Stereo Vision Odometry Using Deep Learning

Vicknesh Balabaskaran
Systems Design Engineering
University of Waterloo
vbalabaskaran@uwaterloo.ca

Harish Krishnamoorthy Murali
Systems Design Engineering
University of Waterloo
hkrishna@uwaterloo.ca

Pranav Kumar Ayee Goundar Venkatesan
Systems Design Engineering
University of Waterloo
pkayeego@uwaterloo.ca

Abstract

This study presents a deep learning-based approach for stereo-visual odometry estimation. While conventional methods dominate the field, research on deep learning applications remains scarce. Our proposed algorithm is evaluated using the KITTI stereo odometry dataset and compared against traditional monocular and traditional stereo-visual odometry techniques to assess its efficacy. Results demonstrate the superiority of our approach over traditional methods, highlighting its potential for advancing stereo-visual odometry methodologies.

1 Introduction

Visual odometry (VO) plays a crucial role in robotics and computer vision, providing essential insights into a camera's motion relative to its environment. Traditionally, VO methodologies have been classified into geometric and learning-based approaches, each offering unique perspectives on understanding camera motion dynamics.

Geometric methods rely on feature extraction, stereo matching, and triangulation, serving as the foundation of visual odometry. These techniques excel in extracting sparse features from images and estimating camera motion through geometric relationships. However, they often face challenges in dynamic environments and regions lacking distinctive textures.

The emergence of learning-based approaches, particularly leveraging deep learning techniques, has revolutionized visual odometry. By employing neural networks to learn motion patterns from raw image data directly, these methods eliminate the need for manual feature engineering, offering scalability and robustness across diverse environments. Supervised and unsupervised learning methods further extend the capabilities of monocular visual odometry, presenting opportunities for accurate motion estimation without reliance on ground truth labels.

Despite significant progress in monocular visual odometry with deep learning, the application of these techniques in stereo-visual odometry remains relatively underexplored. This paper addresses this gap by exploring deep-learning methodologies for stereo-visual odometry. Capitalizing on the inherent advantages of stereo imagery, such as enhanced depth perception and scene comprehension, this study seeks to enhance the accuracy and reliability of visual odometry systems. Through this endeavor, contributions are made to advancing stereo-visual odometry techniques, facilitating the development of robust and efficient navigation systems for various real-world applications.

The paper is organized as follows. Section two discusses the literature review. Section three analyses the dataset employed for stereo vision odometry. Section four describes the different methodologies employed in this paper. Section five analyses the results of performing the traditional and deep learning methods. Section six concludes the paper. Section Seven discusses the Future scope of this research. Lastly, Section Eight states the contribution of different authors in this research.

2 Literature Review

This section provides an overview of methodologies utilized in visual odometry (VO) research, distinguishing between two primary approaches: geometry-based and learning-based methods. These classifications illuminate the varied strategies employed to comprehend and quantify camera motion dynamics.

2.1 Based on Geometry

Traditional visual odometry methods have historically relied on geometric principles to estimate camera motion by analyzing image sequences. These methods typically fall into two main categories: feature-based and direct methods. Feature-based approaches, such as MonoSLAM [2] and ORB-SLAM 2 [6], extract and track sparse visual features across consecutive frames, facilitating motion estimation through feature matching and triangulation. While computationally less expensive, these methods struggle in textureless regions. In contrast, direct methods like DTAM [7] and SVO [3] avoid feature extraction and directly optimize photometric error functions to estimate camera pose and depth. This approach makes them robust in textureless regions but computationally intensive. Despite their effectiveness, they often encounter challenges in dynamic environments and necessitate manual intervention for initialization and error correction.

2.2 Based on Learning

Conversely, learning-based monocular visual odometry utilizes deep learning techniques to regress camera motion directly from image sequences. Methods like DeepVO [12] and UnDeepVO [13] employ deep neural networks to learn motion patterns from raw image data, eliminating the need for manual feature engineering. Trained on large-scale datasets, these models capture complex motion dynamics and generalize well across diverse environments, offering enhanced scalability and robustness compared to feature-based approaches.

