Burst Imaging for Light-Constrained Structure-From-Motion

Ahalya Ravendran, Mitch Bryson, Donald G. Dansereau

***Abstract*—Images captured under extremely low light conditions are noise-limited, which can cause existing robotic vision algorithms to fail. In this paper we develop an image processing technique for aiding 3D reconstruction from images acquired in low light conditions. Our technique, based on burst photography, uses direct methods for image registration within bursts of short exposure time images to improve the robustness and accuracy of feature-based structure-from-motion (SfM). We demonstrate improved SfM performance in challenging light-constrained scenes, including quantitative evaluations that show improved feature performance and camera pose estimates. Additionally, we show that our method converges more frequently to correct reconstructions than the state-of-the-art. Our method is a significant step towards allowing robots to operate in low light conditions, with potential applications to robots operating in environments such as underground mines and night time operation.**

***Index Terms*—Computer Vision for Automation, SLAM**

I. Introduction

CURRENT and emerging robots use vision sensors for a broad range of tasks including simultaneous localisation and mapping (SLAM), navigation, pose estimation, depth estimation and 3D reconstruction. State-of-the-art methods for structure-from-motion (SfM) and vision-based reconstruction (e.g. [1], [2]) perform well under good lighting conditions but fail to reconstruct in low light for tasks such as autonomous driving, drone surveillance and underground mining. When using light-constrained images, state-of-the-art methods [1] yield incorrect 3D shape estimation, inaccurate camera trajectories and in some scenes even fail to converge for light-constrained SfM. This results from low signal-to-noise ratio (SNR) in images, where true features are not detected and/or spurious features are detected and matched with features in other images for reconstruction.

Burst photography is an established mobile photography technique which uses a series of consecutive frames with small exposure time to produce a single image with an improved SNR upon merging [3]–[6]. Burst imaging has been shown to improve the SNR in images without the need for additional ambient light sources [4]. However, prior works and on-going developments on burst imaging are steered towards mobile photography, where the primary objective is to produce convincing content for human visual perception while accounting for camera motion due to handshake and scene motion within a single burst.

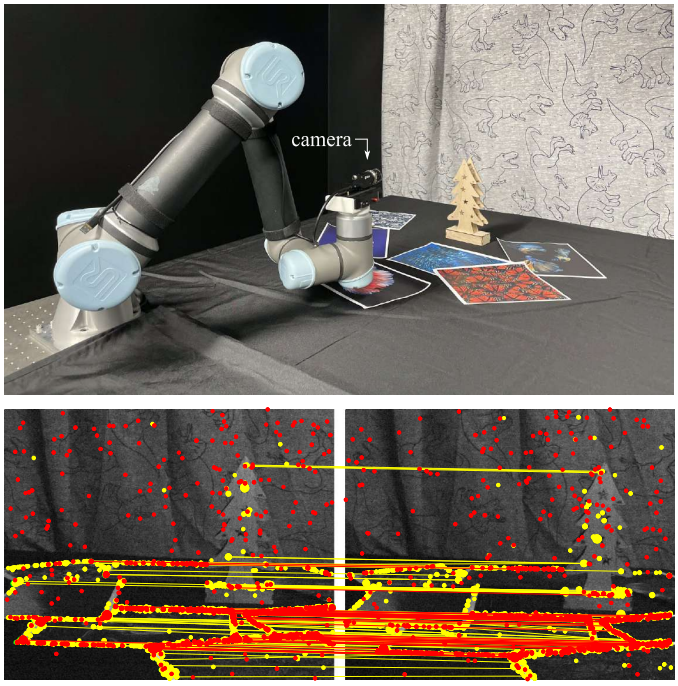


Fig. 1. Visualisation of light-constrained scene. (top) Robot arm-mounted machine vision camera; (bottom) yellow: 110 matching pairs using our proposed merging approach; red: 43 matching pairs between consecutive conventional noisy images which fails to converge for reconstruction. It is important to note that 36% of input images of this particular scene fails to find any matches in conventional noisy approach where as all input images are registered using our method for reconstruction. Our approach provides more stable set of true features and fewer spurious features for motion dependent applications such as SfM

We propose to reconstruct light-constrained scenes by adapting burst imaging for reconstruction. We describe an imaging pipeline that captures multiple bursts of images and accounts for motion variation within each burst using hierarchical tile-based alignment. We temporally merge each aligned burst to a single image using a voting scheme and then use these merged images in feature-based SfM to reconstruct scenes (see Fig. 1). We also consider adding spatial filtering separately on temporally denoised images to evaluate hybrid denoising.

Our key contributions are:

- We establish the viability of using burst imaging to improve robotic vision in low light, and provide a set of recommendations for adopting this approach in recon-

Manuscript received: August, 12, 2021; Revised October, 05, 2021; Accepted December, 03, 2021.

This paper was recommended for publication by Editor Cesar C. Lerma upon evaluation of the Associate Editor and Reviewers' comments.

The authors are with the Australian Centre for Field Robotics (ACFR), School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney and with the Sydney Institute for Robotics and Intelligent Systems, 2006 NSW, Australia. ahalya.ravendran, mitch.bryson, donald.dansereau@sydney.edu.au

Digital Object Identifier (DOI): see top of this page.

struction tasks such as SfM;
- We evaluate different approaches to burst imaging in robotics applications and show that burst capture with merge offers significant advantages in both computational requirements and performance; we offer a theoretical explanation and experimental evidence showing why this is so; and
- We demonstrate the proposed method improving low-light SfM by yielding more precise 3D points, more true features, fewer spurious features, more precise camera trajectories and an ability to operate in lower light than was previously possible.

