High-ambient, Super-resolution Depth Imaging with a SPAD Imager via Frame Re-alignment

Germán Mora-Martín[1], Abderrahim Halimi[2], Robert K. Henderson[1],
Jonathan Leach[2], Istvan Gyongy[1]
[1]The University of Edinburgh, Institute for Integrated Micro and Nano Systems, Edinburgh, U.K.
[2]Heriot-Watt University, Institute of Photonics and Quantum Sciences, Edinburgh, U.K.
german.mora@ed.ac.uk

***Abstract*—We enhance the lateral resolution of direct time-of-flight (dToF) depth images by combining multiple frames subject to camera motion. Experimental results, based on synthetic as well as real dToF data, are presented to illustrate the effectiveness of the approach.**

I. INTRODUCTION

Direct time-of-flight (dToF) sensors in image sensor format avoid the need for optical scanning, produce camera-type intensity data to accompany depth, and promise increased robustness to environmental conditions such as rain [1]. However, array sizes tend to be relatively small due to the large pixels in such architectures, leading to low lateral resolutions when using flood illumination. Whilst indirect time-of-flight sensors offer larger arrays, multi-path interference problems can arise and the range is typically lower [2].

We here present a method that trades frame rate for increased lateral resolution in a dToF device. It is inspired by schemes already used for RGB sensors in smartphones and cameras, which exploit the user's natural hand shake [3], or a mechanism in the sensor, to capture sub-exposures which are spatially shifted with respect to one another. These sub-exposures are then re-aligned and merged to produce a super-resolution image. The idea here is to take this processing a step further, and use the sensor motion, as estimated from RGB or monochrome intensity data, to re-align and combine dToF depth frames for increased resolution in x,y.

II. SUPER-RESOLUTION ALGORITHM

To demonstrate the approach, we use an example from the Middlebury dataset [4] and run the algorithm for different signal-to-background ratios SBR (ratio of signal photons to background photons) and different levels of random lateral shifts and rotations (representing different amounts of camera shake) to increase the lateral resolution by a factor of 4. To create the input frames, we mimic a SPAD operated in a hybrid imaging modality [5] (Fig. 1a), creating synthetic 256×256 intensity, and 64×64 depth data, in a time-interleaved manner. A total of 50 sub-exposures (50 intensity, 50 depth) are generated. We then estimate the geometric transformation between consecutive intensity frames, and apply the inverse of this transformation to the depth frames (upscaled to 256×256) to align the depth data. The second step in the processing (Fig. 1b), considers the set of depth values, across all the realigned depth frames, corresponding to a given pixel position, and takes a local mean within the dominant cluster of values. As a final step, a dilation operation is performed to compensate for the coarse lateral sampling in depth mode, and sharpen the features of the final depth image (Fig. 1c).

Fig. 2 shows the resulting super-resolved depth images for different SBR and shake levels and includes the 64×64 input depth map and the 256×256 reference depth map. The root mean square error (RMSE) is used as a metric to compare the similarity between the ground truth and the output depth maps. Low SBR depth maps are extremely noisy since most of the ToF peaks are buried in the temporal histogram. However, the super-resolved version shows reduced noise and recovers the main features present in the image. At higher SBR, depth maps become less noisy, providing a better reconstruction of the scene, as indicated by the lower RMSE values. When no lateral shifts or rotations are present between sub-exposures, the super-resolved image is equivalent to a nearest-neighbours interpolation, in other words the algorithm becomes ineffective. On the other hand, excessive amount of shake moves successive depth frames out of the original field-of-view of the scene, again leading to an unsatisfactory reconstruction of the scene. This is reflected in the RMSE values shown in Fig. 2: at high SBR the RMSE of the super-resolved image being minimised when there is "low to medium" shaking.

