

Vision for Multiple or Moving Cameras

Laboratory Evaluation

Maria Fernanda Herrera Perez

Master Programme in Image Processing and Computer Vision
Image

Universidad Autónoma de Madrid

May 2020

Introduction

The current project aims to obtain a 3D reconstruction of a scene by applying three main stages, thus, the description of the reconstruction process is described also into three different sections, which include camera calibration, scene definition according to feature points analysis and 3D points cloud reconstruction and refinement according to multiple views of the scene taken from the same camera.

Section 1. Camera intrinsic parameters

Camera calibration was obtained for a Samsung Galaxy S20 FE camera according to the image of two checkboards of different resolutions projected in computer screens of different sizes. Accordingly, two internal parameters matrixes A and A' were obtained from seven views of each checkboard. The views obtained from the bigger checkboard are shown in figure 1 and views obtained from the smaller checkboard are shown in figure 2. The details for both calibrations are summarized in table 1.

Additional information may be extracted from internal parameters matrixes as follows.

Test for Square Pixels

It can be determined if the pixels of the camera are rectangular or not by considering the skew coefficient s from A where

$$\text{and } s = f_x \tan(\alpha)$$

Then, the angle between image plane axes that determines the skew coefficient is

$$\alpha = \arctan\left(\frac{s}{f_x}\right)$$

If this angle is 0, then it can be determined that pixels are defined as rectangles. Furthermore, if the values for f_x and f_y are equal, it can also be determined that the pixels are square.

Principal point

The principal point defined as the camera center projection into the image plane is defined by c_x and c_y in A where

Orthogonal axes

Orthogonality can be verified by the zero-skew constraint, which specifies that x and y axes in the image plane are orthogonal. This also implies that pixels are square.

Details	Checkerboard 1080	Checkerboard 720
Size in Millimeters	300	160
Image Resolution	4032 width x 3024 height	4032 width x 3024 height
Internal Parameters Matrix	$A = \begin{bmatrix} 3134.39 & 13.76 & 2018.74 \\ 0 & 3121.19 & 1463.49 \\ 0 & 0 & 1 \end{bmatrix}$	$A' = \begin{bmatrix} 3170.83 & -15.047 & 2086.61 \\ 0 & 3182.80 & 1445.64 \\ 0 & 0 & 1 \end{bmatrix}$
Skew Angle in Radians	$a = \arctan\left(\frac{s}{f_x}\right) = \arctan\left(\frac{13.76}{3134.39}\right) = 0.0044$	$a = \arctan\left(\frac{s}{f_x}\right) = \arctan\left(\frac{-15.047}{3170.83}\right) = -0.0047$
Skew Angle in Degrees	$a = 0.2515$	$a = -0.2719$
Pixel Square	Pixels can be considered almost rectangular by a deviation of 0.2515 degrees from the 90 degrees that define a square. Besides, defines a smaller difference from f_x of $3134.39/3121.19 = 0.004\%$, which also suggest that the pixels can be considered as square.	Pixels can be considered almost rectangular by a deviation of 0.2719 degrees from the 90 degrees that define a rectangle. Besides, defines a smaller difference from f_x of $3170.83/3182.80 = 0.0037\%$, which also suggest that the pixels can be considered as square
Principal Point	$cx = 2018.74, cy = 1463.49$	$cx = 2086.61, cy = 1445.64$
Axes Orthogonality	x and y axes of the image plane are considered almost orthogonal, except for a difference of 0.2515 degrees determined by the skew coefficient	x and y axes of the image plane are considered almost orthogonal, except for a difference of 0.2719 degrees determined by the skew coefficient

Table 1. Details for two obtained calibrations according to checkerboards of different sizes.

As it can be observed, similar results and conclusions are and must be derived for camera matrixes obtained from different checkerboards because they define the same camera. However, the parameters A (highlighted in table 1), obtained from the larger checkerboard are the ones used in further sections. The reasoning of this selection is that a larger checkerboard allows to reduce errors coming from size measurement, which can increase overall accuracy.

