A learning method to optimize depth accuracy and frame rate for Time of Flight camera

Yan Chen¹, Keyuan Qian², *, Xuanye Cheng³, Jingkun Zhou¹

¹Department of Electronic Engineering, Tsinghua University, Beijing, China
²Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
³SenseTime Research, SenseTime, Shenzhen, China

Corresponding author’s e-mail: qianky@sz.tsinghua.edu.cn

Abstract. Time of Flight cameras attract increasing attention due to their ability to robustly predict depth values of the scene. The principle of actively emitting modulated signals and receiving sampled signals causes that the artifacts generated by motion blur seriously affect quality of imaging and the power consumption of the light source. The problem of shorter exposure time is that the raw measurements have a lower signal-to-noise ratio. Reducing the four frames of raw measurements for depth reconstruction to two frames can make it difficult to effectively suppress noise from the background infrared signal and capacitive components. However, shorter exposure time and reducing four frames of raw data to two frames will greatly reduce the motion blur of the time-of-flight camera with a higher frame rate and lower power consumption. We propose a method based on deep learning and make it possible to reliably recover the depth information even under extremely low signal-to-noise ratio and the noise of electronic components for alleviating the motion blur problem. The experiment results fully demonstrate the reliability of our solution. Our algorithms and data will be published in the future.

1. Introduction

With the popularity of new technologies such as autonomous driving and 3D face recognition, conventional two-dimensional images have failed to meet the demands of people’s life. Therefore, the acquisition of depth information has become a significant research task. Conventional depth cameras are mainly divided into three categories, binocular cameras, structured light cameras, and time-of-flight cameras. Among them, the time-of-flight (ToF) camera has attracted more and more attention due to its straightforward physical principles and its robust ability to acquire depth information. At present, mature products have come out, such as Kinect V2 [1]. Furthermore, ToF cameras are also active in a variety of research areas, such as gesture recognition [2], 3D reconstruction [3], robotics [4], and human pose estimation. Continuous-wave ToF camera emits an infrared signal modulated by a specific frequency to illuminate a measured scene and the sensor of ToF camera detects part of signal reflected by target object in the scene. ToF camera can collect continuous four-frame raw measurements by demodulating sampled signals and we can obtain time of flight of emitted signals to recover depth information of the scene from these four-frame raw measurements.

Compared with the conventional RGB camera, the ToF camera is required to actively emit continuous modulated illumination signals during the exposure time of acquiring depth image due to the principle of depth reconstruction. The emission of the modulated signal tends to occupy a large proportion in power consumption of the ToF camera imaging. Therefore, reducing the emission power
of the modulated signal or reducing the exposure time tends to greatly optimize the total power consumption of the ToF camera, which also provides the possibility for ToF cameras to integrate on low-power products such as mobile phones. However, the following problem is that reducing exposure time makes the signal-to-noise ratio (SNR) of raw data lower and raw data with low SNR tends to have low reliability of the depth generated by the ToF camera. In addition, in order to eliminate the influence of hardware devices such as capacitance gain and ambient light, ToF cameras often use four continuous frames of raw data to generate a depth map and each two of them are made subtraction to eliminate capacitance gain and ambient light effects. It can improve the accuracy of the final depth map, but the problems that this causes are decreasing frame rates and more severe motion blur.

Depth reconstruction from low SNR raw data and eliminating the capacitance gain and the ambient light effect from the pipeline of depth estimation with two-frame raw data become critical factors of achieving low motion blur, low power consumption, high frame rate, and high image quality in depth acquisition. A large number of recent studies have proved that a high-quality depth map can be obtained from ToF raw collected in a very low illumination environment by deep learning method [5]. Inspired by this, we devised a convolutional neural network to generate high-precision depth maps with low SNR ToF raw data. We extract two frames with a phase interval of $\pi/2$ from four-frame raw data required for generating one depth map as the input of convolutional neural network and recover corresponding high-quality depth map with trained model. We validate the reliability of our approach in scenarios that were acquired with motion and low exposure settings. Experiments present that although lower exposure time makes the SNR of raw data lower and two frames of raw data as inputs cannot eliminate ambient light and capacitance gain, our method is able to stably recover high quality depth information, effectively suppress blur motion and significantly improve the frame rate.