It is of interest to compare the super-resolving performance of the algorithm presented here with other popular methods. In this paper, we consider nearest-neighbours interpolation, bicubic interpolation, guided filtering [6] (using intensity image as a guide) and our algorithm. Fig. 3 shows examples of super-resolved depth maps at a medium-low level of shaking and different SBR, with close-ups to appreciate details of the reconstruction. The RMSE values of our approach are generally lower than those of the other algorithms, especially when the SBR is low. This in part due to a reduction in the effect of noise from background photons, resulting from multiple depth frames being combined. When the SBR becomes higher, the noise in depth maps largely disappears and the observed differences in RMSE are reduced. The study thus indicates the benefits of our approach, especially when used under strong ambient conditions.

We repeated the above study by using 25 sub-exposures instead of the original 50. The processing (and acquisition) time is halved, and the results show similar RMSE for high SBR and a low level of shaking. However, the algorithm fails to reconstruct with the same

detail at low SBR and large levels of shaking due to the reduced amount of depth data for each super-resolved pixel position, leading to inaccurate depth estimates.

A similar study investigates the performance of the algorithm when upscaling by a factor of 8 (from 64×64 to 512×512) is targeted. As with the previous study, the algorithm is able to upscale with a fine level of detail at high SBRs and medium levels of shaking. Fig. 4 shows a comparison of super-resolved depth maps under a "medium-low" level of shaking and different SBRs for ×4 and ×8 upscaling. When increasing the resolution by a factor of 8, the algorithm is seen to reduce the ambient noise more efficiently at low SBR values (as indicated by lower RMSE), at the expense of four times slower processing.

To demonstrate the approach using real data, we consider a 3D-stacked dToF SPAD with in-pixel histogramming [7], which has maximum frame rates of > 1 kFPS [5]. The sensor is built into a portable prototype camera system, featuring an integrated 850nm laser source providing flash illumination, and receiver optics giving 20º diagonal field of view.

The results of the processing are depicted in Fig. 5, for the case when a sequence of 25 intensity and 25 depth frames are captured, each with 5 ms exposure, with the camera intentionally shaken in a moderate way. Fig. 5a shows the first intensity frame from this sequence, and Fig. 5c the subsequent depth frame. The latter is seen to be relatively noisy, as a consequence of the high ambient level that the data was captured in. Indeed, Fig. 5b, showing the underlying histogram for a selected depth point, indicates a SBR of less than 0.1. Fig. 5d depicts the output of the processing, following depth frame re-alignment and merging. The level of apparent noise has been reduced considerably, and the outlines of the objects show improved resolution. Fig. 5e and 5f show point clouds, based on the single depth frame, and the super-resolved depth frame, respectively. Again, the super-resolved version is seen to result in a cleaner, more detailed image.

The equivalent frame rate of the super-resolved depth frame is 4 FPS, which can be increased by reducing the number of frames re-aligned or the exposure time in each frame.

III. CONCLUSIONS

This paper presents an approach which trades frame rate for increased lateral resolution, providing evidence of its effectiveness compared with other popular upscaling techniques, especially at low SBR conditions. One of the advantages of the approach, compared with the conventional technique of intensity-guided upscaling, is that the intensity data is only used for motion estimation here, so could potentially be replaced by data from inertial sensors (or the motion estimated from the depth data itself).

Acknowledgements— This research was supported by EPSRC via grants EP/M01326X/1, EP/S001638/1 and DSTL Dasa project DSTLX1000147844. AH would like to acknowledge the support of the UK's Royal Academy of Engineering (RF/201718/17128). The authors are grateful to STMicroelectronics and the ENIAC-POLIS project for chip fabrication.

References—

[1] Fersch et al., GeMiC 2016
[2] Buratto et al., Sensors 2021, 21(6), 1962
[3] Wronski et al., SIGGRAPH 2019
[4] D. Scharstein et al., CVPR 2007
[5] Gyongy et al., IISW 2019
[6] He et al., IEEE, 35(6), 2013
[7] Hutchings et al., IEEE JSSC 2020

