Vision for Multiple and Moving Cameras

Abel Kahsay Gebreslassie

Section 1: Obtention of the intrinsic parameters of a camera

1. For the first checkerboard image with a resolution of 1080 pixels, it's size on my screen is 19cm x 19cm. Images captured with my phone have a resolution of 3096x4128 pixels. Using the six images below and selecting 9 corner points for approximation for my camera intrinsics,

$$A = \begin{bmatrix} 3306.0 & 14.1 & 1541.8 \\ 0 & 3305.5 & 2039.0 \\ 0 & 0 & 1 \end{bmatrix}$$

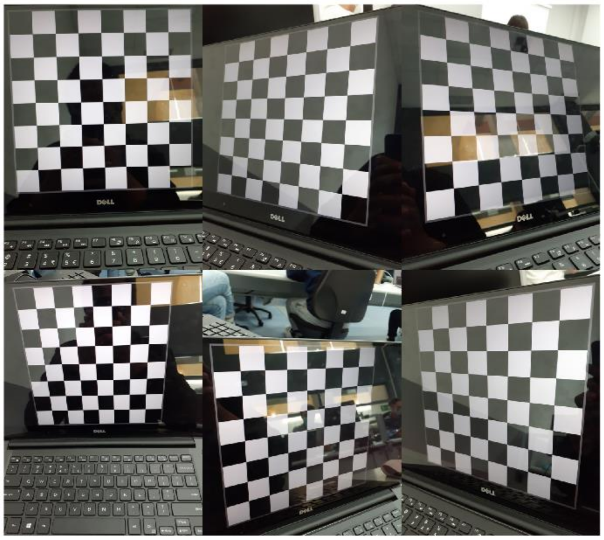


Figure 1 Images of checkerboard(1080p) used for camera calibration

The principal point is at $(u_0,v_0) = (1541.8, 2039.0)$ while center of the image plane is $(\frac{W}{2}, \frac{H}{2}) = (\frac{3096}{2}, \frac{4128}{2}) = (1548, 2064)$ and the degree of coincidence between them is $\tan^{-1}\left(\frac{2039-2064}{1541.8-1548}\right) = 1.3277$ degrees.

The image planes axes of my camera are orthogonal because $\theta = \cos^{-1}\left(\frac{-c}{a}\right) = \cos^{-1}\left(\frac{k_U \cos\theta}{k_U}\right) = \cos^{-1}\left(\frac{14.1}{3306.0}\right) = 90.244331$, that is the skew angle between the two is $\alpha = \tan^{-1}\left(\frac{-c}{a}\right) = \tan^{-1}\left(\frac{k_U \tan\alpha}{k_U}\right) = \tan^{-1}\left(\frac{14.1}{3306.0}\right) = -0.244329$. However, this doesn't guarantee the pixels are square because if we have different scaling factors for both dimensions then the pixels are rectangular instead of square. For my camera, the **focal scaling factors are equal**, f_x=3306.0≈f_y=3305.5 and **the image plane axis are**

orthogonal ($\theta = 90.2 \approx 90$), hence the **camera pixels are square**. The small differences and deviations are mainly due imperfections when selecting the point correspondences and numerical inaccuracies when estimating the intrinsic parameters from those correspondences. That is, if we were to repeat selection of points for images belonging to the 1080p checkerboard we would get slightly different values for the intrinsic matrix due to slight difference in coordinate of manually selected points.

2. The checkerboard with 720 pixels resolution has dimensions of 9.9cm x 9.9cm on my screen and camera picture resolution is the same as 1080 pixels one, that is 3096x4128

$$A' = \begin{bmatrix} 3532.2 & -3.65 & 1635.4 \\ 0 & 3527.1 & 2184.9 \\ 0 & 0 & 1 \end{bmatrix}$$

The intrinsic matrix of a camera is independent of the image being captured and relative position (rotation and translation) of the camera. Therefore, theoretically we expect both intrinsic matrix computations to yield the same value (A=A'). Nevertheless, as mentioned earlier there are manual point selection and numerical approximation errors that result in some discrepancy.

