# On the Design of Robust and Reliable Vision Front-Ends for Visually Degraded Environments

Vikrant Shah, Jagatpreet Nir, Pushyami Kaveti, and Hanumant Singh
Field Robotics Lab, Northeastern University, Boston, USA
Email: {shah.vi, nir.j, kaveti.p, ha.singh}@northeastern.edu

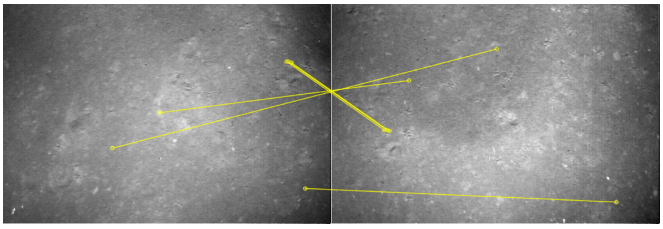
***Abstract*—Many of the well-known field robotic applications viz. Visual Simultaneous Localization and Mapping (SLAM), Structure for Motion (SfM), and Image Mosaicing, heavily rely on robust and reliable feature-based vision front ends. The job of a vision front end is to provide correspondence information between different camera views which is then directly fed into a bundle adjustment step. Although the atomic steps involved in constructing a vision front end are fairly well-known, many of the popular design choices do not perform reliably in a variety of Visually Degraded Environments(VDEs) characterized by low texture and contrast, unevenly lit and low-overlap imagery. In this paper, we develop novel metrics and quantitative analysis methods which can measure the impact of image pre-processing steps such as Contrast Limited Adaptive Histogram Equalization (CLAHE) on the improvement of vision front-end outputs. Our novel metrics, systematic methodology, and quantitative analysis can guide the selection amongst competing design choices to develop a reliable and robust vision front end that can operate in a wider range of situations including VDEs. We showcase that CLAHE improves the saliency of features and feature track length on a visually degraded underwater mud dataset, resulting in a substantial increase in correspondence information for the SfM solution. Finally, we perform an end-to-end SfM analysis that shows reduced accumulated drift over the long term and improved accuracy.**

***Index Terms*—Robust visual perception, Perception failure, Low texture and contrast, SLAM, CLAHE**

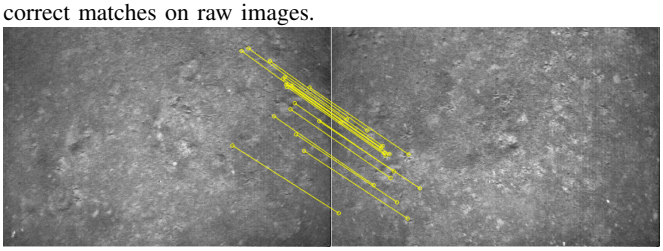
## I. Introduction

As the robotic community moves towards robust perception, robots must operate reliably for the long term in a variety of visual conditions which range from well-textured, well-lit, high contrast scenes to Visually Degraded Environments (VDEs) as illustrated in Fig. 1. Feature-based techniques [1] have shown to be effective in obtaining accurate and robust registration estimates for field robotic applications including visual Simultaneous Localization And Mapping (SLAM), mosaicking, and Structure from Motion (SfM). While executing visual algorithms two components make up the algorithmic core. A salient point feature detection and matching between pairs of images is followed by feature tracking over several consecutive image frames in time. Most of the feature matching algorithms are optimized for urban and natural landscapes which are characterized by 3D structure, high contrast, and rich textured surroundings. In such scenes, feature tracking with popular descriptors such as SIFT [2], SURF, and ORB has been shown to work reliably, both indoors and outdoors.

However, the reliability of the popular feature detectors and descriptors drops significantly in VDEs, which are characterized by low contrast, low texture, and uneven or low lighting. Sometimes, these visually degraded datasets work well with standard algorithmic parameters, but in many cases, the result is a partial or complete failure. Some of the failure cases can be remedied by tweaking the parameters associated with saliency detection and feature tracking while others require implementing more drastic steps like pre-processing images with CLAHE [3] or switching to completely different descriptors such as Zernike Moments [4]. In this paper, we develop a systematic approach for analyzing and quantitatively evaluating the performance of feature tracking in VDEs by defining a metric that can predict the performance of a feature detector-descriptor combination for an environment. Thus, we understand and explain the characteristics that impact pairwise image registration and recommend ways to improve feature tracking to reduce long-term drift. The main contributions



(a) Underwater seafloor mud image pair: SIFT features fail to find correct matches on raw images.



