Stereo Matching Algorithm

Subjects: Computer Science, Artificial Intelligence

Contributor: Zhenhua Quan , Bin Wu , Liang Luo

With the advancement of artificial intelligence technology and computer hardware, the stereo matching algorithm has been widely researched and applied in the field of image processing. In scenarios such as robot navigation and autonomous driving, stereo matching algorithms are used to assist robots in acquiring depth information about the surrounding environment, thereby improving the robot's ability for autonomous navigation during self-driving.

stereo matching deep learning attention mechanism

1. Introduction

The binocular stereo vision system captures images using a pair of cameras. The frame rate of the cameras is selected based on the speed of the scene, and high- and low-resolution images are acquired accordingly. Preprocessing techniques such as filtering are applied to the images before performing stereo matching [1]. Stereo matching utilizes the disparity information obtained from the differences between the two images to calculate depth and other related information. With the advancements in artificial intelligence technology and computer hardware, stereo matching algorithms have been extensively researched and applied. In scenarios such as robot navigation [2] or autonomous driving [3], stereo matching algorithms can assist robots in obtaining depth information about the surrounding environment, thereby enhancing their autonomous navigation capabilities.

In 2002, Scharstein et al. [4] proposed a four-step framework for stereo matching, which includes cost computation, cost aggregation, disparity calculation, and disparity optimization. This framework has been widely adopted and remains in use to this day. Traditional stereo matching algorithms can be categorized into local matching, global matching, and semi-global matching based on the pixel range they process. Local matching algorithms compare a small neighborhood around each pixel in one image with the corresponding region in the other image. The most similar region is identified, and its center point is considered as the matching point. Global matching algorithms compute a cost map between a pair of stereo images and use techniques such as dynamic programming or energy-based algorithms to find an optimal path in the cost map, which represents the matching point of each pixel in the other image, resulting in a disparity map. On the other hand, semi-global matching algorithms, which are computationally efficient, calculate the cost between each pixel in the stereo image pair and all possible matching points in the other image. Cost aggregation techniques are then applied to aggregate the costs in four directions, resulting in the cost sum for each pixel in each direction. Finally, the disparity map is computed based on these cost sums. However, these traditional stereo matching algorithms suffer from limitations such as poor performance in complex scenes, slow computation speed, and low matching accuracy.

With the improvement of computing power and the increase in data volume, convolutional neural networks have brought more possibilities for solving the problem of stereo matching and have gradually become a research focus [5]. Deep learning techniques optimize algorithm performance through a large amount of image data [6], achieving optimal results by autonomously learning and optimizing representations, thereby improving the accuracy and robustness of stereo matching. The stereo matching algorithm based on deep learning extracts multiple features through multi-layer convolution for cost computation and uses regularization methods for cost aggregation to optimize the cost volume, thereby reducing the mismatch rate and improving the matching speed.

2. Stereo Matching Algorithm

A mathematical model is constructed for the global algorithm to build an energy function, utilizing methods such as Newton's method and gradient descent to minimize the energy function and find the optimal matching solution. Shahbazi et al. [7] from the University of Calgary employed the non-parametric census transform as the data term and used the geometric features of intrinsic curves as the smoothness term to minimize the energy function, thereby improving the problem of a high mismatch rate caused by image occlusion. Zhou Jiali et al. [8] from Zhejiang University of Technology based their approach on conventional graph cut algorithms. They corrected the matching region based on the labels of the matching blocks and spatial geometric information, continuously updating the selected labels to find the labels that minimize the global energy. By applying the left–right consistency criterion and mean filtering to refine the disparity map, they achieved higher sub-pixel-level accuracy in the matching disparity. The matching accuracy in the edge and occlusion regions of the image pair was significantly improved.