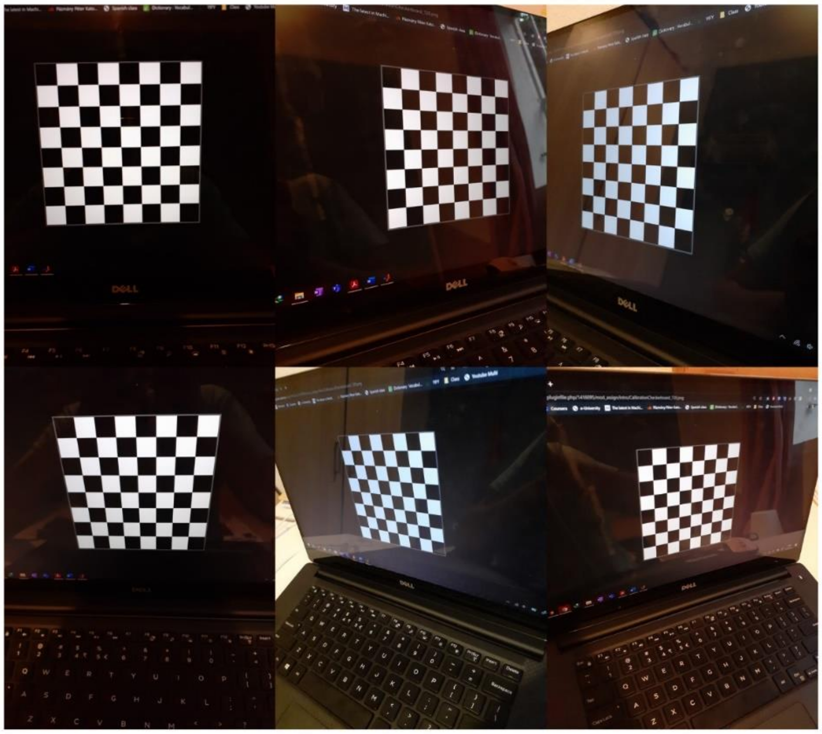


Figure 2Images of checkerboard(720p) used for camera calibration

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

For this task, a scene consisting of books was arranged in the library, with artificial lighting due to lack of adequate natural lighting in the area, and five pictures from different angles were captured.

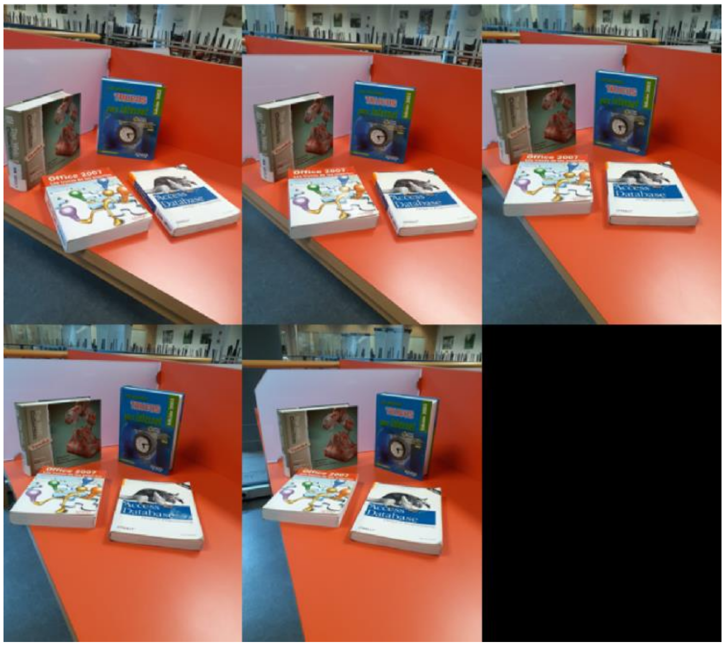


Figure 3multiple images of scene for local matching

Different detector and descriptor combinations were evaluated in the pair of images and both qualitative and quantitative measures were taken to evaluate 'goodness' of matches. In order to have a fair comparison, parameters were kept constant for all combinations. In addition, different image pairs were considered to evaluate which pairs are best for local feature matching. The pairs for matching consisted of consecutive images and others that are two, three and four images apart.

i. *DoH detector and SIFT descriptor*

This combination yielded the least number of matches between all pairs of images. From the detections (see sample image below) it can be observed DoH mainly focuses on corners(saddle points). There are several detected points in the background and this is due to the edges and corners of the chairs in the background will be identified as points of interest by detectors, this was true for all detectors and can be observed in the following sub sections as well.