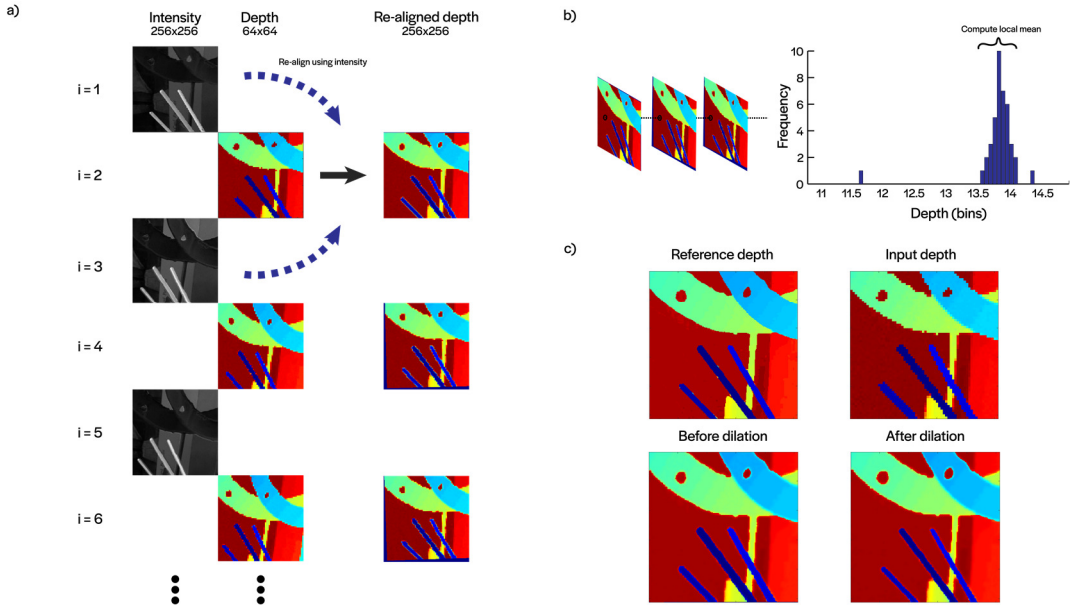

Figure 1. Depth super-resolution: a) intensity and depth data is captured in alternating frames; the similarity transformation between intensity frames is estimated, and used to re-align depth frames. b) the mean of the depth frames is computed, ignoring outlier points. c) reconstructed depth before and after dilation compared to the reference and input depth maps

