Stereo Vision SLAM with SuperPoint and SuperGlue

Si-Won Yoon, Soon-Yong Park

Graduate School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, South Korea,
(sivvon0308, sypark)@knu.ac.kr

Keywords: 2-view stereo matching, SLAM, Mapping, Stereo Camera

Abstract

This paper presents a method for stereo visual odometry and mapping that integrates VINS-Fusion-based visual odometry estimation with deep learning techniques for camera pose tracking and stereo image matching. Traditional approaches in the VINS-Fusion relied on classical methods for feature extraction and matching, which often resulted in inaccuracies in triangulation-based 3D position estimation. These inaccuracies could be mitigated by incorporating IMU-based position estimation, which yielded more accurate odometry estimates compared to using stereo camera only in three-dimensional space. Consequently, the original VINS-stereo algorithm necessitated a tightly-coupled integration of IMU sensor measurements with estimated visual odometry.
To address these challenges, our work proposes replacing the traditional feature extraction method used in VINS-Fusion, the Shi-Tomasi (Good Features to Track) technique, with feature extraction via the SuperPoint deep network. This approach has demonstrated promising experimental results. Additionally, we have applied deep learning models to the matching of feature points that project the same three-dimensional point to pixel coordinates in different images. Instead of using the KLT optical flow algorithm previously employed by VINS-Fusion, our proposed method utilizes SuperGlue, a deep graph neural network for graph matching, to improve image tracking and stereo image matching performance. The performance of the proposed algorithm is evaluated using the publicly available EuRoC dataset, providing a comparison with existing algorithms.

1. Introduction

Simultaneous Localization and Mapping (SLAM) (Dissanayake et al. in 2001) plays a pivotal role in a variety of applications including unmanned vehicles, agricultural robots, and AR/VR systems. This technology leverages an array of sensors such as cameras, IMUs, LiDAR, and GPS to enable robots to map their environments and ascertain their precise locations. SLAM technology that is particularly effective in indoor or urban areas where GPS signals are unreliable, makes it indispensable for advanced robotics navigation and mapping. This technology is crucial in environments where traditional GPS navigation is either ineffective or impractical. Consequently, methods that utilize only camera sensor for state estimation are garnering significant interest in this field.
The simplicity and accessibility of data collection with single-camera based Visual-SLAM, which does not require synchronization, are reasons why it is becoming increasingly prominent. However, using just a monocular camera for odometry estimation does not enable for the recovery of metric scale in three-dimensional space. Thereby complicating the estimation of the robot travel distance and rendering it challenging to implement in the actual robotic vision systems. A stereo vision system, on the other hand, capitalizes on the ability to simultaneously observe the same point in space from two cameras, enabling depth estimation through triangulation.
Nonetheless, if feature extraction within stereo images and stereo image matching are not executed precisely, the accuracy of depth estimation can be compromised, leading to imprecise recovery of metric scale and inaccurate estimation of the shape of trajectory.
To address these limitations, many robotic vision systems often employ auxiliary sensors capable of estimating the robot's position in three-dimensional space or measuring the distance to surrounding objects. These include RGB-D cameras, inertial measurement units (IMUs), 2D or 3D LiDAR, and GPS. The existing Visual-Inertial System (VINS) (Qin et al., 2018) and stereo vision SLAM algorithms also demonstrate that the use of an IMU, in conjunction with cameras, significantly reduces the error in robot odometry estimation compared to using stereo cameras alone. Therefore, achieving accurate trajectory estimation using only stereo cameras remains a significant challenge.
In the existing VINS-Fusion system that utilizes stereo cameras, the method for feature point extraction from stereo images employs the Shi-Tomasi (Good Features to Track) method (Shi et al., 1994), while the technique for matching feature points between images uses the KLT sparse optical flow algorithm (Lucas et al., 1981). These methods are general and traditional image processing techniques that computed by pixel intensity information within the image. This paper proposes enhancements to the method of feature point extraction within images and the method of matching feature points that project the same 3D world coordinates to 2D pixel coordinates in different images. To achieve this, we aim to apply techniques that utilize deep neural networks rather than conventional image processing methods. Therefore, what we propose in this paper is to replace the feature point extraction algorithm with the deep feature and descriptor detector known as the SuperPoint (DeTone et al., 2018), and the feature matching between images with the deep feature matcher known as the SuperGlue (Sarlin et al., 2020), to estimate visual odometry better.
We conducted experiments on (1) replacing only the feature point extraction method and (2) replacing both the feature point extraction and matching methods. In both scenarios, we observed superior experimental results compared to the existing VINS-Fusion system. The deep networks, SuperPoint and SuperGlue, pose challenges for real-time application within the VINS algorithm, thus our proposed system is not a real-time visual SLAM system like the current VINS-Fusion system. The proposed algorithm has been evaluated using the publicly available EuRoC dataset (Burri et al., 2016), allowing for a performance comparison with the traditional VINS algorithms in terms of odometry estimation.