Our approach assumes that the level of noise present in images is not so great as to make information unrecoverable (e.g. the underlying signal being suppressed by quantization noise). Additionally, we assume that there exists sufficient overlap between subsequent sets of images bursts for registration, which is commonly available for mobile platforms moving at moderate speeds relative to their environment.

To validate our method, we mounted a monocular machine vision camera on a UR5e robotic arm, and captured burst imagery of real scenes in an illumination-controlled environment over different exposure times. The images are captured over 20 scenes with objects of various size, shape and colour as 22 bursts for each scene, yielding 6160 images in 880 bursts. We are releasing a dataset containing raw Bayer images, captured over the exposure time of 1ms and 0.1ms for a repeated trajectory. Code and dataset are available at https://roboticimaging.org/Projects/BurstSfM/.

To evaluate, we compare our results against existing alternative approaches depicted in Fig. 2: burst without merge which includes every image in each burst, and conventional single-image capture. We demonstrate our proposed method outperforms the alternative approaches by reconstructing more true features with high feature localization accuracy and fewer spurious features as shown in Fig. 1. We also show our method yields accurate estimation of camera pose trajectory while minimising the failure cases of convergence compared to the existing alternative approaches.

This work opens the way for a broad range of applications in which low light commonly complicates vision such as autonomous driving and delivery drones.

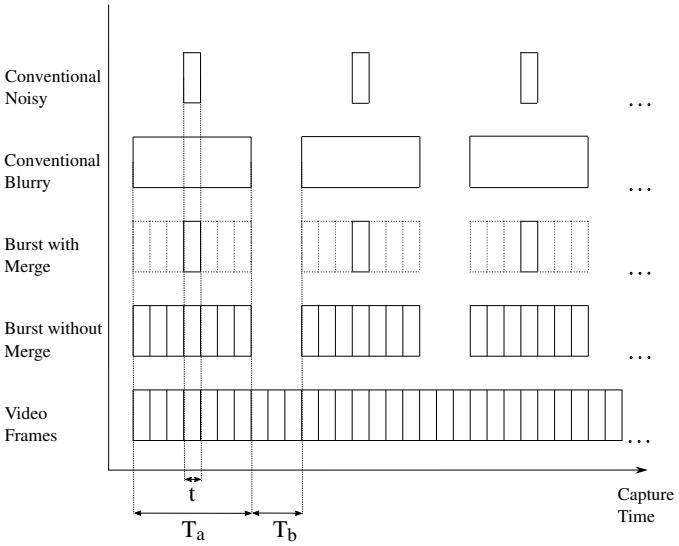


Fig. 2. Alternative exposure schemes (top to bottom) – Conventional noisy: Capturing and processing individual frames over an extended trajectory is a common approach in robotics, and results in noisy imagery in low-light conditions. Conventional blurry: increasing the exposure duration for each frame gathers more light at the expense of increased motion blur and loss of information. Burst with merge: the proposed approach captures bursts of frames then aligns and merges them to one frame per burst, improving SNR without losing information to blur, and decreasing overall computational load. Burst without merge: skipping the merge step directly ingests all burst images into the SfM pipeline; we show this to be less effective than merging in terms of both performance and computational expense. Video: finally, video captures all possible frames, at a high computational expense and with poor low-light performance. T_a is duration of an N-image burst, t is the duration of a single frame, and T_b is the delay between bursts.

II. Related Work

When optimising camera exposure settings for low SNR images, there exist many trade-offs in the final image quality. Conventional cameras can increase SNR by either widening the aperture which reduces depth of field or increasing the exposure time, which for dynamic scenes, increases motion blur.

Considering alternative exposure schemes, depicted in Fig. 2, capturing images as a sequence over a long period of time yields a video. However, it is computationally intensive to process images at full video rates and in low light, each of these frames have low SNR.

A number of existing computational imaging approaches successfully break discussed imaging trade-offs and take advantages of the alternative exposure schemes as demonstrated in coded aperture [7], flexible depth of field photography [8], flutter shutter [9] and motion-invariant photography [10]. Unfortunately they require hardware modification to imaging sensors/cameras whereas our focus is on using existing monocular cameras to get quality images in low light for robotic vision.

Learning-based techniques show promising results for denoising [11]–[13], deblurring [14]–[16], dehazing [17]–[19] and image enhancement [20]–[22]. However learning-based approaches make no guarantee of generalising beyond their training domains [23]–[25] and compared with the proposed non-learning-based approach, these require collection of appropriately scaled datasets and corresponding training times. Learning-based approaches also present limitations around interpretability and transparency [26], [27].

Single-image based approaches are fundamentally limited by the amount of information in a conventional single image. They also have a tendency to produce visually pleasing results and not necessarily accurate information through learned image priors [28]. This affects their reliability for applications such as reconstruction and tracking.

Hasinoff et al. [29] exploit the time slice advantage and use burst imaging as a solution for night photography by capturing multiple underexposed images and merging them to get a single higher SNR image. This is implemented as night sight in Google Pixel mobile phones [4], astrophotography mode [5], multi-frame super resolution [6] and using learning approaches [30]. It has also been adapted as quanta burst photography [31] for single-photon sensors to produce high

quality images. This and other previous work in mobile photography has focused mainly on producing visually pleasing still images from bursts with little camera motion. In this work we adapt burst imaging for 3D reconstruction with extended camera motion in low-light scenes.