Within monocular learning-based visual odometry, a further classification can be made between supervised and unsupervised methods. Supervised methods, exemplified by DeepVO, require ground truth motion labels during training to supervise network learning, enabling accurate motion estimation but necessitating labeled training data. In contrast, unsupervised methods, such as UnDeepVO, D3VO [11], and GeoNet [14], learn motion estimation from unlabeled image sequences by enforcing geometric and photometric consistency constraints between consecutive frames. Although they do not require ground truth labels, they may sacrifice accuracy.

Despite significant advancements in monocular visual odometry leveraging deep learning techniques, the utilization of deep learning approaches specifically for stereo visual odometry remains relatively underexplored. "StereoVO: Learning Stereo Visual Odometry Approach Based on Optical Flow and Depth Information" stands out as a pioneering work in this domain, introducing a novel deep learning framework explicitly tailored for stereo odometry, bridging the gap between traditional stereo methods and modern deep learning techniques.

3 About the Dataset

The KITTI dataset [4] is widely used in computer vision, especially for research related to autonomous driving. It offers various challenges, covering object detection, tracking, and understanding scenes. Developed jointly by the Karlsruhe Institute of Technology and the Toyota Technological Institute in Chicago, this dataset is derived from a platform built for autonomous driving.

The KITTI dataset stands out because of its wide variety of sensor data, including stereo cameras, LiDAR, and GPS/INS sensors. Collected over multiple days in urban areas of Germany, it comprises over 200,000 stereo images, point clouds, and precise GPS/INS data, providing detailed location and orientation information.

The dataset's strengths lie in its accuracy and size. Its sensors capture data with high precision, enabling precise object detection and tracking. Moreover, with over 50 GB of data, including stereo images, point clouds, and GPS/INS data, it offers ample resources for training deep neural networks, which excel with large datasets. Split into various categories like object detection, tracking, and scene understanding, each section offers unique challenges for researchers to evaluate algorithms and compare results.

This paper uses a subset of the KITTI dataset called "SLAM Evaluation 2012". This dataset comprises a sequence of images, ranging between approximately 800 to 4000 frames, captured simultaneously by both left and right cameras. Supplementary sensors, including Velodyne, GPS, and IMU, are utilized to capture Ground Truth, which is available in the poses.txt file. Additionally, the dataset includes calibration data for the cameras provided in a calib.txt file, facilitating the determination of the distance between the two cameras at approximately 0.38m. A times.txt file accompanies the dataset, providing timestamps for each captured image.

4 Methodology

Stereo vision odometry traditionally relies on established techniques such as feature detection, stereo matching, triangulation, and motion estimation. These methods utilize handcrafted features and algorithms to extract disparity information and estimate depth and camera motion from stereo image pairs. In contrast, deep learning-based approaches revolutionize stereo vision odometry by employing neural networks to learn feature representations from data automatically. In the subsequent sections of this paper, we delve into the methodologies used for calculating stereo vision odometry, covering both traditional and deep learning-based approaches.