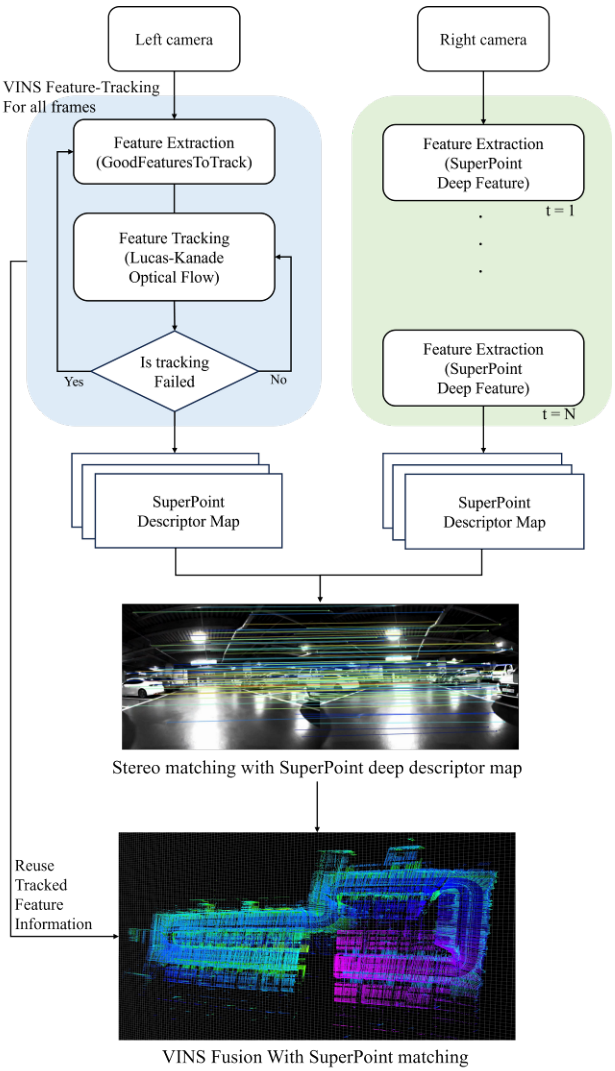

Figure 1. The flow diagram of the proposed stereo vision SLAM with SuperPoint deep feature and descriptor map.

RMSE (m) / EuRoC dataset	OURS (scale aligned)	OURS (scale not aligned)	Original VINS-Fusion RMSE
MH_01_easy	0.087923	0.107956	0.54
MH_02_easy	0.071845	0.072851	0.46
MH_03_medium	0.138786	0.150034	0.33
V1_01_easy	0.175089	0.197012	0.55

Table 1. The Root Mean Squared Error (RMSE) of EuRoC dataset with and without scale alignment. (VINS+SuperPoint)

While implemented as a non-realtime system, the proposed algorithm demonstrates robust performance in estimating visual odometry, not only in terms of the accuracy of the shape of trajectory but also in recovery of the 3D meteric scale. Therefore, it appears universally applicable to vision-based location estimation systems and also to autonomous driving systems in mobility platforms such as automobiles, mobile robots and drones. Future research will consider implementing the system in real-time. The structure of our paper is as follows: Chapter 2 briefly introduces related work similar to this paper. Chapter 3 describes how the proposed method was applied. Chapter 4 presents the experiments conducted. Finally, Chapter 5 outlines the conclusions.

