Sensing and Perception – Naghim Ibragimov

Introduction

This coursework shows a 3D reconstruction system that demonstrates how the resulting model can support a robotic manipulation task. The system follows a standard sfm workflow: SIFT feature extraction, feature matching with the NNDR test, Essential matrix estimation, pose recovery, and linear triangulation to generate a working solution.

Camera intrinsics were obtained using checkerboard, though self-captured data proved to be very unreliable due to noise and bad camera quality. Therefore, cleaner multi-view datasets (Temple, Dino) from the Middlebury were used. For the robotic application, a denser COLMAP mesh was imported into MuJoCo because the sparse coursework reconstruction was not stable enough for grasping task.

Methodology

2.1 Image Capture and Calibration

Images were collected and a checkerboard procedure was used to calibrate the camera and obtain the parameters. Calibration was one of the most difficult steps: variations between photo/video frames and background noise. For this reason, the Temple and Dino Middlebury datasets were used, as they provide clean, consistent multi-view sequences suitable for SfM.

2.2 Feature Detection and Matching

SIFT was used to detect crucial keypoints. Each point is represented by a 128-D descriptor built from local gradient distributions. Matches were obtained using the Nearest-Neighbour Distance Ratio

$$\text{NNDR} = \frac{d_1}{d_2},$$

which removes unecesary things. RANSAC filtering was applied to maximize geometric consistency and remove outliers before motion estimation.

2.3 Structure-from-Motion (SfM)

Given matched points, the matrix was estimated by enforcing this constraint

$$x'^T E x = 0,$$

and decomposed to recover the relative rotation R and translation t. Triangulation was performed using linear formulation learned in class.

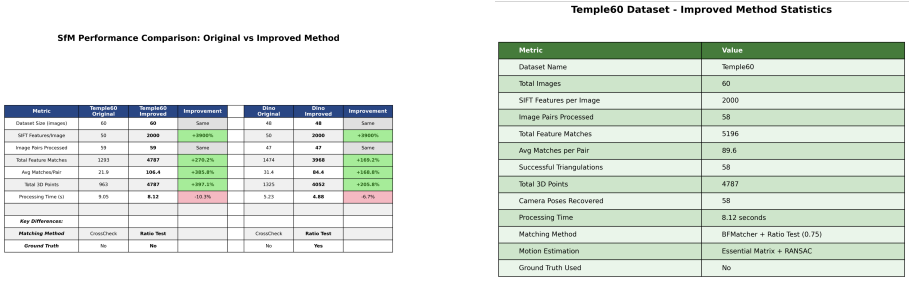
$$x \times (PX) = 0,$$

Repeating this across image pairs yields a good reconstruction, which is later used in the robotic simulation stage.

Results and Analysis

Structure-from-Motion implementations were tested: the original coursework version (50 SIFT features, cross-check matching) and an improved version (2000 SIFT features, ratio-test matching, heavier filtering). Both use the same geometric model matrix estimation and linear triangulation.

With too few features, the original code produced few keypoints and it was just unstable. (Fig. 1). The improved version generated far more consistent correspondences and produced a compact reconstruction. Reconstructed points increased from 963→4787 (Temple60) and 1325→4052 (Dino).



SfM Performance Comparison: Original vs Improved Method

Metric	Temple60 Original	Temple60 Improved	Improvement		Dino Original	Dino Improved	Improvement
Dataset Size (images)	60	**60**	Same		48	**48**	Same
SIFT Features/Image	50	**2000**	+3900%		50	**2000**	+3900%
Image Pairs Processed	59	**59**	Same		47	**47**	Same
Total Feature Matches	1293	**4787**	+270.2%		1474	**3968**	+169.2%
Avg Matches/Pair	21.9	**106.4**	+385.8%		31.4	**84.4**	+168.8%
Total 3D Points	963	**4787**	+397.1%		1325	**4052**	+205.8%
Processing Time (s)	9.05	**8.12**	-10.3%		5.23	**4.88**	-6.7%
Key Differences:							
Matching Method	CrossCheck	**Ratio Test**			CrossCheck	**Ratio Test**	
Ground Truth	No	**No**			No	**Yes**	

Temple60 Dataset - Improved Method Statistics

Metric	Value
Dataset Name	Temple60
Total Images	60
SIFT Features per Image	2000
Image Pairs Processed	58
Total Feature Matches	5196
Avg Matches per Pair	89.6
Successful Triangulations	58
Total 3D Points	4787
Camera Poses Recovered	58
Processing Time	8.12 seconds
Matching Method	BFMatcher + Ratio Test (0.75)
Motion Estimation	Essential Matrix + RANSAC
Ground Truth Used	No

Figure 1: Original (left) vs improved (right) Temple60 reconstruction.