SIFT descriptor and matching on the other hand matches more points in the books with vertical alignment while detections on the horizontal plane are barely matched. Furthermore, almost all of the detections in the background chairs aren't matched. This is likely due to the chairs being located far away and the repetitive pattern of chair legs. One point of interest in the chair legs is matched with a point in a book (yellow arrow in fig. 4), the area around the point consists of dense chair legs and somehow somewhat resembles to the matched area in the book. Another mismatch is the one marked with an orange arrow, even though it's a mismatch the descriptors are likely to resemble because both points belong to cherry wooden color edges while the third mismatched marked with green arrow makes less sense.

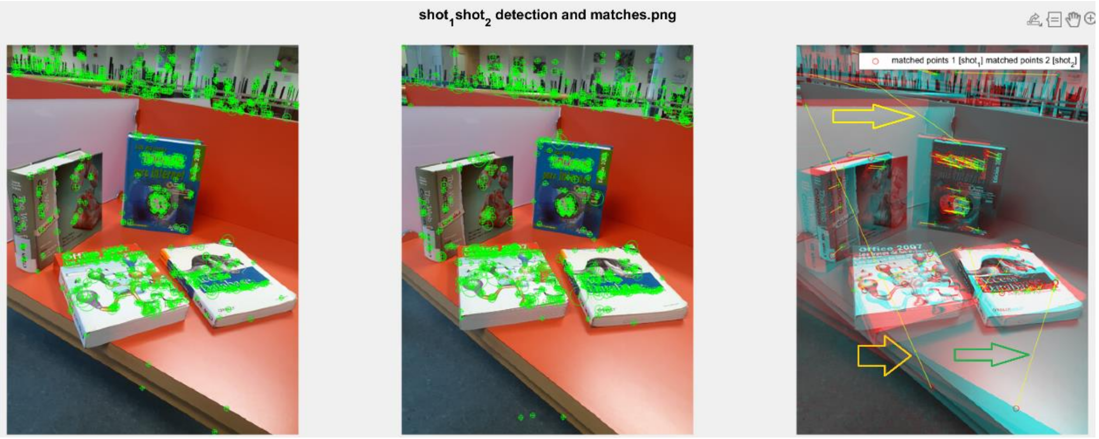

shot$_1$shot$_2$ detection and matches.png
matched points 1 [shot$_1$] matched points 2 [shot$_2$]

Figure 4DoH+SIFT matching for images 1 and 2

Homogrphies from matched points and epipolar line from fundamental matrix estimation are acceptable.

NB: Only sample images of interest are shown here. More images are available for all detector descriptor combinations are available in shots/detector_descriptor directory.

ii. SIFT detector and DSP-SIFT descriptor

This combination produces the second largest number of matching points. SIFT detector produces candidates in most areas of the image. There are no detections in the table as it consists of plain orange and white colors that have no texture in them, except for the table corners and strong illumination change points which are detected. Point of interest(POI) matching with DSP-SIFT is higher because we have a lot more detections. Most detections in the background areas aren't matched due to similar reasons mentioned earlier. Local feature matching for image pairs that are further apart gives a smaller number of matches and is prone to more wrong matches as well. This behavior was observed for all detector-descriptor combinations as can be seen in table 1. Homographies, fundamental matrix and epipolar lines are visually satisfactory as well as quantitatively. The fundamental matrix have a relatively higher number of inlier points, next to KAZE-KAZA combination.