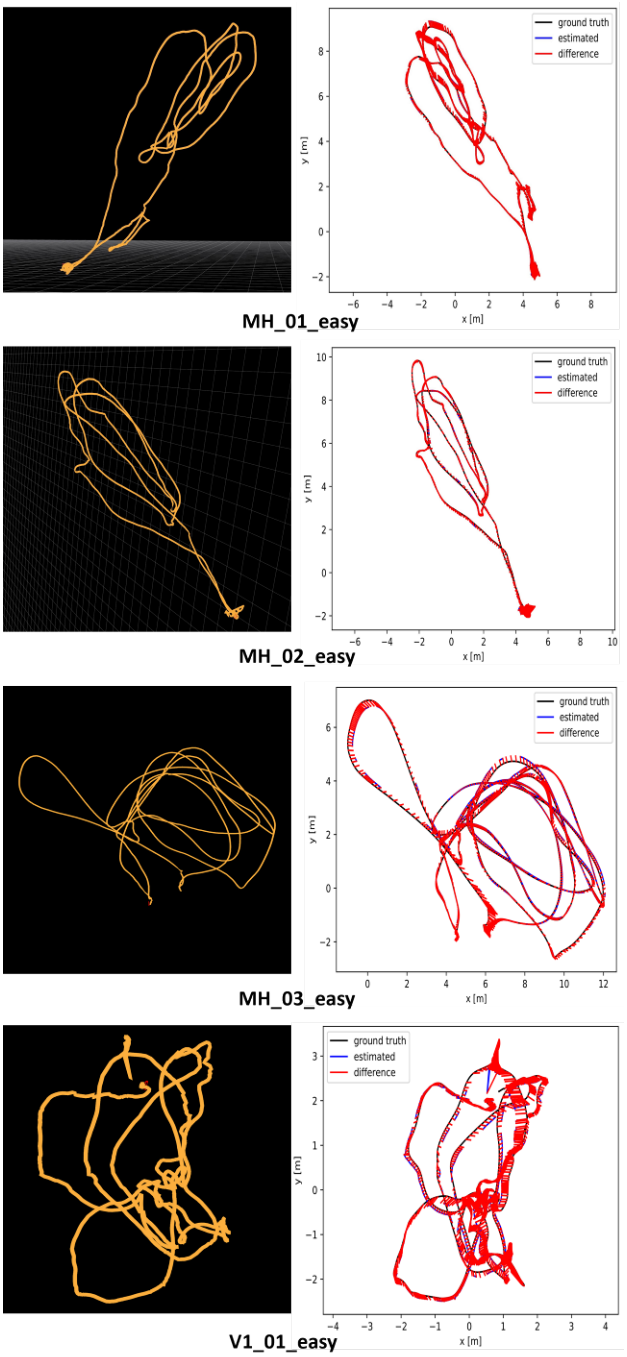

Figure 2. The estimated trajectory of EuRoC MAV dataset by the proposed algorithm(left), The comparision of EuRoC MAV dataset groundtruth and estimated trajectory(right).

2. Related work

In this section, we briefly introduce related research on estimating the visual odometry of mobile robots using stereo camera images.

2.1 Epipolar geometry for stereo image triangulation

A point in three-dimensional space projects onto different pixel coordinates in two-dimensional stereo images captured from different viewpoints. If the geometric relationship between the two cameras, described rotation and translation by the Essential matrix or the Fundamental matrix, is known, one can estimate information about the coordinates in three-dimensional space

using the projected points by pixel coordinates in the two images. Thus, by this principle, stereo imaging can be used to estimate the depth in three-dimensional space, which cannot be determined from monocular images.