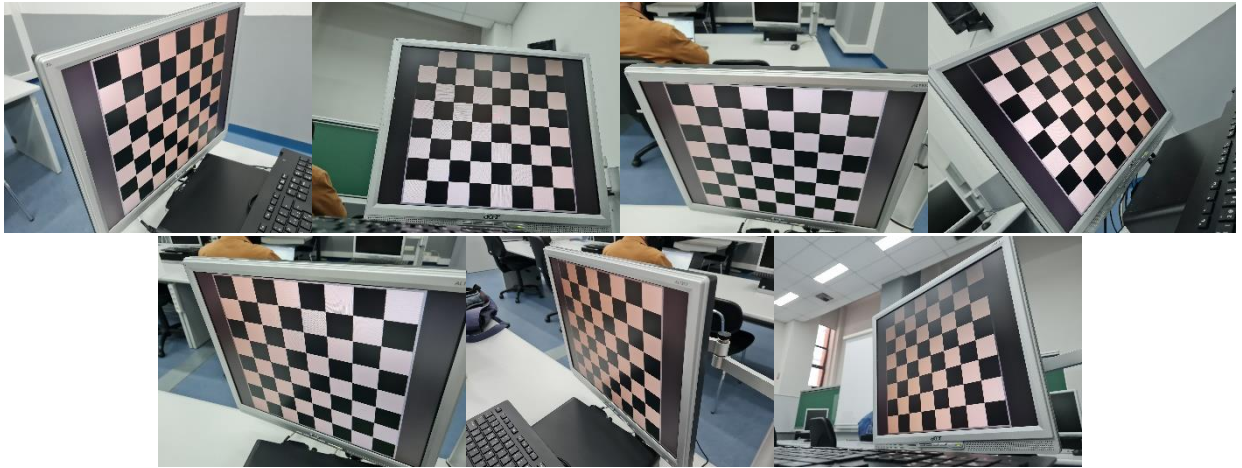


Figure 1. Views obtained for camera calibration according to Checkerboard 1080.

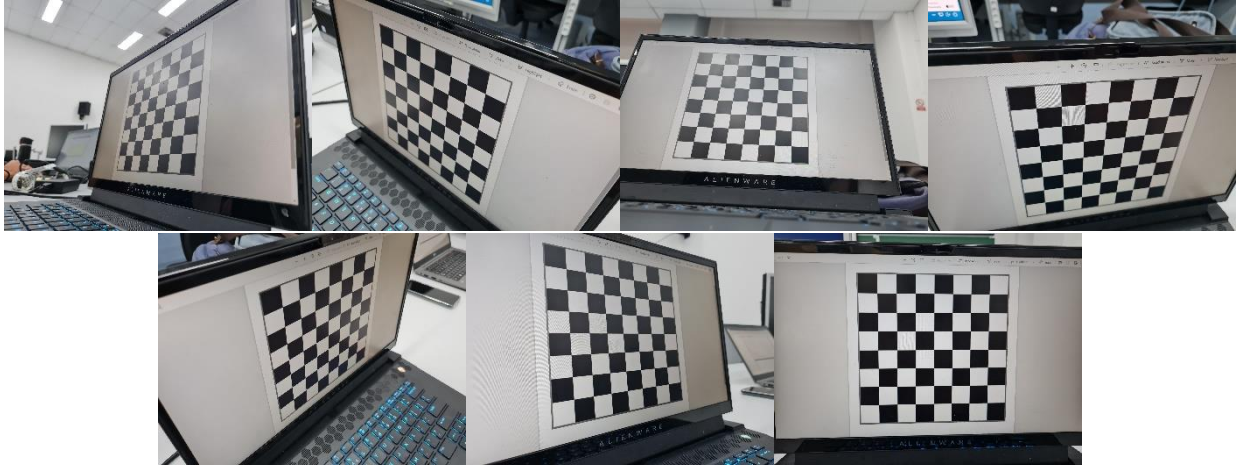


Figure 2. Views obtained for camera calibration according to Checkerboard 720.

Section 2: Feature Point Matching

Definition of the scene

The scene includes different objects that are well separated to be able to easily differentiate objects during the 3D reconstruction described in section 3. Each of the included objects defines a different texture within it to facilitate feature points detection. Besides, the objects are placed at different depths within the scene and different objects have different sizes. This scenario may make it difficult to identify feature points coming from far or/and small objects. To capture the scene, pan and tilt movement of the camera was used with a reduced angular displacement intended to keep as much matching points among views as possible. Three sets of views in the vertical axis were taken where each set contains six views in the horizontal axis. A total of 18 views were taken, where the cameras distribution can be intuited from the view's dataset shown in figure 3.

