Depth video super-resolution using a high-speed time-of-flight sensor

Proceedings Volume 12438, AI and Optical Data Sciences IV; 124380J (2023)
https://doi.org/10.1117/12.2647260

PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Depth video super-resolution using a high-speed time-of-flight sensor

Germán Mora-Martín, Stirling Scholes, Alice Ruget, Robert Henderson, Jonathan Leach, et al.

Germán Mora-Martín, Stirling Scholes, Alice Ruget, Robert Henderson, Jonathan Leach, Istvan Gyongy, "Depth video super-resolution using a high-speed time-of-flight sensor," Proc. SPIE 12438, AI and Optical Data Sciences IV, 124380J (15 March 2023); doi: 10.1117/12.2647260

SPIE.

Event: SPIE OPTO, 2023, San Francisco, California, United States

Depth video super-resolution using a high-speed time-of-flight sensor

Germán Mora-Martín*[a], Stirling Scholes[b], Alice Ruget[b], Robert Henderson[a], Jonathan Leach[b], and Istvan Gyongy[a]

[a]School of Engineering, Institute for Integrated Micro and Nano Systems, The University of Edinburgh, Edinburgh EH9 3FF, UK
[b]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK
*Corresponding author: german.mora@ed.ac.uk

ABSTRACT

Three-dimensional (3D) imaging captures depth information from a given scene and is used in a wide range of fields like industrial environments, smartphones and autonomous driving, among others. This paper summarises the results of a depth video super-resolution scheme that is tailored for single-photon avalanche diode (SPAD) image sensors, which produces 3D maps at frame rates $>$ 100 FPS (32×64 pixels). Consecutive frames are used to super-resolve and denoise depth maps via 3D convolutional neural networks with an upscaling factor of 4. Due to the lack of noise-free, high-resolution depth maps captured with high-speed cameras, the neural network is trained with synthetic data using Unreal Engine, which is later processed to resemble the data outputted by a SPAD sensor. The model is then tested with different video sequences captured with a high-speed SPAD dToF, which processes frames at >30 frames per second. The super-resolved data shows a significant reduction in noise and presents enhanced edge details in objects. We believe these results are relevant to improve the accuracy of object detection in autonomous driving cars for collision avoidance or AR/VR systems.

Keywords: 3D-imaging, Time-of-flight, SPAD, LiDAR, Super-resolution, Neural Networks.

1. INTRODUCTION

3D-imaging has applications in sectors such as augmented reality (AR), autonomous driving or robotics.[1,2] Time-of-Flight (ToF) is a common way to do 3D-imaging, which is based on illuminating a scene with a pulsed light source and measuring the time for the photons to return.[3] In particular, direct ToF sensors (dToF) with Single-Photon Avalanche Diode (SPAD) technology are preferred over indirect ToF in long-range applications for their robustness against background photons and multipath issues.[4] However, SPAD dToF image sensors tend to have low lateral resolution (LR), which can potentially become a challenge in applications like object detection for autonomous driving. Neural networks for object detection have been used with SPAD data trying to overcome the resolution problem, and despite presenting promising results, these approaches are presented for short range or they are not general enough.[5,6]

The goal of this work is to find a way to improve the quality of depth maps by super-resolving (SR) and denoising depth maps with the use of a state-of-the-art method. Most of these methods are focused on using RBG or grayscale data, with a minority involving depth information.[7] Recently, deep learning methods have overperformed any previous approaches and have established themselves as state-of-the-art for SR.[8] The majority of SR work focuses on the upscaling of a single LR map to obtain a high-resolution (HR) output.[9] On the other hand, video super-resolution schemes exploit the temporal information from successive LR frames to produce an HR map, in exchange of frame rate.[10] These schemes are often a combination of an inter-frame alignment network along with a feature extraction network to produce an HR map.[11] Other techniques without alignment feature 3D convolution or recurrent neural networks to exploit spatio-temporal information directly.[12]

Video super-resolution schemes have been designed for RGB data and they have not yet been optimised for the usage of depth frames. Some research has been conducted using depth, but these assume a static scene with a moving camera.[13] Therefore, this paper summarises results regarding an effective video super-resolution

AI and Optical Data Sciences IV, edited by Bahram Jalali, Ken-ichi Kitayama,
Proc. of SPIE Vol. 12438, 124380J · © 2023 SPIE
0277-786X · doi: 10.1117/12.2647260

method using SPAD dToF data, as presented in reference.[14] Synthetic data is generated to train and test the super-resolving and denoising method and its performance is evaluated using both synthetic and real data.