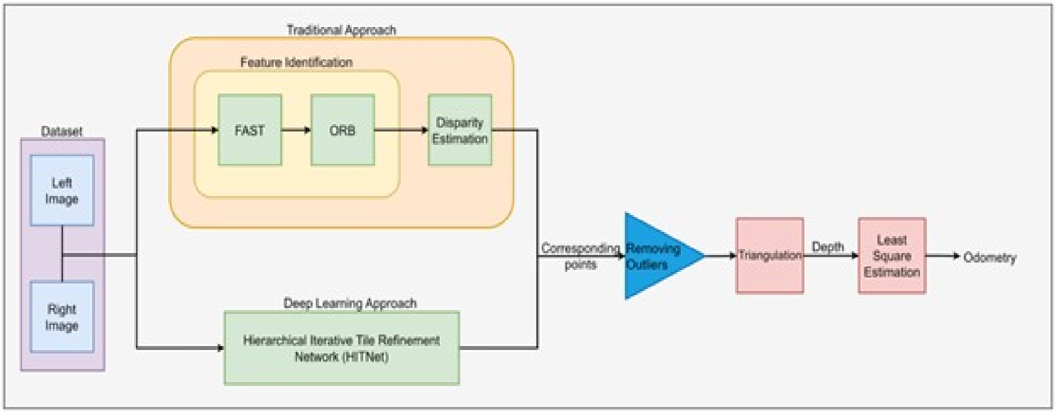


Figure 1: Flow Chart for Stereo Vision Odometry

4.1 Traditional Stereo Vision Odometry

The traditional method of stereo vision odometry relies on established techniques such as feature detection, stereo matching, and motion estimation to extract depth information and estimate camera motion from stereo image pairs. Below, we will provide a detailed breakdown of the steps involved in the traditional method.

1. **Stereo Image Pair Acquisition:** The process initiates acquiring stereo image pairs using a stereo camera setup. Stereo cameras have a known baseline distance between them to capture images from slightly different viewpoints, simulating human binocular vision. This paper uses the KITTI dataset, as section 3 mentioned.
2. **Feature Detection and Matching:** Traditional methods use feature-based techniques to extract distinctive points or keypoints from the stereo image pairs. In this paper, FAST (Features from Accelerated Segment Test) [8] and ORB (Oriented Rotated BRIEF) [9] are employed for this purpose. Initially, the FAST algorithm uses the stereo pair images as input data to detect keypoints in both images and then match corresponding keypoints

between the stereo pairs. ORB algorithm then uses the stereo pair images and the previously obtained keypoints as input data and computes the updated keypoints along with descriptors. ORB is an efficient alternative to SIFT [5] or SURF [1] algorithms for feature extraction, computation cost, and matching performance.

3. **Disparity Estimation:** Once corresponding keypoints are identified, the next step involves estimating the disparity map between the stereo image pairs. OpenCV is employed for this purpose. It compares image patches or windows between the left and right images to determine the disparity values for each pixel.

4. **Triangulation:** Using the estimated disparity information, triangulation is performed to reconstruct the 3D structure of the scene. Triangulation involves determining the 3D coordinates of points in the scene by intersecting corresponding rays originating from the stereo camera positions with the estimated disparity values. Equation 1 determines the depth of each pixel in the image.

   $$Z = \frac{Bf}{\delta} \tag{1}$$

   Where Z is the depth of the 2D pixel from the image plane in the 3D world coordinate system, B represents the baseline distance between the stereo pair cameras, f represents the focal length of both cameras and δ denotes the disparity value of each pixel. This process results in generating a 3D point cloud representing the observed environment.

5. **Motion Estimation:** Once the 3D point cloud is generated for consecutive frames, motion estimation techniques are applied to estimate the camera or vehicle motion between frames. In this paper, Least Square optimization [10] is employed to compute the odometry of the vehicle. This method compares the 3D point clouds from successive frames to estimate the relative pose change, thus providing odometry information. Additionally, the least square method is employed multiple times for each frame, and the error in each time is computed. Ultimately, the output with the minimum error is selected as the result.

Figure 1 describes the entire process followed by the traditional approach in obtaining the odometry from the stereo pair images. By following this methodology, stereo vision odometry using traditional methods can achieve accurate motion estimation. This approach relies on well-established techniques such as FAST and ORB to extract disparity information and estimate camera motion from stereo image pairs.

