REQUIREMENT ANALYSIS FOR PERCEPTION ON ASSISTANT ROBOTS IN MULTI-MODAL ENVIRONMENT CONDITIONS

Xiaozhou Luo and Marco Sewtz

Institute of Robotics and Mechatronics, German Aerospace Center (DLR), 82234 Wessling, Germany,
xiaozhou.luo@dlr.de, marco.sewtz@dlr.de

ABSTRACT

This work presents the results of the preliminary analysis for realizing a robust multi-modal perception framework on mobile platforms at the German Aerospace Center (DLR). Within the evaluation, our focus is directed towards investigating robust approaches for localization, place recognition, and navigation in the visual domain. In addition to the analytical examination of promising state-of-the-art methods, an experimental study is carried out based on real-world datasets from mission-related environments. With this, the prevailing environmental properties are evaluated to identify the best-suited visual abstraction and characterization frameworks. In the end, we summarize our findings and realizations in recommendations for improving situation awareness within the considered projects.

Key words: CV, Feature, Perception, VO, SLAM

1. INTRODUCTION

Human spaceflight beyond low earth orbit has gained more and more interest in recent times. Especially in long-running missions and exploration of extraterrestrial bodies, robotic assistance would significantly increase research capabilities and provide a valuable aid to human operators. To further intensify research efforts in humanoid service and assistance robots, the German Aerospace Center (DLR) is currently conducting two research projects, Surface Avatar (SurfA) and SMiLE, with different fields of focus. In order to reduce the workload for human operators, robots have to be equipped with autonomous features. Therefore, the main requirement here is to provide a robust and accurate working localization and perception system. Especially in collaboration with on-site human operators and other participating robots, establishing situation awareness is essential to ensure operational safety.

In this work, we analyze mission-related environments for the development of application-specific multi-modal perception systems. Hereby, our focus is directed towards investigating robust features with particular attention to technical characteristics and internal boundaries of participating robotic systems. In addition, we target to identify the spatial distribution of perceptional properties. Therefore, an experimental evaluation is conducted with state-of-the-art visual feature extraction frameworks in mission-specific environmental settings.

2. BACKGROUND

2.1. Mission Setup

Starting with the space domain, SurfA, in cooperation with the European Space Agency (ESA), is conceived as a technological demonstrator focusing on man-machine collaboration based on different levels of autonomy. Within the mission, an astronaut on board the International Space Station (ISS) controls the actions of ground-based robots in an extraterrestrial environment either by executing task-level commands or taking direct control in teleoperation mode. Therefore, various robotic platforms are involved, including smaller exploration units with limited processing resources, conventional rovers for planetary exploration, and humanoid-like robots for complex manipulation. Apart from the space domain, SMiLE explores the possibility of integrating robotic assistants in health and elderly care. Different levels of autonomy are investigated, including granting direct control capabilities for medical staff to initiate first aid actions in case of emergencies. The platforms used in this project vary from motorized wheelchairs with a robotic arm to humanoid-like robots as deployed in SurfA.

To fulfill the research objectives, robots have to position themselves in the mission environment reliably. As all participating robots are equipped with RGB or RGB-D cameras, passive visual sensors are the means of choice for primary localization and navigation tasks.

2.2. Perception in the Visual Domain

Visual images have long been utilized for several purposes, as it provides a significant amount of information. Especially in recent times, there has been a growing interest in visual-based approaches since they provide a robust and cost-efficient alternative to active systems, including infrared sensors and laser scanners. Starting with the introduction of Visual Odometry (VO) in the 1980s, the