(b) SIFT features are able to find 17 good matches when operating on images preprocessed with CLAHE.

Fig. 1: An example pair of images from the underwater seafloor dataset of 62 images where the popular SIFT algorithm fails to register is shown in (a). The image pair, with low texture, non-uniform lighting, and low lateral overlap, highlights the challenges of image registration for visually degraded datasets. (b) shows pre-processing with Contrast Limited Adaptive Histogram Equalization (CLAHE) allows SIFT to successfully find matches and register the images.

| | Feature Detector | → Feature Descriptor | → Matching | → Geometric Check |
|---|---|---|---|---|
| **Purpose** | Find salient pts. | Describe pts. uniquely | Match across images | Remove outliers with RANSAC |
| **Options** | Harris, FAST, DOG, Hessian Affine | rBrief, HOG, Zernike Moments | Brute force, KNN tree based | Homography(4 pt), Essential matrix(5 pt), Fundamental matrix(8 pt) |
| **Attributes** | Tiling, Non-max suppression | | Cross matching, Lowe ratio | |
| **Key parameters** | N-No. of features T-Threshold | | D-Min. distance R-Lowe Ratio | n-Iterations, t-Inlier threshold, I-actual Inliers/N |
| **Effect low contrast** | T↓→ N↑ | | D↓, R ↑ | I↓ → n↑ |

TABLE I: This table presents the sequence of steps for most of the feature-based vision front-ends. The last two rows of the table highlight the tuning parameters and their effect on the output at each step. A well-designed and well-tuned front-end provides ample well-distributed and true correspondence information for bundle adjustment. Tuning parameters for long sequences of VDE datasets is challenging.

from this paper are:

1) We define a novel Descriptor Second Closest Distance Distribution (DSCDD) metric that defines the saliency of features in a single image. Therefore, it helps in making the right decisions at the start of the design process to make a reliable and robust image registration technique and front-end for large datasets.
2) We systematically analyze and explain the impact of pre-processing steps such as CLAHE with different aspects of feature detection.
3) We showcase the impact of CLAHE on multi-view constraints in VDEs by analyzing the end-to-end SfM performance of feature detector-descriptor combinations and quantify the improvement in accuracy.

## II. Background and Problem Setup

Sparse feature-based visual navigation methods are most prevalent in practice because they are computationally efficient and enable real-time solutions. The front-end of these pipelines can be summarized as a sequence of steps shown in Table I. The first step is to find distinct key points on the images which correspond to points in the real world. The primary function of this step is to detect salient features that are unique and easy to match. For visual navigation applications, these features must be well distributed over the scene to avoid degenerate cases and improve the numerical stability of the algorithms. Next, the feature descriptors are computed to encode a patch around the keypoints that describes the appearance of the patch. The important characteristics of feature descriptors for robust matching are invariance to rotation, intensity, contrast, lighting, and scale changes. The data association step then matches features between pairs of images and involves a geometric check to remove outliers that do not belong to the homography or projective models depending on the scene.

Feature detectors and descriptors play a vital role in visual navigation applications. During the feature detection stage, the default parameters often fail to find enough features forcing us to reduce detection thresholds. This results in detecting a lot of 'bad' features in addition to good features. The effect of these 'bad' features is that many of them encode to similar descriptors. Thus, in the matching stage, the minimum distance threshold must be reduced and the Lowe ratio threshold must be increased to find true matches. However, these thresholds further exacerbate the problem as they in turn allow a lot more false matches to get through. In the RANSAC step, this manifests as a lower fraction of true inliers making it harder to distinguish the real inliers from the whole set. In VDEs the effect of all these parameters is intertwined and leads to complicated relationships to the final performance requiring fine tuning for each application.

In this work, we aim to understand the characteristics of visually degraded environments and how they affect the performance of feature detectors and descriptors in image registration tasks. A lot of literature such as [5] has been devoted to comparing feature detectors and descriptors including analysis on low contrast thermal images [6]. In these works, the evaluation is based on pair-wise affine transformation analysis which does not correlate to complete system performance. Other studies have analyzed performance by accurately controlling movement of the cameras or the target, and even analyzed performance on long sequences for a SLAM algorithm [7]. None of the existing works systematically evaluates the performance of detectors and descriptors in VDEs as a complete system. We remedy this with a complete end-to-end analysis of feature tracking in VDEs for monocular SFM applications in section III-C.

