Semantic Temporal Constrained Pose Estimation using Structure-from-Motion

Narayanan Elavathur Ranganatha
UC San Diego
nelavathurranganatha@ucsd.edu

Saqib Azim
UC San Diego
sazim@ucsd.edu

Mehul Arora
UC San Diego
mearora@ucsd.edu

Mahesh Kumar
UC San Diego
ar223@ucsd.edu

1. Problem Statement

The objective of this project is to accurately estimate the 6D poses (position and orientation) of a monocular camera moving in an environment. We present an approach for visual pose estimation using the Structure from Motion (SfM) technique with temporally constrained frame matching and semantic assistance in the context of autonomous driving scenarios. We address the challenge of pose estimation in dynamic scene environments, which can introduce errors due to incorrect matching in the reconstruction of 3D scenes and the estimated trajectory using the SfM algorithm. Specifically, we use visual data from outdoor driving scenarios such as the KITTI dataset [2] to evaluate our approach since accurate estimation of the car's pose in dynamic environments is crucial for autonomous driving applications. Our method contributes to this field by providing reliable and precise car pose information, thus advancing the development of autonomous driving systems.

2. Method Description

We improve upon the Pixel-Perfect SfM [3] pipeline, a keypoint-based approach for reconstructing sparse 3D structure from image observations. Most SfM algorithms utilize keypoints extracted from images to perform matching and estimate the camera pose and 3D world points. However, in dynamic environments, estimating camera poses and scene structure is challenging due to incorrect matching caused by dynamic keypoints from objects such as cars, vehicles, people, and riders. Here, dynamic class objects can be moving objects or objects that can move (e.g., stationary cars and people). To address this issue, we incorporate HRNet [4], an off-the-shelf semantic-segmentation model, into the SfM pipeline. HRNet enables the segmentation of dynamic objects, allowing us to exclusively utilize static world points for pose estimation. Additionally, the Pixel-Perfect SfM pipeline performs exhaustive frame matching using all images, which is redundant for driving scenarios since distant frames have a low probability of sharing common world points. In contrast, we introduce a temporal constraint that selectively uses nearby past and future frames. This constraint significantly reduces computational requirements and improves tracking accuracy by eliminating matching with far-away frames. Our approach outperforms the Pixel-Perfect SfM [3] baseline when evaluated against the KITTI 360 benchmark dataset [2]. We achieve a 97.8% improvement in localization accuracy and reduce the overall processing time by approximately 25 minutes.

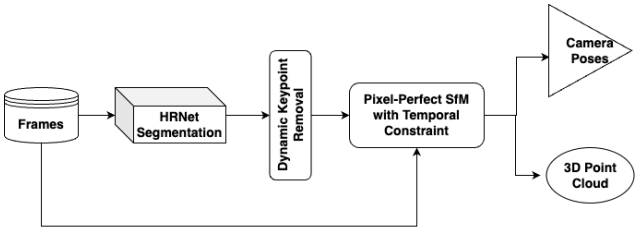

Figure 1. Figure showing our camera pose estimation pipeline.

2.1. Dynamic Segmentation using HRNet

We utilize hierarchical multi-scale attention mechanisms for efficient segmentation and identification of dynamic objects. This approach efficiently leverages multi-scale image information by introducing a hierarchical attention module as shown in Figure 2. The module operates on feature maps from various convolutional neural network (CNNs) levels, capturing contextual dependencies across scales. It combines global and local contextual information to improve object understanding and discrimination based on size and shape. The hierarchical attention module employs both top-down and bottom-up attention mechanisms to iteratively refine feature representations. The top-down mechanism aggregates high-level contextual information, while the bottom-up mechanism captures fine-grained details. To accomplish this, we employ an HRNet model pretrained on the Cityscapes road dataset [1].

Following the semantic segmentation using HRNet on each frame, we proceed to identify and exclude keypoints associated with dynamic objects to ensure accurate pose estimation. Specifically, we exclude keypoints belonging to the following dynamic classes: Cars, vehicles of all types,

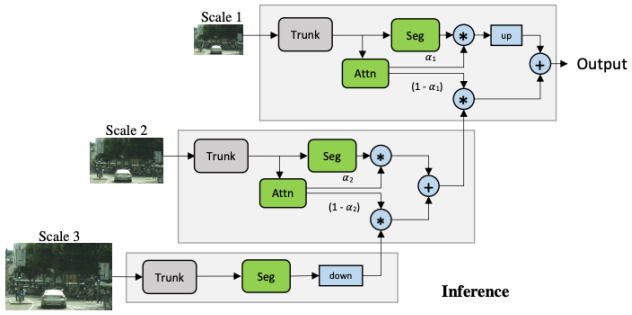

Figure 2. Figure showing Hierarchical attention architecture. Inference is performed in a hierarchical manner to combine multiple scales of predictions. Lower scale attention determines the contribution of the next higher scale. (source: [4])