2.2 VINS-based Vision Processing Front End

In the foundational VINS algorithm of the proposed system, feature points existing in each frame are tracked using the KLT sparse optical flow algorithm. In addition to the tracked feature points via optical flow, each image maintains a minimum count of approximately 100 to 300 feature points. To this end, new corner image feature points are detected using the Shi-Tomasi algorithm. This feature detector sets a minimum pixel distance between adjacent feature points to prevent the detection of too many features at close proximity. These detected image feature points exist as 2D pixel coordinates. They are then projected onto the unit sphere after being undistorted. Subsequently, outliers are eliminated by using the fundamental matrix model.
This process also plays a critical role in the optimization phase (Pose Graph Optimization) of the VINS algorithm. The VINS algorithm optimizes the entire trajectory by selecting keyframes and optimizing the world coordinates position of robot at the keyframe moments, rather than at every moment, using a loop closure module. This is closely related to the criteria for selecting keyframes, which are directly tied to the feature points within the image that are tracked and extracted.
In VINS front-end module, there are two primary criteria for selecting keyframes. The first is the average parallax between the previous keyframe and the current keyframe, which is related to the translation of the tracked image feature points. If the average parallax of the tracked feature points exceeds a threshold, that frame is designated as a new keyframe. The second criterion is related to the number of tracked feature points within the image. The greater the number of tracked feature points, the higher the quality of tracking is considered to be. Therefore, if the tracking quality falls below a certain level, it is determined that the frame represents a transition to a new scene and is treated as a new keyframe. Thus, it can be seen that accurately extracting and tracking feature points within images plays a pivotal role throughout the VINS-fusion process in correctly estimating visual odometry.
In the case of VINS-mono system, after undergoing the aforementioned front-end process, the limitation of being unable to recover depth using only a monocular camera is overcome by employing a tightly-coupled method that integrates visual odometry with an IMU sensor data. However, since our goal is to estimate visual odometry in the front-end of the visual SLAM algorithm using only stereo cameras and not an IMU, our objective is to estimate scale solely through stereo camera by tracking mono image and triangulating stereo images. Consequently, correctly aligning stereo images captured by stereo cameras is crucial for accurately recovering metric scale. For an image processing perspective, matching stereo images and tracking monocular images are fundamentally similar processes. In both cases, it is essential to identify and accurately match feature points that project the same 3D world coordinates to 2D pixel coordinates within the images.

2.3 SuperPoint and SuperGlue : Image Feature Extraction and Matching

The SuperPoint deep learning network is a self-supervised manner using the MS-COCO image dataset which is initially pre-trained by Synthetic Shapes dataset. Unlike conventional patch-based neural networks, this model is capable of generating feature point and descriptor map for the entire image.