## III. Methodology, Results, and Analysis

### A. Descriptor Second Closest Distance Distribution metric

To build a systematic understanding, we first try to define a VDE with a metric, so that given an image one might be able to say whether it is a VDE or even how much of a VDE. After a thorough literature review, the closest we come to such a definition are contrast metrics [8] like root-mean-square (RMS), box filter, bilateral filter, Global Contrast Factor (GCF), etc. Even though these metrics capture the contrast and local texture of imagery, none of them are good predictors of feature tracking performance.

We bridge this gap by proposing a novel Descriptor Second Closest Distance Distribution (DSCDD) metric demonstrated in Fig. 2 which describes VDEs as a combination of detector-descriptor-environment. This metric quantifies the visual saliency of the detected keypoints by computing the distribution of the descriptor distances to their second-best matches within an image. The DSCDD metric is inspired by the concept of Lowe ratio [2] used for filtering good matches. The spirit of this metric is based on an idea expressed in Shi and Tomasi's seminal work [9] that "a good feature is one that can be tracked well, so that the selection criterion is optimal by construction." Intuitively, a visually salient feature in the image must be different from other detected features. Therefore, in general, the distance of the descriptor of a feature

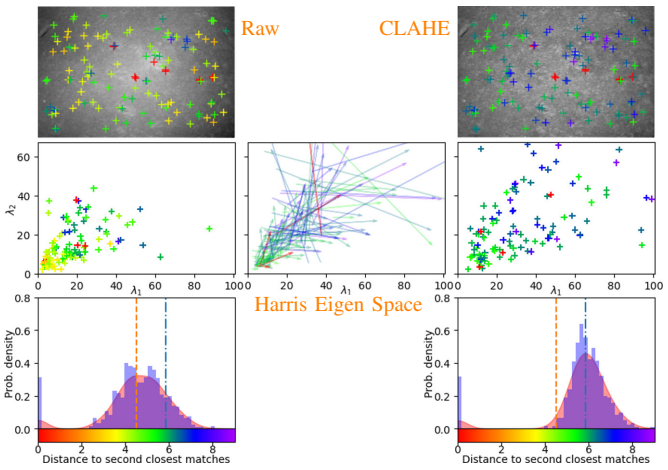



Fig. 2: This figure demonstrates how CLAHE helps Harris-Zernike features. The top left image shows the top 125 of 669 features on the raw underwater seafloor image color-coded by the distance to their second closest matches (third row). The top right image shows the performance of the same features on the CLAHE image. The Harris corner detector works in the eigenspace of the second-moment matrix. The middle plots show how the eigenvalues move towards the top right of the plot when using CLAHE images. The bottom plots show the proposed Descriptor Second Closest Distance (DSCDD) metric as a way to quantify the visual saliency of features of an image. Applying CLAHE moves the distribution to the right, implying that the descriptors become more 'unique'. The orange and green vertical lines show the two peaks.

to its second closest match represents the degree of saliency. A DSCDD distribution heavily weighted to the right indicates an abundance of unique features for tracking, while one weighted to the left predicts poor tracking performance.

### B. Improving Visual Saliency of Keypoints using CLAHE

CLAHE [3] is a contrast enhancement technique that takes care of both the global and local contrast by adaptively clipping the contrast to an upper threshold setting. This clip limit enables CLAHE to amplify the signal without adding noise.

The analysis in Fig. 2 quantitatively shows how CLAHE helps improve feature matching. Fig. 2 shows Harris-Zernike features on an underwater seafloor image. The Harris corner detector [10] works in the eigenspace of the second-moment matrix of intensities. CLAHE moves these eigenvalues up and to the right making existing features easier to detect and pushes new features above the threshold. The effect is also clear when we see the DSCDD plots; for the CLAHE processed image, the distribution moves to the right which indicates an improvement in visual saliency. Thus CLAHE helps by improving existing features and enabling the detection of new features. As we show later, this especially helps when working with low overlap underwater images.