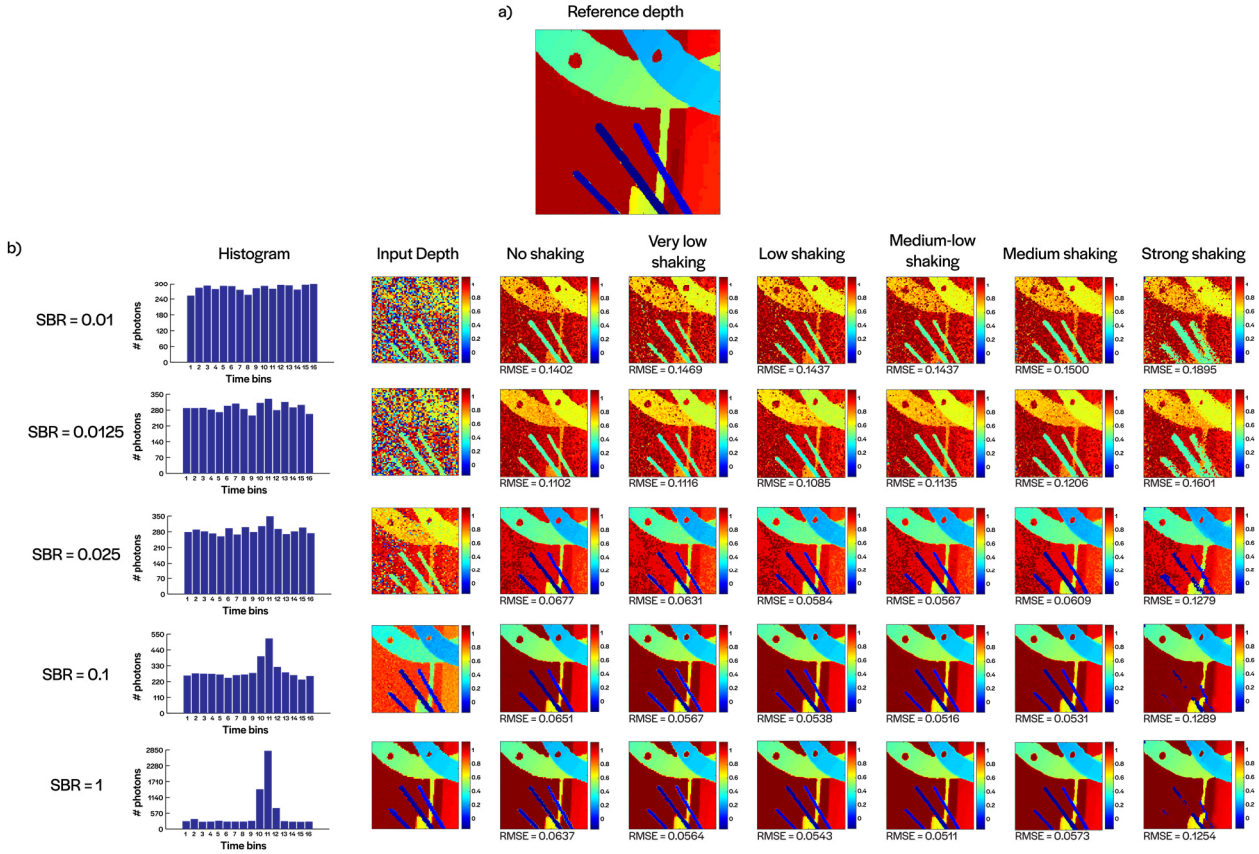

Figure 2. Results of processing for different SBR and shake levels: a) reference 256×256 depth map b) example histogram, first input depth map (64×64) and final reconstructed depth maps (256×256). All depth maps are normalised to values between 0-1. A randomised geometric transformation is applied to each input frame, with respect to the previous frame, in terms of a rotation of magnitude×rand degrees, and translations in x and y by magnitude×rand pixels (on the upscaled grid), where rand is a random number between -1 and 1 and magnitude = 0 , 0.05, 0.1 0.25, 0.5 and 2, for "no shaking", "very low shaking", "low shaking", "medium-low shaking", "medium shaking", and "strong shaking", respectively.

