IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

BuFF: Burst Feature Finder for Light-Constrained 3D Reconstruction

Ahalya Ravendran[1], Mitch Bryson[2], Donald G. Dansereau[2]

***Abstract*— Robots operating in low-light conditions with conventional cameras face significant challenges due to the low signal-to-noise ratio in the images. Previous work has demonstrated the use of burst-imaging techniques to partially overcome this issue. This study proposes a novel feature finder that enhances vision-based reconstruction under extremely low-light conditions. The approach locates features with well-defined scale and apparent motion within each burst by jointly searching in a scale-slope space. We demonstrate improved performance in feature detection, camera pose estimation and reconstruction compared to state-of-the-art feature extractors on conventional and burst-merged images. This work opens avenues for robotic applications where low-light conditions often pose difficulties such as disaster recovery and drone delivery at night.**

I. Introduction

Integration of vision sensors has transformed the field of robotics, enabling robots to perform precise simultaneous localisation and mapping (SLAM) [1]–[3], structure-from-motion (SfM) [4]–[6], and depth estimation [7], [8]. However, state-of-the-art methods face difficulties operating in low-light conditions due to the low signal-to-noise ratio (SNR) in captured images. While carrying a light source is one solution, it is not always feasible or desirable, particularly in applications such as nocturnal animal behavioral studies or on smaller, weight- and power-restricted platforms. Thus, it is necessary to develop low-light image processing techniques for monocular cameras without any additional hardware modifications.

Burst imaging was introduced as a mobile photography technique to enable low-light photography on Pixel mobile phones [9]. This is achieved by aligning frames within a burst to a common image and merging them temporally to obtain a higher SNR image [9], [10]. Such burst-merged images demonstrate high SNR which is favorable for vision-based reconstruction, without the need for additional light sources [9]. Adapting burst imaging for robotics, burst-based SfM [11] established the viability of using burst imaging for reconstruction under extremely low-light conditions.

Prior works and ongoing developments in burst imaging [12]–[14] have focused on generating temporally merged images from bursts. However, extracting features from merged images tends to produce localisation errors due to pairwise misalignment. This is because of the limited information available from only two images at a time. When adapting burst imaging for robotics, the study on robotic burst imaging [11] failed to take advantage of the dynamics of robotic platforms. In this paper, we propose addressing these challenges by directly locating features within bursts. This approach mitigates problems associated with pairwise misalignment and missed opportunities in leveraging robotic platform dynamics to understand motion behaviors within individual bursts.

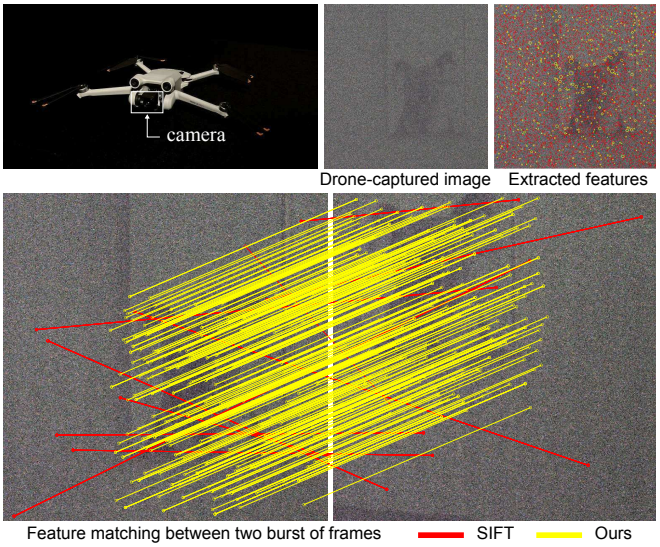


Fig. 1. Feature matching in low light: a commercial drone (top-left) captures imagery that is too noisy for conventional 3D reconstruction in low light (top-middle). This is because of the excessive spurious feature detection (top-right, red) and low-quality feature matches (bottom, red) offered by conventional features like SIFT. The proposed BuFF feature yields fewer, higher-quality features, resulting in many more correctly matched pairs (yellow). In this work, we show that our feature extractor enables 3D reconstruction in previously prohibitive low-light conditions.

We describe a robotic burst as a sequence of multiple images captured along a trajectory of a robot with small motions between each frame. We locate blob features with well-defined scale and apparent motion within each burst. To detect these blob features, we create a motion stack by shifting and summing frames based on their apparent motion. We then apply a difference of Gaussians (DoG) scale filter and describe the features using histograms of edge orientations. This descriptor operates on motion-filtered images, yielding an SNR improvement. By capturing multiple bursts and identifying features within each burst, we enable 3D reconstruction.

[1]Ahalya Ravendran is with the Imaging and Computer Vision Research Group, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, Australia ahalya.ravendran@data61.csiro.au

[2]Donald G. Dansereau and Mitch Bryson are with the Australian Centre for Robotics (ACFR), School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, 2006, NSW, Australia. mitch.bryson, donald.dansereau@sydney.edu.au

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Our key contributions are:

- We introduce burst feature finder (BuFF), a 2D + time feature detector and descriptor that finds features with well defined scale and apparent motion within a robotic burst,
- We propose the approximation of apparent feature motion as either 1D or 2D linear segments under typical robotic platform dynamics, enabling critical refinements relative to prior work on hand-held imagery, and
- We establish variations of BuFF matched to these apparent motion types and demonstrate it significantly outperforming state-of-the-art feature extractors applied to conventional and burst imagery in low-SNR scenes.

