ToF 센서와 스테레오카메라의 융합을 이용한 3 차원 데이터 복원

정석우, 이윤주, 이경택*
한국전자기술연구원

3D depth map estimation using ToF-Stereo sensor fusion

Sukwoo Jung, Yunju Lee, KyungTaek Lee*
Korea Electronics Technology Institute

swjung@keti.re.kr, topgun@keti.re.kr, *ktechlee@keti.re.kr

Abstract

High quality depth maps are required in various applications including robotics, computer vision, and autonomous driving. In order to obtain precise depth information, there has been an increasing interest in the combination of different depth sensors. Depth information can be acquired real-time by Time-of-Flight (ToF) cameras and stereo cameras. There are various Depth-Stereo fusing methods to overcome the limitations of a single depth sensor, including Conditional Cost Volume Normalization algorithm. The CCVNorm framework is generic and closely integrated with the cost volume component that is commonly utilized in stereo matching neural networks. In this paper, we give an overview over methods for the fusion of depth and passive stereo data. We experimentally evaluate the performance and robustness of various Depth-Stereo fusion methods with the KITTI Stereo and Depth Completion datasets.

I. Introduction

The accurate 3D data acquisition has been actively researched since its important role in robotics, computer vision, autonomous driving, etc. Various techniques have been researched to estimation depth precisely. The most commonly used sensors for depth estimation are Time of Flight (ToF) cameras and Stereo cameras.

Stereo camera works well on textured scenes and has a high lateral resolution due to the readily developed high resolution mega pixel cameras. However, there are limitations at occlusion boundaries and ambiguous textures. Also, due to the numerous pixels that have to be calculated, the computational costs of stereo matching algorithms are normally high.

ToF camera on the other hand delivers images at high frame rates and is invariant to the surface texture. ToF sensor computes depth by sending infrared signal measuring the phase shift of the reflected light signals. ToF camera is quite compact, can estimate depth in real-time and is not very sensitive to the texture of the scene. However, it has a limited resolution, limited accuracy, systematic error, and it is sensitive to the illumination of the scene.

The major drawback of these sensors is that they only work in a limited situation and lack the robustness which is required in most applications. Therefore, many researchers combine them to create a fusion sensor to overcome the limitations of the single depth sensor. Several works [1-4] have studied how to fuse depth-stereo modalities in order to obtain more accurate and denser depth map.

An acquisition system composed by a ToF camera and a stereo camera is proposed in [5]. The two subsystems are merged with a depth probability distribution function. In [6] the ToF depth measurement is converted into a disparity map, and then used as an initial condition for a stereo matching algorithm. In [7] information from the two sensors are combined by fusing data on 3D probabilistic occupancy grids. Kim et al. [8] proposed a fusion algorithm to estimate the patchlets using information from a ToF camera and a stereo camera.

II. Depth-Stereo Fusion Methodology

We used a method extending the stereo matching network by Input Fusion to incorporate the geometric information from sparse depth data with the RGB images for learning joint feature representations. Also, Conditional Cost Volume Normalization (CCVNorm) [12] is used to regularize cost volume optimization in depth measurements. CCVNorm is used to encode the sparse ToF information into the features of 4D cost volume.

The stereo matching network used in the CCVNorm method is based on the work of GC-Net [9] and is composed of four primary components which are in line with the typical pipeline of stereo matching algorithms [10]. First, the deep feature extracted from a rectified left-right stereo pair is learned to compute the cost of stereo matching. A cost volume is then constructed by aggregating the deep features extracted from the left-image with their corresponding ones from the right-image across each disparity level. To be detailed, the cost volume actually includes all the potential matches across stereo images and hence serves as a searching space matching. Next, a sequence of 3D convolutional operations is applied for cost volume regularization and the final disparity estimation is carried out by regression with respect to the output volume of 3D-CNN along the dimension.

In the cost computation stage of stereo matching network, both left and right images of a stereo pair are passed through layers of convolutions for extracting features. In order to enrich the representation by jointly reasoning on appearance and geometry information from RGB images and ToF data, the CCVNorm method used Input Fusion that simply concatenates stereo images with their corresponding sparse ToF depth maps. We form the two sparse depth maps corresponding to stereo images by reprojecting the depth map sweep to both left and right image coordinates with triangulation for converting depth values into disparity ones. In addition to the Input Fusion, CCVNorm method used the information of sparse depth points into the cost regularization step of stereo matching network, learning to reduce the searching space of matching and resolve ambiguities.

We also implemented various existing Depth-Stereo fusion algorithms to compare the performance.

III. Experiments and Conclusion

We evaluate various Depth-Stereo fusion methods with KITTI datasets. KITTI Stereo 2015 dataset [13] is commonly used for evaluating stereo matching algorithms. It contains 200 stereo pairs for each of training and testing set, where the images are in size of 1242 × 375. KITTI Depth Completion dataset [14] collects semi-dense ground truth of LiDAR depth map by aggregating 11 consecutive LiDAR sweeps together, with roughly 30% pixels annotated. The dataset consists of 43k image pairs for training, 3k for validation, and 1k for testing.

We adopt standard metrics in stereo matching and depth estimation respectively for the two datasets. On KTTI Stereo, we follow its development kit to compute the percentage of disparity error that is greater than 1, 2 and 3 pixels away from the ground truth. On KITTI Depth Completion, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used.

