CMPT 888 Assignment 2: Incremental Structure-from-Motion

Colin J Brown, Instructor: Ping Tan

March 24, 2015

1 Introduction and Method

The purpose of this assignment was write a system that can generate a 3D point-cloud representation of a scene using a squence of photos taken from mulitple views on a calibrated camera. In order to accomplish this, a set of SIFT feature points are first identified in each image. Corresponding features are then extracted in neighbouring image pairs.

For the first pair of images, the 5-point algorithm [2] is applied the feature pairs to find the Essential matrix, E, between the two cameras. E encodes the relative transform between the corresponding feature points (in normalized coordinates) in each view. RANSAC [1] is used to find a set of point-pairs with few outliers in order to generate an accurate E. The rotation matrix, R, and translation vector, t, between reference frames are then extracted from E, ensuring that the majority of points are in front of both cameras. (Points behind each camera are discarded). Note that the inital camera is defined to be at the origin, facing in the negative Z direction. With the first two cameras fixed, matching point-pairs can then be triangulated in 3D space.

For the third image, the position and orientation of the camera is found in 3D space using points which were matched for 3D reconstruction in the previous step. Again, RANSAC is used to mitigate the effect of outliers. Using this camera and the camera from the previous image, a set of new points, not yet reconstructed, can be localized in 3D. This step is then repeated across the remaining images.

2 Experiments and Discussion

The structure-from-motion algorithm was tested on two datasets, the fountain and the cathedral, both containing a sequence of photos pointed roughly at walls with distinctive structural features. Figure 1 shows the first two images in the fountain dataset with epipolar lines between selected corresponding features.

Various stages of reconstruction of the fountain images is shown in Figure 2. While points from the first image pair give a rough representation of the shapes

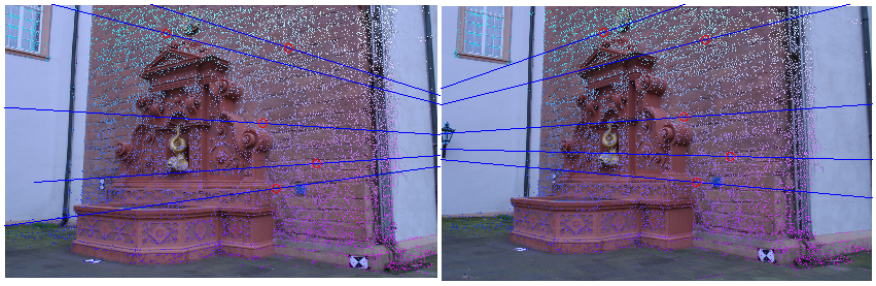

Figure 1: Scene from two views with SIFT features plotted as coloured points. Epipolar lines (blue) between images are plotted through selected features (red).

in the scene, there are many gaps, epecially in places occluded from the first two camera positions as seen on the left side of the fountain. Using more images for reconstruction at a wider range of angles clearly increases the level of detail and fills in many of the gaps. The positions and orientations of the camera corresponding to each image is shown in Figure 3.

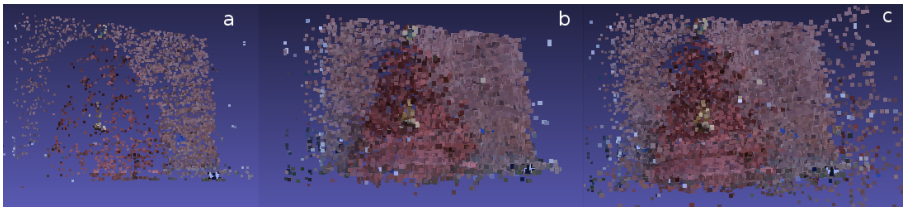

Figure 2: Fountain scene, reconstructed from two (a), six (b) and eleven (c) images.

In Figure 4, the reconstruction of the cathedral is shown. The larger span images (i.e. with less overlap) and greater number of occluded regions from image to image makes this scene more challenging. Due to this, the reconstruction of this scene is clearly worse than that for the fountain.

In general, both results show many outlying points and some mis-aligned structures. These artefacts would likely benefit from stricter outlier rejection thresholds (i.e. when constructing E) and also the application of bundle adjustment.

References

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

[2] David Nistér. An efficient solution to the five-point relative pose problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(6):756–770, 2004.

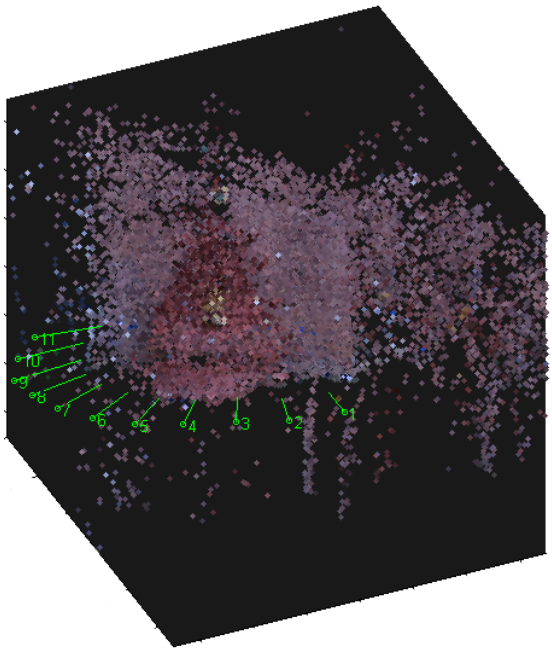

Figure 3: Camera positions and orientations reconstructed for each image of the fountain.

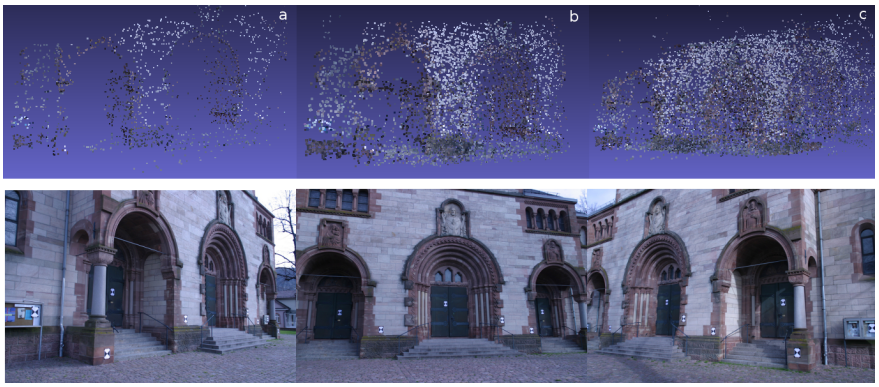

Figure 4: Cathedral scene, reconstructed from two (a), six (b) and 15 (c) images. Bottom row shows three characteristic images sampled from the entire set.