We validate our method using real burst imagery collected by a robotic arm. Code and dataset are available at `https://roboticimaging.org/Projects/BuFF/`.

To evaluate, we compare our results against state-of-the-art feature extractors including scale invariant feature transform (SIFT) [15], SuperPoint [16] and repeatable and reliable detector and descriptor (R2D2) [17] on conventional images and burst-merged images for SfM. We demonstrate that our proposed feature finder locates more true features and fewer spurious features in light-constrained scenes as shown in Figure 1. We demonstrate an overall improvement in reconstruction performance using the widely available COLMAP SfM software [4]. We also show that our method yields more complete 3D models with increased 3D points, more accurate camera pose estimates, and improved inlier matches with high precision compared to alternative approaches. We enable COLMAP to successfully converge for scenes where it was previously unable to do so due to low SNR.

This work opens the way for a broad range of robotics applications in low light, such as disaster recovery and drone delivery at night.

II. Related Work

A. Robotic Burst Imaging

Robotic burst imaging [11] captures multiple bursts of images with small motion variation along a trajectory of a robot using a conventional monocular camera. Unlike burst imaging [9], [10] that focuses on generating visually appealing images for the human visual system while accounting for handshake motion [12]–[14], robotic burst imaging deals with generating reliable estimates for downstream tasks such as 3D reconstruction. Robotic bursts also operate on lower light conditions compared to hand-held burst photography. In this work, we capture multiple robotic bursts similar to burst-based SfM [11]. Instead of merging images in a burst to improve the SNR of the resulting image for 2D feature extraction, we directly extract blob features in robotic bursts. This approach allows us to overcome the challenges posed by pairwise misalignment by leveraging robotic platform dynamics.

B. Feature Detection

For visual reconstruction, feature-based methods rely on establishing accurate and reliable pixel-level correspondences across multiple images. In robotics, the widely used technique for feature extraction is SIFT [15]. SIFT uses DoG filters at various scales to create a multi-scale pyramid, in which it searches for local extrema to identify scale-invariant blob features. SIFT then describes these features based on their histogram of gradients. Compared to other 2D feature detectors like speeded up robust features (SURF) [18] and ORB [19], SIFT demonstrates higher tolerance to noise [20].

Drawing inspiration from SIFT and light field features (LiFF) [21], we find blob features with well-defined scale and apparent motion by directly searching in a robotic burst. Our key insight is that while SIFT locates blobs with well-defined scales and locations in the 2D image plane and LiFF deals with identifying blobs with well-defined scales and locations in 3D space on light-field images, our approach leverages the use of multiple 2D images captured rapidly by a monocular camera. By incorporating burst frames, which include both 2D spatial and temporal data, we are able to locate and describe blobs with well-defined scales and apparent motion on the 2D image plane.

C. Feature Description

SIFT finds correspondences across other images by measuring the similarity between descriptors and computes feature matches for 3D reconstruction [15]. SIFT descriptors exhibit more robust performance and matching scores than other classical feature extractors [18], [19] in the presence of geometric changes for well-lit images [20]. In our work, we describe each feature, similar to SIFT, based on the histogram of edge orientations around its local neighborhood. Since these descriptors are computed at a stage where images have a higher SNR, we show them to have better accuracy and selectivity advantages over conventional methods.

D. 3D Feature Extraction

Employing SIFT for video feature extraction, Scovanner et al. [22] extends the SIFT descriptor to a higher dimension while relying on sequential frames from videos. Another approach [23] constructs a spatio-temporal space to locate features using explicit motion modeling via flow methods. However, these 3D methods exhibit limited performance [24], as they depend on individual pixels within low SNR images to identify local extrema. This leads to misalignment errors particularly within low-light robotic bursts. Our approach draws a parallel to the space-sweep technique widely adopted in computer vision [25], [26] for high SNR images. However, diverging from the conventional path of employing sequential frames naïvely, we leverage the typical motion characteristics of the robotic platform in designing the search space with higher SNR images for feature extraction.

E. Learning-based Feature Extraction

In this work, we focus on designing a physics-based approach to feature extraction. This is because capturing

Fig. 2. (left): In 3D scenes under general 6-DOF platform motion, features in a burst exhibit apparent motion that is well approximated by line segments, even under platform rotation; (right): In the special case of a platform moving orthogonal to the principal axis of the camera, the apparent motion follows parallel line segments. We exploit these observations to search either 2D or 1D spaces of linear apparent motion to detect BuFF features.

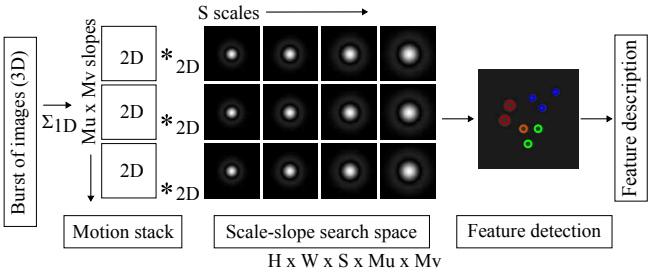


Fig. 3. The proposed feature detection architecture: The N-frame input burst is converted to an $M_u \times M_v$ motion stack by shifting and summing frames across a range of putative apparent motions ("slopes"). This is reduced to an M-frame stack for 1D apparent motion. Each frame of the motion stack passes through a DoG scale-space filter, and features are detected as extrema in the resulting joint scale-slope search space. Each feature is described using a SIFT-style histogram of gradients applied to its corresponding motion-stack image, rather than the input frames. The resulting features have distinct location, scale and apparent motion, and exhibit high precision, recall, and matching performance.

large amounts of low-light images for training purposes can be expensive, and learning-based methods offer no guarantee of generalizing beyond their training domains [27], [28]. We compare our physics-based approach to learning-based methods during evaluation.

