CS/RBE 549 Computer Vision : Fall 2019
Project Report

Feature based 3D - 2D Visual Odometry

Team Members

Baladhurgesh Balagurusamy Paramasivan
Madhan Suresh Babu
Nikhil Jonnavithula
Sabhari Natarajan

Abstract

In this ever developing world of technology, mobile robotics have gained huge popularity. In the near future we would see a fleet of driverless cars on the roads. One of the important aspects of mobile robotics, especially the autonomous ones is its ability to estimate its own motion i.e. odometry. Using imaging devices such as cameras to estimate the motion has been greatly researched on as it uses simple sensors and limited resources. This kind of odometry is termed as Visual Odometry

In this project we implement feature based Stereo Visual Odometry (VO). This algorithm works based on tracking features between Frame at Time 'T' and 'T+1'. As, we are using a stereo camera, the 3D world coordinates of the features can be calculated. The coordinates of these features are used to find the pose of Frame 'T+1' wr.t Frame 'T'. The pose is incrementally calculated every time instant as combined to build the final odometry of the mobile robot.

An improvised algorithm was also tested to improve the accuracy of the visual odometry. The Stereo VO algorithm is tested on the widely used KITTI dataset (project by Karlsruhe Institute of Technology and Toyota Technological Institute). Using semantic segmentation for improving Stereo VO has been discussed briefly in the project. The Stereo VO generated using this algorithm are also shown in comparison to the ground truth provided in the KITTI dataset.

Table of Contents

List of Figures

1. Introduction

The most traditional method to get odometry is to use wheel encoders. This is a simple method, but these can only be used in ground vehicles. Even in ground vehicles, these provide inaccurate readings whenever there is wheel slippage due to muddy or loose soil kind of terrains. These inaccuracies build up over time and the odometry estimate drifts proportionally to the distance travelled. Another method commonly used for odometry is the GPS. This performs well in conditions where the signals can be received. For cases like indoor, underwater navigation where signals cannot be received this fails to help.

VO comes to the rescue here. Research suggests that VO are more accurate than wheel encoders (relative position error 0.1 - 2 %) as these are not affected by wheel slips. VO can be used to complement other sensors like Inertial Measurement Units (IMUs), SONAR, LIDAR. In regions with lack of GPS, VO can play a crucial role.

1.1 Problem Definition and Objective

VO is the process of estimating the egomotion of an agent by examining the changes that motion induces on the images of its single or multiple cameras. It is divided into 2 types: Monocular VO and Stereo VO. As the name suggests, Monocular VO is the process of calculating odometry of vehicle using one camera and Stereo VO is by using two cameras. The advantage of using Stereo VO is that we can calculate the scale of the odometry.

Fig. 1.1 Monocular VO Setup (Left) and Stereo VO Setup (Right)

Our objective in this project is to “Determine 6-D pose of the camera by using a stream of stereo image sequences”.

1.2 Assumptions

Before proceeding with Visual Odometry process, few assumptions were considered:

- Sufficient Illumination on the environment, to detect and track features.
- Dominance of static scene over moving objects, so that we calculate odometry with static frame of reference.
- Enough texture to allow apparent motion to be extracted, there should be enough features to detect in image, we cannot find enough texture on infinity corridor with the same color, or scene with full of snow.
- There should be sufficient overlap between subsequent frames so that we will be able to track features from 1 frame to another.

Fig. 1.2 Scenarios not ideal for VO

2. Methodology

2.1 Stereo VO Pipeline

Following flowchart shows the various steps followed by us for implementing the Stereo Visual Odometry. Each of these sections is explained in detail in the upcoming sections.