Our work also considers more challenging imagery than what is typically employed in existing burst photography work [4], [5]. Photography applications must produce imagery that is pleasing for human consumption, meaning relatively high SNR is required. Robotics applications, on the other hand, can tolerate lower-quality imagery so long as it is informative. Another key difference to photography is that robotics typically covers extended areas over long trajectories, meaning multiple bursts of images become necessary.

In this work, we adapt burst imaging for 3D reconstruction in low light. We do this by combining feature-based SfM and burst photography, exploiting the advantages of each. By capturing rapid successions of frames, we enable the use of direct methods for image registration [32], exploiting the small camera motion between frames to yield a strong SNR advantage with moderate computational expense. After extracting features from the merged burst images, we apply feature-based SfM [1] to handle large camera motions between bursts. Feature-based methods handle large camera motions [33] while benefiting from the improved feature quality associated with the burst-merged images.

Because our method compresses each burst into a single denoised image, total computation is lower during reconstruction compared with an approach employing all measured frames. Furthermore, the higher SNR in our imagery can reduce the numbers of spurious features, again lowering computational requirements during reconstruction as there are fewer outliers to detect and reject.

III. Burst-Based SfM

Here we review salient aspects of burst photography [4], and explain its integration into an SfM pipeline. As in [4], we capture multiple images, establish dense correspondence between them, then merge the aligned stack with a temporal voting scheme. Taking this direct approach within the burst exploits the temporal coherence between frames, manifesting as relatively little illumination variation and occlusion/disocclusion within a burst. We then extract features from each merged burst to deal with appearance changes associated with the larger timescale and translation between bursts. In this work we further introduce and evaluate spatial Wiener and bilateral filtering.

The following outlines image acquisition, alignment, merging, and integration into a reconstruction pipeline. Further practical considerations are discussed in Sec. V.

A. Image Acquisition for Reconstruction

As depicted in Fig. 2, we capture multiple bursts over a trajectory, with each burst containing N frames and each frame having exposure time t, taking time T_a per burst with a delay T_b between bursts.

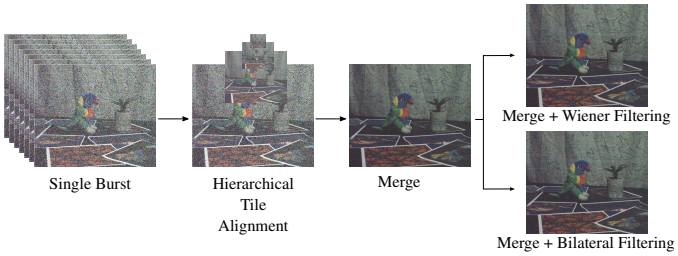


Fig. 3. We improve the noise in the captured images by aligning every image in a burst with the chosen reference image of the burst. We merge the aligned images in temporal direction with a voting scheme to avoid misalignment. We use Wiener and bilateral filtering on temporally merged images. We give the outputs of this pipeline to COLMAP as three different inputs: burst with merge, burst with merge and Wiener filtering and burst with merge and bilateral filtering for reconstruction.

Capturing a greater number of images per burst increases the resulting SNR in merged images, provided there is a sufficient image overlap. However, more images are also more computationally expensive for capturing, buffering and processing, so there is a trade-off in selecting the number of images in a burst between quality and computation. This is application-dependent: in our experiments, we use bursts of N=7 frames based on empirical evaluation. – see further discussion in Sec. V.

B. Hierarchical Tile Alignment

To address non-uniform apparent motion between the frames in a burst, we perform coarse-to-fine alignment on multi-level Gaussian pyramids of single-channel images, as in [4]. We use a patch of pixels as a tile and using the initial estimates from the coarser scale, we compute pairwise-tile-based alignment at each pyramid level by minimizing the distance measure between the common tile and the corresponding candidate tile of each alternate frame within a burst [34].

We empirically tune the parameters of the alignment process to strike a balance between processing time and tolerance to motion between the frames. Practical guidance for tuning alignment parameters is provided in Sec. V.

C. Robust Temporal Merge

We merge aligned images following the pairwise method presented in [4]. To increase robustness to motion, this employs temporal filtering with a per-image contribution in the frequency domain:

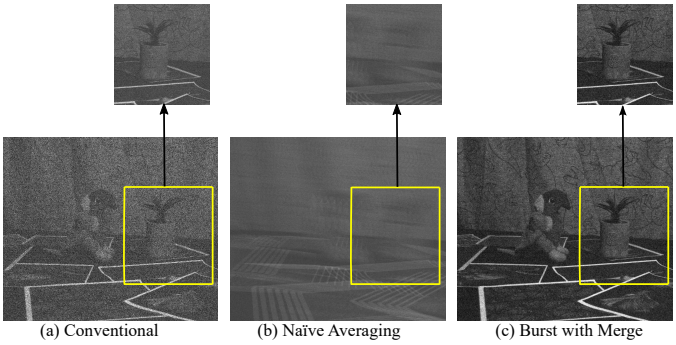
$$\overline{T_R}(\omega) = \frac{1}{N}\sum_{z=1}^{N} T_z(\omega) + A_z(\omega)[T_R(\omega) - T_z(\omega)], \qquad (1)$$

where N is the number of frames within a burst, $T_R(\omega)$ is reference frame, $T_z(\omega)$ is frame to be added to the estimate, and $\overline{T_R}(\omega)$ is the updated estimate.