4.2 Traditional Monocular Odometry

Monocular odometry is a technique used in computer vision and robotics to estimate the motion of a camera using only a single camera sensor. It relies on extracting features from consecutive frames captured by the camera and then computing the camera's motion by tracking these features over time. Unlike stereo visual odometry, which utilizes information from two or more cameras to estimate depth and motion, monocular odometry operates with a single camera. This makes it more lightweight and cost-effective in terms of hardware requirements, as it doesn't necessitate additional cameras. However, monocular odometry typically struggles with accurately estimating scale and depth due to the lack of stereo information, leading to potential drift in longer-term motion estimation compared to stereo methods.

4.3 Deep Learning Method

4.3.1 Initial steps

The deep learning method for stereo vision odometry harnesses neural networks to learn feature representations from stereo image pairs automatically. These networks are trained end-to-end to directly estimate disparity maps, eliminating the need for explicit feature engineering.

The stereo images from the KITTI dataset have been utilized as input. Later, a deep learning approach is used to obtain the disparity between the stereo image pairs. This paper employs the HITNet model, a neural network architecture for real-time stereo matching.

4.3.2 HITNet model

HITNet is a neural network model designed for stereo-matching tasks, particularly in the context of stereo-vision odometry. It leverages the planar geometry of the scene as an inductive prior in its network design, guiding stereo predictions using predicted tiles. HITNet employs a tile-based method to decide if each pixel lies on a plane, facilitating learning.

Additionally, HITNet maintains a running stereo prediction at full resolution, optimizing memory efficiency while enabling accurate depth estimation for megapixel images. HITNet combines feature extraction and disparity estimation between stereo image pairs, which multiple algorithms have done using traditional methods.

4.3.3 Estimating Odometry

After obtaining the disparity map using the HITNet model explained in Section 4.3.2, the obtained disparity map is used to obtain the depth for each pixel from the stereo pair images using triangulation. Finally, Least Square optimization estimates the camera or vehicle motion between frames. The steps in estimating the odometry after estimating the disparity follow the same method as in Steps 4 and 5 of traditional methods in Section 4.1.

Figure 1 describes the entire process followed by the deep learning approach in obtaining the odometry from the stereo pair images. By following this methodology, stereo vision odometry using deep learning methods can achieve accurate motion estimation. This approach adopts techniques such as HITNet to extract disparity information and estimate camera motion from stereo image pairs.