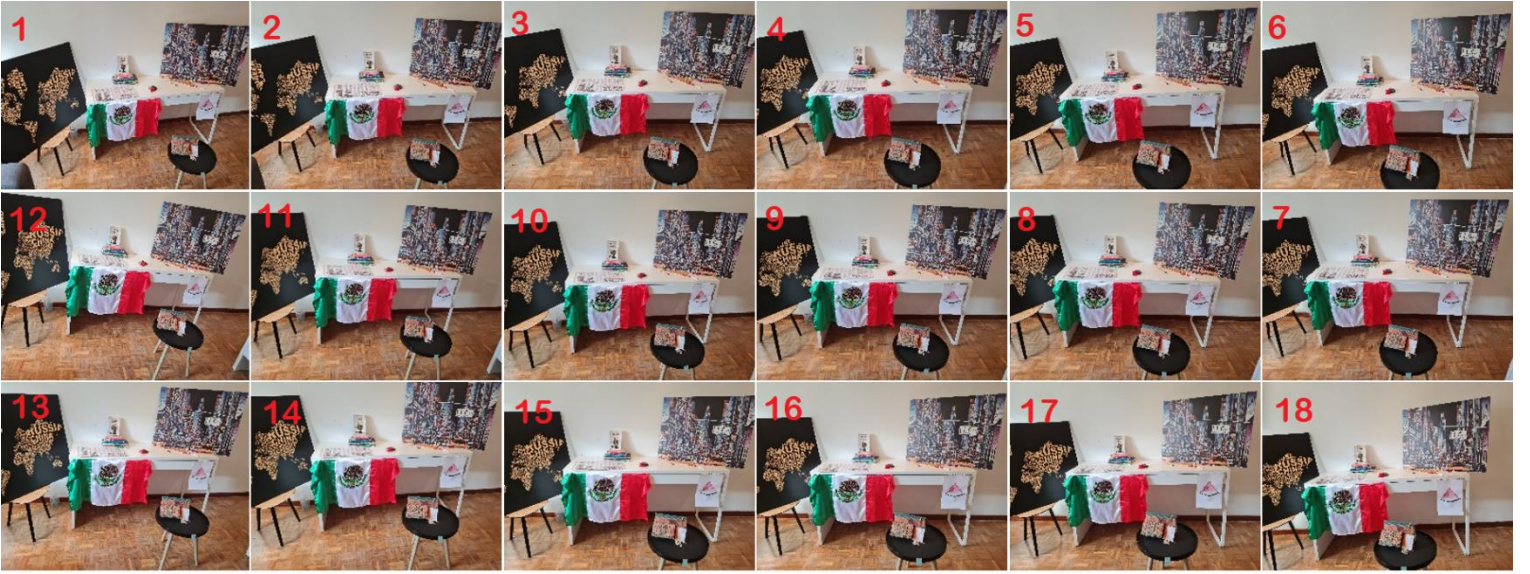


Figure 3. Arrangement of views obtained for the scene.

Detection, description and matching of feature points

To evaluate the performance of different feature point detectors and descriptors on the previously defined dataset, 11 pair of views were selected by considering mainly the farthest cameras (camera 1 and camera 18) and one of the central cameras (camera 9). Besides, four combinations of detector + descriptors were tested across all the view's pairs. This approach resulted in 44 triplets defined by the detector + descriptor pairs DoH + SIFT, SURF + SURF, KAZE + KAZE, SIFT + DSP-SIFT against camera pairs C01-C06, C01-C11, C01-C16, C13-C18, C08-C18, C03-C18, C01-C18, C08-C09, C09-C10, C07-C09, C06-C09. The parameters used for detector+descriptor pairs are 10 scales, 3 octaves, initial sigma equal to 1.6 and 300 points, (for DSP-SIFT, the setting was 15 number of sampled scales, 1/6 for smallest scale and 7 for largest scale) and matching parameters include maximum ratio of 0.8 and SSD metric.

First, the point correspondences for each triplet were obtained. Then, the homography and two homography transformations were obtained for each triplet. Then, the fundamental matrix for each triplet was estimated by excluding outliers through random-sample consensus (RANSAC) and the quality of this fundamental matrix was evaluated by analyzing the absolute and relative number of inliers or points used to generate the fundamental matrix, where the relative number refers to the percentage of inliers obtained from total matched points. Finally, the quality of the fundamental matrix was further evaluated visually according to the epipolar lines that are generated.