A_z is the contribution of each frame, found as

$$A_z(\omega) = \frac{|D_z(\omega)|^2}{|D_z(\omega)|^2 + c\sigma^2}, \qquad (2)$$

where $D_z(\omega)$ = $T_R(\omega)$ - $T_z(\omega)$, σ^2 is the noise variance and c is the degree of contribution that increases noise reduction



(a) Conventional (b) Naïve Averaging (c) Burst with Merge

Fig. 4. Application of alternative exposure schemes: (a) Single frame captured over an exposure time of t ms. (b) Multiple frames captured over T_a ms and merged naïvely – this is representative of a long exposure scheme with motion blur. (c) The proposed merging approach yields a clearer image without blur.

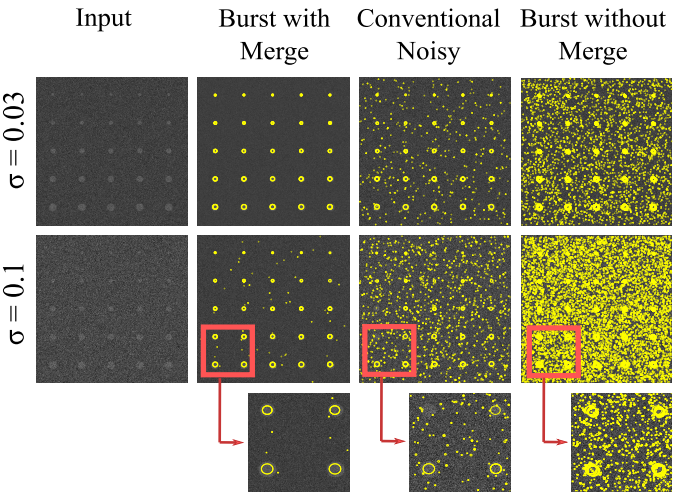


Fig. 5. Detected features on synthetic images using SIFT, presented at two noise levels σ. At the lower noise level (top row), our method performs well with no spurious features, while conventional noisy method and burst without merge generates more spurious features. At higher noise (bottom row), while our method detects all true features, conventional noisy approach detects more spurious features and fewer true features than ours and burst without merge registers overwhelming spurious features.

at the expense of robustness to misalignment. Burst imaging does not degrade image quality for fast motion but the quality gain decreases as overlap between images decreases.

Outliers are rejected because these present a large difference D_z between the reference and alternative frames, placing a higher weight on the reference frame. Ringing artifacts that are commonly associated with frequency-domain filters are avoided with a raised cosine windowing approach – see [4] for further detail.

As depicted in Fig. 3, we also evaluate addition of a noise shaping Wiener filter and an edge-preserving bilateral filter for further noise reduction.

Fig. 4 illustrates typical performance of the align-and-merge approach, compared with conventional capture which yields noisy imagery, and naïve averaging which is sensitive to motion. Burst with merge shows much higher SNR, ideally improving SNR, the ratio between the power of signal and noise by $\sqrt{N}$ for N-image bursts [35].

D. Reconstruction Pipeline

We use COLMAP [1], an end-to-end feature-based state-of-the-art reconstruction pipeline to extract, match and geometrically verify features between sparse camera poses. We demonstrate sparse reconstruction by triangulating scene points and refining via bundle adjustment. We evaluate performance following reconstruction and feature performance metrics laid out in [36], [37].

IV. Results

In the following we first evaluate improvements in feature detection and accuracy afforded by our method in noisy imagery using synthetic image sequences. By employing synthetic scenes we offer greater control over noise and scene content similar to prior quantitative feature evaluations [37], [38]. Then we quantitatively evaluate our method in an SfM pipeline, comparing against conventional image capture and direct use of all images in the burst. We consider both 3D reconstruction performance and camera trajectory accuracy.

A. Feature Performance in Noise

We generate a synthetic set of images with known feature locations to demonstrate feature performance in different noise levels. We use synthetic noise which correspond closely to the noise levels of the captured imagery appearing in the following sections.

The synthetic images, shown in Fig. 5, have 25 disks at varying scales. Every image has an apparent motion of 12 or fewer pixels between each other. The contrast between the disks and background is 0.1. We introduce moderate noise with variance 0.03 (top) and strong noise with variance 0.1 (bottom).

We extract scale invariant feature transform (SIFT) [39] features at a peak threshold of 0.015. It is evident from Fig. 5 that burst with merge performs better than burst without merge and conventional noisy with no spurious features for the moderate noise case (σ = 0.03). While our method extracts all true features in strong noise, the conventional method fails to extract all true features and burst without merge is overwhelmed by spurious features.

We demonstrate the measurement of true positive (TP) rate and false positive (FP) count quantitatively on these synthetic images with varying noise levels and peak detection thresholds as shown in Fig. 6 for both the proposed and conventional methods. The top row shows the TP rate and FP count at peak detection thresholds of 0.006 and 0.01 for a range of noise levels.

An appropriate peak threshold gives negligible FP count and higher TP rate at lower noise level. As the amount of noise increases, our method outperforms the conventional method yielding more true positive features and fewer false positives.