The matching cost convolutional neural network (MC-CNN) was proposed by Zbontar et al. [9] of the University of Ljubljana. In the cost-computation stage, a convolutional neural network is used to train on input image patches and annotated images, obtaining matching cost values. These values are then optimized using a cross-cost aggregation method, incorporating a left–right consistency check and bilateral filtering for refining the disparity map. Seki et al. [10] from Toshiba Corporation introduced the SGMNet algorithm, which utilizes a convolutional neural network to automatically learn penalty parameters, eliminating the manual adjustment process in traditional SGM algorithms. This algorithm divides the disparity transition along the scanline into positive and negative disparity transitions based on different occlusion relationships between objects, ensuring good disparity prediction even in pathological regions. However, non-end-to-end stereo matching algorithms often fail to meet practical requirements due to their high time complexity. They remain within the traditional stereo matching framework and require significant time and effort for parameter adjustments, such as filter size and matching window size.

The end-to-end stereo matching algorithm learns the features in the input images adaptively through a deep learning model, eliminating the need for manual feature design and selection. The model is more adaptable to different scenes and details. Mayer et al. [11] constructed a virtual synthetic dataset and proposed the end-to-end network DispNetC. This network introduced an autoencoder–decoder structure, taking in left and right images and outputting a disparity map without any post-processing steps. Xu et al. [12] from the University of Science and Technology of China extracted features of different resolutions using a shared feature pyramid network. They

designed three same-scale aggregation modules to optimize different resolution features and proposed the adaptive aggregation network (AANet), which fuses features through a cross-scale aggregation module to address the issues of large parameter and computational requirements in deep stereo matching networks, thereby improving algorithm efficiency. Vladimir et al. [13] from Google introduced HITNet, a convolutional neural network for real-time stereo matching. It infers disparity through fast multi-resolution initialization and transformation, without incurring significant costs. By propagating information across levels, it reduces algorithm complexity while improving matching accuracy. Tang Haifeng et al. [14] from Inner Mongolia University proposed DFFNet, an end-to-end stereo matching network that incorporates dense feature fusion. They utilized multiple residual modules to construct a feature pyramid network, capturing multi-scale contextual information with a low parameter count. The network enhanced its matching capability in complex regions such as weak texture areas and edges through dense fusion modules and mixed attention modules, thereby improving the extraction of useful information.

References

1. Mur-Artal, R.; Tardos, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. IEEE Trans. Robot. 2017, 33, 1255–1262.
2. Goldberg, S.B.; Maimone, M.W.; Matthies, L. Stereo vision and rover navigation software for planetary exploration. In Proceedings of the IEEE Aerospace Conference, Big Sky, MT, USA, 9–16 March 2002; IEEE: Piscataway, NJ, USA, 2002.
3. Li, H.; Xu, C.; Xiao, Q.; Xu, X. Visual navigation of an autonomous robot using white line recognition. In Proceedings of the IEEE International Conference on Robotics and Automation, Taipei, Taiwan, 14–19 September 2003; IEEE: Piscataway, NJ, USA, 2003.
4. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. 2002, 47, 7–42.
5. Ning, I. Research on Target Pose Measurement Technology Based on Monocular Vision. Ph.D. Thesis, Beijing Institute of Technology, Beijing, China, 2016.
6. Knöbelreiter, P.; Pock, T. Learned collaborative stereo refinement. Int. J. Comput. Vis. 2021, 129, 2565–2582.
7. Shahbazi, M.; Sohn, G.; Théau, J. High-density stereo image matching using intrinsic curves. ISPRS J. Photogramm. Remote Sens. 2018, 146, 373–388.
8. Zhou, J.; Yu, C.; Chao, W. Binocular stereo matching algorithm based on labeled matching region correction. Pattern Recognit. Artif. Intell. 2020, 33, 11.
9. Zbontar, J.; LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. 2016, 17, 2287–2318.

10. Seki, A.; Pollefeys, M. Sgm-nets: Semi-global matching with neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
11. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
12. 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, 13–19 June 2022.
13. Tankovich, V.; Hane, C.; Zhang, Y.; Kowdle, A.; Fanello, S.; Bouaziz, S. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
14. Tang, H. Research on Optimization of End-to-End Binocular Stereo Matching Algorithm Based on Convolutional Neural Network. Ph.D. Thesis, Inner Mongolia University, Hohhot, China, 2022.