2. Related Work
ToF cameras can estimate depth information from the phase shift of backscattered signal with respect to the emitting amplitude-modulated infrared illumination signal as a promising solution when 3D information is required. However, during the process of obtaining depth values, a lot of significant problems should be solved. When measuring depth of large environments where the distances exceed a certain range determined by the modulation frequency, many different depth values can be decoded from the same measurement and distance will be ambiguous. An approach of using different modulation frequencies to unwrap high-frequency phases with their lower-frequency counterpart was proposed by Dorrington et al. [6]. Due to global illumination, not only the direct reflected signals but also indirect reflected signals can be record by a single pixel. This results in severe multipath interference (MPI) distortion of captured depth maps. A lot of previous work attempted to address this problem, such as Wu et al. [7] exploited time profile to separate direct component and other different components from light transport for correcting MPI errors and Su et al. [8] proposed a deep convolution neural network to generate depth from ToF raw measurements and effectively resolved MPI and phase unwrapping problems.

Another serious distortion of ToF camera imaging that should be corrected is motion blur resulted from physical motion of camera or object in scene. ToF camera obtains the depth values by measuring phase shift between emitted signal and reflected signal and a non-linear transformation is performed to calculate depth value. Thus, distance values are calculated incorrected and uneven errors occurs when phase images values do not align to each other, caused by motion. This unique characters of ToF cameras imaging can cause that the method of resolving conventional color motion blur is unable to be applied to address motion blur of ToF camera. Hussmann et al. [9] proposed a method to detect single direction motion based on his analysis of motion blur resulted from a single direction motion of a conveyor belt. Linder et al. [10] introduced another motion artifacts compensation technique based on optical flow. However, this algorithm can only work on some special simple cases. Lee et al. [11] tried to recover distance information from remaining measurements that were not corrupted by motion when examining that motion occurred during integration time.
3. Depth Estimation Principle

Fig. 1 illustrates the depth reconstruction principle of ToF cameras. An intensity modulated near infrared signal is emitted to target object and the reflected signal is captured by ToF sensor. The depth values can be obtained by measuring the phase difference between radiated signal \( s \) and reflected signal \( r \). The reference signal \( s \) can be expressed as follow with the assumption that modulated angular frequency is \( \omega = 2\pi f \) and aptitude is normalized,

\[
s(t) = \cos(\omega t)
\]

The aptitude of optical signal attenuates during light traveling, thus we can describe the reflected signal as:

\[
r(t) = k + a \cos(\omega t + \varphi)
\]

where \( a \) is the aptitude after optical signal attenuation, \( \varphi \) is phase shift related to the distance of objects, and \( k \) is the offset. The correlation signal \( c(\tau) \) can be obtained from emitted signal and received signal:

\[
c(\tau) = r(t) \otimes s(t)
\]

\[
= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} r(t) \cdot s(t + \tau) dt
\]

\[
= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} [k + a \cos(\omega t - \varphi)] \cdot [\cos(\omega t + \omega \tau)] dt
\]

\[
= \frac{a}{2} \cdot \cos(\omega \tau + \varphi)
\]

The correlation signal can be sampled four times in one period and the phase delay for \( \tau_i \) \((i = 0, \ldots, 4)\) equals to \( 0^\circ, 90^\circ, 180^\circ \) and \( 270^\circ \). According to these four samples, we can calculate the correlation amplitude \( a \) and phase shift \( \varphi \) as follow:

\[
a = \frac{1}{2} \sqrt{[c(\tau_3) - c(\tau_1)]^2 + [c(\tau_0) - c(\tau_2)]}
\]

\[
\varphi = \arctan \frac{c(\tau_3) - c(\tau_1)}{c(\tau_0) - c(\tau_2)}
\]

Distance can be derived:

\[
d = \frac{c\varphi}{4\pi f}
\]
4. Method

4.1. Synthetic dataset
For network training, we used Unity to generate a batch of Synthetic dataset. As shown in Fig. 2, we first model some virtual scenes, such as bedrooms, gymnasiums, etc. Then we utilize the 3D virtual camera of unity to collect the depth map of the corresponding scene. Then, according to the principle of ToF imaging described in Sec. 3 and Eq. 1-6, four frames of raw data can be obtained from corresponding depth map, and random Gaussian noise is added to the image. We set $f$ in Eq. 6 to 6MHz and generated a total of 20,000 sets of data to train the model. As shown in Figure 3, each set of data contains a depth map and four corresponding raw measurements.

