Vision for Multiple or Moving Cameras
Final Lab Evaluation

Subham Shome
subham.shome@estudiante.uam.es
IPCV 2022-24
Escuela Politécnica Superior — Universidad Autónoma de Madrid

Introduction

The objective of the suggested approach is to acquire a three-dimensional reconstruction of an object. To achieve this, images of the object were captured using a chosen camera and subsequently calibrated. Sufficient images of the object were taken to ensure an appropriate number of views. This facilitated the extraction and matching of feature points across different views, as well as the computation of the fundamental matrix between views. The final step involved generating a three-dimensional point cloud reconstruction and representing the object's geometric components within this point cloud.

I. Obtention of the intrinsic parameters of a camera

A camera's intrinsic (or internal) parameter matrix A or K is given as follows:

$$K = \begin{bmatrix} F_x & S & C_x \\ 0 & F_y & C_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \alpha & -\alpha cos\theta & U_0 \\ 0 & \beta/sin\theta & V_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

where, F_x and F_y are the horizontal and vertical focal lengths of the camera respectively, (C_x, C_y) is the principal point, S is the skew of the camera and α is the aspect ratio.

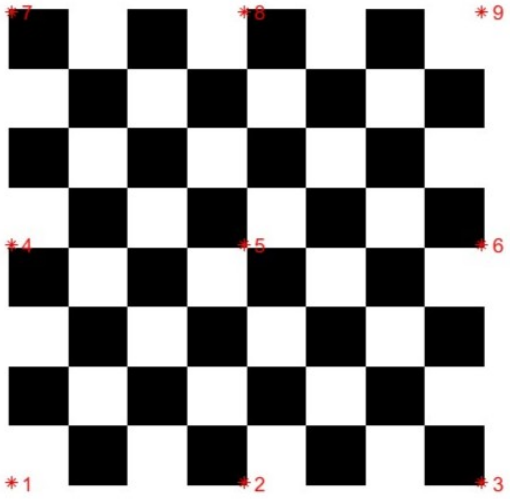


Fig. 1: Initial Checkerboard with marked points

In this project, we are supposed to calibrate our cameras and choose one of the internal parameter matrices for the full 3D Scene Reconstruction. Figure 1 shows the initial image with marked points.

A. Camera Calibration: Image 1

It was required to know a couple of values before starting off with the images. They are as follows:

1) Size (in mm) of the checkerboard (1080p) in the screen: **90 × 90**
2) Resolution of the image (in pixels): **4608 × 2592**

A set of 9 images were used to obtain the internal parameters matrix A. It is as follows:

$$A = \begin{bmatrix} 3645.552449 & -10.126372 & 2289.652099 \\ 0 & 3633.830078 & 1354.472463 \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

Figure 2 shows Zhang's homography for the set of 9 images of image 1.

Fig. 2: Zhang's Homography - Image 1

1) *Are the pixels of your camera square?*
   From 2, we can see that F_x is almost equal to F_y with the presence of some noise. Moreover, the ratio
   $$\frac{F_x}{F_y} = \frac{3645.552449}{3633.830078} = 1.0032259$$
   is almost equal to 1. So, we can say that **the camera has square pixels**.
2) *Which is the degree of coincidence between the principal point and the center of the image plane?*
   The image dimensions are 4608 × 2592 pixels. By analyzing the internal parameters matrix A, it can be observed that the value for C_x is approximately half of the image width (4608), which is approximately 2304. Similarly, the value for C_y is close to half of the image height (2592), approximately 1296, with some noise. Thus, **the principal point (2289, 1354) is very close to the center of the image plane at (2304, 1296)**, with a slight offset in the coordinates due to some noise.
3) *Are the axes of the image plane orthogonal?*
   In order to assess the skewness (θ or S), we have

information about α and $-\alpha cos\theta = -10.126372$. Solving for θ yields a value of $90.159153°$, which is very close to $90°$. Therefore, we can confidently state that **the image planes are almost orthogonal to each other**.