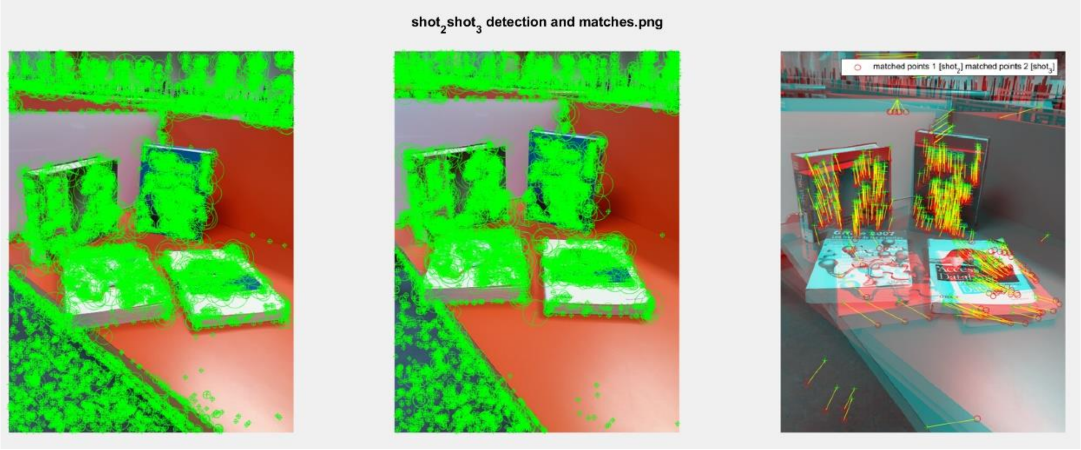

shot$_2$shot$_3$ detection and matches.png
matched points 1 [shot$_2$] matched points 2 [shot$_3$]

Figure 5SIFT + DSP_SIFT matching for images 2 and 3

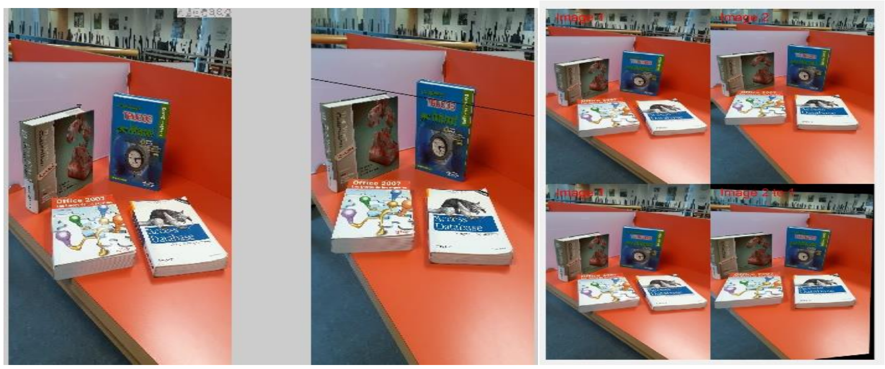


Figure 6Homogray and epipolar line for images 2 and 3

iii. SURF(detector and descriptor)

The SURF detector seems to favor POI at a larger scale because most of the detections produced have relatively larger scales. There are more detections are in the books that have horizontal alignment(white books). Detections look similar to those produced by SIFT even though they are smaller in number. The matching rate is also high but there are many wrong local feature matchings at the same time. The POIs in the background are the main source of wrong matches. Homographies from matches are of average quality and epipolar lines for pairs that are far apart are of low quality.

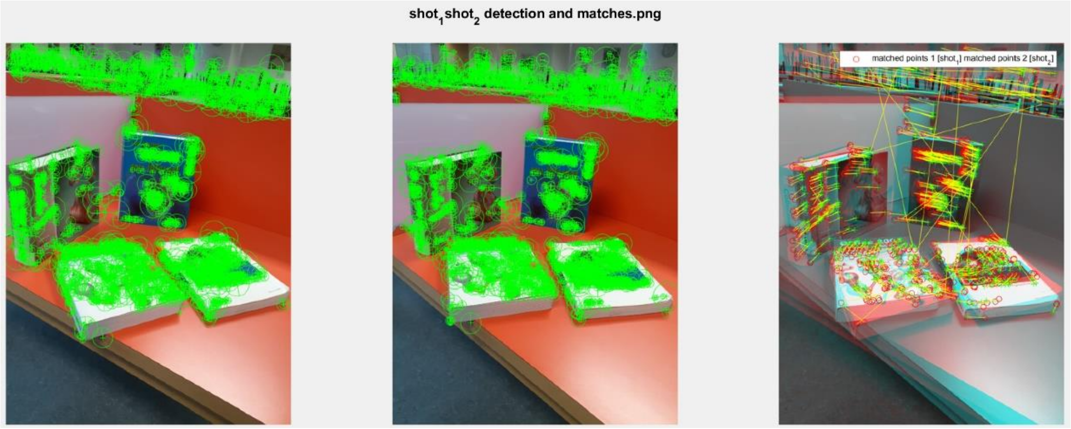

shot$_1$shot$_2$ detection and matches.png
○ matched points 1 [shot$_1$] matched points 2 [shot$_2$]