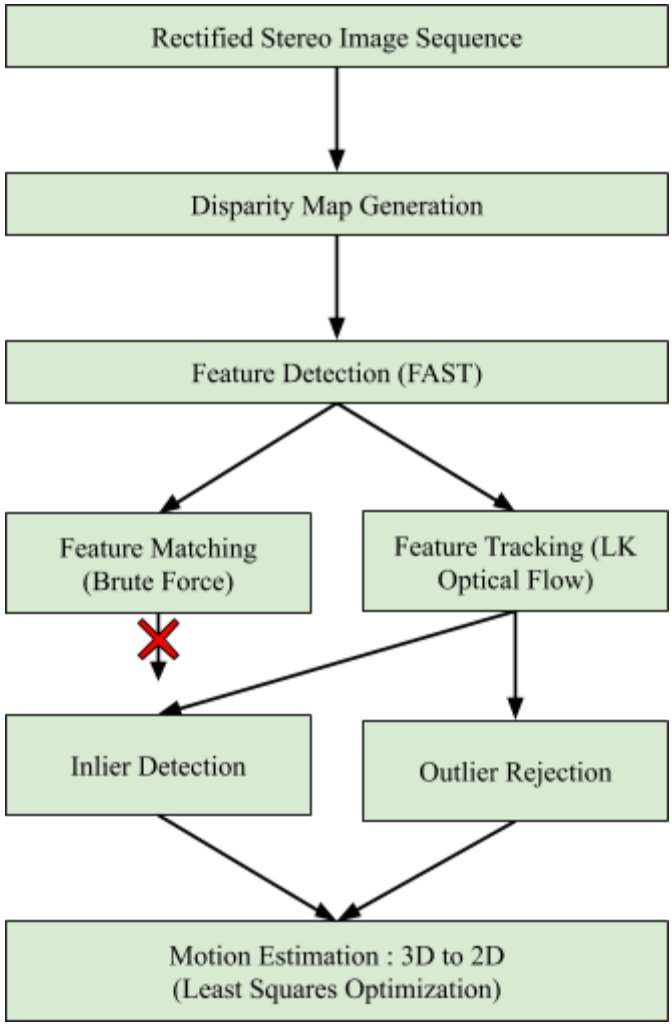

Fig. 2.1 Stereo VO Pipeline

2.2 Disparity Map Generation

For a point closer to the camera, the difference between the pixel coordinates in Left and Right frame will be higher than for a point far away from the camera as shown in Figure 2.1. This difference is known as disparity. In an undistorted and rectified set of images this difference in pixel coordinates will be only along the 'x' direction.

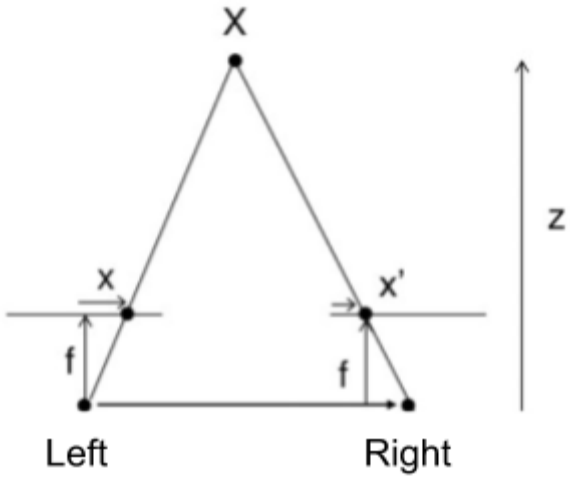

Fig. 2.2 Projection of 3D point in Image

Hence, we used the Block Matching Algorithm with a window size of 15x15 to find the disparity map at a given time 'T'. This algorithm takes a window from left image, slides (only along 'x') it onto the right image. It computes the Sum of Absolute Differences (SAD) for each disparity value and returns the value with least SAD. The left camera is used as the reference.

Further, we build a disparity mask from the disparity map, which will be used during the feature detection process. At higher depths stereo degenerates to a monocular case i.e. loses its depth perception. So, using features at higher depths will lead to inaccurate results. Therefore, we use a depth constraint of 3m to 25m to build the mask. Also, we trim small amount of pixel from all four sides, because, either these sections of image will not have good features or the same feature will not be available in the next frame. Figure 2.3 shows the disparity map and disparity mask.

