Vision for Multiple Moving Cameras
Final Project
3D reconstruction of a 3D object

Tamás **Bukits** tamas.bukits@estudiante.uam.es

Abstract

The ultimate goal of the given task is to achieve a 3D depiction of a selected object or scene. This involves the selection and calibration of a camera, choosing a suitable object or scene based on certain guidelines, selecting an appropriate number of views of the object, extracting and matching feature points across the views, computing the fundamental matrix between views, obtaining a 3D points cloud reconstruction, and ultimately depicting the geometric components of the object over this points cloud.

1 Section 1: Obtention of the intrinsic parameters of a camera

Camera calibration is the process of determining the intrinsic and extrinsic parameters of a camera, which are necessary for accurately mapping 3D points in the world to their corresponding 2D points on the camera sensor or image plane. The camera calibration process typically involves capturing a set of calibration images or a calibration pattern with known geometry. The calibration pattern can be a checkerboard seen on the figure below.

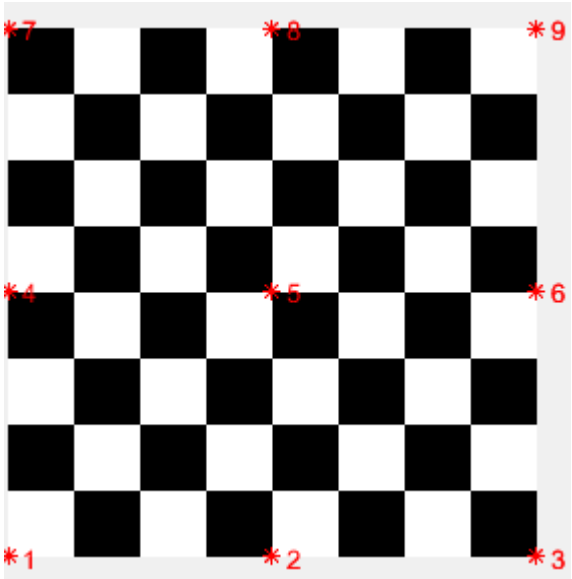

Figure 1: Real checker board with marked points

K is the camera intrinsic matrix, which is represented using a 3x3 matrix.

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \alpha & -\alpha\cos(\theta) & u_0 \\ 0 & \frac{\beta}{\sin(\theta)} & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

The parameters in the matrix are:

- ◇ f_x: The focal length of the camera in the x-direction
- ◇ f_y: The focal length of the camera in the y-direction
- ◇ s: Skew factor between x and y directions
- ◇ c_x: The x-coordinate of the camera's principal point
- ◇ c_y: The y-coordinate of the camera's principal point

1.1 Camera calibration part A

Here a checkerboard is used for the camera calibration.

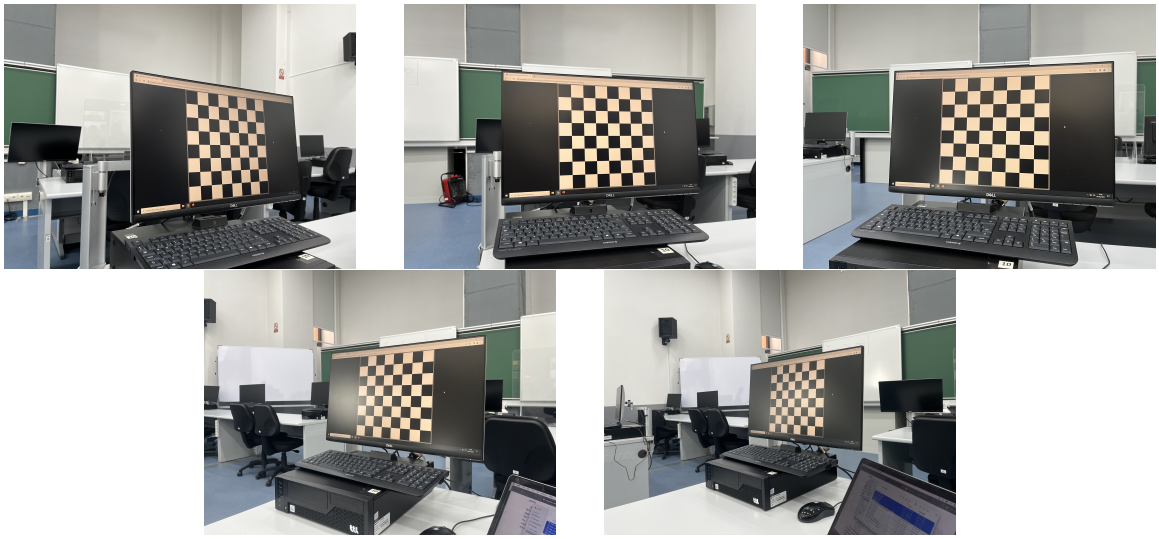

Figure 2: Set of images of the screen checkerboard that have been used for calibration

Some data about the calibration:

- ◇ Size of the checkerboard (1080P) in milimeters on the screen is 260x260mm.
- ◇ The resolution of the captured images (in pixels) = 4032 x 3024.
- ◇ Internal parameters (matrix A) of the camera:

$$A = \begin{bmatrix} 3086 & 4.7 & 1949 \\ 0 & 3057 & 1448 \\ 0 & 0 & 1 \end{bmatrix}$$

Analysis of the following three aspects:

- ◇ **Are the pixels of your camera square?**
  To calculate the aspect ratio of a camera given its intrinsic parameters, we can simply divide the focal length in the y-direction by the focal length in the x-direction. The aspect ratio is usually denoted by r and is given by: $r = \frac{f_y}{f_x} = 1.009$ which is close to 1 (with some noise) which means $f_x = f_y$. Hence, we can say that the camera has square pixels.