Figure 7SURF detection and matching for images 1 and 2

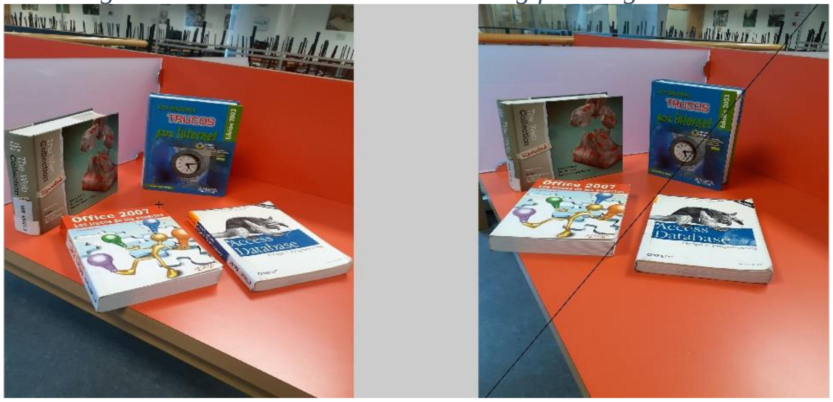


Figure 8Epipolar line for images 1 and 4

iv. *KAZE(detector and descriptor)*

KAZE results in the largest number of detections and matchings from all the combinations experimented with. Almost all lines between matched points have similar orientation that is similar to the motion between the images. We can easily observe the mismatched points in the background areas but in dense matches, for instance in the books, makes it hard to see mismatched points. Homographies and epipolar lines estimated are adequate.

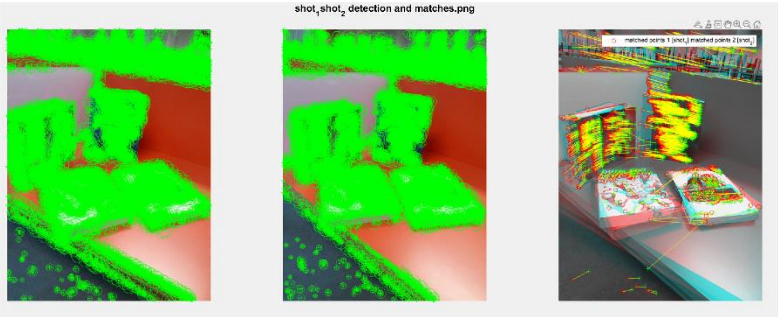

shot$_1$shot$_2$ detection and matches.png
matched points 1 (shot$_1$) matched points 2 (shot$_2$)

Figure 9KAZE detection and matching for images 1 and 2

As mentioned earlier, consecutive image, images that are two steps apart(shot_2 and shot_4), images that are three steps apart(shot_1 and shot_4) and images that are 4 images apart(sho_1 and shot_1) were used as pairs to evaluate the detector-descriptor combinations and determine which pairing results in better matching