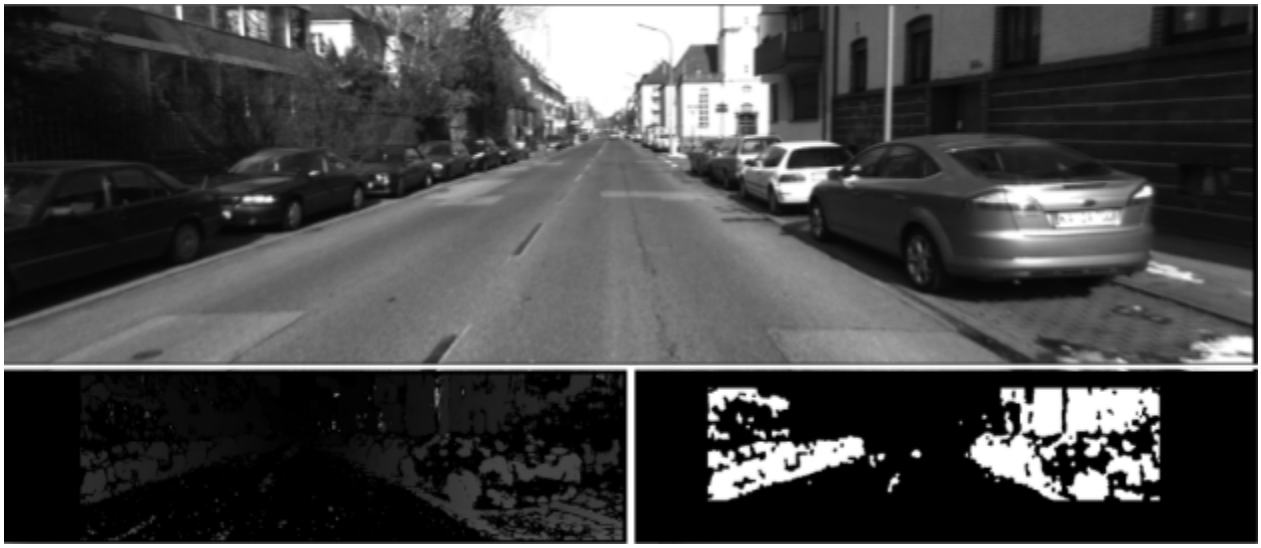

Fig 2.3 (Clockwise from top) Input Left Image, Disparity Mask, Disparity Map

2.3 Feature Detection

In order to perform VO we need to identify key points or features in each frame so that they could be used as a reference for the motion estimation of the robot. To identify the features in each image we used 'Features from accelerated segment test' (FAST) algorithm. FAST is a corner detection algorithm.

FAST features identified ~8000 features in our road scene images. We only require the features for which we can find the 3D coordinate and also doesn't have larger depths. This is where the disparity mask found in previous step is used. This gives ~1400 features. Here the features are dense in some regions and scarce on other. Using features very close to each other will not give us accurate results.

So we add a step called feature binning, which divides the image into grids (20x20 windows) for feature detection, and select only one best feature from each bin. This resulted in ~200 evenly spread out features in each frame.

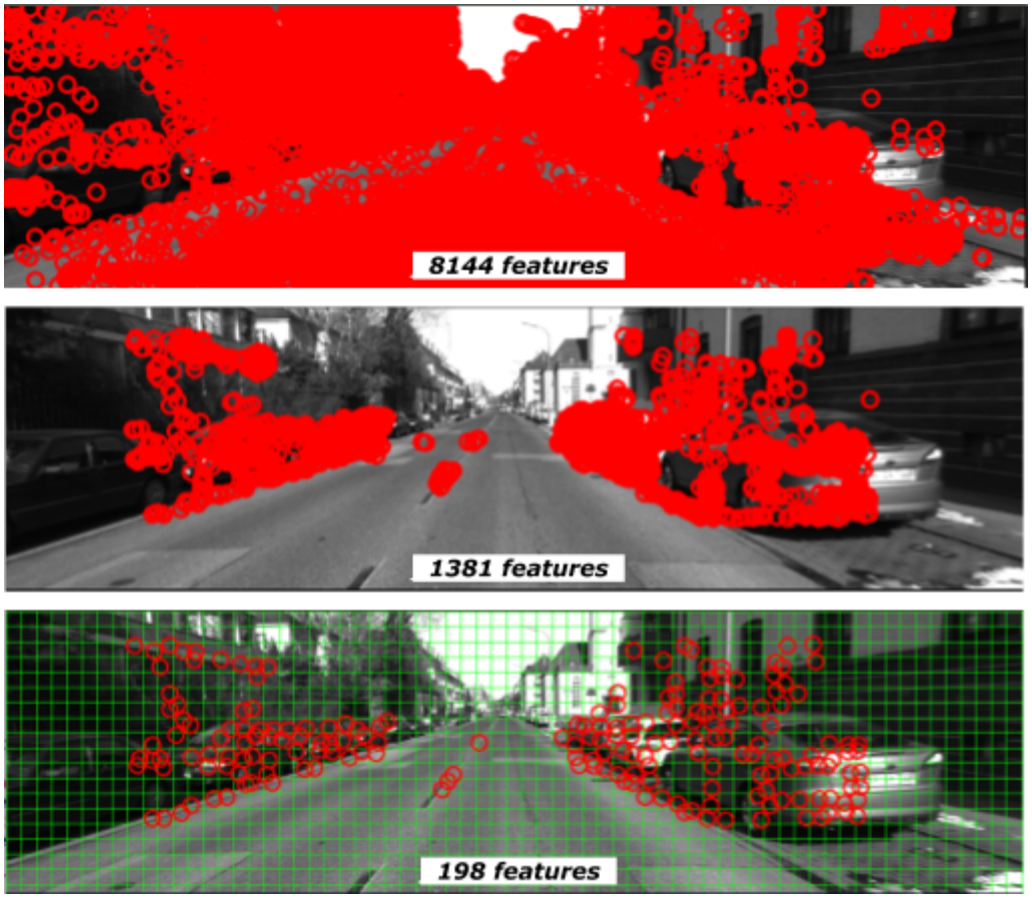

Fig. 2.4 FAST features detected : Full Image, Adding Disparity Mask, Adding Feature Binning

2.4 Feature Matching and Tracking

2.4.1 Feature Matching

Feature Matching as the name suggests is the process where features in two different image are matched, in this case between Frame at Time ‘T’ and ‘T+1’. For this the features in both the frames are found suing the previous method. To find matches, we used the Brute Force matcher that which takes the descriptor of one feature in time ‘T’ and is matched with all other features time ‘T+1’ using some distance calculation and the closest one is returned. This feature matching method was computationally very expensive due to comparison between every pair of features, hence was not used for correlating the features.