The improvement comes primarily from stronger feature detection (Fig. 2). Lowe's ratio test,

$$\frac{d_1}{d_2} < 0.75,$$

removes weird correspondences and increases the RANSAC inlier set.

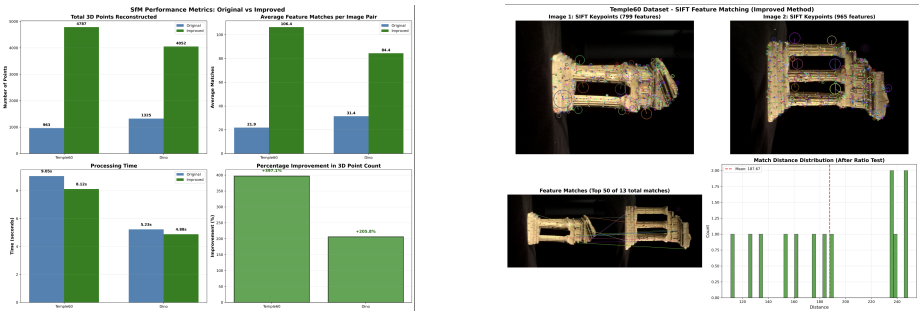


Figure 2: Original (left) vs improved (right) SIFT matches.

Triangulation. Given projection matrices

$$P_1 = K[I|0], \quad P_2 = K[R|t],$$

each of these satisfies the cross-product constraint

$$x \times (PX) = 0,$$

which forms a homogeneous equation

$$AX = 0,$$

solved via SVD. The original system produced unstable R, t, making A rank-deficient. The improved matching stabilised estimation, giving well-conditioned matrices and physically really good depth.

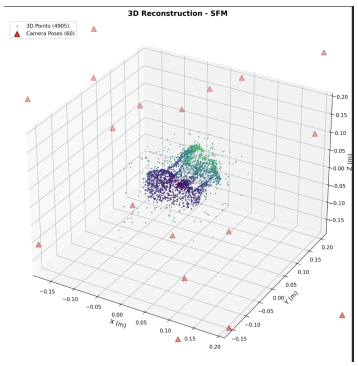


Figure 3: Final improved 3D reconstruction (Temple60).

Robotic Application in MuJoCo

A denser COLMAP mesh was used for simulation, as the current SfM output contained noise, missing surfaces. Due to thin structures in the object, a simple transparent collision shell was added to ensure stable robot contact.

The robot model was a UR5 with a Robotiq S-model gripper [1]. Using MuJoCo's inverse kinematics solver, the end-effector was able to approach, push, and reposition the reconstructed object (Fig. 6). Grasping remained unreliable due to gripper jitter, irregular geometry, and noisy contact normals.

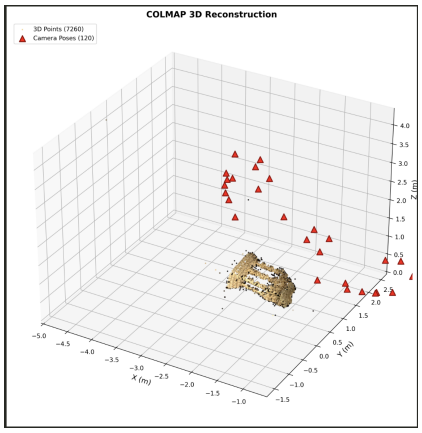


Figure 4: *
(a) COLMAP mesh.

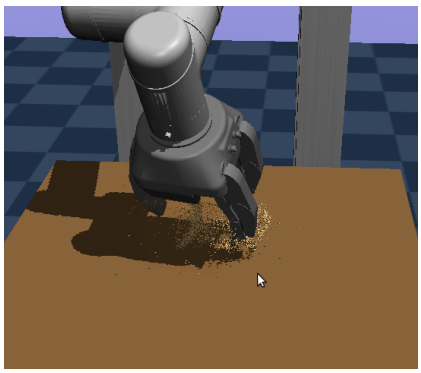


Figure 5: *
(b) MuJoCo with UR5 + gripper.

Figure 6: Reconstruction-to-simulation pipeline.

References

[1] Universal Robots UR5 + Robotiq S-Model Gripper (2025). Available at: `https://roboti.us/forum/index.php?resources/universal-robots-ur5-robotiq-s-model-3-finger` (Accessed: 27 November 2025).

[2] Seitz, S.M. et al. (2006). 'Multi-View Stereo Benchmark'. Available at: `https://vision.middlebury.edu/mview/` (Accessed: 27 November 2025).