III. Burst Feature Finder

In this section, we extract blob features with well-defined scale and apparent motion on low-light robotic burst and explain the integration of BuFF into an SfM pipeline.

A. Apparent Motion within a Robotic Burst

Capturing burst images using a camera mounted on a moving robotic platform exhibits fundamentally different motion profiles compared to hand-held burst photography. While handshake on handheld cameras contains significant high-frequency components, robotic platform dynamics are usually dominated by higher inertia and thus smoother instantaneous motion. This results in smooth and, for sufficiently fast bursts, locally linear apparent motion.

For a robotic burst [11] that captures multiple images with small motion variation along a trajectory of a robot using a conventional monocular camera, we introduce two versions of the feature finder, BuFF 2D and BuFF 1D, each tailored to address specific scenarios. BuFF 2D is designed to extract features from bursts characterized by 2D linear apparent motion as shown in Figure 2 (left). In this case, our motion hypothesis is that the local motion between frames in a burst can be approximated as locally linear in two dimensions. This behavior aligns with instances where the robotic platform, equipped with a front-facing camera, either approaches or moves away from the scene. This model also suits scenarios involving drones taking off or landing, utilizing a downward-facing camera.

Conversely, BuFF 1D serves as a subset design aimed at extracting features from bursts exhibiting 1D linear apparent motion. This occurs commonly when the motion of the robot is perpendicular to the principal axis of the camera. Here, the extracted features conform to a 1D linear motion pattern, which may occur either in parallel or in opposite directions as shown in Figure 2 (right). Reasons for the design choices are described in subsection III-E.

B. Motion Stack

BuFF identifies 2D + time features in a robotic burst, as shown in Figure 3. BuFF 2D searches through location, scale, and a 2D apparent motion space to build the 5D search space and find burst features at unique locations (ϕ), scales (σ), and apparent motions (λ). We pass all the images in a burst, i.e., N frames, through a shift-sum motion filter to generate a motion-filtered image H_M as in,

$$H_M(\phi)|_\lambda = \frac{1}{N}\sum_{n=1}^{N} I_n(u-\lambda_u\cdot(n-k), v-\lambda_v\cdot(n-k)), \quad (1)$$

where I_n is the n^{th} image in a burst and k is the index of the middle frame. $\lambda = [\lambda_u, \lambda_v]$ describes the apparent motion of the pixels in the horizontal and vertical directions and $\phi = [u, v]$ represents the pixel position in the horizontal and vertical directions, respectively.

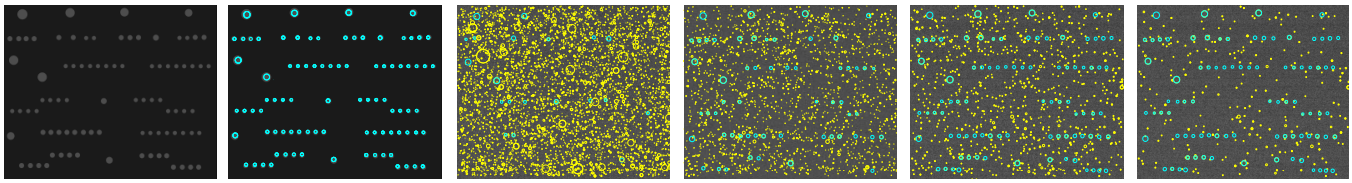
The motion filter shifts and sums each image in the burst according to the apparent motions determined by each λ. This process results in a stack of motion-filtered images, where each image corresponds to a unique apparent motion value. This technique is similar to the design of a focal stack [29].

C. Scale Slope Search Space

We extend the search space of SIFT by jointly searching for extrema across multiple scales and apparent motions. To achieve this, we convolve each motion-filtered image, computed by the motion stack, with DoG filters F_S over multiple scales. This generates a scale-slope search space. We find extrema in this joint 5D search space, D_{5D} as in

$$D_{5D}(\phi, \lambda, \sigma) = H_M(\phi)|_\lambda * F_S(\phi)|_\sigma. \quad (2)$$

Finding features on motion-filtered images rejects spurious features from noisy images. In the special case of BuFF



Test target	Gold standard (SIFT under bright illumination)	Conventional noisy TPR 0.3464 FPC 3875.2	Burst with merge TPR 0.7128 FPC 2558.8	BuFF 2D (Ours) TPR 0.9298 FPC 566.7	BuFF 1D (Ours) TPR 0.9298 FPC 469.8

Fig. 4. Validation with a printed test target: (left to right) Test target with disks of varying scales to demonstrate feature detection. SIFT finds features well under bright illumination (gold standard) but suffers from many false positive count (FPC) (yellow) and reduced true positive rate (TPR) (cyan) in low light. Employing burst imaging prior to applying SIFT improves performance, but BuFF 2D and 1D show much greater performance due to the joint scale-slope search – see also Figure 5.