2.4.2 Feature Tracking

Another method to co-relate the features, is feature tracking. Here, we find features in time ‘T’ and it is tracked onto time ‘T+1’ using the Lucas–Kanade Optical Flow. This is much faster than feature matching as number of computations are only equal to the number of features found. This method assumes that the flow is essentially constant in a local neighbourhood of the pixel/feature under consideration, and solves the basic optical flow equations for all the pixels in that neighbourhood, by the least squares criterion. It is also less sensitive to image noise than pointwise methods. Feature tracking is done only on the left frame. From the ~200 features tracked ~40 best features tracked are selected for further processing.

Fig. 2.5 Features tracked from Frame ‘T’ to ‘T+1’

2.5 Determining Static Features

In a typical road scene there are both static or stationary objects and dynamic objects. The odometry of the vehicle would be accurate only if we calculate its translation and rotation with respect to a set of stationary points as this is equivalent to finding the motion of the robot with respect to a stationary world frame. Hence, we extract a set of stationary point from the given set of features that were tracked and use it for motion estimation of the robot.

The basic principle used here is that for a pair of static points, the euclidean distance (d) between then at "t" should be the same as that of at "t+1". If d is not same, then it is either:

1. An error in 3D coordinate estimation of at least one of the two features.
2. One of the two features is dynamic.

We performed this step in two different methods :

1. Inlier Detection : Selection best features
2. Outlier Rejection : Rejecting bad features

2.5.1 Inlier Detection

This method makes use of graph theory and is popularly known as the Maximum Clique problem. In this method the idea is to determine the set of points that are all stationary with respect to each other. The Inlier Detection methods involves the following steps :

2.5.1.1 Calculating the Pairwise Absolute Distance Error Matrix :

	f_1	f_2	f_3	...	f_n
f_1	0	e_{12}	e_{13}	...	e_{1n}
f_2	e_{21}	0	e_{23}	...	e_{2n}
f_3	e_{31}	e_{32}	0	...	e_{3n}
⋮	⋮	⋮	⋮	0	⋮
f_n	e_{n1}	e_{n2}	e_{n3}	...	0

Fig 2.6 Pairwise Absolute Distance Error Matrix

For the selected features, 3D coordinates are found using the disparity value and camera parameters (provided by KITTI).

Using the 3D coordinates of the features from time T and T+1 we calculate the distance between every pair of features in both the time instants (d) and use them to find the change in distance between the features over time. These changes in distances are the entries of the Pairwise Absolute Distance Error Matrix , i.e., $e_{12} = | d_{12}^{t} - d_{12}^{t+1}|$.

2.5.1.2 Graph formation and Adjacency matrix calculation :

Next, we form a graph of the features with the number of nodes in the graph equal to the number of keypoints that were tracked. Then the Adjacency matrix of the graph (indicates the connections between nodes in the graph) is calculated by thresholding the Pairwise Absolute distance error matrix, i.e., nodes 1 and 2 are connected if e_{12} in the Pairwise Absolute Distance Error matrix is less than a threshold value. The thresholding operation results in a graph that has the stationary nodes connected to each other. The best feature is the one which has maximum number of nodes connected to it i.e. the row which has maximum values less than the threshold. If there are multiple choices available the first one in the matrix is selected

2.5.1.3 Finding the largest fully connected subgraph :

Since the graph connections indicate stationarity of features with respect to each other, finding the largest subgraph that is fully connected results in a set of points that are all stationary with respect to each other based on the best feature found. The result is also called as the Maximum Clique, where a clique is a fully connected subgraph. The points of the maximum clique are used in the Optimization step for motion estimation of the robot.

2.5.2 Outlier Rejection

The inlier detection method relies heavily on the value of threshold to determine the stationarity of the nodes. Choosing a high value for the threshold would lead to incorrect results and choosing a very small value might not find stationary points at all. So in order to avoid the dependence on an arbitrary value of threshold to identify stationary points, we performed the Outlier detection method, which directly uses the Pairwise Absolute Distance Error matrix instead of thresholding.