Our implementation is based on PyTorch and follows the training of GC-Net to get loss for disparity estimation. All the methods run on a personal computer with a Core i7 processor and a NVIDIA RTX3080 GPU.

We tested the implemented five ToF-Stereo fusion algorithms: 1) Semi-Global Matching (SGM) – performs poorly on uniform textures such as the whiteboard, common to most stereo algorithms; 2) Naive Fusion method (NF) – performs slightly better than SGM but still has trouble filling in erroneous stereo estimates; 3) Diffusion based method (DB) – performs significantly better than both the above methods while preserving disparity discontinuities and retaining accurate disparity measurements; 4) Neighborhood Support error method (NS) – performs well even with the monocular image and range information, but suffers in regions where nearby range information doesn't exit; 5) Conditional Cost Volume Normalization method (CCVNorm) – performs well in overall dataset, but the computational cost is higher than the other methods.

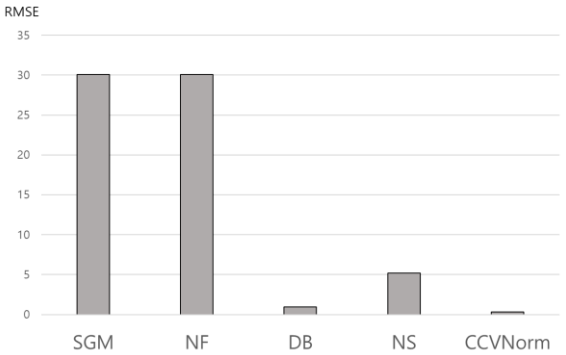

Figure 1. Average RMSE for the tested Depth-Stereo fusion algorithms

The average Root-mean-square error is shown in Fig. 1. SGM and NF shows the largest error as 30.07, and 30.06 RMSE. The error of DB and NS is 0.94, and 5.19 RMSE each, showing better result than the above algorithms. CCVNorm shows the best performance with 0.39 RMSE.

In this paper, we compared the performance of various Depth-Stereo fusion methods on both the KITTI Stereo and Depth Completion datasets. We are planning to obtain the datasets with the latest commercial sensors. Several tests will be taken with the given datasets. We will find and optimize the proper method for the ToF-Stereo fusion algorithm.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00103, Development of 5G-based 3D spatial scanning device technology for virtual space composition)

References

[1] R. Nair et al., "A survey on Time-of-Flight stereo fusion," in Time-of-Flight and Depth Imaging. Sensors Algorithms and Applications, Springer-Verlag, 2013, vol. 8200, pp. 105-127.

[2] R. Nair, F. Lenzen, S. Meister, H. Schäfer, C. Garbe, and D. Kondermann, "High accuracy TOF and stereo sensor fusion at interactive rates," in ECCV 2012 Ws/Demos, Part II. LNCS, Springer, Heidelberg, 2012, vol. 7584, pp. 1-11.

[3] J. Zhu, L. Wang, R. Yang, J. E. Davis, and Z. Pan, "Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps," IEEE Trans. Pattern Anal. Mach. Intell., 2011, vol. 33, no. 7, pp. 1400–1414.

[4] G. D. Evangelidis, M. Hansard, and R. Horaud, "Fusion of Range and Stereo Data for High-Resolution Scene-Modeling," IEEE Trans. Pattern Anal. Mach. Intell., 2015, vol. 37, no. 11, pp. 2178–2192.

[5] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[6] V. Gandhi, J. Čech, and R. Horaud, "High-resolution depth maps based on TOF-stereo fusion," in Proceedings - IEEE International Conference on Robotics and Automation, 2012, pp. 4742–4749.

[7] C. D. Mutto, P. Zanuttigh, and G. Cortelazzo, "A probabilistic approach to tof and stereo data fusion," in 3DPVT, 2010.

[8] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun, "Multi-view image and ToF sensor fusion for dense 3D reconstruction," in IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, 2009, pp. 1542− 1546.

[9] G. Agresti, L. Minto, G. Marin, and P. Zanuttigh, "Deep Learning for Confidence Information in Stereo and ToF Data Fusion," in Proceedings IEEE International Conference on Computer Vision Workshops, ICCVW, 2017, vol. 2018-January, pp. 697–705.

[10] G. Agresti, L. Minto, G. Marin, and P. Zanuttigh, "Stereo and ToF data fusion by learning from synthetic data," Inf. Fusion, 2019, vol. 49, pp. 161–173.

[11] X. Cheng, Y. Zhong, Y. Dai, P. Ji, and H. Li, "Noise-Aware Unsupervised Deep Lidar-Stereo Fusion," in Proceedings IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6339–6348.

[12] T. H. Wang, H. N. Hu, C. H. Lin, Y. H. Tsai, W. C. Chiu, and M. Sun, "3D LiDAR and Stereo Fusion using Stereo Matching Network with Conditional Cost Volume Normalization," in IEEE International Conference on Intelligent Robots and Systems, 2019, pp. 5895–5902.

[13] M. Menze and A. Geiger, "Object Scene Flow for Autonomous Vehicles," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3061-3070.

[14] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity Invariant CNNs," in International Conference on 3D Vision (3DV), 2017, pp. 11-20.