Table 2 describes a summary of the quantity of inliers obtained for each triplet. The following conclusions can be derived by analyzing to this summary:

- The pair of views defined by camera 13 and 18 is the pair of views that, when averaging over the percentage of inliers obtained for all the detector + descriptor pairs, results in the lowest percentage of inliers from the triplet dataset (~10.04%). This result is highlighted in the table with red color. A reasoning for this result is that this view's pair define views that are far away from each other.
- The pair of views defined by camera 8 and 9 is the pair of views that, when averaging over the percentage of inliers obtained for all the detector + descriptor pairs, results in the highest percentage of inliers from the triplet dataset (~24.17%). This result is highlighted in the table with pink color. A reasoning for this result is that this view's pair define the closest views and thus, the matching points are less noisy.
- The pair of views defined by camera 6 and 9 is also highlighted in the table in orange color as the median of all the percentage of inliers for each pair of views (~13.4%). By observing at this row in the table, it can be observed that views that have a small variation in vertical and horizontal directions use a similar percentage of inliers to compute the fundamental matrix, even when different detector+descriptor pairs are used.
- The triplets defined according to the pairs of detector and descriptor DoH+SIFT and SURF+SURF use less points (~800) than SIFT+DSP-SIFT (~1400) and much less points than KAZE+KAZE (~4600) to obtain the fundamental matrix across all the different pair of views.
- The detector+descriptor pair that has the highest percentage of inliers across all the views is DoH+SIFT (~17.44% highlighted in yellow color), while the one with lowest inlier percentage is SURF + SURF (~12.45%). Both KAZE+KAZE and SIFT+DSP-SIFT has a similar percentage of inliers (~14.6%).
- The standard deviation of the percentage of inliers obtained for each detector+descriptor pair across all views was obtained. All detector+descriptor pairs describe a similar standard deviation (~5.5%), except for SIFT+DSP-SIFT, which describes the smallest standard deviation of 2.6%. This implies that the last pairs is the more consistent across different views (highlighted in green color).

<i>TRIPLET</i>	DoH + SIFT		SURF + SURF		KAZE + KAZE		SIFT + DSP-SIFT		Relative total
	<i>Absolute</i>	<i>Relative</i>	<i>Absolute</i>	<i>Relative</i>	<i>Absolute</i>	<i>Relative</i>	<i>Absolute</i>	<i>Relative</i>	
<i>C01-C06</i>	191	14.28%	372	8.82%	1459	9.28%	743	13.96%	11.58% +/- 2.94%
<i>C01-C11</i>	354	13.65%	331	8.77%	2139	9.57%	929	9.62%	10.4% +/- 2.2%
<i>C01-C16</i>	374	19.24%	467	10.61%	2166	12.83%	991	16.88%	14.89% +/- 3.89%
<i>C13-C18</i>	270	10.63%	300	6.33%	2453	10.38%	814	12.82%	10.04% +/- 2.7 %
<i>C08-C18</i>	728	14.27%	811	13.79%	4396	13.9%	1326	14.44%	14.1% +/- 3.1 %
<i>C03-C18</i>	607	16.19%	346	7.07%	2791	10.59%	881	12.38%	11.56% +/- 3.79%
<i>C01-C18</i>	188	15.28%	251	7.32%	1918	13.35%	697	17.21%	13.29% +/- 4.28%
<i>C08-C09</i>	2358	27.37%	2205	25.06%	11785	25.16%	3077	19.09%	24.17% +/- 3.55 %
<i>C09-C10</i>	2085	23.97%	1657	18.80%	9225	20.31%	2648	16.13%	19.8% +/- 3.27 %
<i>C07-C09</i>	1710	23.35%	1384	17.37%	8357	20.53%	2098	15.97%	19.31% +/- 3.3 %
<i>C06-C09</i>	714	13.59%	847	12.97%	4400	13.69%	1355	13.30%	13.39% +/- 3.2%
Total	870.8 +/- 794.23	17.44 % +/- 5.3%	815.5 +/- 657.3	12.45 % +/- 5.9%	4644.5 +/- 3522.9	14.51 % +/- 5.22%	1414.5 +/- 824.01	14.71 % +/- 2.6%	

Table 2. Absolute and relative number of inliers obtained for all the triplets, for each pair of views across all detector+descriptor pairs and por each detector+descriptor pair across all views.