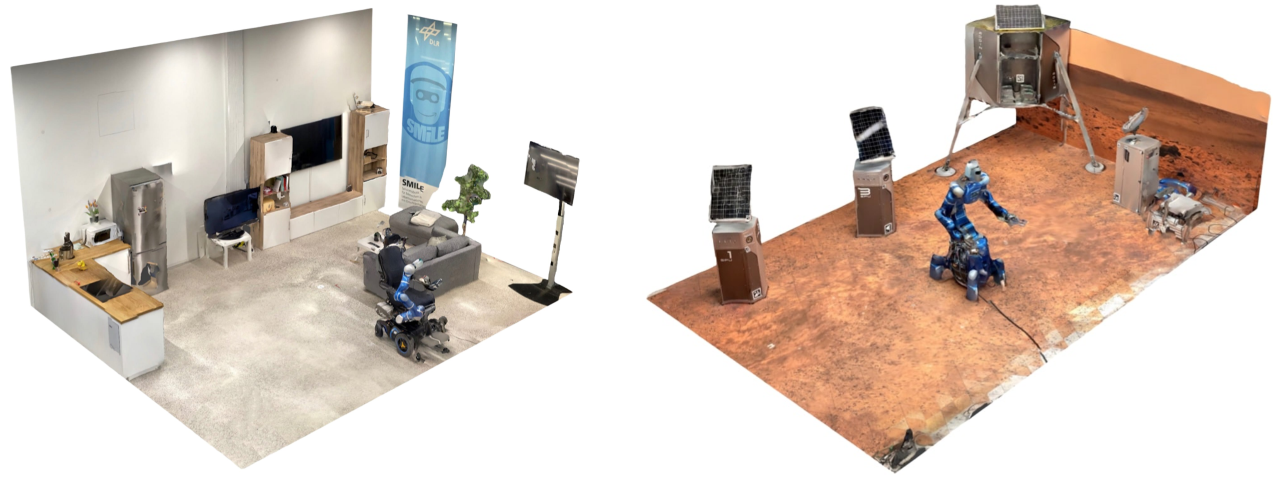


Figure 1: Overview of the mission-related environmental conditions within our contemplated projects. On the left-hand side, SMiLE contains a typical urban housing setting including a kitchen and living room assembly. In contrast, the SurfA scenario consists of three Smart Payload Units (SPUs) and the RODIN lander on the right-hand side.

ego-motion of an agent can be incrementally estimated using only the information from a single attached camera. Taking it a step further, Simultaneous Localization and Mapping (SLAM) is a process in which a robot is required to localize itself in an unknown environment while incrementally constructing a map of its surroundings. Thus, it focuses on establishing a globally consistent estimate of the robot's trajectory inside the generated map by revisiting and recognizing already mapped regions. In terms of visual-only approaches, we selected the feature-based method for estimating relative motion in image sequences. Unlike the direct method based on optical flow, information from the original image is compressed to selected regions of interest for further processing steps. Within this approach, only a fraction of the original data has to be saved to generate an adequate surroundings model for localization and navigation purposes, therefore reducing the required hardware specifications.

In the early days of computer vision and feature detection, there was no consensus how a proper performance evaluation framework should look like. Along with the development of stable feature extraction algorithms with invariant transformations, Mikolajczyk et al. introduced the first comparative performance parameters for standardized and conclusive comparison between image processing algorithms in [1] and [2]. Several mentionable benchmarking studies were carried out in [3] and [4] based on these parameters. While the latter authors concluded their evaluation with behavior verification in practical situations, preceding studies evaluated the methods on datasets comprising fixed, sparse image sequences. Besides, publications focusing on application-related data are rare across all research branches. A mentionable example is the study by Rondao et al. [5], where a non-cooperative rendezvous scenario in space was simulated. In the field of robotics, studies were carried out dominantly with the target of performance benchmarks in the context of VO and SLAM algorithms. However, we did not find any relevant publications using application-specific data since sophisticated datasets and generic recorded image sequences are primarily used for benchmarking studies.

3. ENVIRONMENTAL CONDITION

In the feature-based approach, visual information of an image is analyzed and abstracted into a collection of regions of interest. Starting from point features, Harris and Stephens introduced the first reliable keypoint detection algorithm in the late 1980s, which was later optimized by Shi and Tomasi in their Good Features To Track (GFTT) [6]. Invariance against scale changes was first introduced in Lowe's Difference of Gaussian (DoG) detector as a part of his Scale-Invariant Feature Transform (SIFT) [7]. His approach was then accelerated by Bay et al. in the Fast-Hessian detector inside their Speeded-up Robust Features (SURF) [8] using box filters. Based on Laplacian of Gaussian (LoG), further improvements considering position accuracy were made in the Center Surround Extrema (CenSurE) [9] algorithm. To extend the application area to mobile platforms and boost on-line processing capabilities, Rosten and Drummon developed Features from Accelerated Segment Test (FAST) [10]. Binary Robust Invariant Scaleable Keypoints (BRISK) [11] further robustified this detector. Rublee et al. provided each feature with a defined orientation as part of their Oriented FAST and Rotated BRIEF (ORB) [12] algorithm. While most of state-of-the-art detectors are based on the Gaussian pyramid, the Maximally Stable Extremal Regions (MSER) [13] detector explores the possible utilization of these characteristic regions.

After identifying stable and transformation-invariant features, each element has to be equipped with a unique signature for comparison and recognition purposes. At first, distribution-based descriptors are used, where the directional values are stored in vectors containing floating-point numbers. In feature matching, Euclidean distance is used for comparison purposes, e.g., in SIFT and SURF.