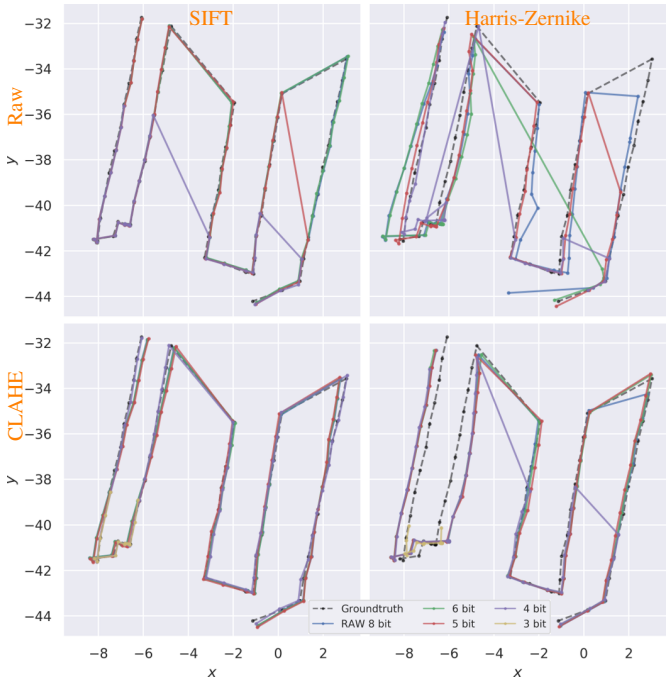



Fig. 3: These plots show the end-to-end SfM results from the underwater seafloor dataset of 62 images. SfM algorithms register all possibles images into a model and then start a new model to try registering the remaining ones. The plots show the all sub-models individually aligned to the groundtruth and combined. The dotted lines shows the groundtruth trajectory and different colors represent different simulated contrast levels by reducing their bit depths. SIFT struggles with lateral registration on the raw images and its performance improves significantly with CLAHE. The accuracy for the Harris-Zernike combination also improves after applying CLAHE.

### C. Analysing accuracy and long term drift with groundtruth

In addition to the direct effects of visually degraded environments described above, there are subtle issues that do not manifest during pairwise registration. Even though the registration algorithms only need 4, 5, or 8 points to define the model between pairs of images, fewer feature points cause instability in long sequences in the form of drift or a complete failure. Typically we would like to track tens or hundreds of well-distributed features based on image resolution. Our experience with underwater and iceberg imaging suggests degradation in a navigation estimate is associated with the lower number of tracked features as well as the stability of their tracks. Thus, for a thorough understanding of feature tracking performance, we perform an end-to-end analysis.

In contrast to prior work [7] that performs analysis on SLAM applications, we chose SfM to reduce the variability in the results due to processor load, multi-threading, loop closure detection, etc. Our methodology with this analysis is to understand feature tracking performance with respect to multiple dimensions viz. datasets from different VDE environments, the impact of contrast enhancement techniques such as CLAHE,

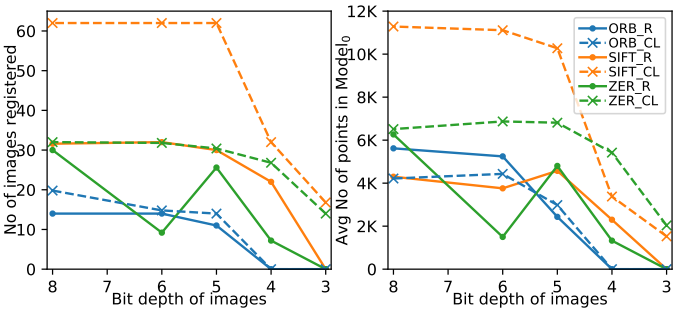



Fig. 4: These plots show the performance of ORB, SIFT and Harris-Zernike against different levels of image contrasts. ORB has the worst performance of the three. SIFT performs poorly on the raw images but its performance improves significantly with CLAHE. CLAHE has a smaller effect on Harris-Zernike's performance since the energy normalization of the Zernike Moment descriptor has a similar effect built-in.