Here we eliminate the features that are the most dynamic or non-stationary. We find the most dynamic feature by finding the total sum of each row in the Pairwise Absolute Distance Error matrix. The row with the largest value (largest error) would be the most dynamic point. We perform elimination of the most dynamic point and remove the feature's row and column in the Pairwise Absolute Distance Error matrix. This elimination process of the most dynamic point is performed until we get to the required number of points for the Optimization step.

2.6 Optimization for Motion Estimation

Using the static set of features from the Inlier detection or outlier rejection, we feed it to the optimization routine. In the optimization routine, we use Levenberg-Marquardt nonlinear least squares minimization for finding the translation and rotation matrix. We are using 3D-2D motion estimation, so our Objective function looks like,

$$\epsilon = \sum_{\mathcal{F}^t, \mathcal{F}^{t+1}} (\mathbf{j_t} - \mathbf{PTw_{t+1}})^2 + (\mathbf{j_{t+1}} - \mathbf{PT^{-1}w_t})^2$$

Where,

- P - Projection matrix (From camera parameters)
- T - Homogeneous Transformation Matrix which needs to estimated
- w_t , w_{t+1} - 3D World point at time instant t and t+1
- j_t , j_{t+1} - 2D feature points at time instant t and t+1

Here the objective function is minimizing the reprojection error. For a feature, TW_{t+1} gives the transformed 3D world point at time 'T', using projection matrix (P) we reproject the 3D point to 2D image frame.The error between the 2D feature point and the reprojected 2D point is calculated. In the same way, the world point at 'T' is reprojected to 2D image frame at time 'T+1'. We are using 10 features for this minimization process.

The Levenberg-Marquardt algorithm combines two minimization methods: the gradient descent method and the Gauss-Newton method. In the gradient descent method, the sum of the squared errors is reduced by updating the parameters in the steepest-descent direction. In the Gauss-Newton method, the sum of the squared errors is reduced by assuming the least squares function is locally quadratic, and finding the minimum of the quadratic. The Levenberg-Marquardt method acts more like a gradient-descent method when the parameters are far from their optimal value, and acts more like the Gauss-Newton method when the parameters are close to their optimal value. This algorithm allows us to find the optimal 6D pose vector in a quicker way as it uses the advantages of both, gradient descent and Gauss- Newton method.

2.7 Improvement in Inlier Detection

As discussed earlier, the Inlier detection method depends on the threshold value selected which can make or break the VO. A large value of threshold would incorrectly classify the dynamic point as a static point. Hence for a more robust method of estimation of the stationary features, we developed a variant of the Inlier detection method that starts with a small value for threshold and makes small increments to the threshold value until we obtain a feature that is static w.r.t ~10 other features. This reduces the probability of classifying dynamic features as static features.

Optical Flow tracking does not perform well when the features in image have moved larger distances. Features at lesser depths move large distances in the image as compared to distant features. So, tracking the features at lesser depths resulted in incorrect tracking. To solve this ,we updated the lower threshold for disparity mask from 3m to 10m. The number of feature

points after feature binning reduced but there were sufficient points for tracking. If the number of features went too low, the lower threshold was reduced dynamically.

This method provided better results compared to the original Inlier detection and outlier rejection method due to identification of a better set stationary points.

2.8 Extended Work : Semantic Segmentation

Estimating the stationary points through the Inlier detection or the Outlier detection works as long as there exists a dominance of static key points over dynamic ones. Our methods fail when the whole view of the cameras are covered by a moving object, for example when a vehicle crosses at an intersection in front of the robot. A better method to identify the stationary points would be to determine the class of each pixel (trees, buildings, other vehicles). When pixels are identified as belonging to certain classes we could then easily conclude that features of the class trees and buildings for example are the stationary points.

Semantic segmentation is the method of classifying each pixel of an image to certain classes. We worked on building the pixel-wise classifier using a popular semantic segmentation architecture called SegNet. It is a deep CNN architecture with an encoder network and a decoder network. The novelty in this architecture is the use of max-pooling indices from the encoders for up-sampling in its corresponding decoder layer. We were able to achieve a global class accuracy of 82.27% .

