conference on computer vision, Springer (2006) 430–443
11. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: Computer Vision (ICCV), 2011 IEEE international conference on, IEEE (2011) 2564–2571
12. Savinov, N., Seki, A., Ladicky, L., Sattler, T., Pollefeys, M.: Quad-networks: unsupervised learning to rank for interest point detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
13. Ono, Y., Trulls, E., Fua, P., Yi, K.M.: Lf-net: Learning local features from images. arXiv preprint arXiv:1805.09662 (2018)
14. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision **60**(1) (2004) 63–86
15. Tuytelaars, T., Mikolajczyk, K., et al.: Local invariant feature detectors: a survey. Foundations and trends® in computer graphics and vision **3**(3) (2008) 177–280
16. Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y., Yang, R.: The apolloscape dataset for autonomous driving. arXiv preprint arXiv:1803.06184 (2018)
17. Strecha, C., Lindner, A., Ali, K., Fua, P.: Training for task specific keypoint detection. In: Joint Pattern Recognition Symposium, Springer (2009) 151–160
18. Verdie, Y., Yi, K., Fua, P., Lepetit, V.: Tilde: A temporally invariant learned detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5279–5288
19. Altwaijry, H., Veit, A., Belongie, S.J., Tech, C.: Learning to detect and match keypoints with deep architectures. In: BMVC. (2016)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

21. Schmid, C., Mohr, R., Bauckhage, C.: Comparing and evaluating interest points. In: Computer Vision, 1998. Sixth International Conference on, IEEE (1998) 230–235
22. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: Conference on Computer Vision and Pattern Recognition (CVPR). Volume 4. (2017) 6
23. Alcantarilla, P.F., Solutions, T.: Fast explicit diffusion for accelerated features in nonlinear scale spaces. IEEE Trans. Patt. Anal. Mach. Intell **34**(7) (2011) 1281–1298
24. Mair, E., Hager, G.D., Burschka, D., Suppa, M., Hirzinger, G.: Adaptive and generic corner detection based on the accelerated segment test. In: European conference on Computer vision, Springer (2010) 183–196
25. Aldana-Iuit, J., Mishkin, D., Chum, O., Matas, J.: In the saddle: chasing fast and repeatable features. In: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE (2016) 675–680
26. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 Year, 1000km: The Oxford RobotCar Dataset. The International Journal of Robotics Research (IJRR) **36**(1) (2017) 3–15
27. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 (2018)