B. Camera Calibration: Image 2

We do the same as we did for the seocnd image, which was the provided 720p image of the checkerboard. The pre-requisite values are as follows:

1) Size (in mm) of the checkerboard (1080p) in the screen: **90** $\times$ **90**
2) Resolution of the image (in pixels): **4608** $\times$ **2592**

The internal parameters matrix A' for another set of 9 images is as follows:

$$A' = \begin{bmatrix} 3608.359131 & -38.406212 & 2330.131226 \\ 0 & 3593.776685 & 1313.569133 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

Figure 3 shows Zhang's homography for the set of 9 images for image 2.

Fig. 3: Zhang's Homography - Image 2

1) *Are the pixels of your camera square?*
From 3, we can see that F_x is almost equal to F_y with the presence of some noise. Moreover, the ratio
$$\frac{F_x}{F_y} = \frac{3608.359131}{3593.776685} = 1.004057$$
is almost equal to 1. So, we can say that **the camera has square pixels**.
2) *Which is the degree of coincidence between the principal point and the center of the image plane?*
The image dimensions are 4608 $\times$ 2592 pixels. By analyzing the internal parameters matrix A, it can be observed that the value for C_x is approximately half of the image width. Similarly, the value for C_y is close to half of the image height with some noise. Thus, **the principal point (2289, 1354) is very close to the center of the image plane at (2330, 1313)**, with a slight offset in the coordinates due to some noise.
3) *Are the axes of the image plane orthogonal?*
In order to assess the skewness (θ or S), we have information about α and $-\alpha cos\theta = -38.406212$. Solving for θ yields a value of $90.609849°$, which is very close to $90°$. Therefore, we can confidently state that **the image planes are almost orthogonal to each other**.

C. Conclusion - Part 1

In this section, the camera was calibrated using two sets of images. It is noted that the calibration using 1080p image of the chekerboard (Image 1) gave better results than the other image. So, the internal parameter matrix of the first image (A) is considered for the further analysis.

In the next section, an original scene will be presented and local matches between different pairs of images from the scene will be calculated.

II. Finding Local Matches Between Several Views of an Object

A. Description of the Scene

Figure 4 shows the montage of 12 images used for obtaining the points of interests (PoIs) using different detectors and descriptors. The images are captured by the same camera (that was calibrated in Section 1), which has a resolution of 4608 $\times$ 2592 pixels. The objects in the scene are (from left to right) A UAM notebook, an iPad and a packet of snacks. The reason for choosing these objects were that they are somewhat planar yet not all of them are fully planar, which gives enough variety on getting the homography and the PoIs. The scene is subjected to various lighting conditions and different textures (screen and books). It is also worth noticing that the first 5 images are subjected to partial reflection which should make it difficult to detect points.

Fig. 4: Input Images for the Original Scene

The images from different views are as follows:

- **Left View:** Images 1 to 4
- **Center View:** Images 5 to 8
- **Right View:** Images 9 to 12

B. Combining various Detectors and Descriptors

The objective of identifying two distant scene images from various options is to obtain a view pair that contains a substantial number of corresponding points. This allows for a reliable initial reconstruction of the two distant views. However, it becomes more challenging to find point matches in views with a wider angle. Therefore, in order to test the capabilities of the available detectors and descriptors, it was decided to select Image 3 and Image 10. This image pair was selected because it does not have a significantly wide angle, and the images exhibit distinct perspectives. The chosen images are suitable for determining an appropriate combination of detector and descriptor.