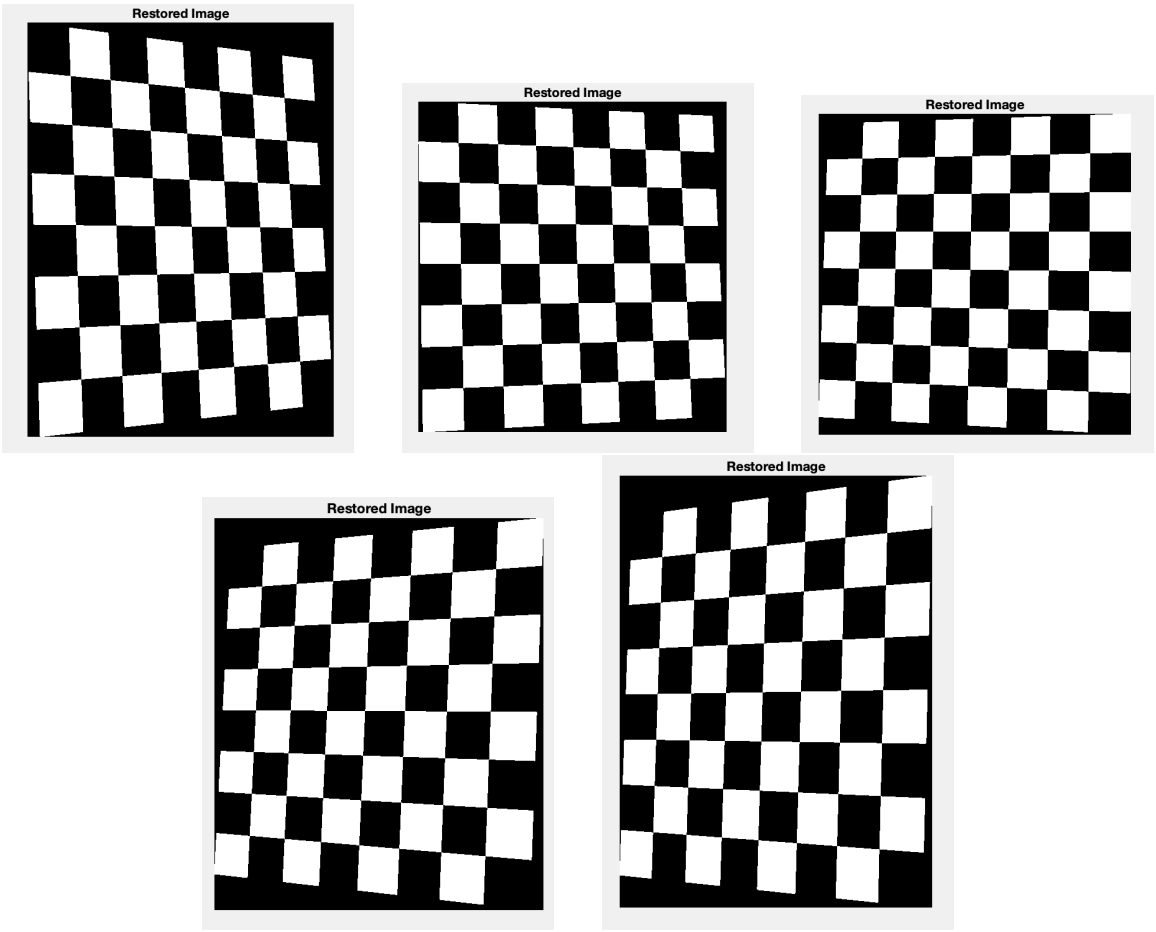

Figure 3: Showing the respective homography images for the checkerboard with 5 different orientations

◇ **Which is the degree of coincidence between the principal point and the center of the image plane?**
If the principal point is coincident with the center of the image plane, then c_x and c_y are equal to half the width and height of the image, respectively. In this case, the degree of coincidence is 100%. However, if the principal point is not coincident with the center of the image plane, then the degree of coincidence will be less than 100%. One way to measure the degree of coincidence is to calculate the distance between the principal point and the center of the image plane and express it as a percentage of the image width (w) or height (h).

$$d_x = \frac{\left|\frac{w}{2} - c_x\right|}{\frac{w}{2}} \times 100\% = 3.3\%$$

$$d_y = \frac{\left|\frac{h}{2} - c_y\right|}{\frac{h}{2}} \times 100\% = 4.2\%$$

$$d = \frac{d_x + d_y}{2} = 3.75\%$$

Overall the average offset is 3.75 which means around with 4% offset is not in the middle of the center according to the noise.

◇ **Are the axes of the image plane orthogonal?** To prove this we have to calculate θ from the skew parameter of the K matrix (s):

$$\beta = \arccos(\frac{s}{\alpha}) = 89.91^\circ$$

in degrees which is close to 90° meaning the planes are orthogonal.

1.2 Camera calibration part B

Here a floor pattern is used for the camera calibration by selecting 9 controlling points marking the pattern which is repeated.