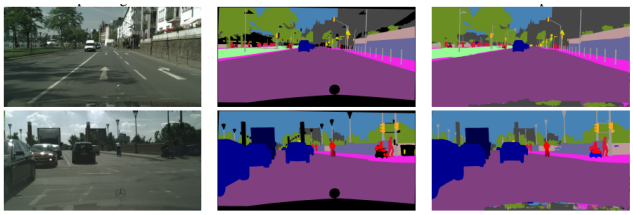

Figure 3. HRNet segmentation results on Cityscapes dataset - Left: Raw Images, Middle: Ground-Truth segmentation, Right: HRNet segmentation (source: [4])

Pedestrians, Persons, and Cyclists. By removing these keypoints from consideration, we enhance the reliability and accuracy of our pose estimation and point cloud reconstruction process.

2.2. Pixel-Perfect Structure from Motion (SfM)

The basic steps of the SfM pipeline include:

- Feature Extraction
- Keypoint Matching
- Reconstruction
- Bundle Adjustment

Keypoint matching is a core component of the SfM pipeline and sparse features are preferred for their efficiency and robustness. The process involves detecting a small number of interest points, computing their visual descriptors, matching them with nearest neighbor search, and verifying the matches with two-view epipolar geometry and the RANSAC algorithm. SfM assumes that the sparse interest points can be reliably detected across views. It selects the features (typically corner points) for each image independently and relies on them for the rest of the reconstruction process. Since the keypoints are independently detected, they aren't localized properly as can be seen in Figure 5. Hence, during bundle adjustment the reprojection error is

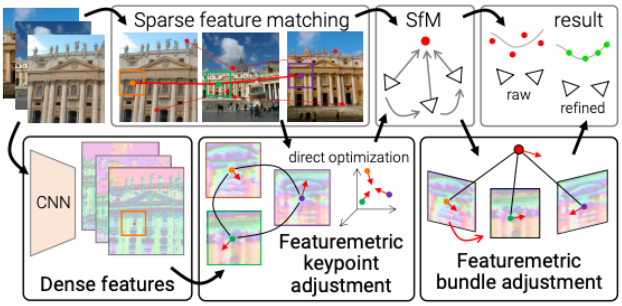

Figure 4. Figure showing the Pixel-Perfect SfM pipeline which performs a two-stage adjustment of keypoints and bundles. The approach first refines the 2D keypoints only from tentative matches by optimizing a direct cost over dense feature maps. The second stage operates after SfM and refines 3D points and poses with a similar featuremetric cost. (source: [3])

minimized w.r.t an erroneous keypoint and this error propagates through the entire SfM pipeline.

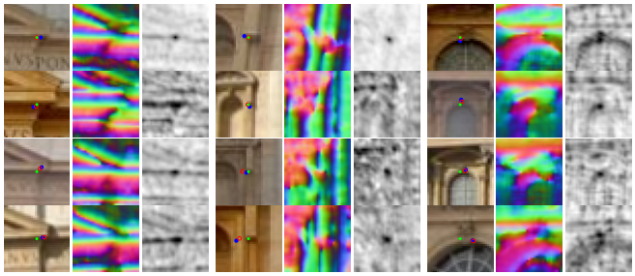

Figure 5. Figure shows points refined with pixel perfect SfM (in green) are consistent across multiple views while those of a standard SfM pipeline (in red) are misaligned because the initial keypoint detections (in blue) are noisy.

The Pixel-Perfect SfM improves the entire SfM pipeline by a two step refinement process. Given the constraints provided by the coarse but global correspondence or initial 3D geometry, it is sufficient if the dense information is only locally accurate and invariant. Thus, we make use of CNNs which exhibit high invariance by capturing large contexts and retain fine local details. The utilization of such deep features replaces the geometric bundle and keypoint adjustments with their feature-metric counterparts. The refinement process first adjusts the keypoints prior to any geometric estimation and subsequently refines points and camera poses as post-processing.

