Structure from Motion: creating a 3D world from an image sequence

Advanced Computer Vision Assignment

pbqk24

6 December 2018

Approach to the Problem

The solution implements the general Structure from Motion pipeline, specifically for 3D terrain point cloud creation, composed of the general steps below for each image pair:

1. Feature point extraction
2. Feature point matching between the images
3. Estimation of the Fundamental matrix through RANSAC on the feature matches
4. Computation of the Essential matrix
5. Computation of relative camera matrix (rotation and translation) of the second camera from the first
6. Camera matrix composition to produce global camera matrix of the first and second camera
7. Triangulation of matched feature points to produce 3D points
8. Creation and plotting of 3D point cloud for visualization

At each part of the pipeline, filtering was performed and the parameters used were tuned in order to improve the output. Firstly, SURF features were extracted from the images with a hessian threshold of 750. Compared to using a lower threshold, this yielded nearly identical results while being significantly faster. For the feature point matching, a ratio test was performed with a ratio of 0.7 to filter out non-unique matches. This improved the set of matches and minimised the number of obviously false matches produced.

In step 3, estimating the Fundamental matrix (F), a distance threshold of 0.3 and a confidence level of 0.999 was used. These parameters ensured that a good F was found: if a higher distance threshold was used the F would be too inaccurate for the rest of the processing, and a lower distance threshold meant very few of the point matches were inliers with F. Only the inliers with F were passed further through the pipeline, ensuring only those points that were accurately mapped between the images influenced the output.

After computing the Essential matrix (E), these inliers were used to compute the geometric relationship [R|t] between the two cameras. As part of this, the inliers were further filtered to only include points which, when plotted in 3D, were in front of both of the cameras. At this point, the resulting [R|t] matrix and the inlier points were checked to ensure good results. This was done through requiring a z translation value in the range $[-0.05, 0.05]$ and at least 50 inliers. This was based on the cameras being mounted on a car, which cannot (normally) translate sideways (in the z axes of the cameras). Also, if the [R|t] matrix resulted in very few inliers the point cloud was likely inaccurate, so a requirement was placed on this. If the output did not meet these requirements the process was repeated from estimating F up to 30 times. After this, the image was skipped. This process drastically improved the quality and coherence of the final point cloud.

Performance Achieved