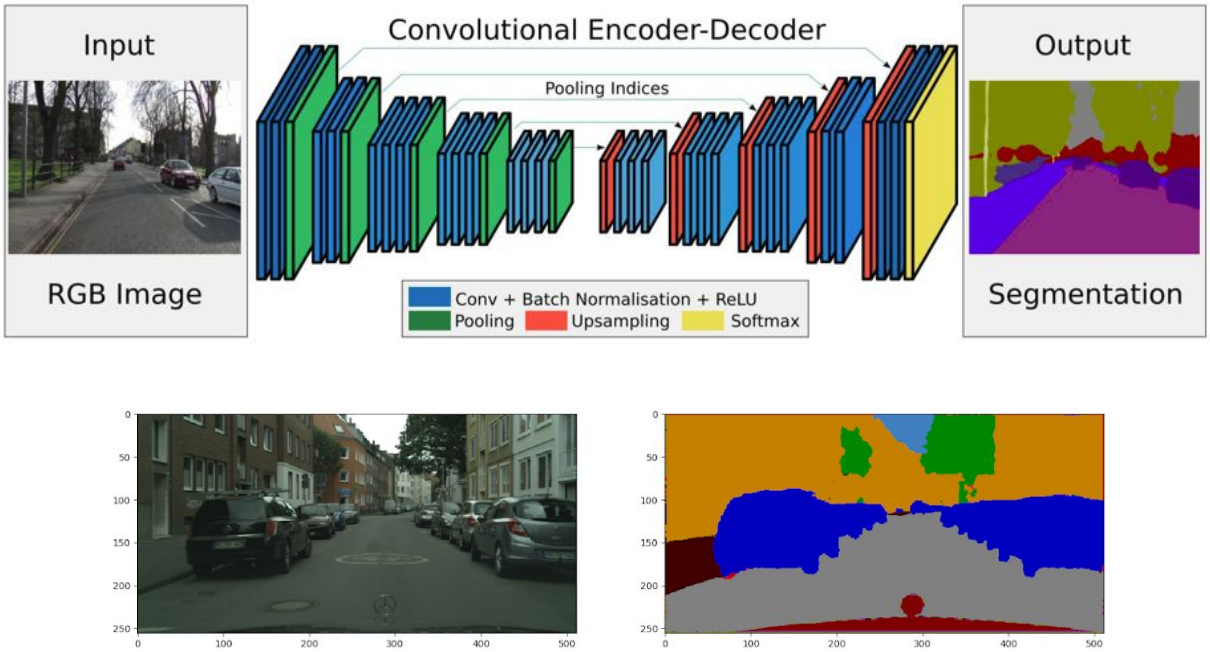

Fig.2.7 SegNet architecture and a test result

We also implemented a modified version of the SegNet that contains *strided 2* convolutional layers instead of the max-pooling layers for performing the down sampling operation, and skip connections from each encoder layer to its corresponding decoder. The modified architecture provided improved segmentation results with a global class accuracy of 86.1%.