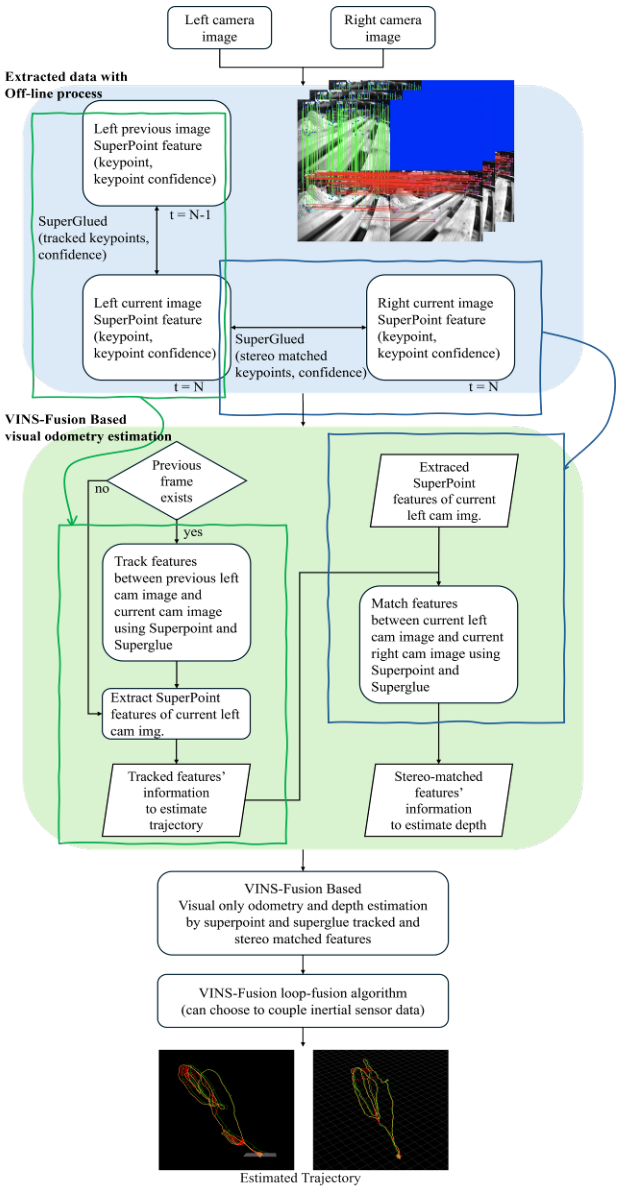

Figure 3. The flow diagram of the proposed stereo vision SLAM with SuperPoint deep feature and descriptor map and SuperGlue feature matching model.

As we will discuss in Chapter 3, our proposed algorithm requires obtaining a feature map and descriptor map that covers the entire image, hence our adoption of the SuperPoint model. Additionally, this model has been designed to be applicable across various fields such as SLAM, SfM, and others, with a capability proposed for real-time operation.
The SuperGlue model functions as a deep middle-end matcher, bridging the gap between the visual SLAM front-end, which acquires features and descriptors, and the back-end optimizer. Composed of an Attentional Graph Neural Network (GNN) and an optimal matching layer, the SuperGlue deep model is specifically designed to accurately match feature points that project the same 3D world coordinates to 2D pixel coordinates in different images with diffent camera viewpoints. This matcher is known to outperform with using SuperPoint keypoint information(feature, feature confidence, descriptor) compared to conventional nearest-neighbor(NN) (Cover et al., 1967) method with other feature detector such as ORB, D2-Net, ContextDesc, and SIFT. So we have opted to replace traditional feature trackers and matchers with the SuperGlue model.

EuRoC dataset	imu (loop closing)	KeyPoint Confidence Threshold	Tracking Confidence Threshold	Stereo Matching Confidence Threshold	error (scale aligned) (m)	error (not scaled) (m)	original VINS	
							stereo only error (m)	stereo+inertial error (m)
MH01	x	0.2	0.8	0.7	**0.04085**	0.045887	0.54	0.166
	o	0.2	0.8	0.7	0.057663	0.05807		
MH02	x	0.2	0.75	0.7	**0.047182**	0.048378	0.46	0.152
	o	0.2	0.8	0.75	0.04279	0.042814		
MH03	x	0.25	0.75	0.7	**0.05205**	0.058069	0.33	0.125
	o	0.25	0.75	0.7	0.056154	0.062685		
V101	x	0.3	0.75	0.75	**0.088619**	0.091797	0.55	0.076
	o	0.3	0.75	0.75	0.11507	0.116896		

Table 2. The Root Mean Squared Error(RMSE) of EuRoC dataset with and without scale alignment. (VINS+SuperPoint+SuperGlue)

3. Approach

In this section, we describe the proposed methodology on how the SuperPoint and SuperGlue models have been integrated into the VINS-fusion algorithm.

From the discussion above, it is evident that efficiently extracting and matching feature points within images is crucial for estimating odometry and trajectory in visual SLAM using stereo cameras. Particularly in cases where visual odometry is estimated without the integration of other sensors such as IMUs, reliance solely on stereo images obtained through stereo cameras is necessary. In this context, both tracking images from one camera of the stereo pair over time, and matching features from stereo images captured simultaneously are crucial for stereo camera vision SLAM. Therefore, we propose shifting from traditional feature extraction and tracking (or stereo matching) methods to those utilizing deep learning techniques.

