Deep Learning Assisted Pose Estimation for Visual Odometry

Shaik Althaf Veluthedath Shajihan
sav4

Vardhan Dongre
vdongre2

1 Motivation

Camera pose estimation is the problem of determining the position and orientation of the camera between image frames using a sequence of consecutive frames. Accurate estimates of the absolute pose of the camera are required for many purposes including AR and SLAM. This has been challenging task to incorporate and many efforts are being made to use learning algorithms in the pipeline for pose estimations. In this study, we particularly look at Visual odometry (VO) for vehicle path tracking. Previous attempts and existing deep learning-based approaches demonstrate limited success in handling different parts of the pipeline as they are unable to outperform the classical processes. [4] improves upon the existing relative camera pose estimation methodology by developing a robust learning-based pipeline inspired from the classical geometry-based processes such as SIFT[5] coupled with RANSAC[2]. In this work, we intend to perform a comparative analysis of the two approaches. **Link to video explaining our work is:** [1]

2 Approach

The basic pipeline of Pose Estimation involves: Keypoint detection, Feature Extraction, Matching Keypoints and Outlier Rejection followed by computation for the geometry. We performed these tasks first using the geometry based classical approach which includes using SIFT based feature tracking and RANSAC for outlier rejection. Based on this, the essential matrix $\mathbf{E} = \mathbf{R}|[\mathbf{t}]_x$ is estimated between subsequent frames of the video, and the pose is recovered and updated along the path.

Superpoint (SP) is a fully convolutional neural network which provides a learning-based alternative to SIFT for keypoint detection and description. The network is trained using Homographic Adaptation which is a self-supervised domain. Adaptation network. For our experiments, we used the SP, which was pre-trained on MS-COCO dataset for keypoint detection and description. SP employs NMS for generating sparse keypoints and we coupled this with RANSAC for outlier rejection following which we compute our essential matrix (E).

Moreover, we also collect our-own dataset using a smartphone attached to the windshield of a car - recording few minutes of a monocular-video of path traversed on a highway and in the city of Champaign-for evaluation. We evaluate the pipelines on the KITTI dataset quantitatively.

3 Implementation

In this section we discuss the implementation details of the two methods.

3.1 Primary Dataset

In this work we have primarily used the KITTI grayscale odometery data. The data required included:

- Raw Data Files (City): This included raw data recordings of stereo sequences in grayscale, they were rectified and undistorted
- Caliberation Data: It contains the Intrinsic Camera matrix sued for data-acquisition
- Ground truth poses: It has the ground-truth 6 DOF poses estimate collected using IMU's and high-precision GPS.

3.2 Self-Collected Data

We also collected video recording using smartphone handheld/attached to windshield of car for few routes along Champaign-Bloomington highway and on our campus. The Ground truth was limited to altitude and

[1] https://uofi.box.com/s/q9lfljyx1342lghkxe30zhr2xqzobz7k

latitude (path) GPS estimate recorded on the phone. Due to the lack of frame synchronized scale and GPS positions, only qualitative comparison is performed for this case. In this work we have primarily used the KITTI grayscale odometery data. The data required included:

3.3 Classical Pipeline: SIFT + RANSAC

Feature detection is performed using SIFT on *frame(t)* and tracked to *frame(t+1)* using Sparse Optical flow algorithm of OpenCV.RANSAC based outlier rejection is used in addtion to the two-way filtering of outliers in Optical flow based tracking. For the classical approach [2]., these keypoints are used for Essential matrix estimation and decomposed to get the pose (*R and T*). It is a non-trivial task to identify the frames in which the vehicle is not moving (eg. traffic stops), the path traversed by the Vehicle is updated based on the scale estimated using Euclidean distance between subsequent poses and thresholding it. Moreover, we come up with a scene based masking approach to further improve the inliers for pose estimation. The basic-masking used the scene knowledge that the sky takes a quarter of top of the frame, and features at far-view and edge regions of frame usually negatively impact the essential matrix estimation. For KITTI dataset Seq-03, masking reduced the RMS error from 14.3 to 2.44 ,as illustrated in Figure 1.

3.4 Deep Learning based Pipeline: SP + RANSAC

Keypoint Detection and Description: For using the model described by [1] we utilized the model architecture in Pytorch as described in this paper. The model's pre-trained weights "superpoint_v1.pth' on the MS-COCO dataset were obtained from the Magicleap's publicly available repository [3]. This pretrained model was then used to obtain the keypoints, descriptors and heatmaps for each frame of the video in the dataset. The framework uses a tracker function to retrieve point tracks of a minimum specified length, we used a minimum length of 2 and modified the code to obtain the tracking points of these matching keypoints across the two consecutive frames for the purpose of creating the predicted trajectory as compared to the ground truth trajectory provided in dataset. We combine the outcomes of this process with RANSAC with different threshold settings and then compute the geometry of the camera. Figure 1 shows the trajectories plotted against ground truth for 3 videos using SIFT and Figure 2 . The video renderings of this process can be found at [4]

[2]'https://github.com/mtszkw/visual-odometry"

[3]"https://github.com/magicleap/SuperPointPretrainedNetwork"