As shown in table I, various combinations of detectors and descriptors were tested for the scene: $DoH + SIFT$, $SURF+SURF$, $KAZE+KAZE$, $SIFT+DSP-SIFT$, $SIFT + SIFT$, $SURF + DSP - SIFT$, $SURF + SIFT$

	DoH + SIFT	SURF + SURF	**KAZE + KAZE**	SIFT + DSP-SIFT	SIFT + SIFT	SURF + DSP-SIFT	SURF + SIFT	DoH + SURF
Number of inliers, calculating the homography transformation between the two views	179	144	**865**	52	116	64	128	219
Number of inliers, calculating the fundamental matrix between the two views	100	102	**858**	52	82	38	75	122
Total number of matched points	199	203	**1716**	104	163	76	149	243

TABLE I: Combination of Detectors and Descriptors and the number of inliers

and $DoH + SURF$. The following parameters were used to test the methods:

- threshold = 0.001 for detection
- MaxRatio = 0.5 for matching
- nscales = 10
- noctaves = 3
- Metric = SSD
- npoints = 350
- sigma0 = 1.6

Out of all the detectors, only a handful of them were found to be performing well on the set of images. $KAZE+KAZE$ was performing the best and found 1716 matching points between the images.

Reasoning for $KAZE+KAZE$ to perform better than others:
KAZE is an algorithm for detecting and describing multi-scale 2D features in nonlinear scale spaces. It has demonstrated superior performance compared to other methods. Unlike those approaches that use Gaussian scale space to detect and describe features at different scales, KAZE employs nonlinear diffusion filtering in a nonlinear scale space. This approach addresses the limitations of Gaussian blurring, which does not respect object boundaries and equally smoothes details and noise. By adapting the blurring locally to the image, KAZE effectively reduces noise while preserving object boundaries. This results in enhanced localization accuracy and distinctiveness of the detected 2D features.

Table II shows the results of $KAZE+KAZE$ on various sets of images. It is worthy to note that the image pair of (6,7) gave the best values, however, from Figure 4, it is seen that they are very close to one another, so there is not a lot of change in perspective. Hence, the image pair (3,10) was selected for further purposes.

After the detections and descriptions were done, the estimated homography matrices $tform12$ and $tform21$ were calculated. They are as follows:

$$tform12 = \begin{bmatrix} 0.495637 & -0.184722 & -0.000221 \\ 0.072905 & 0.766941 & 0.000019 \\ 306.251709 & 193.741806 & 1.000000 \end{bmatrix}$$

$$tform21 = \begin{bmatrix} 1.899579 & 0.348113 & 0.000403 \\ -0.164445 & 1.414661 & -0.000060 \\ -538.761230 & -374.593262 & 1.000000 \end{bmatrix}$$

Finally, the Fundamental Matrix (F) was estimated.

$$F = \begin{bmatrix} 9.914713e-08 & 8.424124e-07 & -0.001161 \\ 1.077317e-07 & -1.584797e-07 & -0.003274 \\ 0.000123 & 0.002345 & 0.999991 \end{bmatrix}$$

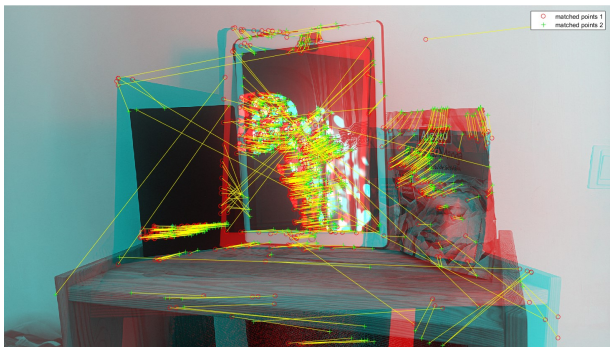


Fig. 5: Detected PoIs for image pair (3,10) using KAZE + KAZE

Image 3 and Image 10 were chosen for their relatively moderate angle, making them suitable for the initial reconstruction. Additionally, these images exhibit significant perspective changes, posing a challenge for various detector and descriptor combinations. Through a comprehensive evaluation, it was determined that the KAZE detector paired with the KAZE descriptor outperformed other combinations. This particular combination proved effective in handling the substantial perspective changes and successfully matched enough points for further analysis.