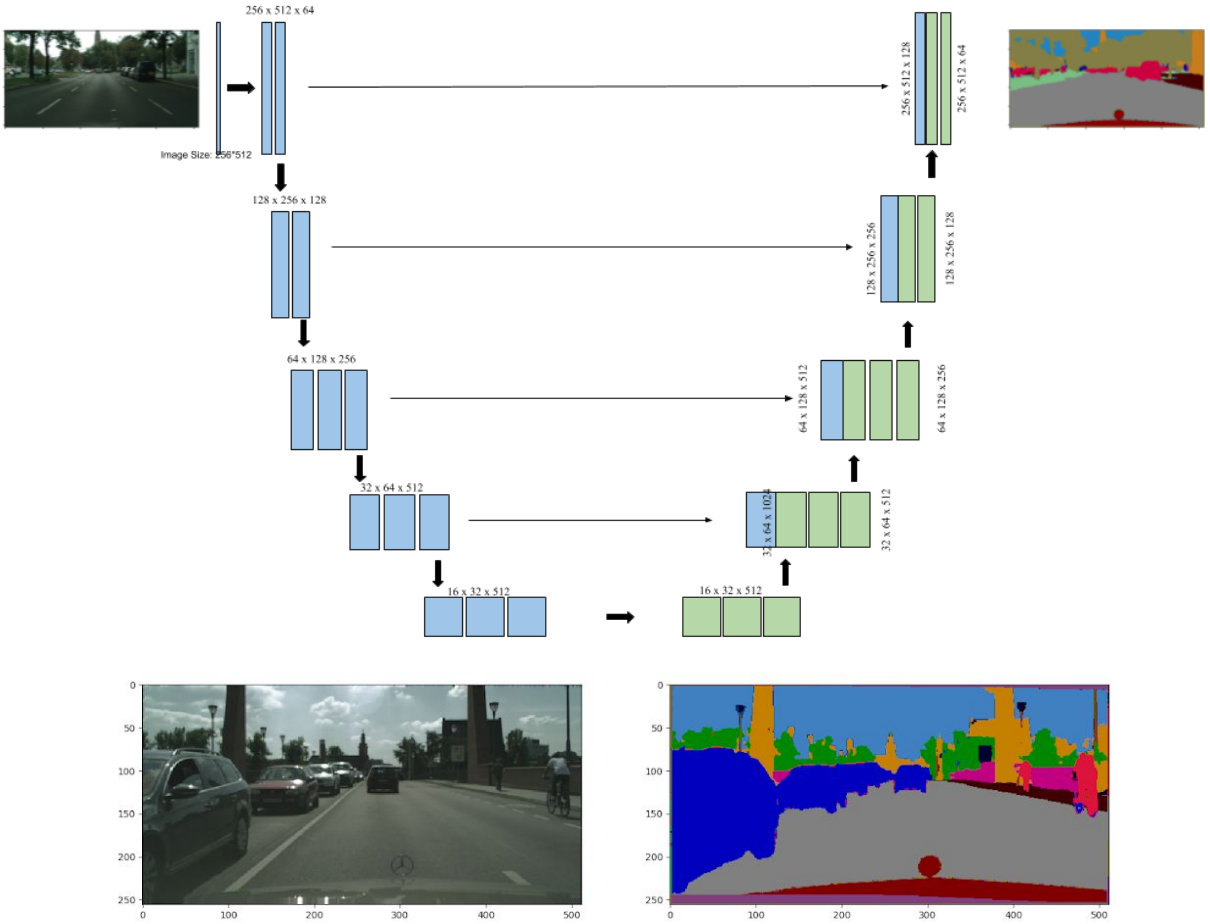

Fig.2.8 Modified SegNet architecture and a test result

3. Results

3.1 Inlier Detection vs. Outlier Rejection

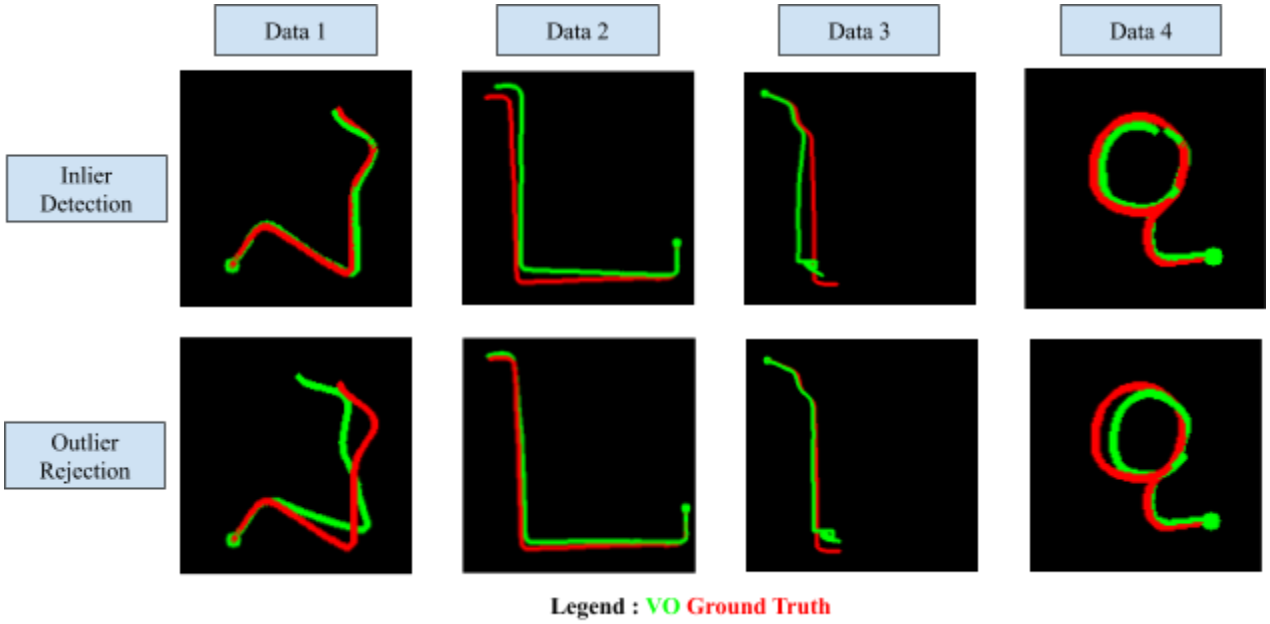

Fig 3.1 Inlier Detection vs. Outlier Rejection Comparison

In data set 1 & 4, we see in this data set, Inlier detection out performs Outlier rejection. It is the first right turn where the Outlier rejection fails and that error is carrier forward throughout. This is because at few instances it had chosen a set of features which belong to a moving vehicle. All the features on the vehicle are static w.r.t each other. When we have less features tracked and more features are from a single dynamic object like a car, this system fails.

In this data set 2 & 3, outlier rejection performs better than Inlier detection, this shows us that both of these methods are not consistently working good on varied datasets. In data set 3 towards the end of VO we see a distortion. This is due to the complete scene being covered by a tram (dynamic object). This was also one of our assumptions that there should be more static features than dynamic ones. When this assumption fails, it leads to incorrect VO results.