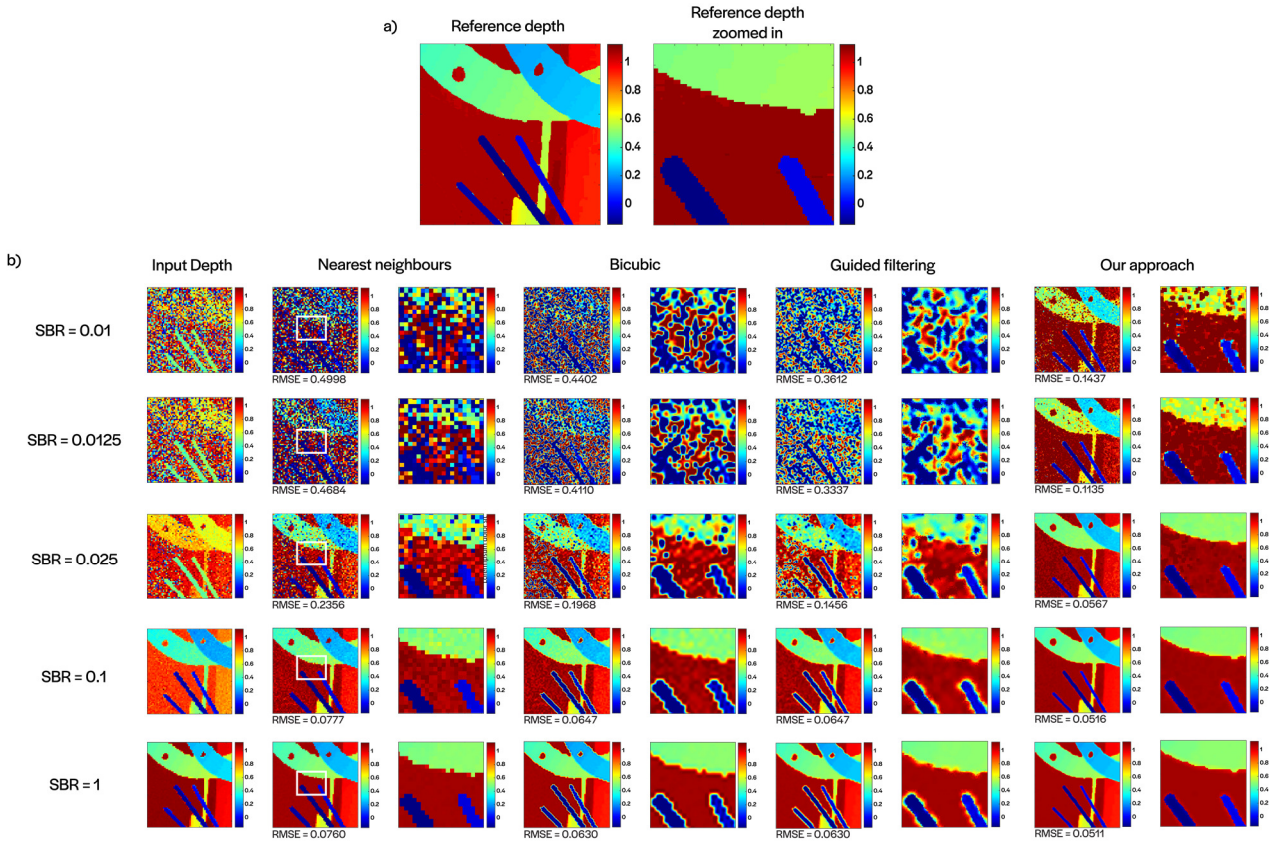

Figure 3. Super-resolution processing for different SBR and medium-low shaking: a) reference 256×256 depth map b) input depth map (64×64) and reconstructed depth maps (256×256) including close-up (on area indicated by white rectangle) for nearest neighbours, bicubic interpolation, guided filtering and our approach. All depth maps are normalised to values between 0-1.

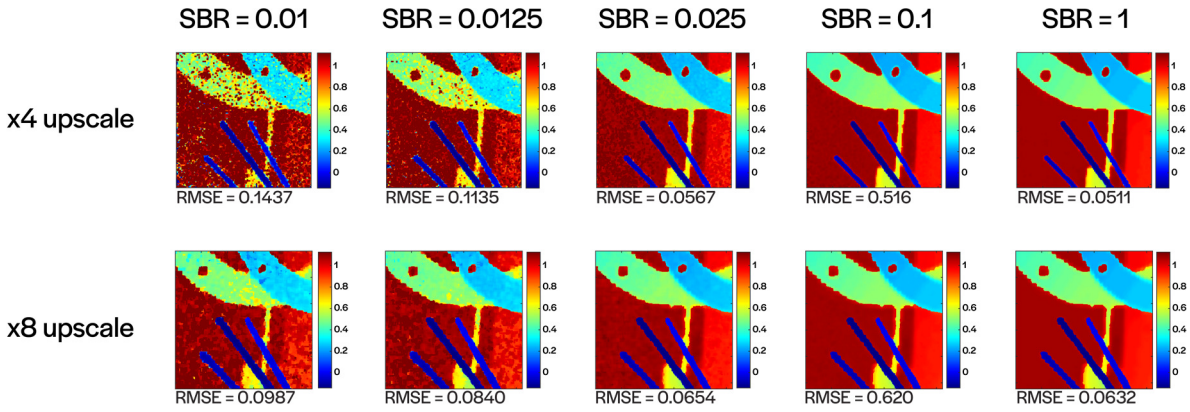

Figure 4. Super-resolution processing for different SBR and medium-low shaking for ×4 and ×8 increase in lateral resolution. All depth maps are normalised to values between 0-1.