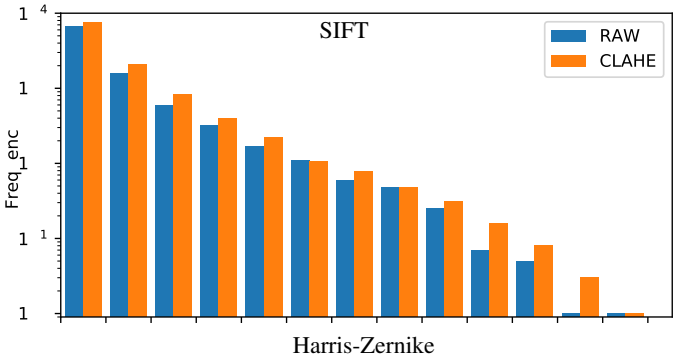



Fig. 5: Comparison of track lengths on the underwater mud dataset with and without CLAHE pre-processing. CLAHE improves the feature track lengths, which signifies stronger multi-view constraints that reduce pose and scale drift. Since Zernike descriptor's demeaning and energy normalization has an effect similar to CLAHE, there is a smaller increase in number of multiview constraints. This small addition of salient keypoints still improves the accuracy of the overall SfM solution in Fig. 3

and different combinations of feature detectors and descriptors.

We present one of the most interesting results from our analysis on the underwater seafloor dataset. This dataset (Fig. 1) was the most challenging since in addition to low texture and low contrast, the images also had non-uniform lighting and low lateral overlap. In fact, unlike all our other datasets, this dataset did not have any prominent edges or features. As we see in Fig. 3 and Fig. 4, SIFT fails to register all the raw images. However, when working with images pre-processed with CLAHE, SIFT successfully registers the trajectory even when the image bit depths are reduced to 5 bits.

Additionally, we analyze the long-term stability of the tracks for each case (Fig. 5). Feature points that can be tracked in multiple views are much stronger constraints as compared to pair-wise constraints for visual navigation algorithms and help in improving the accuracy and decreasing the drift of the navigation estimate. In Fig. 5 we can see that CLAHE improves the long-term trackability of the features which correlates with improvement in end-to-end SFM results in Fig. 3 as well. We have performed similar analysis on other visually degraded datasets of low contrast iceberg and IR images where we see similar improvements of the end-to-end SfM solution when CLAHE is applied.

## IV. Conclusions

The analysis and metrics developed in this paper can be used as a set of tools to quantify and compare competing choices while designing vision front-ends. We laid the foundation to systematically understand visual degradation using a novel DSCDD metric, explained how CLAHE helps feature detectors and descriptors perform better without the complex parameter tuning. Finally, we presented quantitative results on how different feature detectors fare in VDEs and how much CLAHE improves accuracy. This work also demonstrated how a true evaluation of the front end requires an end-to-end analysis of the system, especially in VDEs. Our work here scratches the surface for understanding feature tracking and designing visual front-ends that can operate in VDEs. Future research should explore computationally efficient and robust feature detectors and descriptors for VDEs.

## References


[1] G. Younes, D. Asmar, E. Shammas, and J. Zelek, "Keyframe-based monocular SLAM: design, survey, and future directions," *Rob. Auton. Syst.*, vol. 98, pp. 67–88, 2017.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, 2004.

[3] R. Eustice, O. Pizarro, H. Singh, and J. Howland, "Uwit underwater image toolbox for optical image proc. and mosaicking in matlab," in *Proc. IEEE Int. Symp. Underw. Tech. (02EX556)*, 2002, pp. 141–145.

[4] O. Pizarro and H. Singh, "Toward large-area mosaicing for underwater sci. appls." *IEEE J. Ocean. Eng.*, vol. 28, no. 4, pp. 651–672, 2003.

[5] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005.

[6] T. Mouats, N. Aouf, D. Nam, and S. Vidas, "Performance Evaluation of Feature Detectors and Descriptors Beyond the Visible," *J. Intell. Robot. Syst.*, vol. 92, no. 1, pp. 33–63, 2018.

[7] A. Schmidt and M. Kraft, "The impact of the image feature detector and descriptor choice on visual SLAM accuracy," in *Image Process. Commun. Challenges 6*. Springer, 2015, pp. 203–210.

[8] M. A. Qureshi, A. Beghdadi, and M. Deriche, "Towards the design of a consistent image contrast enhancement evaluation measure," *Signal Process. Image Commun.*, vol. 58, pp. 212–227, 2017.

[9] Jianbo Shi and Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR-94*, 1994.

[10] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Fourth Alvey Vis. Conf.*, 1988, pp. 147–151.