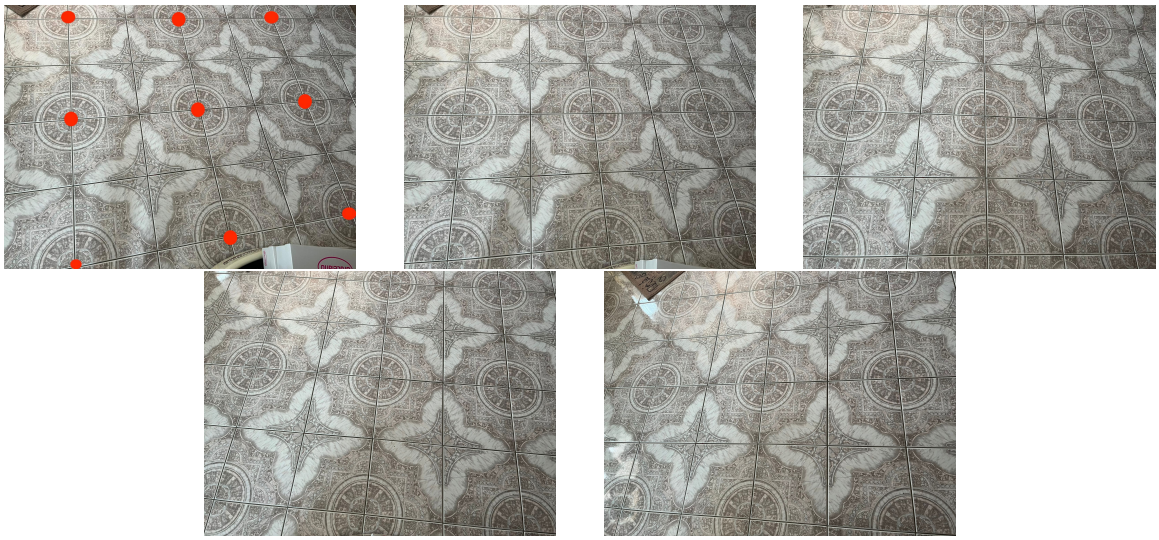

Figure 4: Set of images of the a floor pattern

Some data about the calibration:

◇ Size of the titles of the floor in milimeters overall is 1300x1300mm.

◇ The resolution of the captured images (in pixels) = 4032 x 3024.

◇ Internal parameters (matrix A') of the camera:

$$A' = \begin{bmatrix} 2967 & 7.82 & 2055 \\ 0 & 3001 & 1667 \\ 0 & 0 & 1 \end{bmatrix}$$

Analysis of the following three aspects:

◇ **Are the pixels of your camera square?**
To calculate the aspect ratio of a camera given its intrinsic parameters, we can simply divide the focal length in the y-direction by the focal length in the x-direction. The aspect ratio is usually denoted by r and is given by: $r = \frac{f_y}{f_x} = 0.988$ which is close to 1 (with some noise) which means $f_x = f_y$. Hence, we can say that the camera has square pixels.

- ◇ **Which is the degree of coincidence between the principal point and the center of the image plane?**
  If the principal point is coincident with the center of the image plane, then c_x and c_y are equal to half the width and height of the image, respectively. In this case, the degree of coincidence is 100%. However, if the principal point is not coincident with the center of the image plane, then the degree of coincidence will be less than 100%. One way to measure the degree of coincidence is to calculate the distance between the principal point and the center of the image plane and express it as a percentage of the image width (w) or height (h).

$$d_x = \frac{\left|\frac{w}{2} - c_x\right|}{\frac{w}{2}} \times 100\% = 1.89\%$$

$$d_y = \frac{\left|\frac{h}{2} - c_y\right|}{\frac{h}{2}} \times 100\% = 10.25\%$$

$$d = \frac{d_x + d_y}{2} = 6.07\%$$

  Overall the average offset is 31 which means around with 31% offset is not in the middle of the center according to the noise.

- ◇ **Are the axes of the image plane orthogonal?** To prove this we have to calculate θ from the skew parameter of the K matrix (s):

$$\beta = \arccos(\frac{s}{\alpha}) = 89.84^\circ$$

  in degrees which is close to 90° meaning the planes are orthogonal.

Conclusion: In theory, both A and A' should represent the same intrinsic camera parameters, assuming that the calibration process is accurate and the calibration objects are correctly used since we use the same camera. However, calibration errors, measurement inaccuracies, or limitations of the calibration method can contribute to differences between A and A'. For using the floor as calibration object, the probability of the error by manually selecting the points is higher. Hence, for the subsequent tasks we will be using A matrix obtained by the 1080p checkerboard.