To achieve this, we sought to implement the SuperPoint feature detection model and the SuperGlue feature matching model—both of which are deep learning models known for their robust performance in real-time applications—into the front-end of the VINS-based visual SLAM module. Considering the challenges associated with online integration of the SuperPoint and SuperGlue models into the VINS-based visual SLAM algorithm, we opted to apply them in a non-realtime, offline mode. Future research will explore their implementation in a realtime format.

3.1 Visual SLAM with deep stereo matching :VINS-based visual SLAM algorithm with SuperPoint

We integrated the deep learning methods to visual SLAM in two different ways. The first proposed approach is as follows: Using the SuperPoint model, we extract the right image feature points and descriptor map, replacing only the traditional feature detector for stereo matching with a deep learning model to enhance stereo image matching performance. This method is designed to maximize performance improvements while minimizing changes to the existing algorithm.

We aimed to enhance performance by improving only the stereo matching aspect with minimal changes to the algorithm. Thus, when performing stereo image matching, the left of stereo camera image related to image tracking still use the Shi-Tomasi method employed by the existing VINS algorithm. The image feature points extracted by the Shi-Tomasi method are then represented by the SuperPoint descriptor, derived from the descriptor map obtained through the SuperPoint model.

Within the whole VINS visual SLAM system, the right camera image, used only for stereo matching, has its features extracted by the SuperPoint feature detect deep model. Simultaneously, the descriptor map obtained from the same model is utilized to represent these features as SuperPoint descriptors. When both feature points in the stereo images are represented using SuperPoint descriptors, the employed matching method is the classical 2-NN algorithm. This process is schematically illustrated as a flow diagram in Figure 1.

3.2 Visual SLAM with deep tracking and stereo-matching :VINS-based visual SLAM algorithm with SuperPoint and SuperGlue

Our second proposed approach, like the first, is implemented offline in a non-realtime fashion. This method entails replacing two key modules of the visual SLAM front-end (tracking one of the stereo images and stereo-matching of the stereo images) to deep learning techniques from traditional methods. Given that minimal improvements to only the stereo matching in the first method demonstrated performance enhancements, it is anticipated that replacing all image feature processing in the front-end with deep learning could lead to further significant improvements in performance.

This process is illustrated in Figure 3 as a flow diagram. Both stereo images are processed with feature extraction through the SuperPoint model. During the tracking and stereo-matching phases, the SuperGlue model is employed for graph matching. Traditionally, the VINS algorithm utilized optical flow for image tracking. It is kind of tranditional and relatively inaccurate mathod to deep method, so necessitated an algorithm to estimate the feature positions of the next frame of the left camera based on the previous robot position. However, we have determined that the module of feature extraction and tracking with the SuperPoint and SuperGlue deep models does not rely on prediction of robot motion and can perform accurate tracking regardless to such algorithms. This has enabled a reduction in computational costs, which is anticipated to offer advantages in real-time performance when configured for real-time operation.

4. Experiments

4.1 Visual SLAM with deep stereo matching :VINS-based visual SLAM algorithm with SuperPoint

Among the two methods we proposed, the algorithm that solely improves the stereo matching module operates as follows. We need to implement the algorithm by offline, a process to store and retrieve acquired data was required during the intermediate steps.