The homography and two homography transformations were obtained for each triplet. The visualization of detected points, matching points and homography transformations obtained for the camera pair C13-C18 is shown in figure 4. This pair is shown because is the one with worst percentage of inliers across all detector+descriptor pairs. Evaluating detector+descriptor pair according to inliers information for all the cameras and according to point correspondences and homography transformations for this camera pair allows to verify the quality of matching

points and may allow to hypothesize that enough points would be recovered across all cameras during point cloud reconstruction.

From figure 4, it can be concluded that different detector+descriptor pairs allow to recover similar homography transformations for the camera pair. A reasoning for this result is that the distribution and texture of different objects within the scene allows to properly identify enough matching points for homography computation. However, the quality of the points can also be evaluated by observing the yellow lines linking point correspondences. SIFT+DSP-SIFT describes the less noisy set of matching points since the direction of yellow lines are consistent. On the other hand, KAZE+KAZE shows noisier matched points, since yellow lines cross each other. It can also be observed that the two pairs of detectors return less points than the two last detectors and that SIFT+DSP-SIFT still shows relatively less noisy points against the first pair of detector+descriptors. The homography matrix defined according to SIFT+DSP-SIFT for camera pair C13-C18 is:

$$(1) \quad H = \begin{bmatrix} 1.9923 & 0.3412 & 1.61e^{-4} \\ -0.0319 & 1.4701 & 8.13e^{-5} \\ -804.01 & -1322.6 & 1 \end{bmatrix}$$

Camera 13 / Camera 18				
<i>Detected points for camera 18</i>	<i>Detected points for camera 13</i>	<i>Matched points</i>	<i>Homography transformation from image 13 to 18</i>	<i>Homography transformation from image 18 to 13</i>
DoH + SIFT				
SURF + SURF				
KAZE + KAZE				
SIFT + DSP-SIFT				

Figure 4. Detected points, matched points and homography transformations for all detector+descriptor pair for camera 8 and camera 9 pair.

Finally, the quality of the different fundamental matrixes was further evaluated visually according to the epipolar lines that are generated. Figure 5 shows the results obtained for the triplet described by camera pair C13-C18 and SIFT+DSP-SIFT detector+descriptor pair. It can be observed that the fundamental matrix does not return an epipolar line because this line does not intersect with the object of interest point. However, this solution will only be considered as an initial solution that will be refined by considering other views. The fundamental matrix defined according to SIFT+DSP-SIFT for camera pair C13-C18 is:

$$(2) \quad F = \begin{bmatrix} 5.16e^{-8} & -7.52e^{-7} & 9.61e^{-4} \\ 6.37e^{-7} & 1.07e^{-7} & -0.0017 \\ -0.0013 & 0.0016 & 1 \end{bmatrix}$$

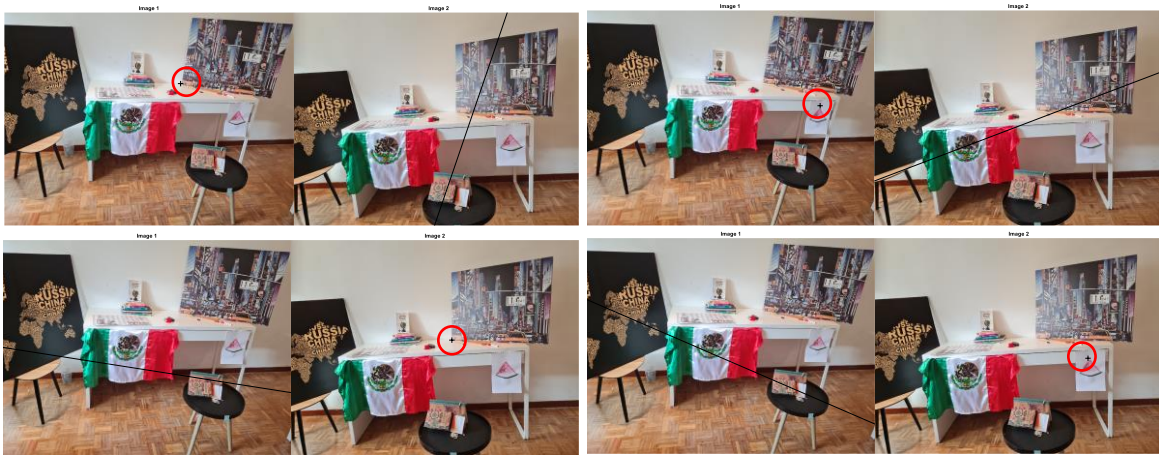


Figure 5. Epipolar lines obtained for camera pair C13-C18 and SIFT+DSP-SIFT from fundamental matrix (2). Camera 18 correspond to image one in each pair of images and camera 13 to image two in each pair.