To correct this, in case of autonomous cars, we can add a non-holonomic constraints i.e. the dominant direction of motion for the vehicle should be forward.

3.2 Improved Inlier Detection

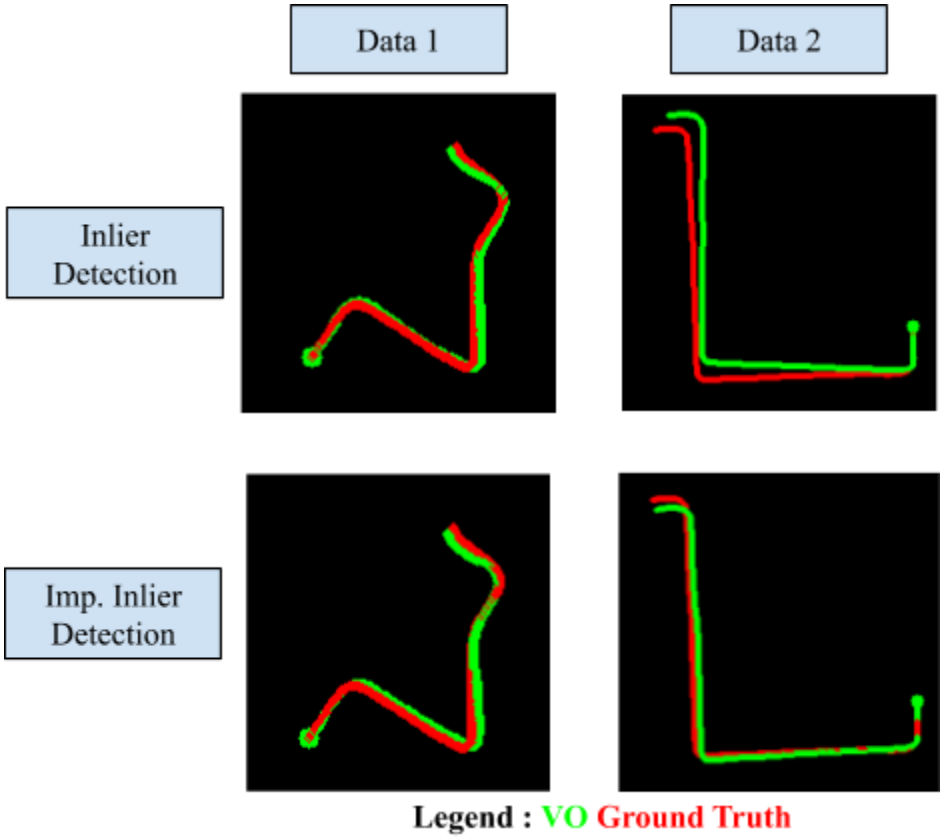

Fig 3.2 Inlier Detection vs. Improved Inlier Detection

Since in our results from Inlier detection and outlier rejection did not work consistently on all the datasets, we tried to improve the Inlier detection method which is explained in Section 2.7. After improvement the algorithm worked better than the previous methods and was also consistently over different data sets.

4. Conclusion

From the methods tested and the results seen we arrived at the following conclusion:

1. Addition of Disparity Mask helped reduce the number of features and combining it with feature binning, it gave us evenly spread out features. Number of features also reduced by 97% (~8000 to ~200), which helped in faster computation time.
2. Feature Tracking is more efficient that Feature Matching in terms of computation time.
3. Inlier Detection and Outlier Rejection gave considerably good performance, but was not consistent over different data sets.
4. The improved Inlier Detection Algorithm performed better than both the previous algorithms and was also consistent over different data sets as it was able to select a better set of static features.

This processing was done on a Core i7 processor and following were the computational time observed for the 3 methods tested:

Computational rate comparison for the VO Algorithms		
Inlier Detection	Outlier Rejection	Improved. Inlier Detection
15 fps	10 fps	14 fps

The computation speed can be further improved if we use GPU for processing.

The further scope for the project include the following:

1. Adding Non-Holonomic constraint
2. Using the Semantic Segmentation results
3. Implementing Bundle Adjustment to reduce drift over time

References

1. KITTI Dataset - http://www.cvlibs.net/datasets/karlsruhe_sequences/
2. Stereo Visual Odometry Blog http://avisingh599.github.io/vision/visual-odometry-full/
3. Visual Odometry Tutorial by Davide Scaramuzza(Univ of Zurich) http://rpg.ifi.uzh.ch/visual_odometry_tutorial.html
4. Segnet - https://arxiv.org/pdf/1511.00561.pdf