Stereo Disparity Estimation for Mobile Robots

Subjects: Computer Science, Artificial Intelligence

Contributor: Thai La , Linh Tao , Chanh Minh Tran , Tho Nguyen Duc , Eiji Kamioka , Phan Xuan Tan , Thai Quang La

Stereo cameras allow mobile robots to perceive depth in their surroundings by capturing two separate images from slightly different perspectives. This is necessary for tasks such as obstacle avoidance, navigation, and spatial mapping.

mobile robots depth estimation stereo camera

1. Introduction

Mobile robots have witnessed a surge in popularity and find versatile applications in numerous fields [1]. One compelling use case for mobile robots is their deployment in hazardous environments, such as automated agriculture and the handling of dangerous materials, where they can replace human workers [2]. However, to ensure optimal performance, it is imperative for mobile robots to swiftly and accurately gauge the geometric attributes of their surroundings, specifically the depth information. Depth estimation plays a pivotal role in enabling mobile robots to excel in various tasks. It empowers these robots with the capability to perform obstacle detection [3], construct detailed environmental maps [4], and facilitate object recognition [5]. One of the potential solutions for depth estimation is stereo matching [6]. Stereo matching is a computer vision technique that simulates human vision by analyzing a pair of 2D images captured from slightly different viewpoints to reconstruct 3D scenes. The primary objective of stereo matching is to establish correspondences between pixels in these input 2D images and, subsequently, to compute the corresponding depth values for each pixel. This process is executed by identifying the disparity, which denotes the horizontal displacement between correspondences in the 2D images [7]. The accurate calculation of this disparity is instrumental in calculating the depth information, thereby empowering mobile robots to navigate, interact with, and operate effectively in their surroundings.

In order to accurately determine the disparity, recent studies have applied deep learning methods and achieved promising results [8]. Particularly, these works first used a convolutional neural network (CNN) to extract features from 2D images, then concatenate them and store the disparity values between them to construct a 4D cost volume ($height \times width \times disparity \times features$). Then, the 4D cost volume is input in a 3D CNN model for regularization into a 3D cost volume ($height \times width \times disparity$). Finally, the predicted disparity is regressed from the cost volume via a softmax operation (σ) [9].

For example, GC-Net [9] proposes to learn the context of cost volume through an encoder–decoder 3D CNN architecture. PSMNet [10] utilizes a feature extractor with a spatial pyramid pooling module and regularizes the cost

volume using a 3D CNN based on stacked hourglass architecture. GA-Net [11] incorporates a 3D CNN with semiglobal matching for cost filtering. These approaches have demonstrated cutting-edge performance in stereo matching. Despite the high accuracy, when applying these methods to mobile robots, which often have low computational power, the computational cost is also a critical challenge.

As reported in [12], PSMNet [10] could only run at approximately 0.16 frames per second (fps) on an NVIDIA Jetson TX2 module. Similarly, although it has been proposed specifically for mobile robots, StereoNet [13] could only provide fewer than 2 fps on the same device. Such performances are far from the requirement for real-time applications in mobile robots, which is often a minimum of 30 fps [14].

Recently, the authors of [12] proposed attention-aware feature aggregation (AAFS) to obtain a better tradeoff between computational time and accuracy for real-time stereo matching on edge devices. The authors reported that AAFS could run at a maximum frame rate of 33 fps on low-budget devices, such as an NVIDIA Jetson TX2 module. However, the accuracy of AAFS is still limited due to the fact that it cannot efficiently exploit the contextual information of stereo images. The reason is that AAFS attempts to not increase the number of feature maps in its cascaded 3D CNN to limit the computational cost. In this case, leveraging the idea of a deep convolutional encoder–decoder, which is intended for dense prediction tasks, is a potential solution. Deep encoder–decoder tasks could reduce the computational cost by compressing the input data, then decoding the compressed data back to the input data dimension [15]. For example, a stacked hourglass based on a deep encoder–decoder consists of hourglass blocks that apply two-stride 3D convolutions to reduce the cost volume size by a factor of four [16]. This allows for an increase in feature dimensions with little impact on computational resources. Then, 3D transposed convolutions are applied to decode the volume to the original dimension.

2. Hourglass 3D CNN for Stereo Disparity Estimation for Mobile Robots

Zbontar et al. originally introduced a CNN-based stereo-matching technique [17] whereby the similarity metric of tiny patch pairings was learned using a convolutional neural network. GCNet [9] was one of the first methods incorporating 4D cost volume, using the soft argmin operation in the disparity regression steps to obtain the best matching disparity. PSMNet [10] introduced a spatial pyramid pooling module and 3D stacked hourglass networks and yielded promising results. The authors of [18] proposed GwcNet, which is a modified 3D stacked hourglass architecture, and a combined 3D cost volume based on group-wise correlation. GA-Net [11] includes a semiglobal aggregation layer and a local guided aggregation layer to replace several 3D convolution layers. To replace the 3D architecture, AANet [19] includes an intrascale and cross-scale cost aggregation algorithm, which can reduce inference time and maintain equivalent accuracy. On the other hand, DeepPruner [20], a coarse-to-fine approach, includes a differentiable PathMatch-based module to estimate the pruned search range of each pixel. Although 4D cost volume-based methods have achieved promising results, they operate at high computational cost and do not accommodate real-time operation on low-budget devices.