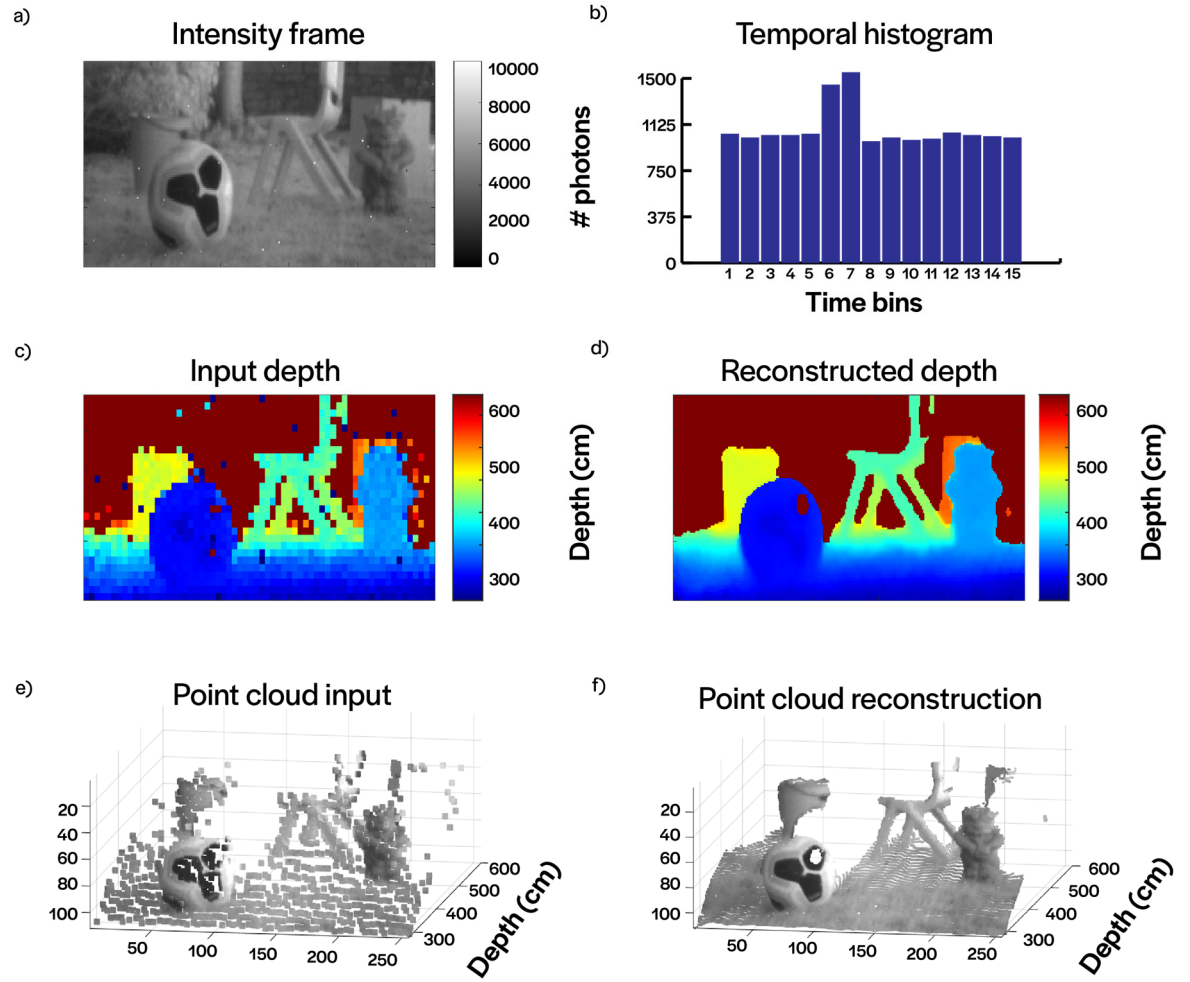

Figure 5. Results of processing for the garden scene: a) single intensity image b) example photon timing histogram, corresponding to nose of gnome. c) single depth frame, obtained by centroiding the raw histogram frame from the sensor d) processed depth frame e) point cloud for depth frame in c with intensity overlaid f) point cloud for depth frame in d with intensity overlaid.
The sensor was operated in flash mode and configured with 4ns bin width, leading to ~10m of unambiguous depth range. A 850nm laser producing 2W, 10ns pulses at 6MHz was used, with a 10nm bandwidth filter being placed in front of the sensor.