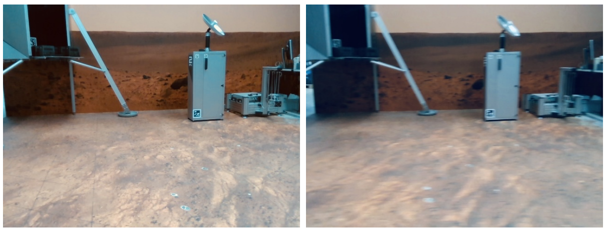


Figure 2: Exemplary images from the SurfA setting depicting the degree of motion blur to be expected within the contemplated projects. The image on the left-hand side illustrates the motion-blur-free case from dataset 11, whereas the image on the right-hand side from dataset 12 shows an increased degree of disturbance.

In pursuit of efficiency, the support region can also be described by correlating specific properties of individual pixel pairs inside the region of interest. Hereby, the characteristics are sampled in a binary string, and Hamming distance is utilized for comparison and matching. Promising patterns were proposed in Binary Robust Independent Elementary Features (BRIEF) [14] and its subsequent advancements ORB and BRISK [11], as well as Fast REtinA Keypoint (FREAK) [15].

Moving a step further, line features extracted by, e.g., Line Segment Detector (LSD) [16] and the corresponding Line Band Descriptor (LBD) [17], are highly suitable for describing the contours of human-built objects since reliable orientation information is automatically included.

4. EXPERIMENTAL EVALUATION

To extend our decision-making basis regarding further developments, we decided to benchmark the abilities of state-of-the-art feature detection and description algorithms in a reasonable frame using mission-related data.

4.1. Benchmark Metrics

For the evaluation, we introduce five performance metrics for an unbiased comparison between detectors and descriptors according to [1] and [2].

Correspondence — Especially for our target of creating a localization and mapping application, features have to be obtained repeatedly in a reliable manner. For the assessment of detectors, correspondence is defined as the number of regions of interest, which could be further utilized for, e.g., tracking purposes. To express it mathematically, the following condition must hold

$$1 - \frac{R_{\mu_A} \cap R_{(\boldsymbol{H}^T \mu_B \boldsymbol{H})}}{R_{\mu_A} \cup R_{(\boldsymbol{H}^T \mu_B \boldsymbol{H})}} < \epsilon_O, \qquad (1)$$

where R_{μ_A} represents the elliptic region A and R_{μ_B} region B. They are classified as corresponding, in case the overlap between R_{μ_A} and R_{μ_B}, when transformed to the reference image using homography relation $\boldsymbol{H}$, surpasses a given threshold value.

Repeatability — While correspondence represents the absolute number of "useful" features, repeatability expresses the same information relatively:

$$\textit{Repeatability} = \frac{\textit{\# Correspondences}}{\textit{\# Features in Image A}} = \frac{C^+}{F_A}. \qquad (2)$$

By expressing it in the relative frame, it is a measure for the precision of the feature extractor.

Matching Score — Apart from evaluating feature detectors from a theoretical perspective, the matching score indicates how well computer algorithms match the obtained regions. Thus, it is a measure for the distinctiveness of the feature support region. The metric is defined as

$$\textit{Matching Score} = \frac{C^+ \cap M^*}{F_A} = \frac{M^+}{F_A}, \qquad (3)$$

where M^+ represents the number of correct matchings, and M^* the number of all matchings. A proper match is achieved if the associated regions of the algorithmically identified match correspond to each other according to Equation 1.

Receiver Operating Characteristics — The characteristics of feature descriptors are benchmarked by the precision-detection-rate relationship, which is based on the number of correct and false matchings obtained from an image pair. The detection rate is defined as

$$\textit{Detection Rate} = \frac{\textit{\# Total Matchings}}{\textit{\# Correspondences}} = \frac{M^*}{C^+}. \qquad (4)$$

On the contrary, precision is calculated as M^+ with respect to the overall number of algorithmically identified matchings:

$$\textit{Precision} = \frac{\textit{\# Correct Matchings}}{\textit{\# Total Matchings}} = \frac{M^+}{M^*}. \qquad (5)$$

Since the characteristic is sensitive to the number of matchings, we have to select an adequate matching strategy. For the following descriptor benchmark, the procedure regarding nearest neighbor distance ratio (NNDR) [2] is selected. Therefore, two regions are matched in case the distance ratio between the first and second nearest falls below a given threshold θ:

$$\frac{\parallel D_B - D_A \parallel}{\parallel D_C - D_A \parallel} < \theta. \qquad (6)$$

This restriction narrows down the number of total matchings. It penalizes descriptors with a high amount of similar matchings, which is unfavorable since the uniqueness of the feature's fingerprint is desired. To improve the clarity, the precision-detection-rate relationship is plotted against each other. The resulting Receiver Operating Characteristics (ROC) is obtained by the variation of θ.