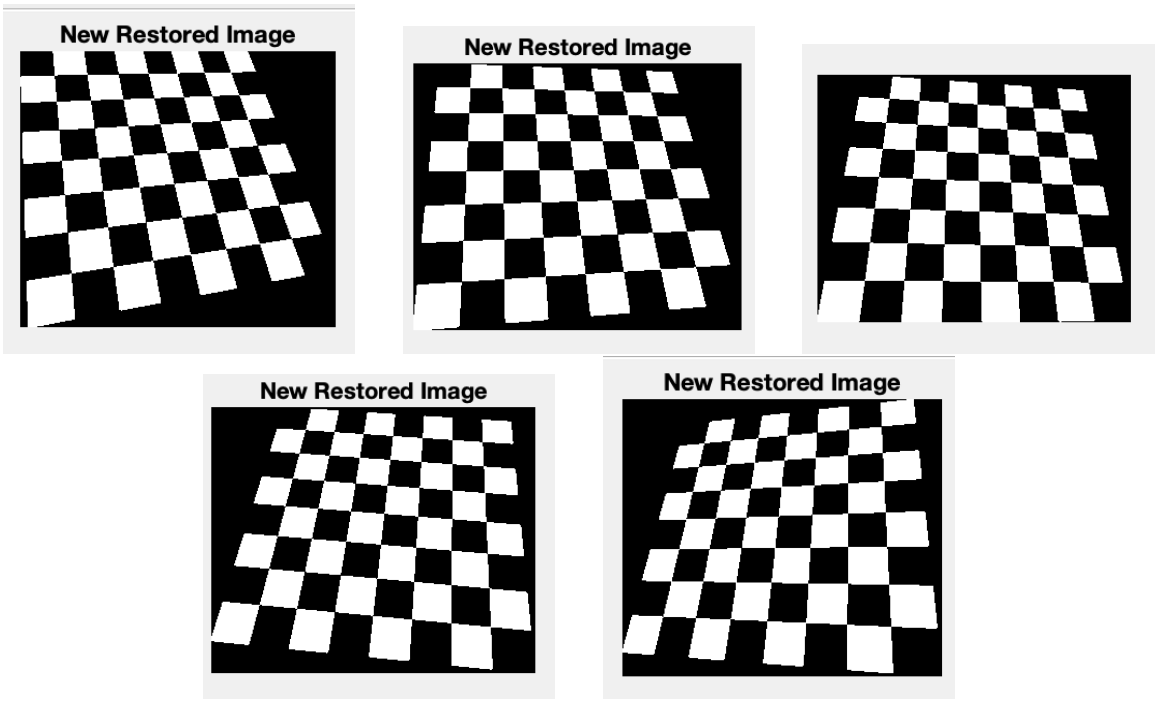

Figure 5: Showing the respective homography images for the the floor with 5 different orientations

2 Section 2: Finding local matches between several views of an object.

2.1 About the scene

The figure 6 shows a montage of images taken at a resolution of 4032x3024 from different perspectives (left, center, and right) of the scene. To enhance computational speed and prevent memory issues, the images have been scaled down by a factor of 0.5, resulting in minimal information loss. The process of capturing these images posed various challenges, including ensuring appropriate lighting conditions and capturing them from angles that allow for sufficient detection and matching of corresponding point correspondences. Moreover, it was crucial to ensure the robustness of the detected points. The captured scene presents additional obstacles such as the presence of reflected light (displayed as green color from the object on the left side) visible in images I1, I2, and I4. These particular images (I4 to I5) present a challenge for the feature point detector and matcher due to the potential difficulties caused by the reflected light, which may hinder the detection and matching of feature points. For the scene an object (image frame) with strong texture elements has been selected.

The different view points and orientations of the scene. Left view: I1, I2, I3. Center view: I4, I5, I6. Right view: I7, I8, I9.

2.2 Detector and Descriptor combination

To identify two distant scene images among various options, the key is to locate a view pair with a substantial number of corresponding points. This approach enables the generation of a solid initial reconstruction based on these two distant views. Nevertheless, it becomes more challenging to find point matches in views that have a wider

Figure 6: Montage about the 3D Scene

angle between them.

As a result, Image pair I4 and I5 were chosen since they exhibit a relatively small angle between them, while still offering different perspectives. This selection was made intentionally to test the capabilities of the available detectors and descriptors. The chosen images provide an appropriate setting for identifying a suitable combination of Detector and Descriptor.

As shown in table 1, six different efficient combinations of detector+descriptor were tested: DoH+SIFT, SURF+SURF, KAZE+KAZE, SIFT+DSP-SIFT, SURF+SIFT, SURF+KAZE. These methods were with the parameters below: threshold = 0.001 for detection; max ratio= 0.5 for point matching; nscales = 10; noctaves=3; Metric = 'SSD'; npoints = 350.

From the the table it can be observed that KAZE+KAZE has outperformed all of the other detector-descriptor combinations and successfully handled the strong changes in the perspective as well compared to detectors like SIFT, DSP-SIFT and SURF. Hence, for further analysis, we will proceed with KAZE+KAZE detector-descriptor combination.

	DoH+SIFT	SURF+SURF	KAZE+KAZE	SIFT+DSP-SIFT	SURF+SIFT	SURF+KAZE
Number of in-liners in calculating the homography transform	203	247	**629**	832	207	102
Number of in-liners in calculating the fundamental matrix	270	370	**773**	719	270	190
Number of corresponding points	540	739	**1546**	1437	540	380

Table 1: Number of in-liners and corresponding points for different detector descriptor combination on the image pair of I4-I5

KAZE has several advantages that make it a popular choice compared to other descriptors and detectors in certain scenarios. KAZE is a multi-scale 2D feature detection and description algorithm in nonlinear scale spaces which allows for better adaptation to different image structures and provides more flexibility in detecting key points at multiple scales. It uses other feature (conductivity) for handling the noise reduction and keeping the natural borders between objects as well which can not be preserved in linear scale space detector-description combinations which use Gaussian for blurring. In the further experiment, more image pairs have been tried out using KAZE+KAZE detector-descriptor to measure its robustness and performance.

	I2-I3	I3-I4	I8-I9	I7-I8	I1-I9	I4-I8	I1-I6	I4-I9	I2-I7	I5-I9	I4-I5
Number of in-liners in calculating the homography transform	1519	382	2754	1122	353	733	354	309	272	262	**559**
Number of in-liners in calculating the fundamental matrix	1426	681	2988	1357	352	1410	563	657	494	409	**773**
Number of corresponding points	2851	1362	5976	2714	703	2820	1126	1313	987	818	**1546**

Table 2: Comparison between the number of in-liners for homography, number of in-liners for the fundamental matrix and number of corresponding points for different selected image pairs.