5 Results

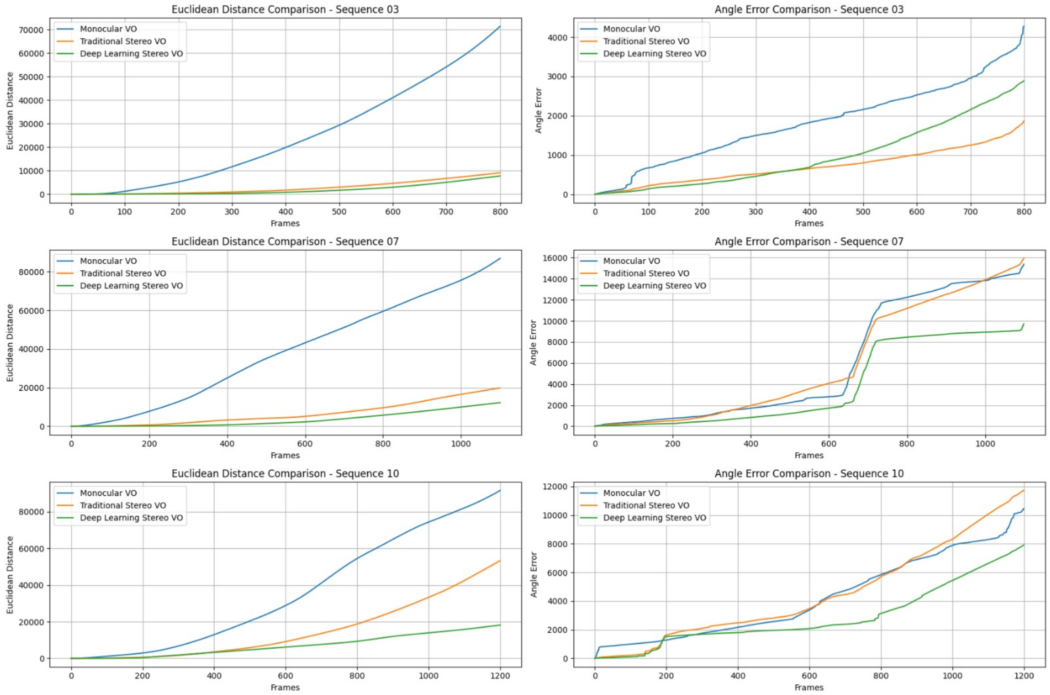


Figure 2: Error Plots

Three different methods are implemented in this research. Namely,

1. Traditional Monocular Visual Odometry
2. Traditional Stereo Visual Odometry
3. Deep learning Stereo Visual Odometry

Table 1: Error Metrics for Different Methods

Test cases (Seq no.)	Traditional Monocular		Traditional SVO		HitNet SVO	
	Euclidean	DevAng	Euclidean	DevAng	Euclidean	DevAng
Test case 1 (03)	71466.3	4276.63	9124.8	1862.8	7409.8	2943.61
Test case 2 (10)	91564.68	10455.04	76670.33	14198.46	34593.59	7362.4
Test case 3 (07)	86935.47	15384.47	21454.03	14312.11	12114.14	9830.39

Table 2: Percentage Improvement in Error Metrics

Test cases (Seq no.)	Traditional SVO		HITNet SVO	
	Euclidean	DevAng	Euclidean	DevAng
Test case 1 (03)	87.23	89.63	56.44	31.17
Test case 2 (10)	16.27	62.22	-35.80	29.58
Test case 3 (07)	75.32	86.07	6.97	36.10

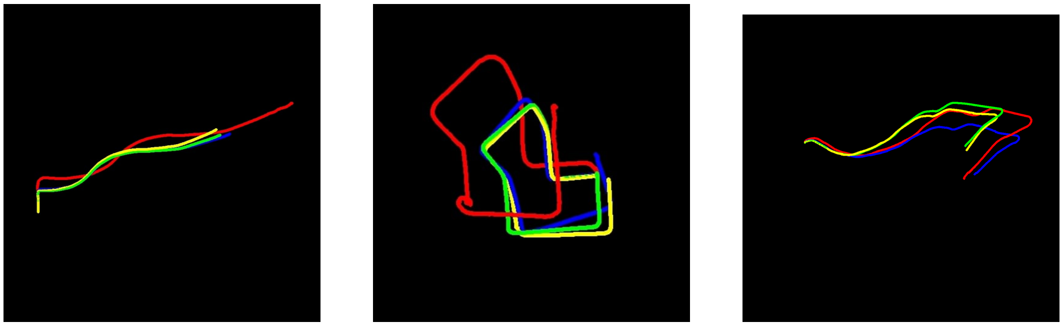
This paper employs the Hierarchical Iterative Tile Refinement Network (HITNet) to detect the disparity between stereo images in the deep learning approach. Subsequently, the identified disparity is leveraged to estimate the depth and derive the vehicle's odometry. Pre-trained weights were utilised for HITNet model in the identification of disparity.

5.1 Implementation details

For the experiments, PyTorch framework was employed in the calculation of odometry. Instead of training the deep learning model, pre-trained models were used in this research to identify odometry. Two different metrics were used to compare the performance of the models, namely Absolute Trajectory Error (ATE) and Angle of Deviation (DevAng).

5.2 Comparision with different methods

Using the evaluations of three different models that was done using test sequences, the deep learning based SVO performs better in terms of reducing the error metrics and seems to be better at approximating the ground truth odometry. For each sequence experimented we plot the odometry resulted from all three methods and the ground truth to visually compare how close the SVOs approximate ground truth in Figure 2, 3. We see that the error metrics are comparatively lesser for HitNet SVO from Table 1. It can also be observed that HITNet improves the performance of the SVO from the percentage improvement in error metrics as shown in Table 2.