2. METHODS

We use Unreal Engine to obtain high-resolution, ground-truth depth maps for the neural network datasets.[15] The dataset comprises RGB sequences (and corresponding depth information) from different realistic scenarios generated at a lateral resolution of 256×128 and FoV of 30° and captured at different frame rates. The generated depth maps from Unreal lack Poisson noise and background photon counts as it commonly appears in single-photon dToF sensors.[16] To overcome this issue, the grayscale version of an RGB frame and its corresponding depth map are used to produce synthesized SPAD data in the shape of temporal photon histograms (with lateral resolution of 64×32 pixels) with different signal-to-background ratios (SBR). Synthetic depth maps are then obtained by using centre-of-mass peak extraction[17] and are normalised between 0 and 1. Depth maps are finally concatenated in groups of $2T_R+1$ frames, where T_R is the temporal radius (T_R prior and posterior frames are used to super-resolve the central frame). Outputs are compared to the ground-truth using metrics like peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM). Figure 1 shows the key steps in the pre-processing of data for the method presented here.

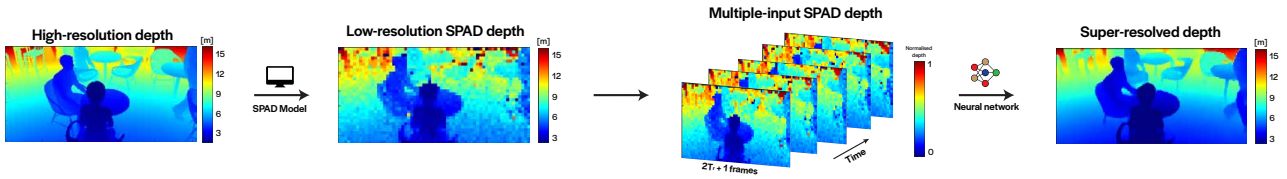

Figure 1. Workflow diagram showing the process of: capturing virtual scenes; converting them into SPAD-like data; grouping them according to a temporal radius T_R (TR = 2 in the example) and using the neural network to super-resolve frames. Adapted figure from reference.[14]

We adapt the structure of the VSR-DUF neural network to perform video super-resolution and denoising of SPAD dToF data.[12] This network comprises blocks of 3D convoluitions that extract spatio-temporal features without the need of extra steps (e.g. frame realignment). The input of the network has a size of $64\times32\times(2T_R+1)$ (here T_R is considered between 0 and 4) and an output that up-scales ×4 both transverse axes (256×128). Training parameters can be found in reference.[14]

3. RESULTS

The method proposed here is evaluated using a model trained with 15.500 examples with a variety of scenarios and SBR levels. Different versions of the network are trained for different T_R (from 0 to 4, which corresponds to using 1 to 9 input frames). The validation and test datasets consist of 3 different sequences of 500 frames each. Figure 2 shows an example frame of the input, ground-truth and the super-resolved depth maps for different radii of a sequence in the test dataset (average SBR of 0.54).

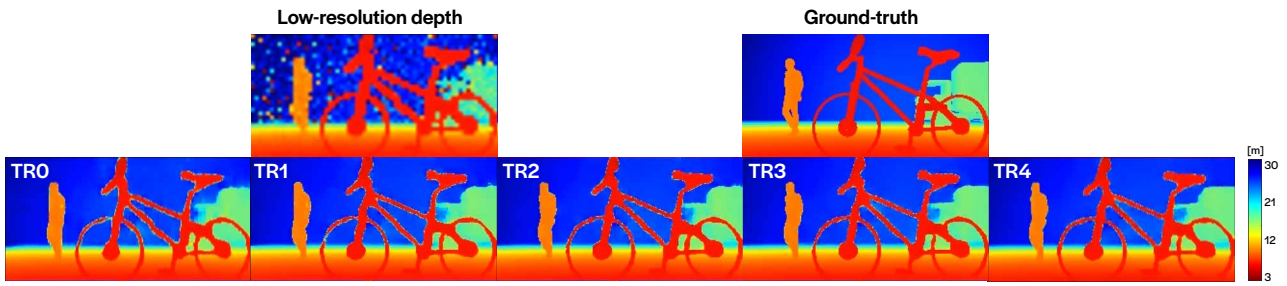

Figure 2. Example frame of the input, ground-truth and the super-resolved depth maps for different radii of a sequence in the test dataset (Scene 2, average SBR of 0.54). Original sequence is taken from reference.[14]