After selecting the KAZE+KAZE as the detector-descriptor combination, it was used to test on different image pairs to measure it's robustness and performance for different image pairs and observed that it works very well for different image pairs and it matches enough number of points even for images with strong perspective changes. Even for the challenging scenario where in images (I2 to I3) we had observed reflection. It was able to detect and match a large number of points. The number of in-liners obtained for different image pairs can be seen in table 2. The estimated homography matrices tform21 and tform12 appears to be correct as there were enough point correspondences. The transformation of image 2 with respect to image 1 can be observed in the left image refer 8 and the transformation of image 1 with respect to image 2 can be observed in right image 8.

Estimated Fundamental matrix: The rank of the obtained fundamental matrix F is 2 (checked using matlab rank function). Since the rank of F is 2 and the epipolar lines are spinning around one point refer 9, the estimation of the fundamental matrix F appears to be correct. The corresponding point should lie on the epipolar line on the second image. Although the image pair of I1-I9 has the largest number of point correspondences the fundamental matrix checked by visually it was not correct since the center of the spinning was outside of the camera area. This is why image pair of I4-I5 was selected.

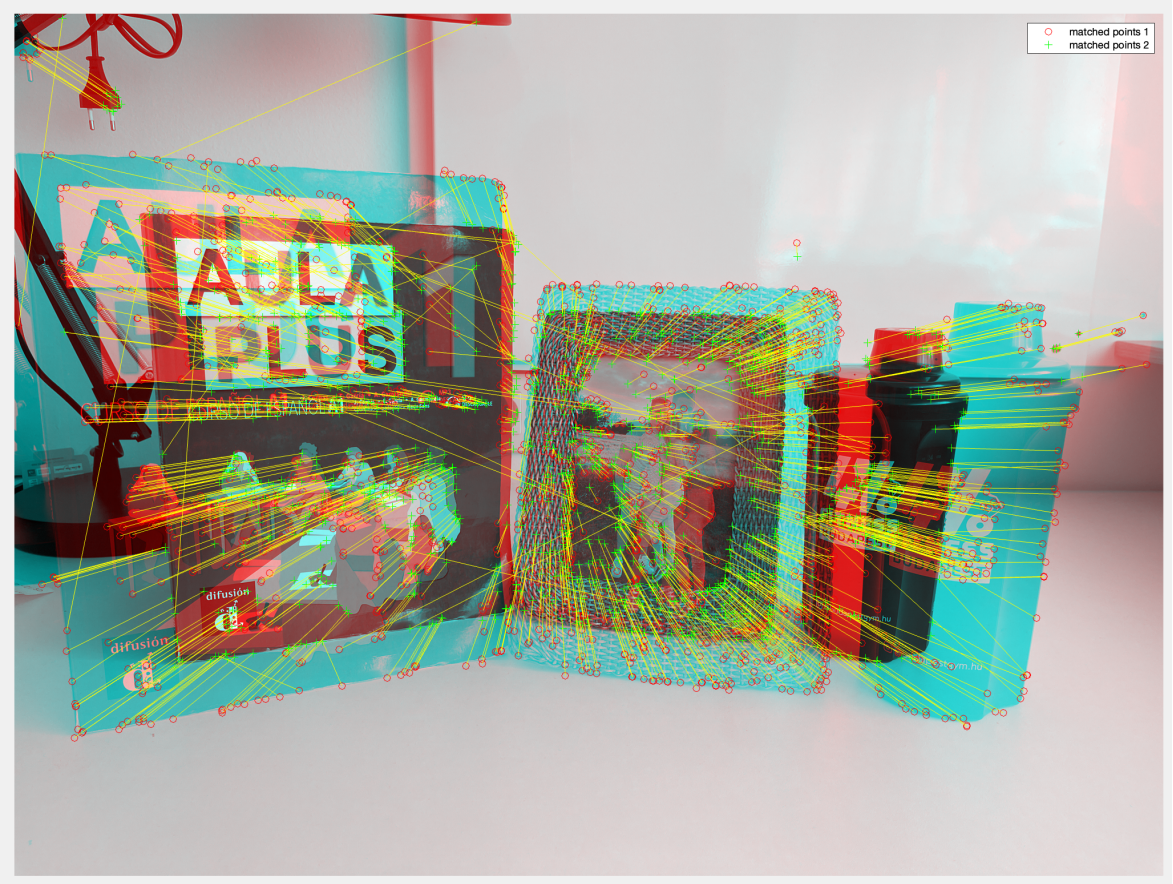

Figure 7: Point correspondences between the image pair I4 and I5.

$$F = \begin{bmatrix} 4.8 - 2e^{-7} & 7.49e^{-5} & 4.953e^{-3} \\ 7.4e^{-5} & 6e^{-7} & -0.0620 \\ 4.95e^{-3} & -0.0645 & 0.9936 \end{bmatrix}$$

The estimated homography matrices which shows the transformation which takes the one to the another image:

$$tform21 = \begin{bmatrix} 1.21 & -0.0083 & 8.38e^{-7} \\ -0.0139 & 1.19 & -2.4e^{-5} \\ -216.8 & -145.9 & 1 \end{bmatrix}$$

$$tform12 = \begin{bmatrix} 0.84 & 0.018 & 1.8e^{-5} \\ 0.0014 & 0.84 & 1.03e^{-5} \\ 167.5 & 115.3 & 1 \end{bmatrix}$$

Figure 8: (Left) Transformation between Image 1 and Image 2 (Right) Transformation between Image 2 and Image 1

Figure 9: Screen captures of the GUI, showing the epipole in the image pair.