Initially, we extract and track the feature points of the left camera image using the VINS method, subsequently storing all tracked and extracted keypoints offline. Originally, these keypoints do not contain descriptor information. We utilize the

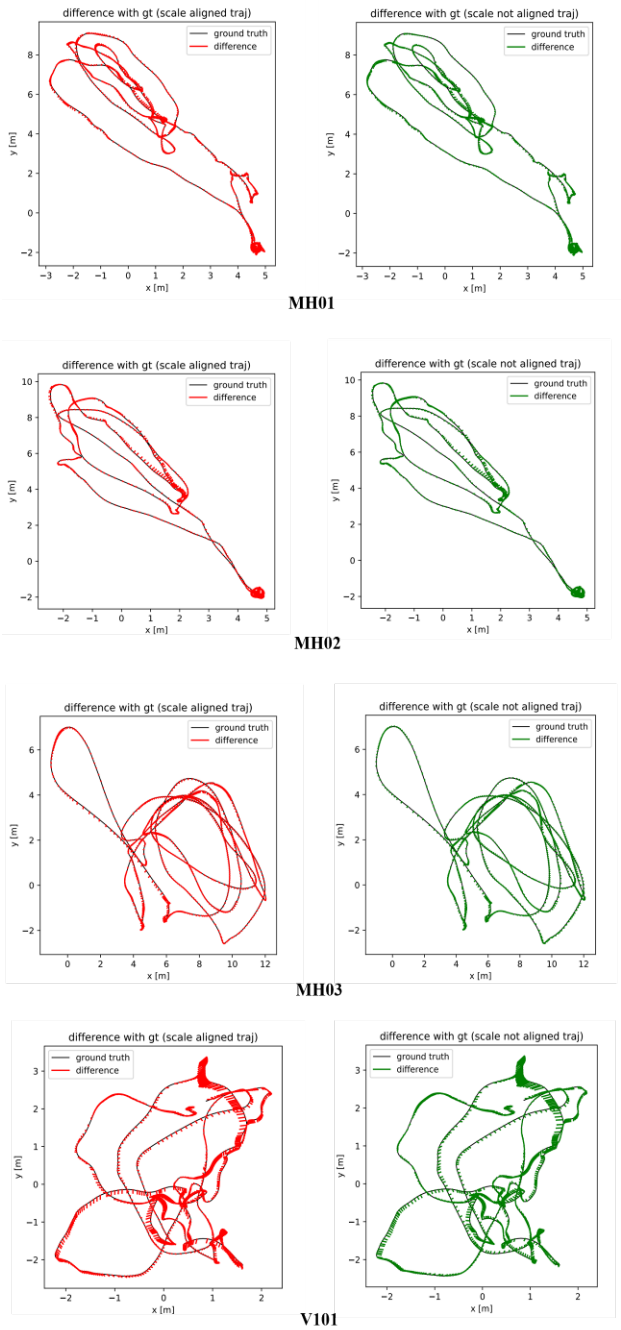

Figure 4. The comparision of EuRoC MAV dataset groundtruth and estimated trajectory. The between groundtruth and estimated trajectory of EuRoC MAV dataset by the proposed algorithm(left:scale-aligned, right:aligned without scaling).

SuperPoint model to append SuperPoint descriptor information to these left camera keypoints. Feature extraction for the right camera is exclusively performed using the SuperPoint model, and the right camera image keypoint data also integrated with SuperPoint descriptor information.

We then match the stereo image keypoints using SuperPoint descriptor matching. Following this, the information is realigned according to the information of tracking stored with the left camera keypoints. Ultimately, only the right camera keypoints are saved, containing the information of stereo-matching relative to the tracked left camera keypoints. This data is then reapplied offline to the front-end module of the VINS-based visual SLAM algorithm.

The experiments were conducted using the EuRoC open dataset to compare performance with the existing VINS algorithm. The results are displayed in Table 1 and Figure 2. The results presented solely represent outcomes without the use of IMU throughout the entire process of the visual odometry estimation algorithm. Accordingly, the performance comparison results for the original VINS-stereo algorithm also reflect scenarios where the IMU was not utilized.

Compared to VINS-stereo, our method demonstrated significantly superior performance. It is common for many SLAM algorithms to evaluate performance by comparing the estimated trajectory, which is scale aligned, against groundtruth data. The categories in the table, 'scale aligned' and 'not scaled', indicate whether the scale aligned algorithm was applied during comparison with GT. As can be seen in the table, our proposed algorithm shows little difference in results between these categories, indicating that the proposed algorithm robustly estimates scale. Therefore, it is clear that the process of estimating scale through triangulation via stereo matching and the performance of visual SLAM algorithm has been improved.