- **Featuremetric Keypoint Adjustment** : The first step of the process is to perform track separation. Tentative tracks are given by the connecting components in the matching graph. The elements in a track observe the same 3D point from different views. Since, there can only be a single projection for the 3D point on a given image

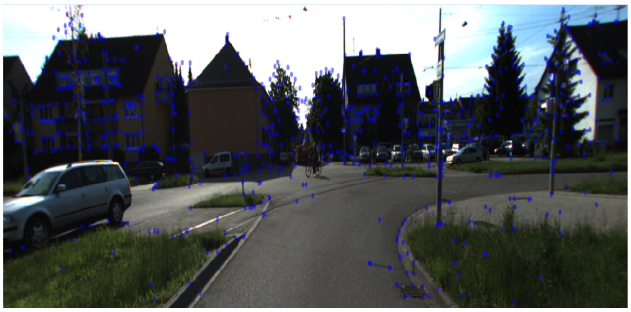

Figure 6. The figure displays the remaining static keypoints (in blue) obtained by passing a frame through the dynamic segmentation module and excluding the dynamic keypoints.

plane we can eliminate the tracks that have multiple keypoints detected on the same image. This step speeds up the optimization and reduces the noise in estimation. The locations of the 2D keypoints belonging to the same track j are adjusted by optimizing the feature metric consistency along tentative matches with the cost:

$$E^j_{FKA} = \sum_{(u,v)\ \epsilon\ M(j)} w_{uv} \| F_{i(u)}[p_u] - F_{i(v)}[p_v] \|_\gamma \quad (1)$$

where w_{uv} is the confidence of the correspondence between (u, v). This allows the optimization process to split tracks that are connected by weak correspondences. Due to the lack of geometric constraints, the points are free to move around in the 3D scene in order to prevent this drift the keypoint with the highest connectivity is fixed and the locations of the other keypoints p_u are constrained w.r.t its initial detection p^0_u such that $\|p_u - p^0_u\| \leq K$. After all tracks have been refined we proceed to geometric estimation and SfM. Despite the sheer number of tentative matches this implementation is quite fast in practice.

- **Featuremetric Bundle Adjustment** : The keypoints here are the projections of the 3D points into the 2D image. For the optimization a reference appearance is defined as the observation closest to the robust mean μ over all initial observations f^j_u of the track and is given by:

$$f^j = \text{argmin}_{f \epsilon \{f^j_u\}} \|\mu^j - f\| \quad (2)$$

$$\mu^j = \text{argmin}_{\mu \epsilon R^D} \sum_{f \epsilon \{f^j_u\}} \|f - \mu\|_\gamma \quad (3)$$

Then we minimize for each track j the difference between the projection and the reference for that track:

$$E_{FBA} = \sum_j \sum_{(i,u)\ \epsilon\ \tau(j)} \|F_i \left[\prod (R_i P_j + t_i, C_i) \right] - f^j \|_\gamma \quad (4)$$

This provides resistance to outliers and accounts for the unknown topology of the feature space.

2.3. Temporal Matching Constraint

The approach used in Pixel-Perfect SfM [3] estimates camera poses and generates a sparse 3D map using exhaustive bundle adjustment by optimizing over all image observations. However, since we are using it for estimating poses in driving scenarios, where frames are sequentially related, we introduce a temporal constraint on keypoint and bundle adjustment. This constraint performs feature matching between a current frame and K neighboring past and future frames. The rationale behind enforcing this constraint is to use local contextual frame information which also reduces erroneous matches between far-away frames that are very unlikely to share common 3D points.

3. Experiments and Results

3.1. Dataset

In our experiments, we conducted a careful selection of a driving scenario and utilized challenging sequences from the KITTI 360 benchmark dataset [2]. Our objective was to evaluate and compare the pose estimation and tracking accuracy of our approach in the presence of dynamic objects, which encompassed cars, riders, people, and other vehicles. To present our findings, we focused on test sequence 00 from the KITTI dataset. Due to compute limitations, we selected a subset of 411 consecutive frames (from frame 2130 to frame 2540) from test sequence 00. This subset was chosen to ensure the inclusion of dynamic objects, as well as variations in camera movement such as straight paths and U-turns.

3.2. Results

In order to evaluate and compare our method, we begin by estimating the scale factor using ground-truth IMU data. Since monocular camera-based pose estimation lacks scale information, we rely on the IMU data to estimate scale. This scale estimation is then integrated into our estimated poses and 3D point cloud. To assess the accuracy of our estimated poses, we employ the absolute translation error (ATE) metric and compare it to the ground-truth poses provided by the KITTI dataset. To evaluate the performance of our approach, we conduct an ablation study that considers different configurations: the baseline model, the addition of a temporal constraint, and the inclusion of semantic information with the temporal constraint. The results of this study are presented in Table 1.