The bottom row of Fig. 6 depicts the performance of both methods in two different noise levels for peak detection thresholds between 0.006 and 0.03. Our method results in fewer false

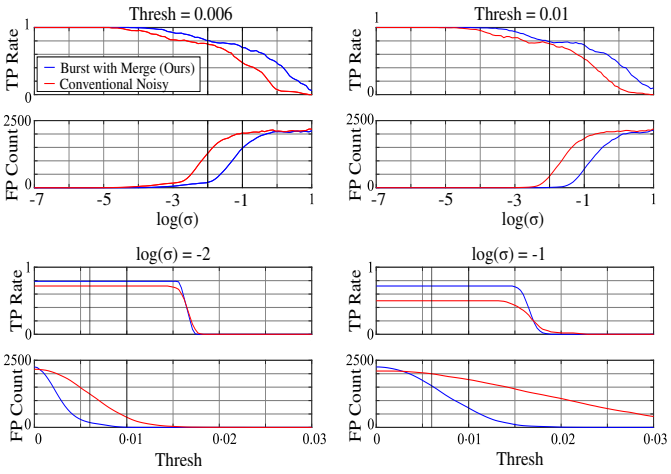


Fig. 6. Noise performance: (top row) Sweeping noise level σ for fixed detection thresholds, our method (blue) shows a higher true positive (TP) rate and lower false positive (FP) count than conventional imaging (red). (bottom row) Sweeping detection threshold, our method delivers a much higher TP rate in high noise and lower FP count for appropriately set threshold than conventional imaging. Overall, our method matches or outperforms conventional method in feature performance in noise.

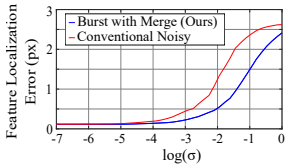


Fig. 7. Feature localization accuracy: Sweeping noise level σ for fixed detection threshold of 0.015, our method (blue) outperforms conventional imaging (red) by yielding accurate feature localization.

positives and more true features. The lower false positive count results in lower overall computational requirements, as fewer putative matches need to be evaluated and rejected.

For a selected peak detection threshold of 0.015, we also evaluated the localization accuracy of the extracted true positive features compared to the conventional approach as depicted in Fig. 7. As noise level increases, our method outperforms the conventional approach by yielding significantly lower feature localisation errors.

B. Reconstruction Accuracy

We demonstrate our method by mounting a monocular machine vision camera FLIR FL3-U3-120S3C-C with an f/2.1 lens on a UR5e robotic arm as shown in Fig. 1. We capture 16 bit raw Bayer images of size 2992 x 2500. We capture 22 bursts of 7 images each for a single trajectory, and repeat the trajectory for 20 scenes composed of objects with different textures, shapes and sizes in an environment with controlled lighting as shown in Fig. 8.

We capture our dataset at 1 ms and 0.1 ms exposure times, adjusting the gain in each case to maximize contrast while avoiding saturation. Apparent motion within each burst is typically about 1/8 of the camera's field of view, with examples of faster motion up to 1/4 of the field of view.

We run COLMAP to generate sparse reconstructions of 20 different light-constrained scenes using our methods: burst

Fig. 8. Examples of captured imagery, showing a diversity of objects, textures, shapes and sizes. The full dataset contains 20 scenes, each with 22 bursts of 7 images.

with merge, burst with merge and Wiener filtering and burst with merge and bilateral filtering. We compare against alternative approaches: burst without merge where we directly use all images in the burst and conventional single-image capture, with the single image corresponding to the centre most frame of the burst, as depicted in Fig. 2.

We build a 4-level coarse to fine hierarchical pyramid with tile size 8 for alignment and use overlapping tiles of same size for pairwise temporal merge throughout the experiments. Further guidelines on selecting appropriate tile size and pyramid levels for robust merging are discussed in Sec. V.

We employ COLMAP with default settings, meaning a constant feature peak threshold of 0.0066 is employed for all images. To more fairly compare methods, we also repeat the experiment with peak thresholds adjusted to suit the different noise levels yielded by each method. Because unmerged images have more noise, they call for a more selective peak threshold. We empirically select peak thresholds of 0.001 and 0.005 for the proposed and conventional methods, respectively, as these yield similar levels of spurious feature detection.

Finally, we repeated the same experiments with COLMAP configured to be more permissive of images with few inlier feature matches, allowing as few as 15. This is useful when dealing with challenging scenes and allowed more images to be successfully incorporated in to the SfM solution.

Following the feature comparison approach in [36], we evaluate reconstruction performance as shown in Table I in terms of numbers of keypoints per image, putative feature matches per image, 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, the mean number of 3D points in the reconstructed models and the mean number of 3D points per image on captured imagery.

In the table, bold highlights the best results, red shows best and second-best results from competing approaches, while green shows best and second-best results from the proposed method.

At a moderate noise level, i.e. for images captured over 1ms, our method shown in green outperforms alternative approaches across all metrics, by reconstructing all images passed for all scenes at default settings. Our method reconstructs hundreds more matches and twice as many 3D points per image compared with alternative approaches.

At a higher noise level, i.e: for images captured over 0.1ms, not all images are reconstructed with default settings, and thus, we evaluate by tuning peak threshold values. The proposed method successfully reconstructed all the scenes with the strongest putative matches per image, inlier matches per image, match ratio, match score and 3D points per image.

On an Intel i7-9700 at 4.70 GHz, our MATLAB burst imaging implementation takes 6.54 seconds to align and merge a monochrome burst of 7 images. We expect this could be accelerated substantially. Following align and merge, we employ COLMAP for reconstruction, extracting features using an NVIDIA GTX770. The complete reconstruction pipeline, including time for align and merge, was fastest using the proposed approach. For a dataset of 154 images in 22 bursts, align, merge, and SfM reconstruction took 3.25 minutes. By contrast, operating on 22 conventional images with no align or merge took 6.32 minutes, and operating on all 154 unmerged images took 49.58 minutes. Our method produces fewer spurious features, resulting in faster overall processing times.