4.2. Network structure and training details
We expect that a specific structure of convolutional neural network can recover high-quality depth information from very noisy raw ToF measurements under the influence of motion blur. In order to achieve this goal, it is required to take two-frame raw measurements with a phase interval of $\pi/2$ as input and corresponding depth map captured under long exposure time as ground truth labels for learning task. Compared with conventional colour image generation, the pixel value of raw ToF measurements is more affected by the geometry of the object and camera settings, and theoretically the depth map has a strict corresponding relation with the four-frame raw ToF measurement. We believe that this task is similar to the image translation task. Inspired by [12], we designed an Unet-based network structure.
Figure 4. Network Architecture

Based on above considerations, we devised a convolutional neural network as illustrated in Figure 4. The size of input data is changed into 320×240 after concat operation. The whole structure consists of a convolutional layer, three downsample layers, six residual blocks and three up-sample layers. Each downsample layer contains a convolutional layer and a LeakyReLU layer with kernel size 4×4 and stride 2. After four-time downsample, the size of photo becomes 20×15×256. And each up-sample layer includes a convolutional layer and a LeakyReLU layer with kernel size 4×4 and stride 1/2. In addition, we added skip connections between the convolutional layer after i-th down-sample layer and the convolutional layer after the 4-i up-sample layer to improve the accuracy of the final result.

In order to complete the training, we used the data set in [5], which contains real ToF measurements in multiple scenarios. Each scene contains ToF raw, confidence and depth maps under 200us, 400us and regular exposure time. To improve the robustness of the model, we flipped all available data for data augmentation. To supervise the output of each up-sample layer, we designed the loss function as shown follow:

\[
\text{Loss} = \frac{1}{4} \sum_{n=1}^{N} |d_{gt} - d_{up1}| + \frac{1}{4} \sum_{n=1}^{N} |d_{gt} - d_{up2}| + \frac{1}{2} \sum_{n=1}^{N} |d_{gt} - d_{out}|\] (7)

We use the Adam optimizer method to optimize our network and set the initial learning rate to 0.002. The network has been trained for 10,000 epochs totally and the learning rate has been updated to half of the previous value every 1000 epochs. In addition, the confidence map of each scene represents the confidence of the corresponding depth information. To further improve the ability of the model to estimate depth information, we use the confidence information of each scene as the weight of the loss function to fine-tune, as shown in Eq.8. In this stage, the network has been trained for a total of 5000 epochs with the initial learning rate 2e-5 and learning rate has been updated to half of the previous value every 1000 epochs.

\[
\text{Loss} = \frac{1}{4} \sum_{n=1}^{N} |d_{gt}^{ij} - d_{up1}^{ij}| + \frac{1}{4} \sum_{n=1}^{N} |d_{gt}^{ij} - d_{up2}^{ij}| + \frac{1}{2} \sum_{n=1}^{N} |d_{gt}^{ij} - d_{out}^{ij}|\] (8)

5. Experiments and Results

The following subsections demonstrate detailed information of our method to optimize depth accuracy and frame rate for ToF camera in two parts. A quantitative analysis of results shows compared with ground truth and depth map captured under low exposure in the first part. And the second part presents the qualitative evaluation of comparison between our results and conventional method adopted in ToF camera.

5.1. Quantitative result

We have trained two models on the ToF raw measurements under 200us and 400us exposure respectively and only taken two of raw measurements as input. In order to validate the reliability and the accuracy of our method, we tested the two models with the corresponding test set offered by [5].
Then we adopt the mean absolute error (MAE) [13] and the structural similarity (SSIM) of depth map generated by our model with ground truth to evaluate our method. For further evaluation, two models with four-frame input have been trained on ToF raw measurements under 200us and 400us exposure respectively. And we can compare the results of our proposed method with that of the traditional ToF camera pipeline and that of model trained on four-frame ToF raw on the test set offered by [5].

| Method               | MAE (cm) | SSIM (%) | MAE (cm) | SSIM (%) |
|----------------------|----------|----------|----------|----------|
| Traditional pipeline | 56.41    | 0.2216   | 32.53    | 0.5180   |
| Models with 4-frame  | 12.19    | 0.8852   | 9.37     | 0.9116   |
| Ours                 | 17.46    | 0.8754   | 12.86    | 0.9009   |

Table 1. MAE(cm) and SSIM(%) for two models with four-frame input on 200us and 400us exposure time. Our results with two-frame measurement captured under 200us as input meet an overall 17.46cm depth error and results with two-frame measurement captured under 400us as input meet an overall 12.86cm depth error. Although the amount of input frame has been reduced from 4 to 2, the accuracy of depth map generated by our method decreases slightly.

| Method                | 200us | 400us | 4000us |
|-----------------------|-------|-------|--------|
| Traditional pipeline  | \     |       | 24fps  |
| Ours with 4-frame     | 39fps | 38fps | \      |
| Ours with 2-frame     | 77fps | 76fps | \      |