3 Section 3: 3D reconstruction and calibration

In section 2, it was determined that there was a significant change in perspective in images I4 and I5, and an adequate number of matched points for initial reconstruction. Consequently, to choose the remaining set of images, a comparison was conducted among various pairs of images, and their corresponding points were recorded, as shown in table 3.

For the purpose of N-view point matching and achieving a robust 3D point cloud reconstruction, two different sequences have been tried out. One of them has the highest number of point correspondence but lowest number of images taken the cameras. This sequence is: I1, I2, I3

The other sequence is a subset of images (I4, I5, I6, I7, I8, I9) was utilized. The identified points of interest in these images can be observed in figure 10. A maximum ratio value of 0.8 was employed for both of the matching, as a higher value was selected to enhance the number of point correspondences. The positive thing here is that with more images we can achieve more realistic 3D scenario but we have to consider that we can have more mismatched points which can have an affect on the error.

For following the report easily, only the second case is discussed and for the other

one the results is provided only.

Figure 10: Row-1(Left to Right) I4,I5,I6, Row-2(Left to Right)I7,I8,I9. Images used for the N-view point matching, and the detected interest points in each of them.

Set of images	Number of matched points
I1, I2, I3, I4, I5, I6, I7, I8, I9	289
I1, I2, I3	**3660**
I1, I2, I3, I4	1654
I4, I5, I6	2379
I4, I5, I6, I7	1752
I7, I8, I9	3121
I2, I3, I4, I5, I6, I7, I8	407
I1, I2, I3, I4, I5, I6	579
I2, I3, I4	2055
I4, I5, I6, I7, I8, I9	**891**

Table 3: Comparison between the correspondence points between different images pairs to find a subset of images from the scene for 3D point cloud reconstruction.

3.1 Initial projective reconstruction

The corresponding reprojection error histogram of initial projective reconstruction of 2 cameras (first, last one) can be seen in figure 11. These correspondences of the two cameras can be seen below.

Figure 11: (Left)I4, (Right)I9; images used for the estimation of the fundamental matrix, with the detected interest points and point matches.

Residual re-projection error. 8 point algorithm: 2877.0181
Pixel error: mean = [-1.62859 -0.96803], std = [36.21340 66.65032]

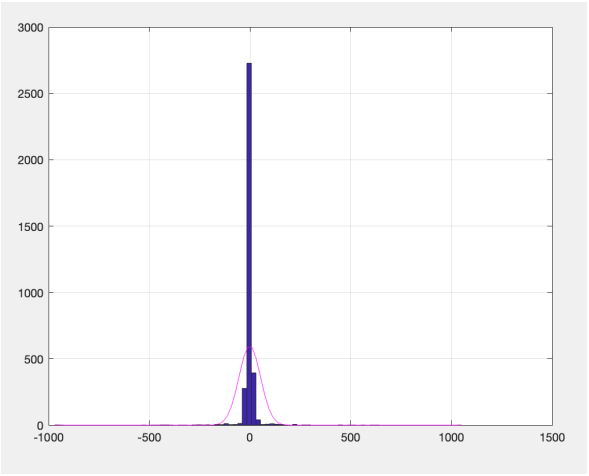

Figure 12: Reprojection error Histogram of initial projective reconstruction.

3.2 Improving the initial reconstruction by Projective Bundle Adjustment.

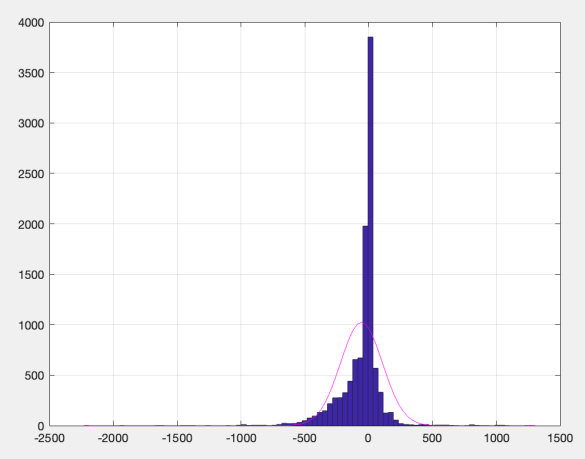

Figure 13: Reprojection error Histogram after re-sectioning step

Residual reprojection error, After resectioning = 30088.2519
Pixel error: mean = [-51.64321 -60.12252], std = [188.06398 136.14988]

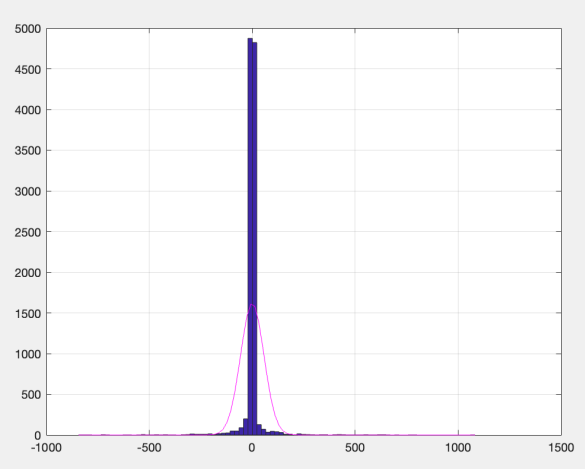

Figure 14: Reprojection error histogram, after the projective bundle adjustment step.