C. Camera Trajectory Accuracy

We evaluate the accuracy of the camera trajectory estimate by using a robotic arm to collect ground truth poses. We aligned the camera poses to the ground truth poses as there is an arbitrary scale factor involved in monocular SfM. The arbitrary scale we use when reporting results is determined by the distance between the first pair of registered images. Fig. 9 shows how our method reconstruct more accurate camera poses than the competing method.

We compute absolute instantaneous error and relative pose error between reconstructed camera poses and ground truth poses for translation and rotation as shown in Table II. The color scheme matches that used in Table I. The green bold values show our method is competitive with moderate noise and outperforms alternative approaches with strong noise by 1cm to 4cm in translation, across all error metrics.

V. Discussion

We have shown that burst imaging can improve the performance of light-constrained SfM while yielding a net reduction in computational requirements.

Although our approach involves increased computation in merging images within a burst, this is offset by the decreased rate of false positive feature detection, reducing the burden of extraneous feature matching and rejection.

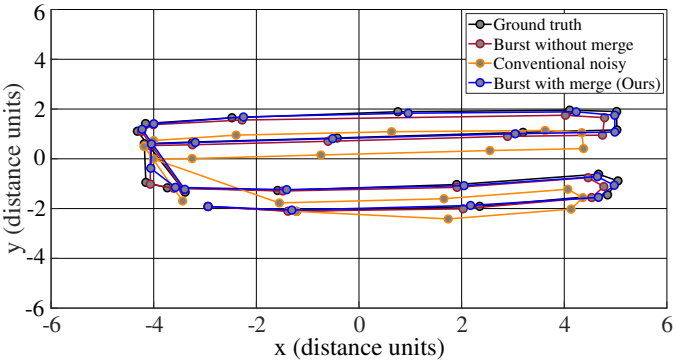


Fig. 9. Camera trajectory for a particular captured scene (scene 16) in distance units, a measure between the first pair of registered images in a reconstruction. Our results reconstruct accurate camera trajectories using all input images; alternative approach that uses reconstructed 94.2% of input noisy images without merge and conventional noisy approach that uses 95.5% produces less accurate trajectory estimates.

The burst approach exhibits multiple trade-offs in acquisition strategy, align and merge parameters, platform motion, and scene content. In the following we offer practical advice for tuning burst imaging for specific robotics applications, based on the experimental findings of this work and on established results from prior works.

- Increase exposure time as high as possible without getting deleterious motion blur [6]. This yields higher-SNR input images and ultimately better reconstruction performance.
- Fixed-pattern noise can easily become the dominant source of noise in capturing low light images. This can be addressed by subtracting an average of multiple dark frames taken with the same gain, sensor temperature, and exposure time as the intended images within a burst [40].
- Increase gain as high as possible without yielding excessive saturation [41]. This amplifies both signal and noise, but is important in overcoming the quantisation limit of the camera.
- Maximize the number of images in each burst to maximise SNR in the merged images. The image count is ultimately limited by availability of compute, and the rate of apparent motion of the scene [3]. Motion relative to the reference frame should not exceed 1/2 of the total frame. For this reason we also recommend limiting apparent motion relative to the reference frame by employing the centre most pose as the reference.
- The image pyramids balance computational effort and quality of alignment via the search window size at each pyramid level. Increase pyramid level count and decrease search window size to avoid local minima – this is especially important for images with strong noise [34]. In our experiments, 3 or more pyramid levels maintaining an image size of at least 1/16 of the original image yielded robust alignment.
- For fast motion within the burst, e.g. 1/4 of the total frame or more, increase the search extent to allow for the larger motion. Reduce the image size at each pyramid level – we found it best to go not lower than 1/16 of the original image size. See [4], [6] for more discussion on motion

TABLE I

AVERAGE PERFORMANCE OF RECONSTRUCTION FOR LIGHT-CONSTRAINED SCENES. SFM: WITH COLMAP'S DEFAULT VALUES, OUR APPROACH OUTPERFORMS CAPTURED IMAGES IN ALL MEASURES, INCLUDING SUCCESSFUL RECONSTRUCTION OF SIGNIFICANTLY MORE REGISTERED IMAGES PER INPUT SCENES, MORE 3D POINTS PER IMAGE, MORE PUTATIVE MATCHES AND INLIER MATCHES PER MATCH PER IMAGE. BY TUNING COLMAP PARAMETERS, OUR APPROACH PERFORM SUCCESSFUL RECONSTRUCTION AT EXTREME LOW LIGHT. THE RESULTS ARE EVALUATED OVER 20 SCENES. BOLD: BEST RESULTS, RED: BEST/SECOND BEST RESULTS FROM CONVENTIONAL APPROACHES; GREEN: BEST/SECOND BEST RESULTS FROM OUR PROPOSED APPROACHES