Computation Time — The last metric aims at the statistical distribution of the computation time. For comparison reasons, time expenditure is normalized for the detection, description, and matching of a single feature entity. It is essential for on-line applications, e.g., the construction of a VO, since the overall computation time is a determining factor for the real-time capability.

Table 1: Hardware properties of the evaluation computer.

Parameter	Specification
Model	Dell Precision 7540
CPU	Intel Core i7-9850H
Clock Rate	4.60 GHz
Memory	4 × 8 GB 2666 MHz DDR4 SDRAM
OS	openSUSE Leap 15.1

4.2. Prerequisite and Preparation

4.2.1. Dataset

Forming the foundation of our examination, the datasets were recorded with multiple Intel RealSense D435i depth-sensing camera systems as in the actual mission setup. For the benchmarking evaluation of visual feature extractors, we only use information from the RGB sensor. The camera settings are directly derived from the participating robots. Thus, the resolution is sized to 480×640 pixels at a frame rate of 15 Hz. The exposure and white balance settings are set to automated mode.

To cover the related environmental conditions as best as possible, we recorded multiple datasets containing major landmarks and representative mission scenarios at two different levels of information content. Table 4 gives an overview of all considered recordings. At first, the image sequences focus on stand-alone distinctive objects. It allows an isolated examination of specific elements of interest with minimized influence from the surrounding scenery. Therefore, the first part of the examination process contains the visual composition of both types of the flooring, SPU, lander in the planetary exploration setting, and the properties of the entire kitchen and living room assembly in the SMiLE-Laboratory. By doing so, it allows us to identify the best-suited feature detection algorithms for every occurring object. The analysis of stand-alone objects is carried out without considering motion blur in the first stage, except for the examination of the flooring. Figure 2 exemplarily illustrates the expected degree of motion blur within our contemplated missions. As a second step, we use image sequences containing realistic mission tasks for the evaluation. Here, we reconstructed five typical scenarios in total. For simplicity, all movements within the setups consist of linear trajectories.

4.2.2. Ground Truth

In order to provide a valuation basis for the performance metrics, the ground truth relation between the evaluated frames has to be established. The planar homography relates to the transformation between any projected points in the reference and target image if a planar surface geometry is achieved. Their relationship is defined as

$$x_{A_i} = \boldsymbol{H} x_{B_i}, \tag{7}$$

where any points in image B can be projected to image A by the application of the homography matrix $\boldsymbol{H}$. For its estimation, we chose the feature-based approach. Keypoints in both frames are detected and matched using a suitable detector-descriptor combination and pre-filtered by the NNDR matching strategy. $\boldsymbol{H}$ is then calculated using the feature-based image alignment estimation method with RANdom SAmple Consensus (RANSAC) for outlier rejection. Hereby, we opted for AKAZE [18] as the detection algorithm since it is unrelated to any of the evaluation candidates. On the descriptor side, BRISK was selected.

Table 2: Feature detectors for experimental evaluation.

Detector	Type	Class	Abbr.
GFTT	Corner	Point	GT
DoG	Blob	Point	DG
Fast-Hessian	Blob	Point	FH
CenSurE	Blob	Point	CS
MSER	Blob	Point	MS
FAST	Blob, corner	Point	FT
BRISK	Blob, corner	Point	BK
ORB	Blob, corner	Point	OB
LSD	Blob	Line Segment	LS

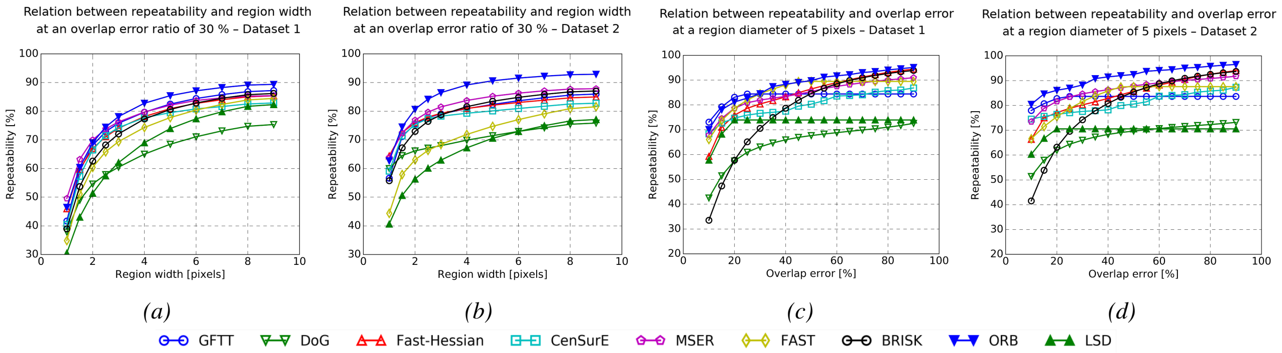
Table 3: Feature descriptors for experimental evaluation.