By utilizing the KAZE detector and matching points in the image pair (3,10), a reliable **fundamental matrix with a rank of 2** was obtained, and the estimated homography matrices were accurate.

Note: Harris affine detector with DSP-SIFT descriptor was tried but was unsuccessful. Although the detected corner points were obtained and were good, the descriptor failed to give proper results and caused an error.

III. 3D Reconstruction and Calibration

This section is mostly based on the residual errors that were calculated after every step. The steps taken were as follows:

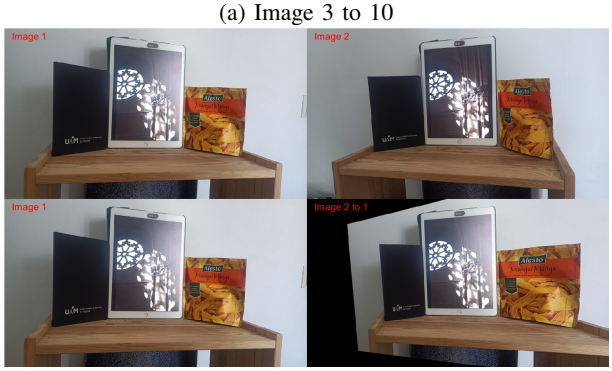
1) 8-point algorithm
2) Resectioning
3) Bundle Adjustment
4) Euclidean reconstruction

From Section 2, it is noted that the image pair (3,10) was used to get the data and the points of interest. Now, several image subset combinations were tested for getting the N-view matching. It was clearly visible that the image subset (3,4,5,6,7,8,9,10) was going to give better results than the rest. This was considered to be the final result.

	Image 1, Image 12	Image 2, Image 11	Image 4, Image 9	Image 5, Image 8	Image 6, Image 7
Number of inliers, calculating the homography transformation between the two views	125	143	399	541	1273
Number of inliers, calculating the fundamental matrix between the two views	83	108	311	569	1168
Total number of matched points	165	203	622	1137	2272

TABLE II: Combination of Image pairs with KAZE + KAZE detector-descriptor combination

(a) Image 3 to 10



(b) Image 10 to 3

Fig. 6: Transformation of image pair (3,10) to each other

Fig. 7: Epipolar Line

However, various other image combinations were tried and the respective residual errors were recorded. Table III shows the results of various image subsets and their residual errors after every step.

Now, the projective reconstruction is performed for the image subset 3-10.

A. *Initial Reconstruction*

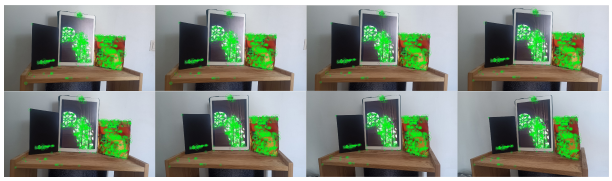


Fig. 8: Final Points for the whole image subset 3-10

The following values of residual reprojection errors were found:

Residual reprojection error, 8 point algorithm = **59.877**
Pixel error: mean = [-0.21586 3.54450]
Pixel error: std = [0.96532 10.30805]

The respective reprojection error (pixels) and histogram is present in Figure 9.

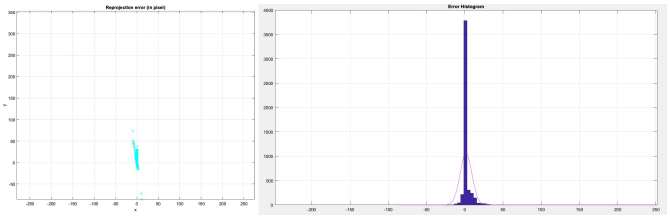


Fig. 9: Error and Histogram, after 8-point algorithm

B. *Improving the Reconstruction - Resectioning and Bundle Adjustment*

The following values of residual reprojection errors which were found after resectioning:

Residual reprojection error after resectioning = **31.4995**
Pixel error: mean = [-0.77070 0.63162]
Pixel error: std = [4.89505 6.16855]

Figure 10 shows the error and histogram after resectioning. Figure 11 shows the error and histogram for the following values of reprojection errors which were found after bundle adjustment:

Reprojection error after Bundle Adjustment = **10.0188**
Pixel error: mean = [-0.00000 -0.00000]
Pixel error: std = [1.92341 4.04230]

Image Set	Minimum Singular Value	Residual Reprojection Errors			
		8-point algorithm	Resectioning	Bundle Adjustment	Euclidean reconstruction
3 to 10	**1.114**	**59.877**	**31.4995**	**10.0188**	**0.54592**
5 to 12	0.48384	6.6634	10.1203	2.6984	539.2793
4 to 9	2.6106	115.3553	295.2872	117.4642	90.9188
1 to 8	0.3722	3.2778	31.102	10.3226	29.1322
1 to 12	0.25011	5.1799	7.3017	2.396	8.6364
4 to 10	1.9821	542.1173	329.6853	31.0668	757.4474

TABLE III: Comparison between residual errors and different image subsets

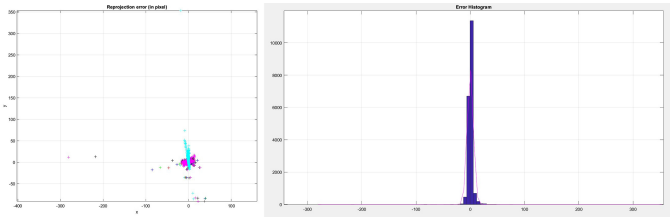


Fig. 10: Error and Histogram, after resectioning

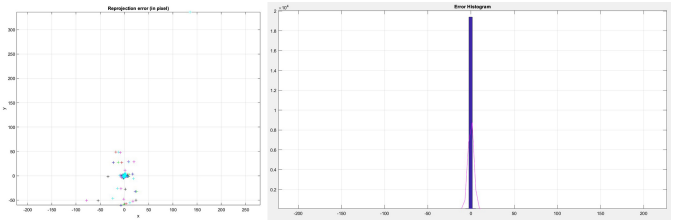


Fig. 11: Error and Histogram, after bundle adjustment

C. Euclidean reconstruction of the Original Scene

The following errors were noticed for the final euclidean reconstruction.

Reprojection error, euclidean reconstruction = **0.54592**
Pixel error: mean = [0.00606 -0.04793]
Pixel error: std = [0.11971 1.03712]

Very low amount of error was achieved for the final reconstruction. Figure 12 depicts the final error histogram and pixels after euclidean reconstruction.

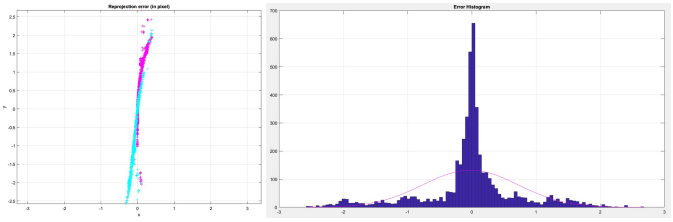


Fig. 12: Error and Histogram, after euclidean reconstruction

Finally, the next two figures depict the N-view matching and final reconstruction of the images from different image subsets (refer to Figure 13) and the final euclidean reconstruction that was acquired with the chosen image subset (3-10).