Pair	Detector-descriptor combination [**inliers /matched points**]				Estimated Homography (SIFT & DSP-SIFT)	Estimated Fundamental matrix (SIFT & DSP-SIFT)
	DoH and SIFT	SIFT and DSP-SIFT	SURF	KAZE		
Shot_1 and shot_2	18/80	206/844	117/710	622/2224	$\begin{matrix} 0.7838 & -0.0481 & -0.0001 \\ -0.0265 & 0.8585 & 0 \\ 175.2 & 84.4224 & 1 \end{matrix}$	$\begin{matrix} 0 & 0 & -0.029 \\ 0 & 0 & 0.0152 \\ 0.0027 & -0.0179 & 0.9997 \end{matrix}$
Shot_2 and shot_3	18/69	122/633	96/614	564/2189	$\begin{matrix} 0.8108 & = 0.0478 & -0.0001 \\ 0.0233 & 0.9597 & 0 \\ 84.9485 & -48.5717 & 1 \end{matrix}$	$\begin{matrix} 0 & 0 & -0.031 \\ 0 & 0 & 0.0153 \\ 0.0028 & 0.0162 & 0.9997 \end{matrix}$
Shot_3 and shot_4	32/105	274/1080	152/812	878/2722	$\begin{matrix} 1.0197 & 0.0086 & -0.0001 \\ 0.0005 & 0.9888 & 0 \\ 2.5094 & -44.0534 & 1 \end{matrix}$	$\begin{matrix} 0 & 0 & 0.0043 \\ 0 & 0 & -0.0278 \\ -0.0043 & 0.0287 & 0.9992 \end{matrix}$
Shot_4 and shot_5	25/145	252/1330	170/898	618/3587	$\begin{matrix} 0.9172 & -0.0119 & 0 \\ -0.0219 & 0.9412 & 0 \\ 158.3683 & 66.2167 & 1 \end{matrix}$	$\begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0.0076 \\ -0.0004 & -0.0090 & 0.9999 \end{matrix}$
Shot_2 and shot_4	14/51	58/228	48/341	265/1239	$\begin{matrix} 0.6450 & -0.0741 & -0.0002 \\ 0.0075 & 0.9046 & 0 \\ 199.4522 & -49.8477 & 1 \end{matrix}$	$\begin{matrix} 0 & 0 & 0.0020 \\ 0 & 0 & -0.0098 \\ -0.0018 & 0.0107 & 0.9999 \end{matrix}$
Shot_1 and shot_4	11/35	40/112	29/228	200/714	$\begin{matrix} 0.5435 & -0.0804 & -0.0002 \\ -0.0207 & 0.8192 & 0 \\ 306.6773 & -1.3684 & 1 \end{matrix}$	$\begin{matrix} 0 & 0 & 0.0020 \\ 0 & 0 & -0.0027 \\ -0.0032 & 0.0015 & 1 \end{matrix}$
Shot_1 and shot_5	9/17	26/65	20/157	115/517	$\begin{matrix} 0.4060 & -0.1176 & -0.0002 \\ -0.0453 & 0.7367 & -0.0001 \\ 464.549 & 83.3690 & 1 \end{matrix}$	$\begin{matrix} 0 & 0 & 0.0014 \\ 0 & 0 & -0.0023 \\ -0.0024 & 0.0015 & 1 \end{matrix}$

Table 1matched points and inliers for different pairs with different detector-descriptor combination

KAZE followed by SIFT and DSP-SIFT has the largest number of matching and inliers for estimated fundamental matrix. Moreover, from the table it can be seen that when image pairs are far apart the number of matching points and inliers for the fundamental matrix significantly decrease. Hence, pairs that are consecutive will produces better results as compared to pairs that are far apart.

Section 3: 3D reconstruction and calibration

According to the results from section 2, KAZE detector gives more matches and inlier points. However, matching n_view_matching for common points in all scene with KAZE followed by 3D reconstructions results in a larger reprojection error of 10^4 order. This is due to the large number of points many of which are likely to be wrong matches. As a result, SIFT and DSP-SIFT was used to produce n_view_matching points and for 3D reconstruction.