[4]https://youtu.be/qyrkGvwfaXs

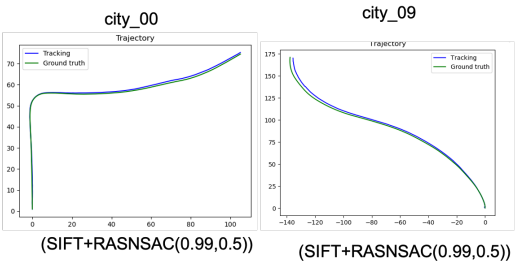


Figure 1: Trajectories predicted using SIFT + RANSAC

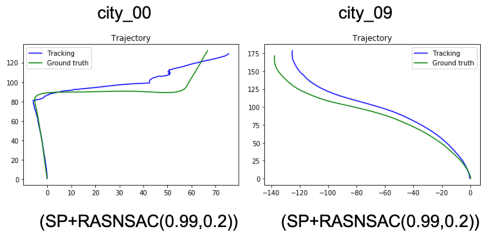


Figure 2: Trajectories predicted using SuperPoint + RANSAC

4 Challenges & Innovation

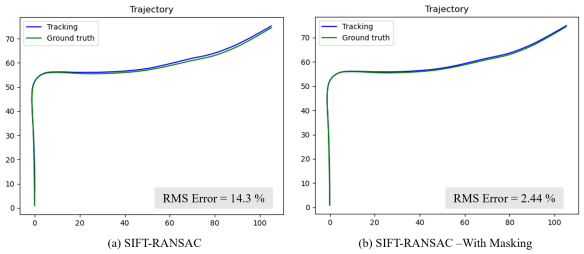
A major challenge of this work was handling the large size of raw odometry data. We propose a new approach to use the knowledge of scene to improve the selection of features being tracked and used for the Essential matrix estimation between frames. The core idea is that we would like to avoid dynamic objects in the scene, and select well-distributed static features in the field-of-view. We got the basic-concept working with a manual-scene based masking, and are still working towards automation using semantic-segmentation as explained :

- Masking of ill-represented regions of the scene: Based on FOV and 1st frame of scene, we mask the sky, edge-regions and regions of glare from windshield on self-recorded dataset. Thereby, filtering out features which deteriorate the quality of Essential matrix estimation.

- Semantic Segmentation for intelligent masking: Although, we were not able to implement this pipeline, our-work looked into using semantic segmentation to classify dynamic objects (eg, cars and traffic) in scene, sky and sun-glare on each-frame. Removing the features falling into these regions is bound to improve the pose estimation, especially, because of the reduced influence of moving cars in traffic-stop signs (where vehicle is static but standard optical-flow indicates motion based on dynamic features in scene).

Seq	Comments	SIFT + RANSAC (No Masking)	SIFT + RANSAC (With-Masking)	SUPERPOINT (No Masking)	SUPERPOINT (With-Masking)
00	City –No traffic	0.62	-	0.75	-
03	Town –with signals +Traffic	14.3	2.44	9.8	5.3
09	'S'-curve - suburbs	2.03	0.629	5.09	0.72

Figure 3: RMS Error comparison for KITTI dataset using Classical and DL approach



(a) SIFT-RANSAC (b) SIFT-RANSAC –With Masking

Figure 4: Path traced for Seq-03 of KITTI dataset without and with masking using SIFT-RANSAC

5 Results & Discussion

Evaluation Metric: Since we get the ground truth (GT) poses for the raw data, we evaluate our two frameworks using thr Root Mean Square Error (RMSE) between the predicted pose and the GT. The RMSE indicates how closely the two poses match. Figure 3 shows the RMSE of the two approaches for three sequences of the KITTI dataset.

Discussion: For our table we present outcomes of three sequence from dataset: 00, 03, 09. These sequences provide variety of scenarios and objects in the scene to us for comparison. Each sequence has unique objects such as pedestrians, traffic, sharp cuts and turns.

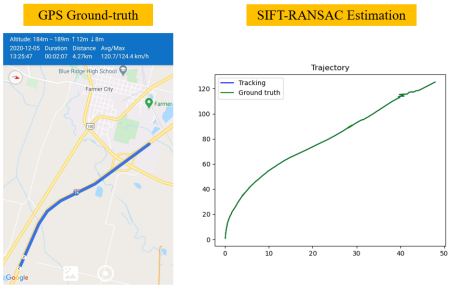


Figure 5: Qualitative comparison for own-data at Bloomington highway with SIFT-RANSAC estimation

On comparing the RMSE of the two frameworks without the use of masks we observe that SuperPoint coupled RANSAC gives results comparable to the classical SIFT + RANSAC method. If we compare column 3 and 4 of the table we observe that by carefully using a mask we can improve the pose estimated by the classical method. If we compare the columns 5 and 6 we observe an improvement in the RMS error after masking, however, these improvements for learning based method couldn't be obtained in all the sequences indicating the need of a more carefully crafted mask for learning based method. Figure 5 shows the result of Pose Estimation using SIFT and RANSAC on the self collected data. We provide a qualitative comparison here with GPS track as ground truth. The plot shows a close resemblance to the GT. We used the second learning based framework on this data and observed an incorrect pose estimation for this data. This outcome can be attributed to the nature of dataset. DL methods require a processed and rectified data for well performance whereas the classical framework showed robustness in handling even raw unsynced and unrectified data. **Github Link** [5]

6 Contribution:

Shaik Althaf Veluthedath Shajihan (sav4): SIFT and RANSAC based VO framework, evaluation on [3] Own-Dataset; Scene-based Masking

Vardhan Dongre (vdongre2): Implementation of the learning based framework SuperPoint + RANSAC [1], to be evaluted on [3]

* *Equal contribution towards Experimentation, Evaluation and Report*

References

[1] DeTone, D., Malisiewicz, T., and Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2018), pp. 224–236.

[2] Fischler, M. A., and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM 24*, 6 (1981), 381–395.

[3] Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2012).

[4] Jau, Y.-Y., Zhu, R., Su, H., and Chandraker, M. Deep keypoint-based camera pose estimation with geometric constraints. *arXiv preprint arXiv:2007.15122* (2020).

[5] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110.

[5] "https://github.com/vardhandongre/CS445_Visual_Odometry"