(a) Plot 1 (03)

(b) Plot 2 (07)

(c) Plot 3 (10)

Figure 3: Odometry plots for different models: Green line represents Ground truth, Red represents Traditional Monocular Odometry, Blue represents Traditional SVO and Yellow represents HITNet (Deep learning based) SVO

6 Conclusion

This paper presents a novel deep learning-based approach to stereo-visual odometry, aiming to improve accuracy and robustness in motion estimation. Our methodology demonstrates significant advancements over traditional techniques by leveraging stereo imagery and employing advanced neural network architectures. Through extensive experimentation on the KITTI dataset, we have shown that our approach achieves superior performance in terms of accuracy and reliability.

7 Future Work

The experiments conducted in this study demonstrate the superiority of our proposed approach over traditional methods. Recently, a novel technique integrating deep learning into Stereo Visual Odometry, termed "StereoVO: Learning Stereo Visual Odometry Approach Based on Optical Flow and Depth Information," has been introduced. Future research endeavors will focus on comparing the outcomes of our methodology with StereoVO to provide a more accurate assessment of our approach's efficacy. Such a comparative analysis will offer valuable insights into both methods' relative strengths and weaknesses, contributing to the advancement of stereo-visual odometry techniques.

8 Contributions

The project work was split equally among the teammates and equal contribution was made by each teammate. Vicknesh worked on the implementation of disparity map calculation using traditional method while Harish and Pranav concentrated on deep learning model selection and implementation. We met as a team on regular basis 3 times a week to touch base on progress made by each teammate and to decide on further steps in the project. All 3 team members were involved in report writing and the project presentation.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. “Surf: Speeded up robust features”. In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I.* Vol. 9. Springer Berlin Heidelberg. 2006, pp. 404–417.

[2] Andrew J Davison et al. “MonoSLAM: Real-Time Single Camera SLAM”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (2007), pp. 1052–1067.

[3] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. “SVO: Fast Semi-Direct Monocular Visual Odometry”. In: *Proceedings of the IEEE International Conference on Robotics and Automation*. 2014.

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[5] David G Lowe. “Object recognition from local scale-invariant features”. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. IEEE. 1999, pp. 1150–1157.

[6] Ra’ul Mur-Artal, Juan D Tard’os, and J. M. M. Montiel. “ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras”. In: *IEEE Transactions on Robotics* 33.5 (2017), pp. 1255–1262.

[7] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. “DTAM: Dense Tracking and Mapping in Real-Time”. In: *Proceedings of the International Conference on Computer Vision*. 2011.

[8] Edward Rosten and Tom Drummond. “Machine learning for high-speed corner detection”. In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I.* Vol. 9. Springer Berlin Heidelberg. 2006, pp. 430–443.

[9] Ethan Rublee et al. “ORB: An efficient alternative to SIFT or SURF”. In: *2011 International conference on computer vision*. IEEE. 2011, pp. 2564–2571.

[10] Jos MF Ten Berge. *Least squares optimization in multivariate analysis*. Leiden: DSWO Press, Leiden University, 1993.

[11] Sen Wang, Ronald Clark, and Hongkai Wen. “D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry”. In: *arXiv preprint arXiv:1905.06316* (2019).

[12] Sen Wang, Ronald Clark, and Hongkai Wen. “DeepVO: Towards End-to-End Visual Odometry with Deep Recurrent Convolutional Neural Networks”. In: *Proceedings of the IEEE International Conference on Robotics and Automation*. 2017.

[13] Sen Wang, Ronald Clark, and Hongkai Wen. “UnDeepVO: Monocular Visual Odometry Through Unsupervised Deep Learning”. In: *Proceedings of the IEEE International Conference on Robotics and Automation*. 2018.

[14] Zhichao Yin et al. “GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1983–1992.