Therefore, some recent studies have focused on lightweight stereo networks based on 4D cost volumes to achieve real-time performance while maintaining competitive accuracy. These methods typically construct and aggregate 3D cost volume at low resolution to significantly reduce computational cost. For instance, StereoNet [13] is an edge-preserving refinement network that utilizes the left images as guidance to recover high-frequency details. Gu et al. [21] proposed a cascade cost volume, which consists of two stages. Cost volume at the early stage is built on a low-resolution feature map. Then, the later stage used the estimated disparity maps from the earlier stage to construct new cost volumes to apply better semantic features. This leads to a remarkable improvement in GPU memory consumption and computation time. AAFS [12] constructs a 4D cost volume by adopting a distance metric (height × width × disparity × 1). A disparity map is then computed at the lowest resolution, and disparity residuals are computed in later stages. However, its 3D CNN cannot exploit the contextual information for cost volume regularization, resulting in a limitation in estimation accuracy.

References

1. Campbell, S.; O'Mahony, N.; Carvalho, A.; Krpalkova, L.; Riordan, D.; Walsh, J. Path Planning Techniques for Mobile Robots A Review. In Proceedings of the 2020 6th International Conference on Mechatronics and Robotics Engineering (ICMRE), Barcelona, Spain, 12–15 February 2020; pp. 12–16.
2. Vasic, M.; Billard, A. Safety Issues in Human-robot Interactions. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 6–10 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 197–204.
3. Wei, B.; Gao, J.; Li, K.; Fan, Y.; Gao, X.; Gao, B. Indoor Mobile Robot Obstacle Detection Based on Linear Structured Light Vision System. In Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics, Bangkok, Thailand, 22–25 February 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 834–839.
4. Amigoni, F.; Caglioti, V. An information-based exploration strategy for environment mapping with mobile robots. Robot. Auton. Syst. 2010, 58, 684–699.
5. Ekvall, S.; Jensfelt, P.; Kragic, D. Integrating Active Mobile Robot Object Recognition and Slam in Natural Environments. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 5792–5797.
6. Murray, D.; Little, J.J. Using Real-time Stereo Vision for Mobile Robot Navigation. Auton. Robot. 2000, 8, 161–171.
7. Scharstein, D.; Szeliski, R.; Zabih, R. A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms. In Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), Kauai, Hawaii, 9–10 December 2001; pp. 131–140.

8. Flynn, J.; Neulander, I.; Philbin, J.; Snavely, N. Deep Stereo: Learning to Predict New Views from the World's Imagery. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5515–5524.
9. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 66–75.
10. Chang, J.R.; Chen, Y.S. Pyramid Stereo Matching Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5410–5418.
11. Zhang, F.; Prisacariu, V.; Yang, R.; Torr, P.H. GA-Net: Guided Aggregation Net for End-to-end Stereo Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 185–194.
12. Chang, J.R.; Chang, P.C.; Chen, Y.S. Attention-Aware Feature Aggregation for Real-Time Stereo Matching on Edge Devices. In Proceedings of the Computer Vision–ACCV 2020; Springer International Publishing: Cham, Switzerland, 2021; pp. 365–380.
13. Khamis, S.; Fanello, S.; Rhemann, C.; Kowdle, A.; Valentin, J.; Izadi, S. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
14. Poppinga, B.; Laue, T. JET-Net: Real-time Object Detection for Mobile Robots. In Proceedings of the RoboCup 2019: Robot World Cup XXIII 23; Springer: Berlin/Heidelberg, Germany, 2019; pp. 227–240.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2481–2495.
16. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VIII 14. Springer: Berlin/Heidelberg, Germany, 2016; pp. 483–499.
17. Zbontar, J.; LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Research 2016, 17, 1–32.
18. Guo, X.; Yang, K.; Yang, W.; Wang, X.; Li, H. Group-wise Correlation Stereo Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3273–3282.

19. Xu, H.; Zhang, J. AANet: Adaptive Aggregation Network for Efficient Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1959–1968.
20. Duggal, S.; Wang, S.; Ma, W.; Hu, R.; Urtasun, R. DeepPruner: Learning Efficient Stereo Matching via Differentiable PatchMatch. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 4383–4392.
21. Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; Tan, P. Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2492–2501.