Reconstruction Modes	Method	No. of images	No. of images pass	% images pass	Key Points/ Image	Putative Matches/ Image	Inlier Matches/ Image	Match Ratio	Precision	Match Score	3D Points	3D Points/ Image
1ms Default	Burst without merge	154	152	98.70	9116	681	629	7.50E-02	0.92	6.90E-02	39790	258
	Conventional noisy	22	21.6	98.18	9106	656	603	7.20E-02	0.92	6.60E-02	8270	376
	Burst with merge	22	22	**100**	10522	866	801	8.20E-02	**0.93**	7.60E-02	11130	506
	Burst with merge + Wiener	22	22	**100**	10300	907	842	8.80E-02	**0.93**	8.20E-02	11387	518
	Burst with merge + bilateral	22	22	**100**	10202	**936**	**868**	**9.20E-02**	**0.93**	**8.50E-02**	11693	**532**
0.1ms Default	Burst without merge	154	89.85	58.34	6187	37	29	5.90E-03	**0.78**	4.60E-03	3456	22
	Conventional noisy	22	13.5	61.36	6187	34	25	5.50E-03	0.74	4.00E-03	382	17
	Burst with merge	22	16.55	**75**	9344	45	33	4.80E-03	0.73	3.50E-03	574	**26**
	Burst with merge + Wiener	22	15	68	10978	47	34	4.20E-03	0.72	3.00E-03	571	**26**
	Burst with merge + bilateral	22	15.8	71.8	7368	**50**	**36**	**6.70E-03**	0.72	**4.90E-03**	566	**26**
0.1ms Permissive Matching	Burst without merge	154	76.9	49.94	6187	37	29	5.90E-03	**0.78**	4.60E-03	2623	17
	Conventional noisy	22	16.85	76.59	6187	34	25	5.50E-03	0.74	4.00E-03	422	19
	Burst with merge	22	21	95.45	9344	45	33	4.80E-03	0.73	3.50E-03	718	34
	Burst with merge + Wiener	22	21.05	95.68	10978	47	34	4.20E-03	0.72	3.00E-03	781	**36**
	Burst with merge + bilateral	22	21.9	**99.55**	7368	**50**	**36**	**6.70E-03**	0.72	**4.90E-03**	780	35
0.1ms Peak Threshold Tuned	Burst without merge	154	85.95	55.81	5431	43	33	7.90E-03	**0.77**	6.10E-03	3826	25
	Conventional noisy	22	16.7	75.9	5431	39	30	7.00E-03	**0.77**	5.50E-03	536	24
	Burst with merge	22	20.9	**95**	5367	50	36	**9.30E-03**	0.72	**6.70E-03**	852	**39**
	Burst with merge + Wiener	22	18.45	83	5977	51	37	8.50E-03	0.73	6.20E-03	792	36
	Burst with merge + bilateral	22	18.7	85	7572	**53**	**38**	6.90E-03	0.72	5.00E-03	731	33
0.1ms Peak Threshold Tuned + Permissive Mapping	Burst without merge	154	50.7	32.92	5431	43	33	7.90E-03	**0.77**	6.10E-03	2152	14
	Conventional noisy	22	19.3	87.73	5431	39	30	7.00E-03	**0.77**	5.50E-03	576	26
	Burst with merge	22	22	**100**	5367	50	36	**9.30E-03**	0.72	**6.70E-03**	864	39
	Burst with merge + Wiener	22	21.9	99.55	5977	51	37	8.50E-03	0.73	6.20E-03	848	39
	Burst with merge + bilateral	22	22	**100**	7572	**53**	**38**	6.90E-03	0.72	5.00E-03	886	**40**

TABLE II

MEAN TRANSLATION ERROR AND MEAN ROTATIONAL ERROR IN CAMERA POSES FOR ALL RECONSTRUCTION METHODS: OUR APPROACH OUTPERFORMS CONVENTIONAL NOISY APPROACH WITH SAME NUMBER OF IMAGES AND PERFORMS COMPETITIVE WITH BURST WITHOUT MERGE WHICH HAS SEVEN TIMES MORE IMAGES THAN OUR APPROACH. BOLD: BEST RESULTS, RED: BEST/SECOND BEST RESULTS FROM CONVENTIONAL APPROACHES; GREEN: BEST/SECOND BEST RESULTS FROM OUR PROPOSED APPROACHES

Reconstruction modes		Absolute Instantaneous Error					Relative Pose Error				
		Default (1ms)	Default (0.1ms)	Perm. Mapping	Peak Thresh.	Peak Thresh.+ Perm. Mapping	Default (1ms)	Default (0.1ms)	Perm. Mapping	Peak Thresh.	Peak Thresh.+ Perm. Mapping
Burst without Merge	trans. (cm)	**0.73**	2.83	2.77	2.46	2.75	2.70	4.28	4.99	3.36	4.66
	rot. (deg)	0.459	0.715	**0.499**	0.771	0.487	**0.004**	0.052	0.035	0.021	0.039
Conventional Noisy	trans. (cm)	0.82	4.51	5.24	5.43	2.92	2.25	3.9	7.15	5.66	4.71
	rot. (deg)	0.505	0.775	1.040	1.155	0.615	0.005	0.032	0.126	0.044	0.091
Burst with Merge	trans. (cm)	0.74	**2.42**	2.25	0.98	2.38	**1.99**	**3.80**	**3.16**	**1.99**	**2.62**
	rot. (deg)	0.452	**0.630**	0.662	**0.341**	0.451	**0.004**	0.030	**0.016**	**0.013**	0.018
Burst with Merge + Wiener	trans. (cm)	0.79	2.82	**2.10**	**0.89**	2.70	2.00	4.27	3.22	2.29	3.96
	rot. (deg)	**0.446**	0.646	0.512	0.421	0.473	**0.004**	**0.015**	**0.016**	**0.013**	0.021
Burst with Merge + Bilateral	trans. (cm)	0.79	2.74	2.81	2.04	**1.49**	2.64	3.87	4.24	2.53	2.71
	rot. (deg).	0.451	0.757	0.645	0.530	**0.440**	**0.004**	0.028	0.028	0.021	**0.017**

robustness with burst imaging.