Descriptor	Datatype	# Elements	Size [Bytes]
SIFT	Float	128	512
SURF	Float	64	256
BRIEF	Binary	256	32
ORB	Binary	256	32
BRISK	Binary	512	64
FREAK	Binary	512	64
LBD	Binary	256	32

In evaluating floor properties, the planar condition is given. Unfortunately, the planar surface assumption in the other image sequences is not satisfied since many objects are visible. Thus, the transformation is not absolutely correct from the mathematical perspective. Nevertheless, in the case of our selected frame rate and relatively small movements between images, the homography relation can be approximately used as the base of evaluation. We performed a visual examination to verify the approximation, where the resulted $\boldsymbol{H}$ was considered sufficient for our analysis.

4.2.3. Hardware and Software Implementation

The calculation of the study-related experiments is performed by a computer with the hardware specification listed in Table 1. On the software side, we set up a benchmarking framework in C++ using OpenCV, which provides the implementation of all evaluated detectors, descriptors, and matching algorithms. In the case of LBD, we utilized a runtime-optimized binary version of the floating-point descriptor. Since the performance of



(a) *(b)* *(c)* *(d)*

Figure 3: Pre-evaluation of repeatability as a function of overlap error threshold and region size.

those algorithms is sensitive to parameter tuning, we use standard OpenCV parameters for each feature extractor to preserve neutrality.

Tables 2 and 3 provide an overview of the participating feature detection and description algorithms. The matching process is realized with the brute-force matcher since it always provides the best possible matching by comparing descriptions of features in the first set with all available descriptors in the second frame.

4.3. Pre-Evaluation of Tuning Parameters

Before conducting the actual study, it is worth discussing the influence of two tuning parameters, which might significantly impact the study's validity. By the nature of our performance metrics, they are closely related to the size of the allowed overlap error ϵ_O. Regions with a sizable surface area have a better chance of scoring a higher overlap value, which automatically improves the correspondence, repeatability, and matching scores. On this account, the comparison between feature detectors only makes sense if we select a common region size, as all of the considered detectors would produce different sizes of meaningful areas around detected features. To choose suitable values for overlap error and region size, we evaluated their relationship to repeatability on datasets containing isolated key elements.

At first, we target the relationship between repeatability and region size by retaining a constant ϵ_O since plenty of reference values can be found in the literature. In [1] the value was set to $40\ \%$. However, we further decreased the maximum acceptable overlap error to $30\ \%$ in compliance with our accuracy requirement. In Figure 3a-b, the results are exemplary plotted for the stand-alone analysis of the SPU and lander. As expected, repeatability increases in parallel to the expansion of region size. Except for FAST and DoG, the slopes of all detectors roughly show a comparable characteristic. They are more sensitive to smaller meaningful regions, where only a below-average performance was achieved. Consequently, we set the region diameter for the upcoming detector benchmark preliminarily to 5 pixels based on the charts. In the case of line segment features, a rectangular line support region is created with the same amount of pixels in its width.

To solidify our proposed selection of both tuning parameters, the sensitivity of the selected pairing is analyzed by calculating the relation between repeatability and ϵ_O. The diameter of the region is kept constant at 5 pixels, and the results are again exemplarily plotted in Figure 3c-d. It is mention-worthy that the BRISK detector is the most sensitive one in our configuration, while the curves of other ones are considerably moderate. Especially for smaller error margins, BRISK shows a more intense performance degradation than all other detectors. In summary, our demand of reaching a minimum of $70\ \%$ overlap between two regions is reasonable and will be used for the upcoming benchmark.

4.4. Feature Detector Benchmark

Within the detector benchmark, every detection algorithm has to be paired with a descriptor for the calculation of the matching score. To provide a neutral evaluation base, we selected FREAK since it is an independent descriptor without native detector pairing. By this means, possible pairing-based synergy effects are minimized. For the description of line segments, we utilized LBD. Table 4 shows the average repeatability and matching score of all participating feature detectors.