Table 2. Frame rate for two models with four-frame input on 200us, 400us, and 4000us exposure time. We calculate the optimization of the frame rate of our method on the EPC660. The traditional method cannot calculate the correct depth map under exposure time 200us or 400us, so we did not calculate the frame rate of these two items. As shown in Tab.2, the learning-based method is utilized to recover the high-quality depth map with two corresponding frames of raw data as input of network. Under the 200us and 400us exposure time, the frame rate based on our method with two-frame input will be increased by nearly three times than the previous frame rate, and the frame rate based on method with four-frame input will be increased by a factor of 1.5 times than the previous frame rate.

5.2. Qualitative result

Fig.5 shows the comparison of depth map quality between our method and traditional method with four frames of raw data captured under 200us and 400us exposure time as input. Besides, the depth maps generated by our method with two-frame raw data captured under 200us and 400us exposure time as input are also included. In order to visually compare the effect of our method with that of traditional methods, we reconstruct the depth maps using the traditional algorithm and our algorithm in the two scenarios of the test set. As shown in Fig.5, our method has better performance in denoising the depth map than the traditional algorithm. And for the results recovered from two frames and four frames, the depth map generated by our algorithm is closer to the ground truth.
6. Conclusion
We use a learning-based approach to reconstruct high-quality depth maps from low-SNR ToF camera raw measurements captured under very low exposure time. To achieve higher frame rates, we utilize two frames of raw data to generate high-quality depth maps according to the ToF imaging principle. For completing the learning process, we collect 20,000 sets of depth map with Unity 3D and calculate the corresponding raw data with random Gaussian noise for data set synthesis. The experiments show that our algorithm not only has higher accuracy of depth map than that of the traditional depth map generation method, but also can increase the frame rate of the ToF camera by 3 times. In the future, we will continuously improve the accuracy of this method on the depth map reconstruction and apply our solution to the motion blur alleviating.

Acknowledgments
Funding: Basic Research Priorities Program of Shenzhen (Grants JCYJ20160608170030295)

References
[1] Fankhauser, P., Bloesch, M., Rodriguez, D., Kaestner, R., Hutter, M., & Siegwart, R. (2015, July). Kinect v2 for mobile robot navigation: Evaluation and modeling. In 2015 International Conference on Advanced Robotics (ICAR) (pp. 388-394). IEEE.
[2] Memo, A., & Zanuttigh, P. (2018). Head-mounted gesture controlled interface for human-computer interaction. Multimedia Tools and Applications, 77(1), 27-53.
[3] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., ... & Fitzgibbon, A. (2011, October). KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In Proceedings of the 24th annual ACM symposium on User interface software and technology (pp. 559-568). ACM.
[4] Hussmann, S., & Liepert, T. (2007, May). Robot vision system based on a 3D-ToF camera. In 2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007 (pp. 1-5). IEEE
[5] Chen, Y., Ren, J., Cheng, X., Qian, K., & Gu, J. (2018). Very Power Efficient Neural Time-of-Flight. arXiv preprint arXiv:1812.08125.

[6] Dorrington, A. A., Godbaz, J. P., Cree, M. J., Payne, A. D., & Streeter, L. V. (2011, January). Separating true range measurements from multi-path and scattering interference in commercial range cameras. In Three-Dimensional Imaging, Interaction, and Measurement (Vol. 7864, p. 786404). International Society for Optics and Photonics.

[7] Wu, D., Velten, A., O’toole, M., Masia, B., Agrawal, A., Dai, Q., & Raskar, R. (2014). Decomposing global light transport using time of flight imaging. International journal of computer vision, 107(2), 123-138.

[8] Su, S., Heide, F., Wetzstein, G., & Heidrich, W. (2018). Deep end-to-end time-of-flight imaging. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6383-6392).

[9] Hussmann, S., Hermanski, A., & Edeler, T. (2011). Real-time motion artifact suppression in tof camera systems. IEEE Transactions on Instrumentation and Measurement, 60(5), 1682-1690.

[10] Lindner, M., & Kolb, A. (2009). Compensation of motion artifacts for time-of-flight cameras. In Dynamic 3D Imaging (pp. 16-27). Springer, Berlin, Heidelberg.

[11] Lee, S. (2014). Time-of-flight depth camera motion blur detection and deblurring. IEEE Signal Processing Letters, 21(6), 663-666.

[12] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).

[13] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.