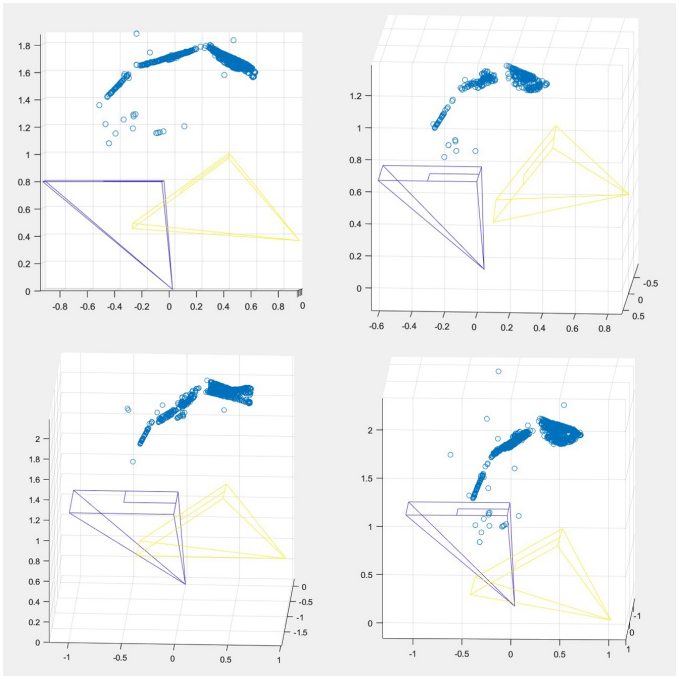


Fig. 13: 4 different euclidean-reconstructions

D. Conclusion - Section 3

The euclidean reconstruction yielded remarkably realistic results, effectively capturing the relative distances and depth of the scene. The scene itself comprised four distinguishable components, namely objects 1,2 and 3, and foreground and background points. Figure 15 demonstrates the accurate preservation of relative distances between these objects. Furthermore, the detected points on the objects remained planar, contributing to the overall fidelity of the reconstruction. The 3D point cloud reconstruction was conducted using a relatively low error (0.54592), which greatly contributed to the quality of the final output. As a result, a realistic 3D point cloud reconstruction was achieved, characterized by minimal errors.

Conclusion

This project was designed to reconstruct a 3D scene based on moving or multiple camera captures of the scene. The proposed problem involved a series of steps, including the selection and calibration of a camera, the identification of a suitable object or scene based on given indications, the selection of an appropriate number of views, the extraction

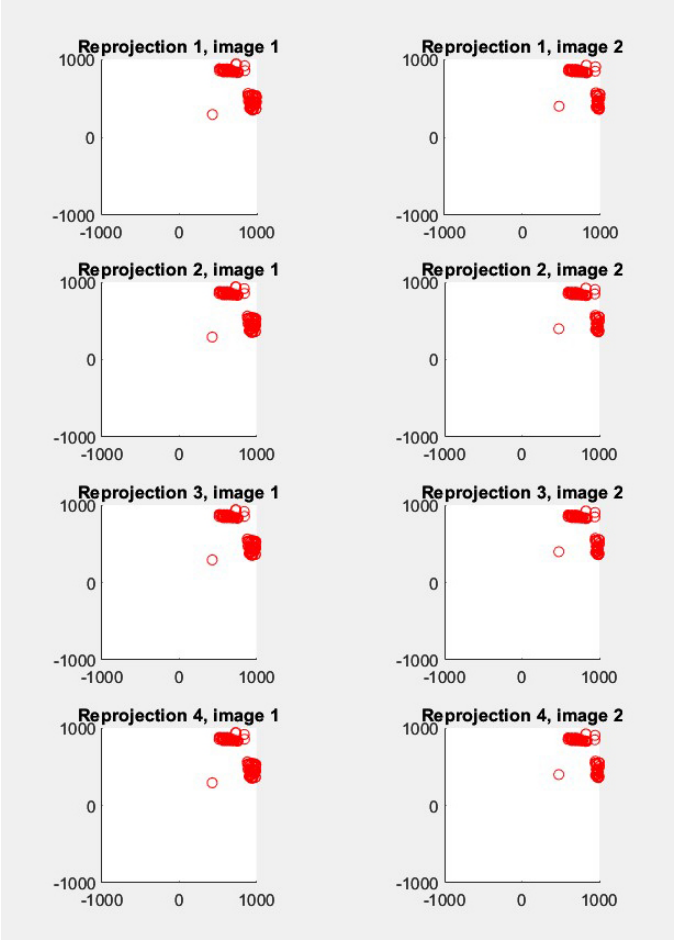
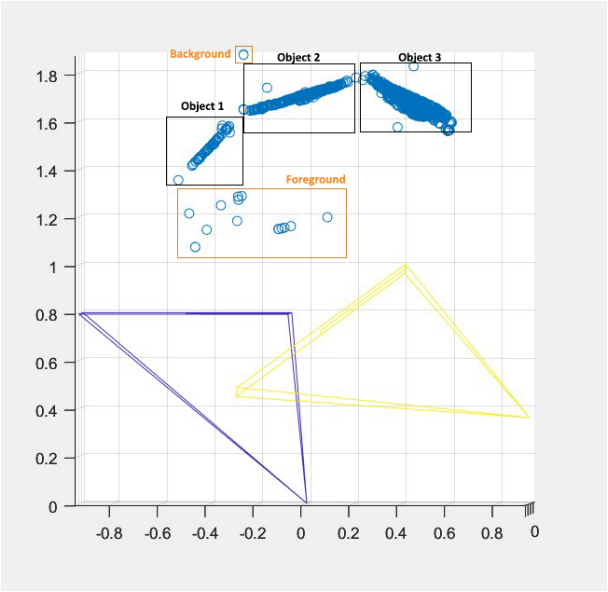