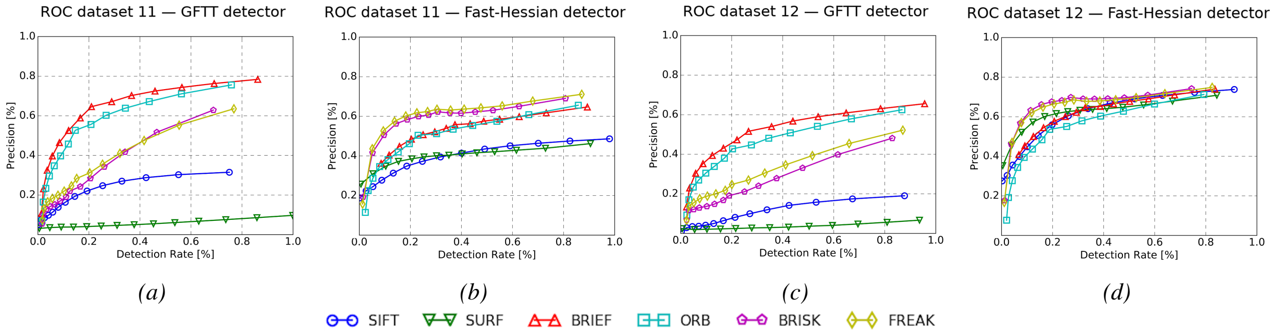
Starting with the analysis of stand-alone objects, it is to say that the findings from both environmental settings coincide with each other. ORB achieved the best performance for repeatability, while the remaining detectors are similar by the means of their robustness, as their repeatability scores are in a considerably narrow spectrum. Within the ranking, DoG and LSD produced the worst results, even though the examined objects are full of straight edges and, therefore, line representatives, especially in the latter case. Conversely, the matching scores have a more varying progression. Under the given circumstances, regions of interest selected by modern blob and traditional corner detectors produced the most promising matching scores. LSD, as the only detector that is not focusing on point features, also reached a ranking in the upper third. The next stand-alone element to be analyzed consists of laboratory-specific flooring. Starting with the motion-blur-free case in datasets 3 and 5, Fast-Hessian and GFTT achieved the best repeatability and matching score. On the other side, the evaluation

Table 4: Results of the detector benchmark averaged over the all image pairs and associated transformations (median). The benchmark suite contains 22 datasets (D.) within 13 different sceneries (S.). An overview of the abbreviations assigned to the detectors is given in Table 2.

	# S.	Environm.	# D.	Repeatability									Matching Score								
				GF	DG	FH	CS	MS	FT	BK	OB	LS	GF	DG	FH	CS	MS	FT	BK	OB	LS
SurfA – Planetary Exploration	1	SPU	1	83	69	82	80	82	78	81	86	74	70	50	68	75	22	62	51	38	65
	2	Lander	2	82	71	82	80	85	75	83	91	71	75	56	71	76	20	64	60	46	65
	3	Floor transl.	3	84	78	91	0	0	76	78	83	75	81	60	83	0	0	70	60	50	67
			4	70	80	91	0	0	55	0	87	74	63	59	72	0	0	49	0	53	68
	4	Floor rotat.	5	82	78	91	0	0	73	76	82	75	80	59	82	0	0	66	58	45	66
			6	74	79	91	0	0	60	0	83	81	69	56	74	0	0	52	0	45	73
	5	Scenario 1	7	84	74	85	85	84	77	84	88	74	79	57	76	82	41	68	61	49	66
			8	82	72	83	82	82	74	82	86	72	73	57	71	78	24	62	56	42	61
	6	Scenario 2	9	83	72	86	85	86	75	83	89	73	76	57	77	81	26	65	61	47	66
			10	79	73	86	83	84	72	82	86	72	68	57	73	78	26	61	58	43	64
	7	Scenario 3	11	84	74	86	87	83	78	84	89	75	80	59	78	84	30	70	63	50	68
			12	78	69	81	80	75	72	81	85	61	65	57	68	76	25	57	55	39	52
SMiLE – Urban Housing	8	Kitchen	13	84	71	83	84	82	76	84	88	81	76	57	72	81	29	65	60	45	71
	9	Living room	14	80	73	84	84	79	75	81	86	78	67	58	72	87	25	61	59	43	65
	10	Floor transl.	15	80	64	0	0	0	75	77	80	0	78	41	0	0	0	70	59	46	0
			16	54	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0
	11	Floor rotat.	17	81	64	0	0	0	77	77	79	0	79	41	0	0	0	72	61	41	0
			18	63	0	0	0	0	0	0	0	0	58	0	0	0	0	0	0	0	0
	12	Scenario 1	19	78	73	83	82	81	70	82	86	77	64	58	71	86	28	58	59	42	65
			20	76	70	80	79	76	68	79	81	74	57	53	64	76	22	52	50	35	58
	13	Scenario 2	21	80	74	85	84	81	72	83	87	78	68	60	73	80	29	61	62	45	68
			22	73	69	78	79	72	66	78	78	72	52	53	62	66	22	48	53	33	56