Frames that are super-resolved with a single image show an improvement in lateral resolution, but lack temporal coherence and introduce temporal artifacts. However, when using multiple inputs, temporal artifacts are reduced significantly thus improving the overall quality of the output. In terms of processing times, the use of multiple inputs naturally slows down the speed of the network due to its increased size, but the maximum temporal radius considered here (TR = 4) is still able to work at video rates (30 FPS). Note that the algorithm works in a rolling fashion, meaning that the first time it needs to capture T_R+1 frames but then only an additional frame is required.

The algorithm is tested experimentally using a high-speed, 64×32 SPAD dToF sensor. The experimental setup uses a 670 nm laser diode (Picoquant LDH-Series, 60 MHz repetition rate) and a 3.3 mm, NA = 0.47 lens (Thorlabs N414TM-A) to illuminate the scene. The imaging is done through a 3.5 mm/f1.4 objective with 25° FoV. Figure 3 depicts short-range examples of depth frames captured under 1ms exposure that have been super-resolved using the method proposed here, with $T_R = 4$. Outdoors examples are available in reference.[14]

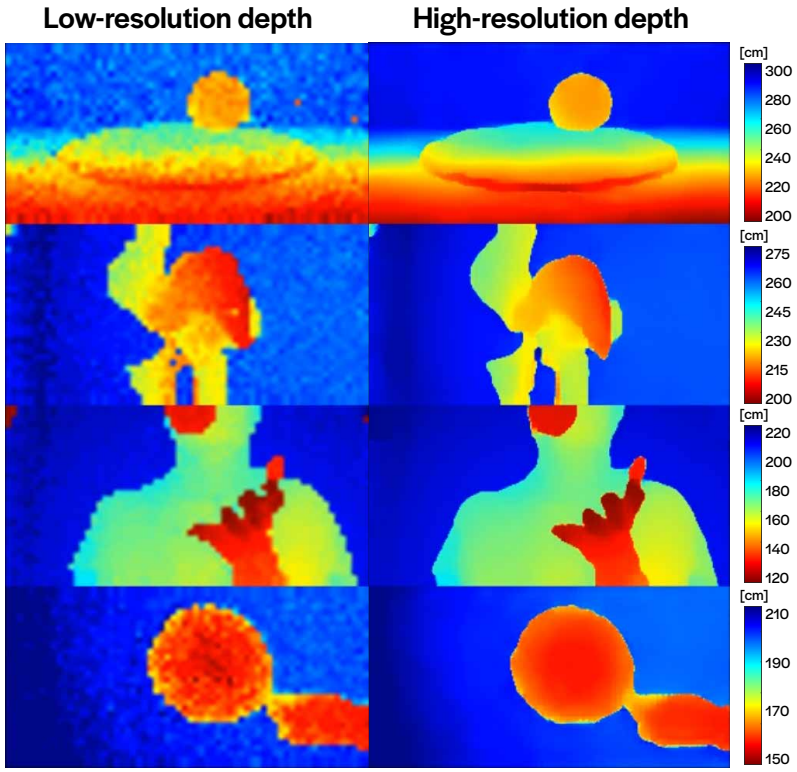

Figure 3. Comparison of low and high-resolution (×4 up-scale) depth maps. Scenes involve a ball dropping into a bowl, a rotating fan, a person juggling and a balloon before bursting. The scenes have been captured using exposure times under the millisecond. Original sequence is taken from reference.[14]

The super-resolved frames show an overall improvement in their quality in terms of smoother surfaces (less noisy data) and better profile definition in objects. The quality of the experimental results highlights the effectiveness of the neural network, which has been trained purely on synthetic data.

4. CONCLUSION

A video super-resolution scheme based on neural networks has been developed and applied to SPAD dToF data, overcoming the transerve resolution limitation of dToF image sensors. A synthetic and diverse dataset targetting common objects and scenarios is created using Unreal and it is processed into SPAD-like data with adjustable noise levels. The results compare the performance of a single-image super-resolution approach with the use of multiple inputs and the latter show a noticeable improvement in the quality of depth maps. Furthermore, the network, which has been trained purely on synthetic data, enhances the quality of depth frames captured with a

real SPAD dToF image sensor. The present network has potential in applications such as collision avoidance in autonomous driving, where accurate depth maps with minimal latency are required. Future work might involve the combination of object detection with super-resolved frames and compare its performance with the one at the native resolution of the sensor.