Reprojection error, After Bundle Adjustment = 3204.4133
Pixel error: mean = [0.39860 0.28263], std = [57.31561 55.89909]

The corresponding reprojection error Histogram, after the projective bundle adjustment step can seen in figure 14.

Justification for the different re-projection error values: The error for the initial reconstruction was calculated using the 8-point algorithm on the data points of the two end cameras. The error value after the initial reconstruction was 2877.0181. After the initial reconstruction (using two end cameras), we did re-sectioning using all the available cameras. Since, we have increased the number of cameras (using all the cameras), the re-projection error will be summed up for all cameras and hence, the error after re-sectioning should increase. We observed the same thing that error increased from 2877.0181 to 30088.2519 after re-sectioning. After performing the bundle adjustment step, the re-projection error must reduce, because we are using the

3D points to calibrate all the cameras, which must minimize the error in all views. However the error did not decrease from 2877.0181. The reason behind it that the additional camera views incorporated in the bundle adjustment process introduced more ambiguous or noisy correspondences, leading to a larger error compared to a smaller subset of images.

3.3 Euclidean reconstruction of the scene

The corresponding reprojection error histogram, for the euclidean reconstruction of the scene can seen in figure 15.

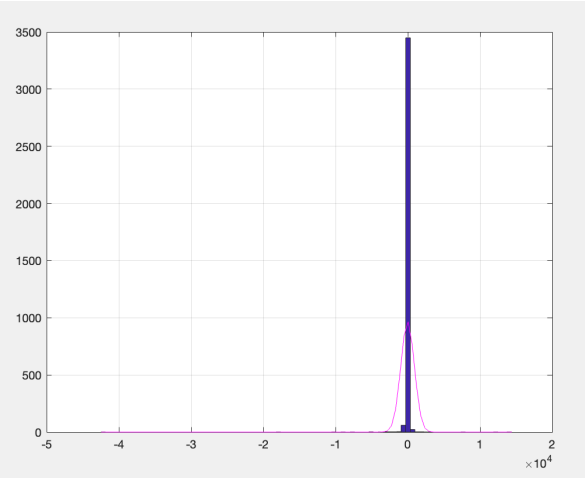

Figure 15: Reprojection error histogram for euclidean reconstruction of the scene

Comments on the Euclidean reconstruction: We have obtained four possible solutions. Solution 4 was the real illustration of the captured scene. The reconstruction seemed accurate visually. The 3D point cloud reconstruction can be observed in figure 16. The different elements in the scene can be distinguished in the 3D world and the visualization of the scene in the 3D world can be observed in the figure 17 with real depth and cameras are on the correct positions. However, the illustration had some pixel errors and the shaker (black object) did not have many matching points.

3.4 Conclusion

In conclusion, we can see that the euclidean reconstruction was very realistic. It illustrates well the relative distances and the depth of the scene. There were 3 distinguishable parts in the scene (book, frame, shaker). With the figures 17, 18 it was illustrated that the relative distances between the objects were well-preserved. The detected points on the objects remained planar. On the book the title and some texts are detected quiet well, the image frame is detected really well because of the strong textures. Using smaller number of images the shaker can be seen as well with some texts.

For the second experiment where smaller number of the cameras have been used the error decreased since we had more good matching points. The shaker is more detected. As the number of the matching was increasing the scene lost the depth and the calibrations of the cameras were not accurate anymore (Figure 18).