shows that the number of correspondence detected by MSER and CenSurE is constantly deficient. Since motion blur is also a common issue within our anticipated hardware setup, we examined the detectors' tolerance by applying this kind of disturbance on top of the already analyzed sceneries. In the planetary exploration case, a general drop in the number of detected correspondences can be observed. Nevertheless, GFTT and LSD can still reach a relatively high number of useful features. The remaining detectors made a dive towards zero but mostly still managed to detect sufficient features for tracking purposes, except for BRISK suffering a total loss of detection capability. The FAST detector also shows a similar behavior, which puts it down at the bottom of the repeatability ranking. Apart from the observations above, the performance ranking is roughly kept in the same order, where Fast-Hessian and GFTT still achieved the highest matchability. It is worth mentioning that LSD showed promising characteristics in dealing with this kind of disturbance. Although mostly resting in the middle field, it achieved the lowest matching score drop among all participants. Following the discoveries and realizations from the Martian environment, the flooring in the SMiLE-Laboratory appeared to be even more challenging for the selected detection algorithms. Unlike in the previous scenario, Fast-Hessian has more significant difficulties finding sufficient interesting points, which eventually results in total loss of tracking. While LSD was ranked in the upper midfield in the Martian scenario, it can only sporadically detect a handful of line features. This is not very surprising since the patterns in this setting are very homogeneous, resembling the characteristic of a noisy image. Encountered with motion blur in datasets 20 and 22, the detection capabilities have been degraded in such a way that GFTT remains the only algorithm that can provide a mentionable amount of features.

In terms of simulated mission tasks, we observed that the general characteristics of the performance metrics are comparable regardless of which scenario is carried out. It is not a surprise because all of the image sequences would comprehensively examine the mission surroundings in our cases. Regarding the matching score, CenSurE outstandingly achieved the highest score, followed by GFTT and Fast-Hessian. FAST-related detectors performed at the same level of performance as LSD and DoG. Remarkably, the drop of matchability in the event of induced movement is most significant in the case of ORB. It can be tolerated by DoG, FAST, and LSD, while all other detectors follow a similar shape of performance degradation. Further, the sensitivity concerning motion blur is not as significant in the isolated inspection of the floor. Comparing the average values, only small fluctuations are observable. This implies that the motion blur sensitivity of the floor is significantly higher than the effect on other included objects. In most cases, it is compensated by additional significant objects in the scene, which feature a more stable behavior.



(a) *(b)* *(c)* *(d)*

Figure 4: Result of the descriptor performance benchmark, illustrated on the example of dataset 11 and 12

4.5. Feature Descriptor Benchmark

After the original image has been simplified into a collection of promising elements, the selection of a suitable descriptor is the next decisive factor to ensure the uniqueness of each key element. In the scope of this study, we only compare the performances of the six introduced point descriptors in Table 3. Similar settings as in the previous feature detection benchmark were used for the tuning parameters. The error threshold is set to 30 %, whereas the size of the meaningful region around a keypoint is not normalized as in the previous examination. Unlike in the detector analysis, where the evaluation process is divided into two different parts, we decided to directly analyze mission-scenario-related datasets in this part of the study. While multiple detectors can be utilized in parallel, descriptors, in general, cannot collaborate with other ones. For this reason, the desired descriptor has to achieve good performances under any circumstances. We had to modify the standard valuation procedure, as it was initially designed for assessing single image pairs. Therefore, we collected results from each image pair during the runtime and constructed the ROC curves by varying the threshold value. The averaging is achieved by summarizing the individual parameters into one key figure, thus eliminating the time factor on its way.

In the evaluation process, we noticed that the description performance follows a general trend, regardless of which scene is chosen. For this reason, we exemplarily analyze its behavior based on mission scenario 3 in the planetary exploration setting in the followings. Starting from the motion-blur-free dataset in Figure 4a-b, the individual performances of the considered descriptors are roughly comparable with each other in the Fast-Hessian case. In the case where GFTT provides the detection basis for the evaluation, the characteristics of individual descriptors deviate considerably from each other. Here, BRIEF and ORB achieved the best results, followed by BRISK and FREAK in between and the floating-point descriptors at the end. In general, it is noticeable that the examined algorithms tend to form groups consisting of two descriptors each. This is presumably due to the utilized methodology for the description process. Thus, as a direct advancement of BRIEF, ORB shows a similar characteristic to its original. In contrast, BRISK behaves in the same manner as FREAK, which shares a lot of theoretical commonalities. With the occurrence of motion blur in Figure 4c-d, the ROC curves are getting flattered compared to the motion-blur-free case but remain in a similar shape. In the case of Fast-Hessian, the characteristics are getting more diffused. The best performances are tied between BRISK and FREAK, followed by the BRIEF-related algorithms in the middle and the floating-point descriptors at the end. All in all, the best results are achieved either by BRIEF-related algorithms or BRISK and FREAK, depending on the utilized feature detector. Surprisingly, SIFT and SURF performed underwhelmingly, as they are ranked among the best description algorithms under normal circumstances.