VI. CONCLUSIONS

We adapted burst photography, commonly used in mobile photography for reconstruction in low light. We enabled the use of direct methods for image registration within the burst and used feature-based SfM to handle the sparsity between bursts for reconstruction. We demonstrated successful reconstruction with decreased failure cases due to non-convergence compared to conventional methods. We also demonstrated improved performance relative to conventional imaging in true features, spurious features, putative matches, inlier matches, 3D points per images and accurate camera pose estimation.

Our method showed faster reconstruction with lower overall computational requirements compared to conventional methods. We expect that in more challenging low light conditions our method can improve the performance of 3D reconstruction and expand the range in which feature-based reconstruction can be applied.

This work is a first step toward solving the problem of 3D reconstruction in low light. For future work we anticipate employing adaptive sampling schemes, in which the parameters of burst capture and processing are dynamically chosen to suit the situation. We also expect the fusion of complementary sensors to yield interesting results.

REFERENCES

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

[2] X. Han, H. Laga, and M. Bennamoun, "Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 5, pp. 1578–1604, 2021.

[3] Z. Liu, L. Yuan, X. Tang, M. Uyttendaele, and J. Sun, "Fast burst images denoising," *ACM Trans. on Graphics (TOG)*, vol. 33, no. 6, pp. 1–9, 2014.

[4] 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 Trans. on Graphics (TOG)*, vol. 35, no. 6, pp. 1–12, 2016.

[5] 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 Trans. on Graphics (TOG)*, vol. 38, no. 6, pp. 1–16, 2019.

[6] 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 Trans. on Graphics (TOG)*, vol. 38, no. 4, pp. 1–18, 2019.

[7] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Trans. on Graphics (TOG)*, vol. 26, no. 3, pp. 70–es, 2007.

[8] H. Nagahara, S. Kuthirummal, C. Zhou, and S. K. Nayar, "Flexible depth of field photography," in *European Conference on Computer Vision (ECCV)*, 2008, pp. 60–73.

[9] R. Raskar, A. Agrawal, and J. Tumblin, "Coded exposure photography: Motion deblurring using fluttered shutter," *ACM Trans. on Graphics (TOG)*, vol. 25, no. 3, p. 795–804, 2006.

[10] A. Levin, P. Sand, T. S. Cho, F. Durand, and W. T. Freeman, "Motion-invariant photography," *ACM Trans. on Graphics (TOG)*, vol. 27, no. 3, pp. 1–9, 2008.

[11] Y. Quan, M. Chen, T. Pang, and H. Ji, "Self2Self with dropout: Learning self-supervised denoising from single image," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1890–1898.

[12] M. Tassano, J. Delon, and T. Veit, "FastDVDnet: Towards real-time deep video denoising without flow estimation," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1354–1363.

[13] H. Yue, C. Cao, L. Liao, R. Chu, and J. Yang, "Supervised raw video denoising with a benchmark dataset on dynamic scenes," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2301–2310.

[14] Y. Nan, Y. Quan, and H. Ji, "Variational-EM-based deep learning for noise-blind image deblurring," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3623–3632.

[15] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3043–3051.

[16] Y. Yuan, W. Su, and D. Ma, "Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3555–3564.

[17] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *Int. Conference on Computer Vision (ICCV)*, 2020, pp. 2157–2167.

[18] Y. Li, Q. Miao, W. Ouyang, Z. Ma, H. Fang, C. Dong, and Y. Quan, "LAP-Net: Level-aware progressive network for image dehazing," in *Int. Conference on Computer Vision (ICCV)*, 2019, pp. 3276–3285.

[19] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Int. Conference on Computer Vision (ICCV)*, 2019, pp. 7314–7323.

[20] Y. S. Chen, Y. C. Wang, M. H. Kao, and Y. Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6306–6314.

[21] V. Sharma, A. Diba, D. Neven, M. S. Brown, L. Van Gool, and R. Stiefelhagen, "Classification-driven dynamic image enhancement," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4033–4041.

[22] R. Wang, Q. Zhang, C. W. Fu, X. Shen, W. S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6849–6857.

[23] 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.

[24] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning; Lessons we've learned," *Int. Journal of Robotics Research (IJRR)*, p. 0278364920987859, 2021.

[25] 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.

[26] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[27] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[28] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Deep burst super-resolution," in *Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9209–9218.

[29] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 553–560.

[30] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2502–2510.

[31] S. Ma, S. Gupta, A. C. Ulku, C. Bruschini, E. Charbon, and M. Gupta, "Quanta burst photography," *ACM Trans. on Graphics (TOG)*, vol. 39, no. 4, 2020.

[32] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 834–849.

[33] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. on Robotics (T-RO)*, vol. 31, no. 5, pp. 1147–1163, 2015.

[34] R. Szeliski and J. Coughlan, "Spline-based image registration," *Int. Journal of Computer Vision (IJCV)*, vol. 22, no. 3, pp. 199–218, 1997.

[35] S. W. Hasinoff, "Photon, Poisson noise," 2014.

[36] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative evaluation of hand-crafted and learned local features," in *Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1482–1491.

[37] 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.

[38] M. Rad, M. Oberweger, and V. Lepetit, "Feature mapping for learning fast and accurate 3D pose inference from synthetic images," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4663–4672.

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

[40] J. R. Janesick, *Scientific charge-coupled devices*. SPIE press, 2001, vol. 83.

[41] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation?" in *Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1163–1170.