4.2 Visual SLAM with deep tracking and stereo-matching :VINS-based visual SLAM algorithm with SuperPoint and SuperGlue

The algorithm that enhances both the image tracking and stereo matching modules also required offline implementation, necessitating the storage and retrieval of data obtained during intermediate steps. However, this approach does not require storing information from the VINS tracking process, thus the data that needs to be saved is comparatively less than the method introduced in section 4.1.. The feature extraction process for both stereo images is conducted using the SuperPoint model. During the tracking and matching phases, graph matching is performed by the SuperGlue model. As a result, the stored information includes the left camera current frame keypoint (feature, feature confidence), left camera previous frame keypoint for tracking (feature, feature confidence), and right camera current frame keypoint for stereo matching (feature, feature confidence). Information regarding whether tracking or stereo matching has occurred is implicitly included through alignment. These data are saved in a single file and integrated into the visual SLAM front-end module.

The experiments were similarly conducted using the EuRoC open dataset to compare performance with the existing VINS algorithm. The results are displayed in Table 2 and Figure 4. During the process of estimating visual odometry through the front-end of visual SLAM, we didn't use the IMU sensor.

We tested cases with and without the use of an IMU during the back-end optimization's loop closing process and found no significant differences in the results. This indicates that the accuracy of visual odometry estimated at the front-end has been improved. Furthermore, it is inferred that the enhancement of the front-end results has led to more precise keyframe selection for back-end optimization process. Performance was compared against both the original VINS algorithm that used only stereo cameras and the one that employed an IMU(stereo+inertial). The findings demonstrate that even without integrating IMU data throughout the VINS process, our method achieves comparable or superior performance to the original VINS stereo camera algorithm combined with an IMU.

5. Conclusion

We proposed enhancements to the image tracking and stereo matching within the front-end of a VINS-based visual SLAM algorithm. Two approaches were suggested and derived enhanced results: (1) improving only the stereo matching, and

(2) enhancing both image tracking and stereo matching. Through experimentations, we confirmed that replacing traditional feature extraction and matching mothods with deep learning extraction and matching methods improves the performances. One is tracking performance to estimate visual odometry through left camera and the other is the scale estimation performance via triangulation following stereo image matching.

In future research, we plan to propose methods for applying deep learning techniques in real-time. We will also experiment to determine if performance can be maintained with fewer frames. In conclusion, we expect that this research will be applicable to SLAM and SfM problems.

Acknowledgements

This work has supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (No. 2021R1A6A1A03043144) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00320, Manipulation and Augmentation for XR in the Real-World Environment).

References

Dissanayake G, Newman P,Clark S,Durrant-WhyteH, Csorba M. 2001. A solution to the simultaneous localization and map building (slam) problem. IEEE Transactions on robotics and automation, 17(3), 229- 241.

T. Qin, P. Li and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," IEEE Transactions on Robotics, Vol. 34, No. 4, pp. 1004-1020, 2018.

J. Shi and C. Tomasi, “Good features to track,” in Proc. IEEE Int. Conf. Pattern Recog., 1994, pp. 593–600.

B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in Proc. Int. Joint Conf. Artif. Intell., Vancouver, Canada, Aug. 1981, pp. 24–28.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In CVPR Workshop on Deep Learning for Visual SLAM, 2018.

M. Burri et al., “The EuRoC micro aerial vehicle datasets,” Int. J. Robot. Res., vol. 35, no. 10, pp. 1157–1163, 2016.

P-E. Sarlin, D. DeTone, T.Malisiewicz and A. Rabinovich, “SuperGlue: Learning Feature Matching with Graph Neural Networks”, CVPR 2020

Cover , T.M. , “ Nearest Neighbor Pattern Classification ” , IEEE Transactions on Information Theory , vol . IT - 13 , No. 1 , Jan. 1967, pp. 27-2