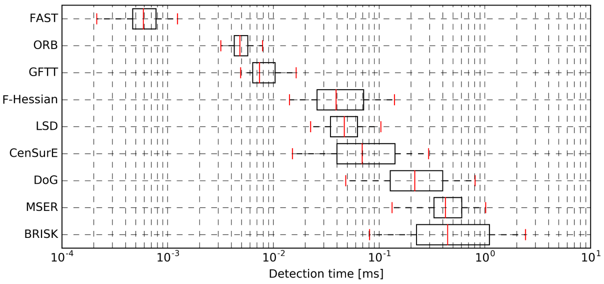


Figure 5: Average detection time per feature.

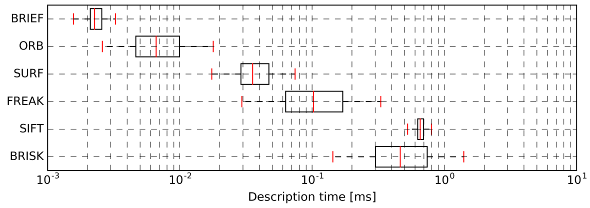


Figure 6: Average description time per feature.

4.6. Computation Time

In this section, the feature extraction algorithms are benchmarked in their computational performance. Therefore, we iterated through all 22 datasets to provide an adequate base for the overall computation time. It is essential to mention that the recorded benchmark values only account for the core tasks of the detection and description process. Other associated tasks inside the feature extraction framework are therefore not included.

Figure 5 displays the statistic distribution of the detection time per feature for the examined detectors. BRISK and MSER scored the slowest median detection time within the considered detectors, while FAST achieved the best score, followed by ORB and GFTT. Analogously, Figure 6 portrays the necessary computation time for the descriptors per feature. As expected, the list is topped with two binary descriptors, while SIFT is the slowest algorithm. On the contrary, BRISK and FREAK are at the lower end of the ranking.

5. CONCLUSION

In this paper, our focus is directed toward assessing possible applications in the field of multi-modal machine perception within the environment of Surface Avatar and SMiLE.

For establishing a localization and navigation framework in the visual domain, feature-based approaches are the superior choice, especially on mobile platforms. A benchmark study was then carried out, in which state-of-the-art feature extraction algorithms are evaluated based on real-world datasets from mission-related environments. In terms of feature detectors, we recommend the combination of multiple algorithms since the performances of feature extraction methods are highly dependent on the scene's photometric characteristics. ORB and Fast-Hessian are the means of choice for general detection tasks, whereas GFTT, LSD, and CenSurE are most suitable for handling individual specific situations in support. On the contrary, feature descriptors are not designed to collaborate with others, resulting in a fixed choice for the entire framework. Thus, we recommend selecting the ORB descriptor, as it provides the best balance between robustness and computational effort. Nevertheless, the utilization of a further descriptor would increase the algorithm's robustness by establishing redundancy, in case computational resources and storage capacity are sufficiently available. Therefore, we recommend the utilization of BRISK or FREAK as an addition since they achieved the best results in the motion-blur-afflicted cases.

REFERENCES

[1] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1, 2005.

[2] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, 2005.

[3] O. Miksik and K. Mikolajczyk, "Evaluation of local detectors and descriptors for fast feature matching," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*.

[4] D. Mukherjee, Q. M. Jonathan Wu, and G. Wang, "A comparative experimental study of image feature detectors and descriptors," *Machine Vision and Applications*, vol. 26, no. 4, 2015.

[5] D. Rondao, N. Aouf, M. A. Richardson, and O. Dubois-Matra, "Benchmarking of local feature detectors and descriptors for multispectral relative navigation in space," *Acta Astronautica*, vol. 172, 2020.

[6] S. Jianbo and Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[8] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer Berlin Heidelberg.

[9] M. Agrawal, K. Konolige, and M. R. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Springer Berlin Heidelberg, 2008.

[10] N. Nain, V. Laxmi, B. Bhadviya, B. M. D, and M. Ahmed, "Fast feature point detector," in *2008 IEEE International Conference on Signal Image Technology and Internet Based Systems*.

[11] S. Leutenegger, M. Chli, and R. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 IEEE International Conference on Computer Vision*.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, 2004.

[14] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg.

[15] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*.

[16] R. Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "Lsd: A line segment detector," *Image Processing On Line*, vol. 2, 2012.

[17] L. Zhang and R. Koch, "An efficient and robust line segment matching approach based on lbd descriptor and pairwise geometric consistency," *Journal of Visual Communication and Image Representation*, vol. 24, 2013.

[18] P. Fernández Alcantarilla, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," in *British Machine Vision Conference (BMVC)*, 2013.