1D, we consider a linear subspace that corresponds to 1D apparent motion, and this reduces the search space to 4D.

BuFF employs similar parameters as SIFT, such as octaves, levels, contrast threshold, and edge threshold. Additionally, BuFF uses predefined motion parameters to consider varying levels of apparent motion within a burst of images. As BuFF finds features in a burst, the number of images in the burst is also a controllable parameter.

D. Descriptor

Similar to SIFT and LiFF, we compute a histogram of edge orientations for each feature which corresponds to a 128-element vector. We describe the features on motion-filtered images which have higher SNR compared to the input images in the burst. This allows robust reconstruction of light-constrained scenes. While we do not directly employ BuFF to deal with dynamic scenes within the scope of this paper, BuFF computes features invariant to apparent motion. This offers inherent depth information from the motion-filtered images in our search space, opening avenues for applications in segmentation and depth selection.

E. Complexity

The speed computation of the BuFF algorithm depends on the number of DoG convolutions. This is because the cost of calculating the motion stack is minor in comparison. For BuFF 1D, where motion primarily occurs in one dimension, the optimal approach is to first construct a motion stack, followed by DoG convolution. This is computationally efficient when the number of burst images is higher than the count of apparent motions. Conversely, for BuFF 2D, designed for scenarios with two-dimensional motion between frames, a strategy of convolving burst images with DoG filters and then forming a motion stack is more computationally effective. This approach is preferable when the number of burst images is less than the square of the apparent motion counts.

By considering the number of burst images as N, apparent motion counts as M, and scales as S, the computational complexity for DoG convolution for BuFF 1D and BuFF 2D can be expressed as $S \times M$ and $S \times M^2$ for constructing a motion stack, followed by DoG convolution, respectively. Alternatively, it can be expressed as $S \times N$ for both approaches when convolving burst images with DoG filters and subsequently forming a motion stack.

In the context of a 10-image robotic burst with 12 scales and 5 apparent motions, employing the motion stack design followed by the DoG stack design proves twice as cost-effective in BuFF 1D compared to the reverse order. On the other hand, in BuFF 2D, opting for the motion stack design prior to the DoG stack design is 2.5 times more resource-intensive compared to starting with the DoG stack design. This is due to the differences in processing requirements – BuFF 1D convolves 5 motion-filtered images with 12 scales, while BuFF 2D convolves 25 motion-filtered images with the same number of scales during the motion stack design.

In terms of algorithmic memory, BuFF 2D demands more memory than BuFF 1D in the volume of $N \times M$. This trend aligns with speed computation. Memory usage can pose limitations depending on the use case and the number of images captured in the burst. This factor is relevant for bursts captured at extremely low light as for optimal feature extraction, this requires more number of images. For applications on constrained-memory resources, employing a motion stack, followed by DoG convolution is the feasible choice for both variations of BuFF. The optimal BuFF design depends on the range of apparent motion between foreground and background and should be tailored to the specific application.

IV. Results

In the following, we first evaluate the performance of our feature detector on a printed 2D test target scene containing features of varying scales. We quantitatively evaluate the performance of BuFF for reconstruction using COLMAP in subsection IV-B and camera pose estimation in subsection IV-C. We compare our method against classical SIFT [15], and learning-based extractors, SuperPoint [16] and R2D2 [17] on conventional noisy images. We also evaluate these feature extractors on burst with merge [11] which employs a technique to perform sequential estimation of motion, compensates for it using pairwise alignment and then performs feature extraction on the burst-merged image. We demonstrate an overall improved performance for light-constrained reconstruction using our method.

A. Feature Performance in Noise

We print a test target of 90 disks at varying scales with known feature locations to demonstrate feature extraction

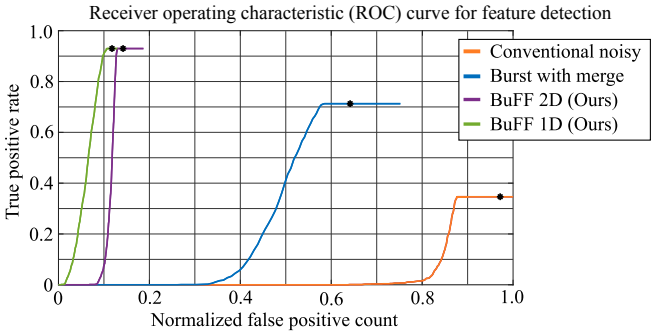


Fig. 5. ROC curves for (orange) SIFT on conventional noisy images, (blue) SIFT on merged burst images, (violet) BuFF with 2D linear local apparent motion, and (green) BuFF with 1D linear local apparent motion. BuFF 1D admits fewer spurious features because of the smaller search space, and both variants of BuFF exhibit higher TPR because of the signal boost associated with joint scale-slope search. We use the ROC curves to select comparable peak thresholds for each method, such that 10% of detected features are false positives.

(see Figure 4, left). The color contrast ratio between the disks and the background is 51:26 where 255 is white.

We use a Basler 1600-60um monocular machine vision camera paired with an f/11 lens to capture bursts of images of the test target. The resulting images are 12-bit monochrome and have a resolution of 1600x1200 pixels. We select an appropriate combination of lux and f-number for our experiment to match the conditions for capturing images using the DJI Phantom Pro drone with an f/2.8, specifically at 1.18 lux with an exposure time of 0.125 ms to emulate a typical low-light drone delivery application.