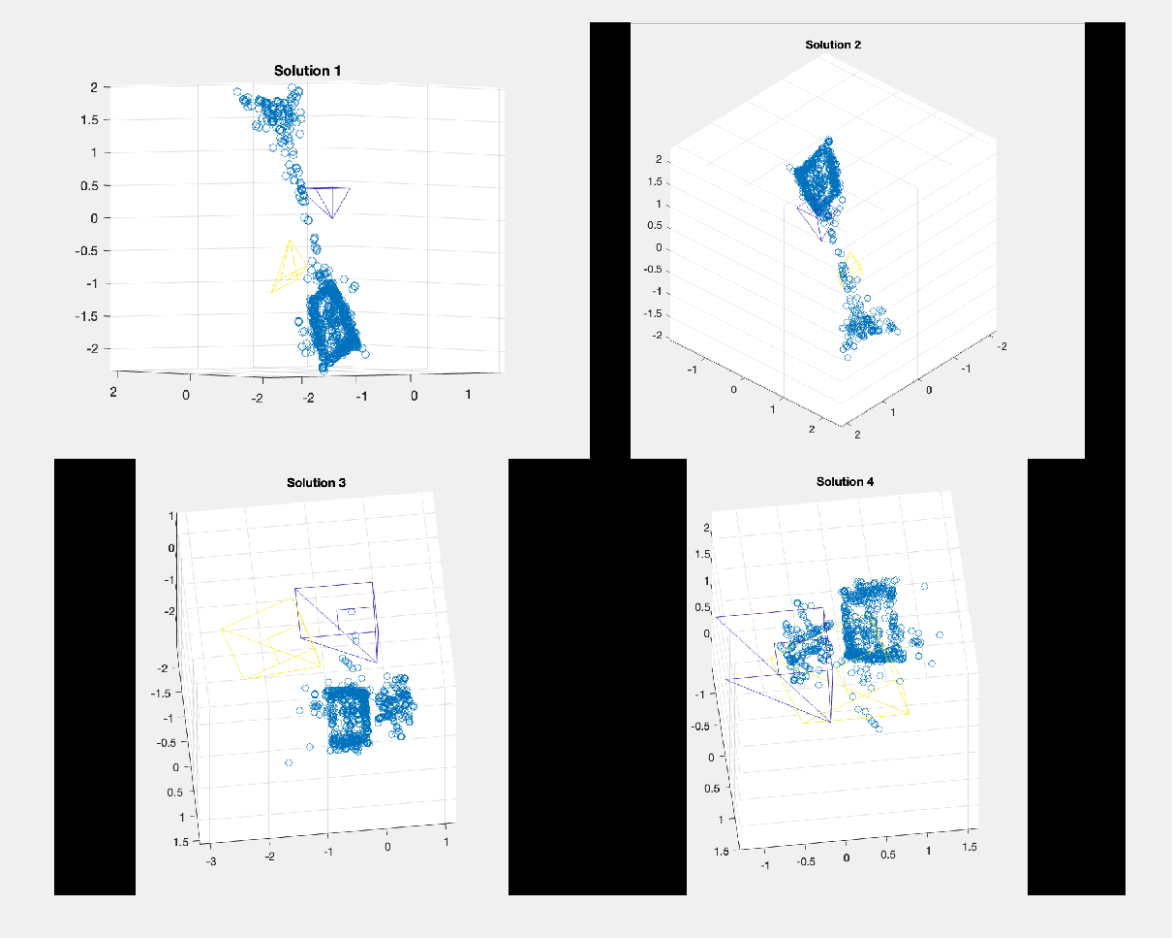

Figure 16: Results showing several viewpoints of the 3D point cloud reconstruction using images from I4 to I9

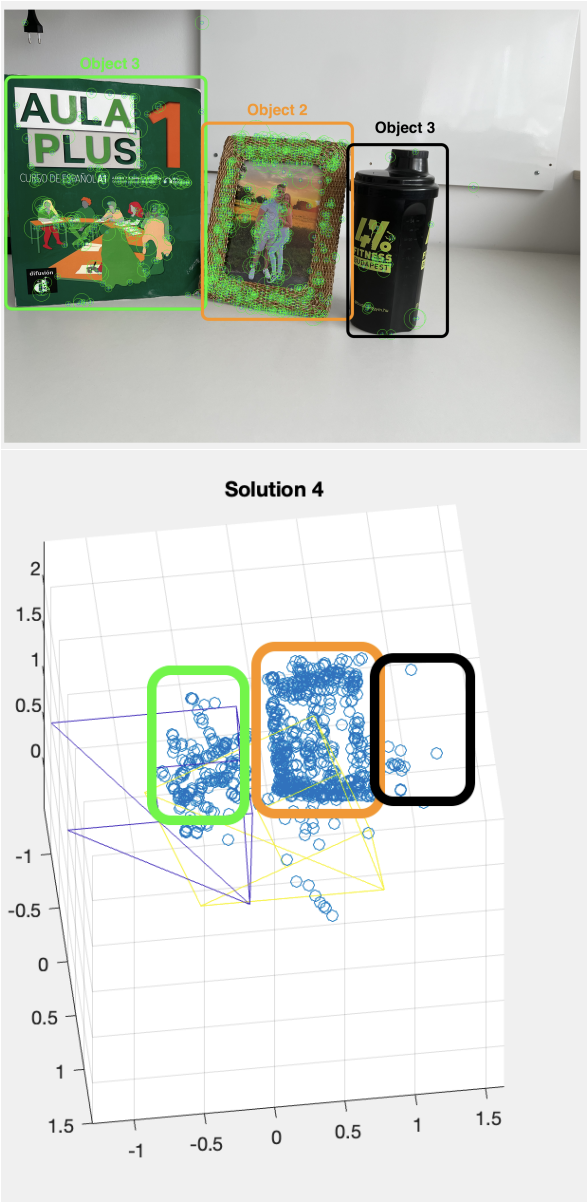

Figure 17: Visualization of the scene in the 3D world. (Top) objects in the scene (Image 6) and (bottom) corresponding objects in the reconstructed 3D point cloud.

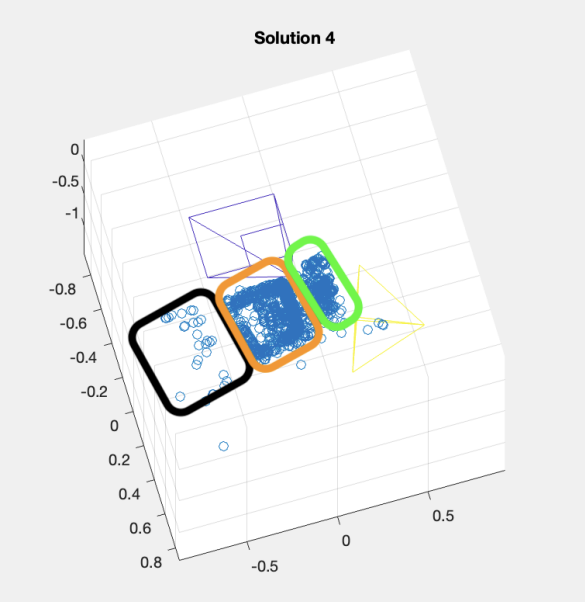

Figure 18: Results showing several viewpoints of the 3D point cloud reconstruction using images from I1 to I3