Our method, which incorporates temporal constraint and dynamic object segmentation, outperforms the baseline SfM system and achieves comparable results to state-of-the-art methods in pose estimation on the KITTI dataset. The reduction in processing time is primarily attributed to the use of selective frame matching, as opposed to the exhaustive frame matching described in the Pixel-Perfect SfM [3]

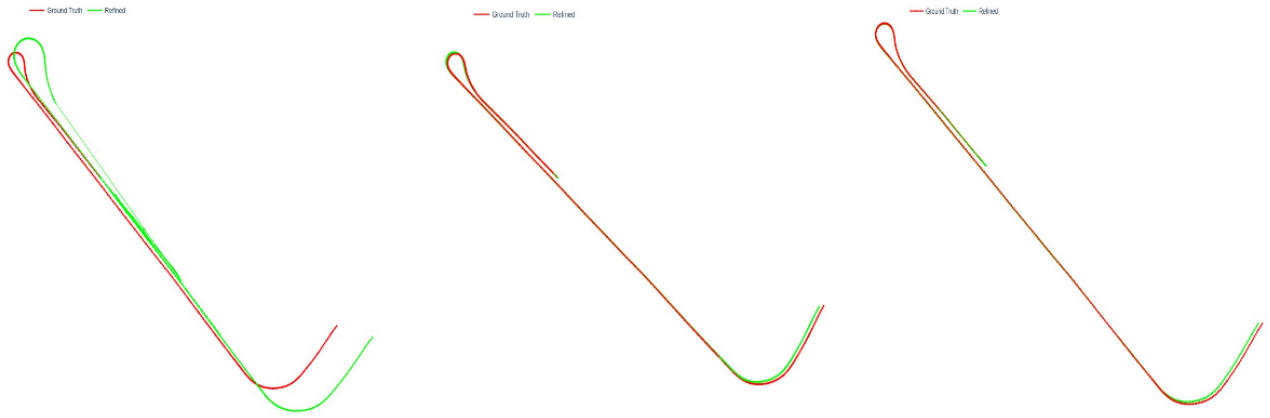

Figure 7. A comparison of ground-truth trajectory (in red) and estimated trajectory (in green) is presented. The left corresponds to the baseline PixSfM model, the middle represents the PixSfM model with temporal constraint, and the right showcases the PixSfM model with both temporal constraint and semantic segmentation.

Method	ATE (in m)	Time (in mins)
Baseline	22.7819	31.47
Baseline + Temporal	0.6351	11.54
Ours	**0.4801**	**6.13**

Table 1. Table showing quantitative results. Our method outperforms the baseline Pixel-Perfect SfM by a considerable margin in terms of both accuracy and processing speed.

paper.

4. Conclusion and Future Works

In conclusion, this research report presents an approach for temporally constrained pose estimation incorporating semantics in driving scenarios. Our approach incorporates dynamic scenes by using the HRNet network for accurate image segmentation and identifying dynamic objects, and exclusively utilizing only static world points for pose estimation. We also introduce a selective usage of nearby frames based on a temporal constraint to enhance localization accuracy while reducing computational requirements. The method contributes to autonomous driving by providing reliable and precise car pose information. Experiments on the KITTI dataset demonstrate promising results, with lower absolute translation error (ATE) values compared to the baseline Pixel-Perfect SfM algorithm. Overall, this report offers an improved approach for accurate and efficient car pose estimation in dynamic environments, advancing autonomous driving systems.

The SfM pipeline estimates poses and a sparse point cloud using prominent keypoints extracted and refined through the Keypoint adjustment process. To transform the sparse road point cloud into a dense point cloud, we fit a plane to these points, as demonstrated in Figure **??**. Although additional improvements can be made to enhance the accuracy of the dense point cloud by utilizing geometric constraints, such enhancements require computational capacity that is currently unavailable. Consequently, we leave these possibilities for future work.

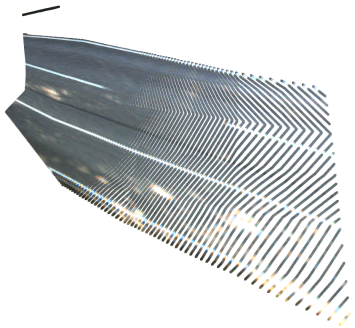

Figure 8. Figure showing the dense road mesh reconstructed from 3D sparse point cloud using geometric constraints.

5. Contribution List

- **Narayanan Elavathur Ranganatha** - Initial Ideation, Code Implementation and debugging, Project development and discussion, Prepared code for submission
- **Saqib Azim** - Initial Ideation, Project development and discussion, Report writing
- **Mehul Arora** - Initial Ideation, Code Implementation and debugging, Project development and discussion, Report writing
- **Mahesh Kumar** - Initial Ideation, Report writing

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding, 2016. 1

[2] 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. 1, 3

[3] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In *ICCV*, 2021. 1, 2, 3

[4] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation, 2020. 1, 2