We capture the test target scene over a 100 ms exposure time, to obtain a high SNR image, which serves as the gold standard for comparison. This gold standard represents capturing the test target scene under bright illumination. We observe fixed pattern noise on all captured images and compensate for it by capturing a bias frame under the same conditions as the images with the lens cap on, and then subtracting the bias frame from the captured images.

We extract features using our method from 100 captured bursts of the same test target scene, each with 10 images for feature evaluation. We compare our method against VLFeat[1] implementation of SIFT on conventional noisy images and burst-merged images as shown in Figure 4. SIFT detects excessive spurious features when applied to conventional images and fails to extract all true features when used on burst-merged images. Both BuFF variants extract most of the true features and fewer spurious features compared to alternative methods.

We examine the average TPR and FPC of the captured images of the test target scene for a sweeping peak threshold, as shown in Figure 5. Conventional methods find fewer true positive features and more spurious features in captured images for all peak threshold, whereas burst-merged images yield more true positive features and fewer spurious features than conventional images. Our method outperforms both

[1]https://www.vlfeat.org/overview/sift.html

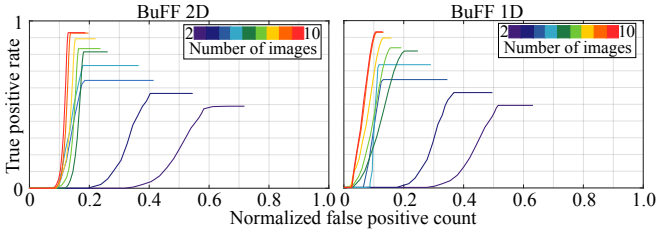


Fig. 6. Impact of number of frames in a burst: Employing more images improves both spurious FPC and TPR. These performance curves vary with camera, scene, and platform motion characteristics. For this scenario, performance saturates at around 10 images per burst.

conventional noisy and burst-merged images by finding the highest overall number of true positive features. As the search space is larger for BuFF 2D compared to BuFF 1D, between our methods, BuFF 1D finds fewer spurious features compared to BuFF 2D.

We select the peak threshold at which each method performs at its best for the input noise level of the test target images with edge threshold 10, and DoG scales covering 6 octaves over 4 levels per octave. Specifically, we choose the peak threshold value at the highest TPR with 10% of total FPC for both feature evaluation and reconstruction performance. While we compute approximate flow estimates using the gold standard scene as apparent motion prior for BuFF during reconstruction, real-time implementation can potentially benefit from leveraging inertial sensor measurements of robotic platforms.

We further compare the performance of our method for different numbers of images in a burst, as shown in Figure 6. As the number of images in a burst increases, the SNR of the motion-filtered images increases by $\sqrt{N}$, improving the search in scale-slope space for our feature finder. There is minimal difference in extracted true features and spurious features after ten images for the captured robotic burst. The number of images needed to find the most true features within a robotic burst depends on the imaging scenario including camera, scene and platform motion dynamics.

B. Reconstruction Accuracy

We evaluate reconstruction performance by mounting the same camera in identical lighting conditions as discussed in subsection IV-A, on a UR5e robotic arm as shown in Figure 7 to capture robotic bursts. We enable the robot to operate

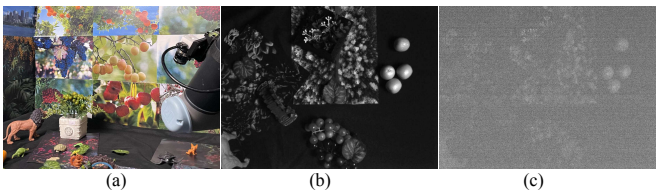


Fig. 7. Examples of captured images: (a) View of a well-lit scene and arm-mounted camera; (b) A gold standard image captured over 100 ms exposure time; and (c) A single noisy image captured over 5 ms at the same camera pose. The full dataset contains 10 trajectories, each with 20 bursts of 10 images.

TABLE I

AVERAGE PERFORMANCE OF RECONSTRUCTION FOR LIGHT-CONSTRAINED SCENES WITH 1D AND 2D APPARENT MOTION.

BOLD: OVERALL BEST RESULTS, UNDERLINE: BEST RESULTS FROM OUR PROPOSED APPROACHES.