ACKNOWLEDGMENTS

This work was supported by EPSRC through grants EP/M01326X/1 and EP/S001638/1. Also it is supported by DSTL Dasa projects DSTLX1000147844 and DSTLX1000147352. The authors are grateful to STMicroelectronics for chip fabrication.

REFERENCES

[1] Chen, X., Ma, H., Wan, J., Li, B., and Xia, T., "Multi-view 3D object detection network for autonomous driving," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* **2017-Janua**, 6526–6534 (2017).

[2] Pan, G., Han, S., Wu, Z., and Wang, Y., "3d face recognition using mapped depth images," in [*2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*], 175–175 (2005).

[3] Horaud, R., Hansard, M., Evangelidis, G., and Ménier, C., "An overview of depth cameras and range scanners based on time-of-flight technologies," *Machine Vision and Applications* **27**, 1005–1020 (Jun 2016).

[4] Henderson, R. K., Johnston, N., M, DellaRocca, F., Chen, H., D. Li, D., Hungerford, G., Hirsch, R., Mcloskey, D., Yip, P., and Birch, D. J. S., "A 192 × 128 time correlated SPAD image sensor in 40-nm CMOS technology," *IEEE Journal of Solid-State Circuits* **54**(7), 1907–1916 (2019).

[5] Turpin, A., Musarra, G., Kapitany, V., Tonolini, F., Lyons, A., Starshynov, I., Villa, F., Conca, E., Fioranelli, F., Smith, R. M., and Faccio, D., "Spatial images from temporal data," *Optica* **7**, 900–905 (Aug 2020).

[6] Martín, G. M., Turpin, A., Ruget, A., Halimi, A., Henderson, R., Leach, J., and Gyongy, I., "High-speed object detection with a single-photon time-of-flight image sensor," *Opt. Express* **29**, 33184–33196 (Oct 2021).

[7] Chen, H., He, X., Qing, L., Wu, Y., Ren, C., Sheriff, R. E., and Zhu, C., "Real-world single image super-resolution: A brief review," *Information Fusion* **79**, 124–145 (2022).

[8] Tang, Q., Cong, R., Sheng, R., He, L., Zhang, D., Zhao, Y., and Kwong, S., "BridgeNet: A joint learning network of depth map super-resolution and monocular depth estimation," in [*Proc. ACM MM*], (2021).

[9] Sun, Z., Lindell, D. B., Solgaard, O., and Wetzstein, G., "Spadnet: deep rgb-spad sensor fusion assisted by monocular depth estimation," *Opt. Express* **28**, 14948–14962 (May 2020).

[10] Liu, H., Ruan, Z., Zhao, P., Dong, C., Shang, F., Liu, Y., Yang, L., and Timofte, R., "Video super-resolution based on deep learning: a comprehensive survey," *Artificial Intelligence Review* (04 2022).

[11] Bare, B., Yan, B., Ma, C., and Li, K., "Real-time video super-resolution via motion convolution kernel estimation," *Neurocomput.* **367**, 236–245 (nov 2019).

[12] Li, Y., Zhu, H., Hou, Q., Wang, J., and Wu, W., "Video super-resolution using multi-scale and non-local feature fusion," *Electronics* **11**(9) (2022).

[13] M. Martín, G., Halimi, A., K. Henderson, R., Leach, J., and Gyongy, I., "High-ambient, super-resolution depth imaging with a spad imager via frame re-alignment," in [*Proceedings of IISW International Image Sensor Workshop*], (September 2021).

[14] Martín, G. M., Scholes, S., Ruget, A., Henderson, R., Leach, J., and Gyongy, I., "Video super-resolution for single-photon lidar," *Opt. Express* (Jan 2023).

[15] Epic Games, "Unreal engine."

[16] Dutton, N. A. W., Gyongy, I., Parmesan, L., and Henderson, R. K., "Single photon counting performance and noise analysis of cmos spad-based image sensors," *Sensors* **16**(7) (2016).

[17] Gyongy, I., Hutchings, S. W., Halimi, A., Tyler, M., Chan, S., Zhu, F., McLaughlin, S., Henderson, R. K., and Leach, J., "High-speed 3D sensing via hybrid-mode imaging and guided upsampling," *Optica* **7**, 1253–1260 (Oct 2020).