This triplet is a sample of the selected detector+descriptor pair, SIFT+DSP-SIFT, which was selected because returns a high number of inliers across all the camera pairs, because has the lowest inlier standard deviation across all camera pairs and because the consistency of the matched points obtained from SIFT+DSP-SIFT pair is the best as described in figure 4.

Section 3: 3D reconstruction and calibration

Detection, description and matching of feature points among N views

All the views captured for the scene were used for the reconstruction and are shown in figure 3. Figure 6 shows the detected points for each view according to SIFT detector and DSP-SIFT descriptor and figure 7 shows the detected and matched points for each view.



Figure 6. Detected points for each of the views taken for the scene according to SIFT+DSP-SIFT.



Figure 7. Matched points for each of the views taken for the scene according to SIFT+DSP-SIFT.

Point Cloud Reconstruction

To obtain an initial fundamental matrix, camera 1 and camera 18 were considered because they are the farthest cameras according to the order in which images were taken. Figure 8 shows these views and the matched points for the views.



Figure 8. Views and matched points obtained for farthest cameras C01 and C18.

Table 3 describes the mean reprojection error, the reprojection histogram and a sample of the obtained scene reconstruction for each of the stages defined to obtain the final Euclidean reconstruction. Initially, the projective reconstruction was obtained by using the fundamental matrix obtained from cameras 1 and 18. Then, resectioning and bundle adjustment was applied to refine fundamental matrix. Finally, internal parameters A of the camera were used to obtain an essential matrix from the fundamental matrix defined using bundle adjustment, and this essential matrix was used to obtain the Euclidean reconstruction.

As it can be observed, mean reprojection error decreases through the stages of projective reconstruction, resectioning and bundle adjustment. Besides, the standard deviation for the same error increases from the first to the second stage but decreases for the third stage. The mean value of the error decreases, while standard deviation increases from first to second stage since intermediate cameras are considered. However, the mean value and standard deviation decreases from the second to the third stage due to the refinement of projection matrixes and 3D points by using bundle adjustment.

Finally, figure 9 allows to verify the obtained Euclidean reconstruction by comparing a frontal view of this reconstruction against camera C01 and the matching points obtained from this view. In this figure, groups of points that represent the same object are differentiated by using different colors in both real scene and reconstructed scene. Additionally, a perspective of the reconstructed scene that would represent a capture from the ceiling is provided in figure 10 to demonstrate that the depth of the different objects in the reconstruction is accurate, this figure describes the same colors for the objects as in figure 9.