Apparent Motion	Method		Convergence Rate	Images Pass %	Keypoints/ Image	Putative Matches/ Image	Inliers Matches/ Image	Match Ratio	Match Score	Precision	3D Points/ Image
1D	Gold standard		1.00	100	2880.96	$1.18\text{x}10^{4}$	$1.17\text{x}10^{4}$	4.16	4.14	0.99	257.40
	Conventional noisy	SIFT	0.00	0	5573.68	$6.84\text{x}10^{1}$	$6.41\text{x}10^{1}$	0.01	0.01	0.87	0.00
		SuperPoint	0.00	0	29.48	$4.50\text{x}10^{-1}$	0	0.02	0.00	0.00	0.00
		R2D2	0.00	0	2000.00	**$1.32\text{x}10^{3}$**	$1.72\text{x}10^{2}$	**0.66**	0.09	0.13	0.00
	Burst with merge	SIFT	0.80	59	2216.14	$3.62\text{x}10^{2}$	$3.32\text{x}10^{2}$	0.19	0.17	0.93	31.60
		SuperPoint	**1.00**	69	154.64	$1.26\text{x}10^{1}$	$7.45\text{x}10^{0}$	0.58	**0.29**	0.56	4.20
		R2D2	0.60	37	2000.0	$1.20\text{x}10^{3}$	$2.18\text{x}10^{2}$	0.60	0.11	0.18	3.20
	BuFF 1D (Ours)		**1.00**	**98**	1386.30	$3.84\text{x}10^{2}$	**$3.75\text{x}10^{2}$**	0.28	0.27	**0.98**	**36.50**
	BuFF 2D (Ours)		**1.00**	92	1635.82	$3.45\text{x}10^{2}$	$3.38\text{x}10^{2}$	0.23	0.23	**0.98**	36.20
2D	Gold standard		1.00	100	3270.16	$1.30\text{x}10^{4}$	$1.29\text{x}10^{4}$	3.89	3.87	0.99	305.50
	Conventional noisy	SIFT	0.00	0	5680.10	$1.84\text{x}10^{1}$	$1.43\text{x}10^{1}$	0.00	0.00	0.86	0.00
		SuperPoint	0.00	0	25.88	$3.38\text{x}10^{-1}$	0	0.01	0.00	0.00	0.00
		R2D2	0.00	0	2000.00	**$1.30\text{x}10^{3}$**	$1.88\text{x}10^{2}$	**0.65**	0.10	0.15	0.00
	Burst with merge	SIFT	0.80	77	2745.74	$4.85\text{x}10^{2}$	$4.57\text{x}10^{2}$	0.18	0.17	0.95	32.20
		SuperPoint	0.40	19	94.40	$2.87\text{x}10^{1}$	$8.15\text{x}10^{0}$	0.24	0.07	0.24	10.84
		R2D2	0.40	41	2000.00	$1.25\text{x}10^{3}$	$1.95\text{x}10^{2}$	0.62	0.10	0.16	3.99
	BuFF 1D (Ours)		0.80	85	2572.40	$5.55\text{x}10^{2}$	$5.43\text{x}10^{2}$	0.21	0.21	0.95	42.20
	BuFF 2D (Ours)		**1.00**	**98**	2143.96	$5.65\text{x}10^{2}$	**$5.51\text{x}10^{2}$**	0.26	**0.26**	**0.98**	**47.20**

in ten different trajectories. Each of these trajectories have 20 bursts, with each burst containing 10 images. Among these trajectories, five demonstrate 2D apparent motion between burst frames. By employing motion constraints, we capture five trajectories with 1D apparent motion between burst frames as described in detail in subsection III-B.

Our scene replicates a forest environment, featuring diverse objects with varied textures, shapes, and sizes. To closely match the noise levels from our earlier printed test target experiment (shown in Figure 4), we capture our dataset over an exposure time of 100ms for gold standard robotic bursts and 5 ms for low-light robotic bursts.

We compute features for reconstruction evaluation from both variants of BuFF using selected peak threshold as explained in Figure 5 and design the motion stack over 7 apparent motion counts in horizontal and vertical directions. We operate all our alternative feature extractors on the common image (i.e., middle image) in a robotic burst as they are single-image based approaches. Burst with merge which employs multiple-frames, similar to our approach also computes sparse features corresponding to the pose of the common image of the burst. For the descriptor, we apply L1 root normalization across all methods, similar to [21] to yield improved matches.

We assume there are no dynamic scenes during reconstruction. This is because we evaluate BuFF using COLMAP SfM reconstruction pipeline which treats features associated with dynamic scenes as outliers and remove them using random sample consensus (RANSAC). We compare the reconstruction performance using the feature comparison approach in [4], as shown in Table I. We evaluate the numbers of keypoints per image, putative feature matches per image, the number of putative matches classified as inliers, match ratio (the proportion of detected features yielding putative matches), precision (the proportion of putative matches yielding inlier matches), matching score (the proportion of detected features yielding inlier matches), and the mean number of 3D points per reconstructed image.

In Table I, we highlight the overall best results in bold and underline that from the variants of our method. Both SIFT and learning-based feature extractors fail to reconstruct light-constrained scenes when using conventionally captured images, regardless of the trajectory followed. Our method outperforms SIFT on burst-merged images by reconstructing more inlier matches, more 3D points with improved match ratio and match score. Although learning-based methods exhibit competitive performance in terms of match ratio and match score, especially R2D2, which is inherently designed to optimize average precision (AP) and excel at finding potential matches, our approach computes higher inlier matches in low-light scenes. Our method also computes more 3D points overall and converges for most of the scenes in both BuFF 1D and BuFF 2D datasets.

Between the two variants of our method, BuFF 2D employs 92% of all images for reconstructing light-constrained scenes on the 1D dataset, displaying competitive putative matches, inlier matches, and 3D points per image while BuFF 1D demonstrates greater match count and overall 3D points. This is because there are fewer spurious feature detection in a limited search space compared to BuFF 2D. When comparing BuFF 2D and BuFF 1D on the 2D dataset, BuFF 1D remains competitive with other methods but shows slightly lower performance than BuFF 2D. This is because of the motion constraints introduced by 1D motion filters.