Fig. 14: Reprojections

(a) Original Scene



(b) Final 3D Reconstruction (top view here)

Fig. 15: Comparison of original image and the final reconstructed result

and matching of feature points across the views, the computation of the fundamental matrix between the views, the generation of a 3D point cloud reconstruction, and finally, the representation of object geometric elements using this point cloud.

The results obtained were of high quality. The 3D reconstruction demonstrated good accuracy and effectively captured the geometric characteristics of the object or scene. The chosen camera was well-calibrated, leading to precise feature point extraction and matching. The computed fundamental matrix accurately represented the relationship between the views, allowing for a reliable 3D point cloud reconstruction. The geometric elements represented over the point cloud exhibited clear and recognizable features. Figure 15 shows the comparison between the original scene and the final 3D reconstruction of the scene.

REFERENCES

[1] Hartley, R. and Zisserman, A. (2000). Multiple View Geometry in Computer Vision. Cambridge University Press

[2] Nonome Tomoaki, Sakaue Fumihiko and Sato Jun, *Super-Resolution 3D Reconstruction from Multiple Cameras*, 2018.

[3] Faugeras, O., Luong, Q., and Papadopoulo, T. (2004). The geometry of multiple images: the laws that govern the formation of multiple images of a scene and some of their applications. MIT press.

[4] Shashua, A. and Werman, M. (1995). Trilinearity. In Proc. ICCV, pages 920–925.

[5] Triggs, B., McLauchlan, P., Hartley, R., and Fitzgibbon, A. (1999). Bundle adjustment - a modern synthesis. In Proc. International Workshop on Vision Algorithms.

[6] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., and Szeliski, R. (2011). Building rome in a day. Commun. ACM, 54(10):105–112.

[7] Longuet-Higgins, H. (1981). A computer algorithm for reconstructing a scene from two projections. Nature, 293:133–135.

[8] S. A. K. Tareen and Z. Saleem, "A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK," 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2018.

[9] Galliani, A., K.Lasinger, and Schindler, K. (2015). Massively parallel multiview stereopsis by surface normal diffusion. In Proc. ICCV.

[10] Alcantarilla, Pablo Fernández, Adrien Bartoli, and Andrew J. Davison. "KAZE features." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012.

[11] S. A. K. Tareen and Z. Saleem, "A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK," 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2018.