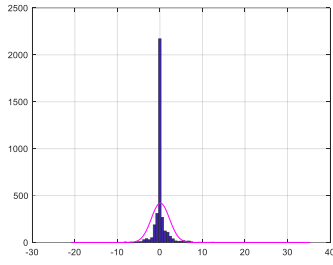
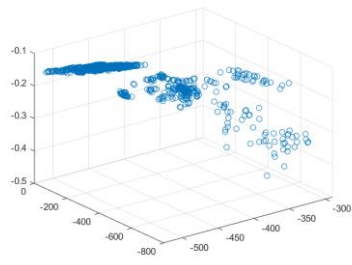
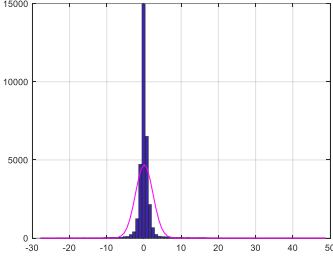
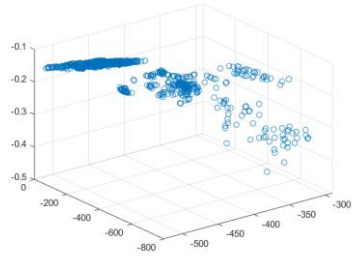
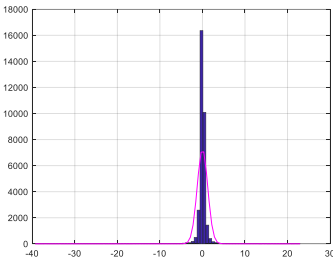
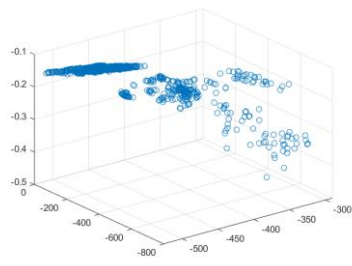
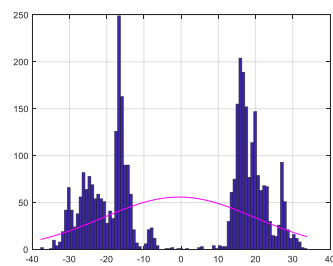
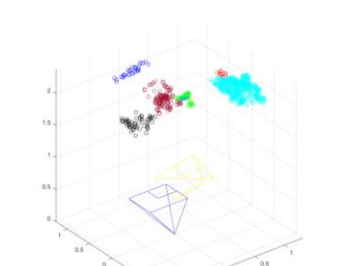
	<i>Reprojection Error</i>	<i>Reprojection Error Histogram</i>	<i>Scene Reconstruction</i>
<i>Initial Projective Reconstruction</i>	<i>Mean [-0.297, 0.515]</i>		
	<i>Std [1.478, 2.557]</i>		
<i>Resectioning</i>	<i>Mean [0.097, 0.049]</i>		
	<i>Std [2.005, 2.637]</i>		
<i>Projective Bundle Adjustment</i>	<i>Mean [0, 0]</i>		
	<i>Std [1.069, 1.372]</i>		
<i>Euclidean Reconstruction</i>	<i>Mean [0.753, -1.858]</i>		
	<i>Std [16.280, 23.730]</i>		

Table 3. Reprojection error and scene reconstructions obtained during the scene reconstruction pipeline.

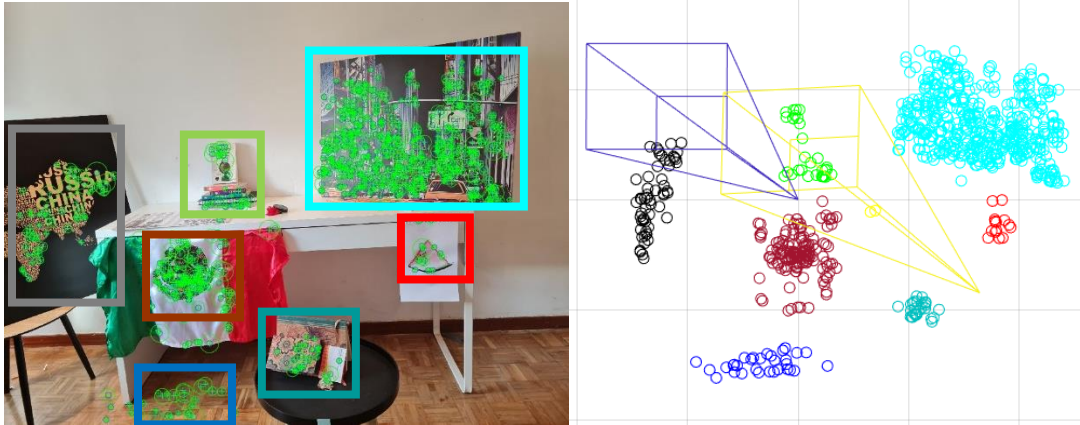


Figure 9. Frontal view of the scene and reconstructed scene.

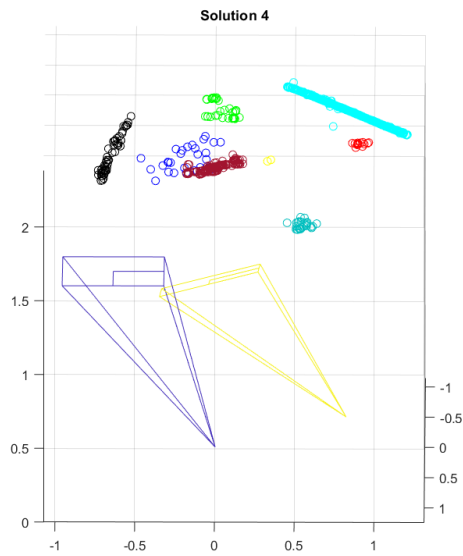


Figure 10. Upper view of the scene.