C. Pose Estimation

We evaluate the accuracy of our camera trajectory estimation by employing a robotic arm to collect precise ground truth poses. We align and scale the camera poses estimated through COLMAP with the ground truth poses as there is an inherent scale ambiguity in monocular SfM. We compute the average absolute trajectory and relative pose

TABLE II

MEAN TRANSLATION ERROR AND MEAN ROTATIONAL ERROR IN CAMERA POSES FOR RECONSTRUCTION METHODS OF LIGHT-CONSTRAINED SCENES WITH 1D AND 2D APPARENT MOTION. BOLD: OVERALL BEST RESULTS, UNDERLINE: BEST RESULTS FROM OUR PROPOSED APPROACHES.

Apparent motion	Method	Absolute trajectory error trans. (cm)	Absolute trajectory error rot. (deg)	Relative pose error trans. (cm)	Relative pose error rot. (deg)
1D	Burst with merge - SIFT	2.84	2.07	4.97	0.73
	Burst with merge - SuperPoint	2.13	2.17	5.12	0.81
	Burst with merge - R2D2	1.71	1.20	3.17	0.04
	BuFF 1D (Ours)	**1.23**	1.24	**2.00**	0.05
	BuFF 2D (Ours)	1.27	**0.62**	2.01	**0.02**
2D	Burst with merge - SIFT	1.80	1.11	3.60	0.04
	Burst with merge - SuperPoint	1.56	1.06	3.71	0.04
	Burst with merge - R2D2	1.48	**0.55**	2.84	0.04
	BuFF 1D (Ours)	1.53	1.17	2.99	0.05
	BuFF 2D (Ours)	**1.44**	0.58	**2.55**	**0.03**

errors for translation and rotation from the poses estimated by COLMAP for all methods across all scenes, as shown in Table II. The BuFF variants outperforms most of the alternative approaches on burst-merged images in translation and rotation. Between the variants, BuFF 2D shows a slight advantage over BuFF 1D for rotation estimates.

We compare the camera pose estimates qualitatively for a scene from the BuFF 2D dataset as shown in Figure 8. The arbitrary scale factor is established based on the distance between the initial pair of registered images. We adjust the inlier rate settings to promote convergence and thereby obtain camera pose estimates using SIFT features on conventional images. However, these estimates remain inaccurate. When applied to burst-merged images, SIFT yields a greater number of reconstructed poses. Both variants of our proposed method outperform alternative approaches by reconstructing the most camera poses. They also provide more accurate camera pose estimates, which closely align with the gold-standard poses achieved using SIFT on high-SNR images.

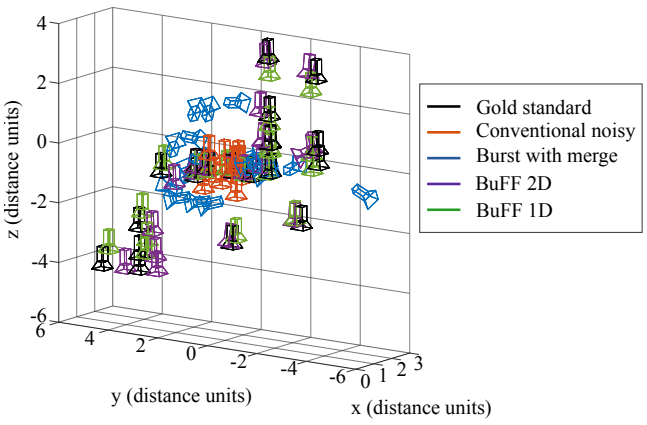


Fig. 8. Camera trajectory estimation: Conventional noisy images using SIFT yields poor camera pose estimates (orange) and fails to converge without relaxing the inlier rate settings. Using burst-merged images with SIFT (cyan) leads to better camera pose estimates than conventional noisy images. However, our proposed BuFF variants, 2D (violet) and 1D (green), outperform both conventional and burst-merged approaches and demonstrate the most accurate pose estimates.

TABLE III

TIME TAKEN FOR RECONSTRUCTION INCLUDING FEATURE EXTRACTION.

Method		Time taken (min)
Conventional noisy	SIFT	15.04
	SuperPoint	0.32
	R2D2	1.99
Burst with merge	SIFT	13.86
	SuperPoint	0.27
	R2D2	1.56
BuFF 2D (Ours)		108.50
BuFF 1D (Ours)		18.42

D. Speed

For a dataset comprising 200 images captured in 20 bursts over a 5 ms duration with strong noise, our MATLAB implementation took 18.42 minutes for detecting and describing burst features and reconstructing the scene using BuFF 1D. When using BuFF 2D, the processing time extended to 108.5 minutes on an Intel i7-9700 processor operating at 4.70 GHz, as detailed in Table III. The speed of our implementation depends on the noise levels in the images; higher noise levels lead to the detection of spurious features, and describing each detected feature requires a considerable amount of time.

The reported time for BuFF 2D and BuFF 1D followed the design of building a motion stack prior to the DoG convolutions, which is the slower of the two options for BuFF 2D and faster for BuFF 1D as discussed in subsection III-E. In comparison, operating the SIFT MATLAB implementation on 20 conventional noisy images took 15.04 minutes, while operating on burst-merged images with align and merge took 13.86 minutes. Learning methods that used pre-trained models and corresponding evaluation showed significantly faster performance. Note, all the learning methods are accelerated with NVIDIA GeForce RTX 3080 Laptop GPU. We expect BuFF implementation to be substantially accelerated with an adaptive design and an optimized compiled language, with potential for further computation improvement through parallel architecture of GPUs.

E. Failure cases

The performance of our method depends on the signal quality present in the captured burst. Having more images

in the burst (N-frames burst) benefits feature detection, but as $\sqrt{N}$, where N is the number of images in a burst. Ultimately, there needs to be enough signal in the images for BuFF to detect features, which is particularly challenging in low-contrast scenes. Our method also struggles to operate on bursts with a wide range of apparent motion between frames. Such conditions result in aliasing features, which can be addressed by using faster burst rates and implementing smoother parameterization techniques to account for variations in apparent motion.

V. Conclusions

We introduced a novel feature extractor that finds blob features with well-defined scale and apparent motion in low-light robotic bursts. By capturing multiple bursts along typical motion trajectories of a robot and locating burst features within each burst, we successfully demonstrated the ability to reconstruct scenes in low-light conditions. We showed higher quality true features, fewer spurious features using our feature finder. We demonstrated improved reconstruction performance with higher inlier matches and better precision compared to state-of-the-art feature extractors on conventional and burst-merged images. We also reconstructed more low-light scenes with a higher number of 3D points with an overall improved convergence rate and accurate camera pose estimates.

This work paves the way for a broad range of applications in which low light commonly poses challenges such as disaster recovery and drone delivery at night. For future work, we anticipate learning-based approaches for burst feature extraction and reconstruction of scenes obscured by visually challenging features like snow.

References

[1] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[2] Z. Teed and J. Deng, "Droid-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 16 558–16 569, 2021.

[3] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "Nice-SLAM: Neural implicit scalable encoding for SLAM," in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 786–12 796.

[4] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[5] D. Rajamohan, J. Kim, M. Garratt, and M. Pickering, "Image based localization under large perspective difference between SfM and SLAM using split sim (3) optimization," *Autonomous Robots*, vol. 46, no. 3, pp. 437–449, 2022.

[6] S. Liu, X. Nie, and R. Hamid, "Depth-guided sparse structure-from-motion for movies and TV shows," in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 980–15 989.

[7] N.-H. Wang, R. Wang, Y.-L. Liu, Y.-H. Huang, Y.-L. Chang, C.-P. Chen, and K. Jou, "Bridging unsupervised and supervised depth from focus via all-in-focus supervision," in *Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 621–12 631.

[8] T.-W. Hui, "RM-Depth: Unsupervised learning of recurrent monocular depth in dynamic scenes," in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1675–1684.

[9] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–12, 2016.

[10] O. Liba, K. Murthy, Y.-T. Tsai, T. Brooks, T. Xue, N. Karnad, Q. He, J. T. Barron, D. Sharlet, R. Geiss, *et al.*, "Handheld mobile photography in very low light," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–16, 2019.

[11] A. Ravendran, M. Bryson, and D. G. Dansereau, "Burst imaging for light-constrained structure-from-motion," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 1040–1047, 2021.

[12] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Krainin, C.-K. Liang, M. Levoy, and P. Milanfar, "Handheld multi-frame super-resolution," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–18, 2019.

[13] G. Bhat, M. Danelljan, F. Yu, L. Van Gool, and R. Timofte, "Deep reparametrization of multi-frame super-resolution and denoising," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 2460–2470.

[14] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3291–3300.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[16] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 224–236.

[17] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2D2: Reliable and repeatable detector and descriptor," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[18] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *European Conference on Computer Vision (ECCV)*. Springer, 2006, pp. 404–417.

[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

[20] M. Bansal, M. Kumar, and M. Kumar, "2D object recognition: A comparative analysis of SIFT, SURF and ORB feature descriptors," *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18 839–18 857, 2021.

[21] D. G. Dansereau, B. Girod, and G. Wetzstein, "LiFF: Light field features in scale and depth," in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8042–8051.

[22] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT descriptor and its application to action recognition," in *ACM International Conference on Multimedia*, 2007, pp. 357–360.

[23] M. Al Ghamdi, L. Zhang, and Y. Gotoh, "Spatio-temporal SIFT and its application to human action classification," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 301–310.

[24] S. Yu, B. Park, J. Park, and J. Jeong, "Joint learning of blind video denoising and optical flow estimation," in *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[25] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Computer Vision and Pattern Recognition (CVPR)*, 1996, pp. 358–363.

[26] X. Zabulis, G. Kordelas, K. Mueller, and A. Smolic, "Increasing the accuracy of the space-sweeping approach to stereo reconstruction, using spherical backprojection surfaces," in *International Conference on Image Processing*, 2006, pp. 2965–2968.

[27] K. Wei, Y. Fu, J. Yang, and H. Huang, "A physics-based noise formation model for extreme low-light raw denoising," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2758–2767.

[28] H. S. Choi, C. Crump, C. Duriez, A. Elmquist, G. Hager, D. Han, F. Hearl, J. Hodgins, A. Jain, F. Leve, C. Li, F. Meier, D. Negrut, L. Righetti, A. Rodriguez, J. Tan, and J. Trinkle, "On the use of simulation in robotics: Opportunities, challenges, and suggestions formoving forward," *Proc. of the National Academy of Sciences of the United States of America*, vol. 118, no. 1, pp. 1–9, 2021.

[29] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Linear volumetric focus for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, pp. 15–1, 2015.