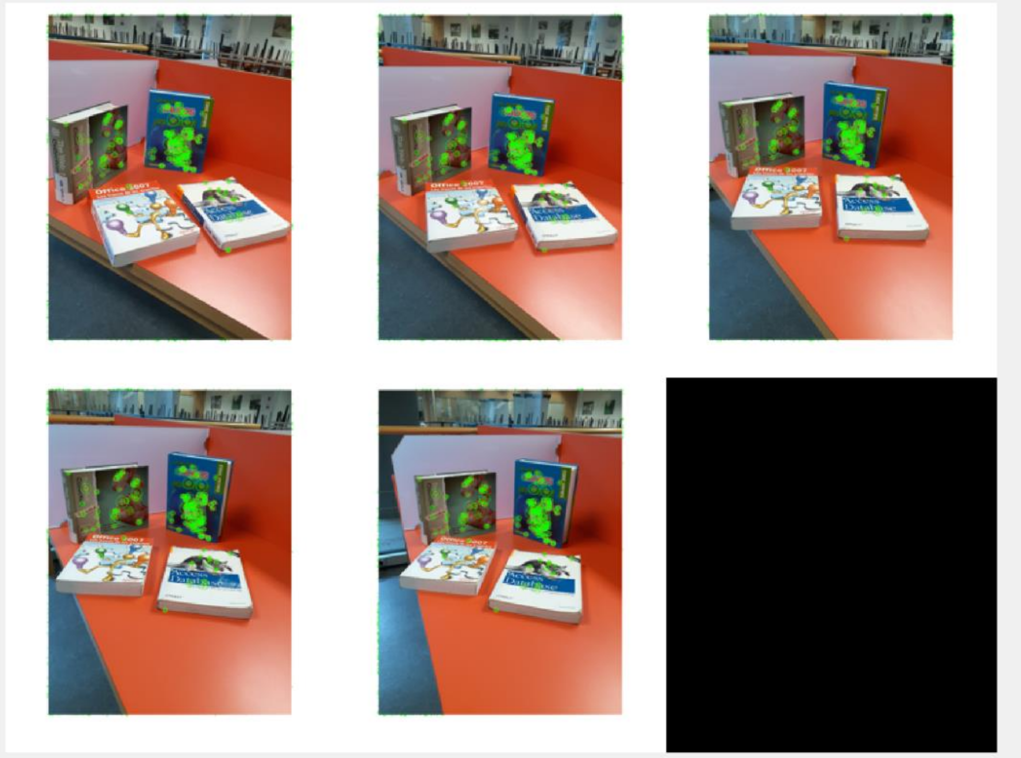


Figure 10matching points in all images(n_view_matching)

Using only two images, first image(shot_1) and last image(shot_5), the mean reprojection error: mean x is 0.17243 and mean y is -0.69236.

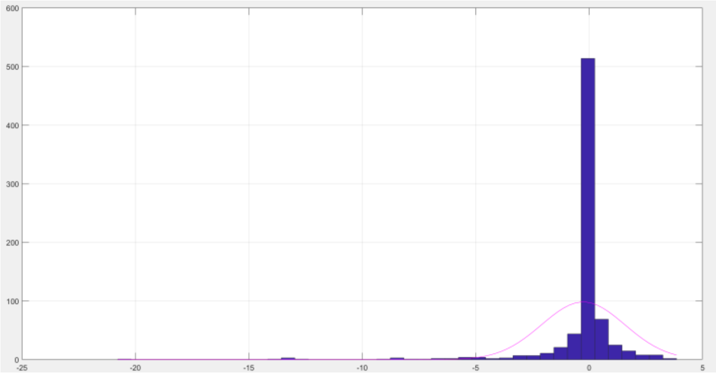


Figure 11Error histogram using 2 cameras

Mean reprojection error after resectioning is: [-0.03502   -0.38704] while after bundle adjustment it drops to zero.

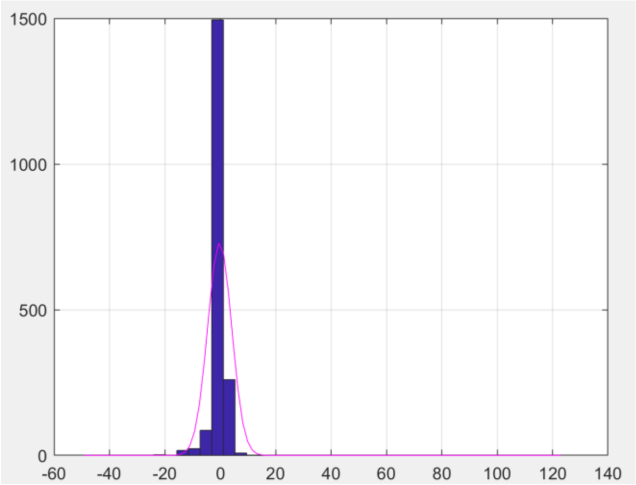


Figure 12Error histogram for resectioning

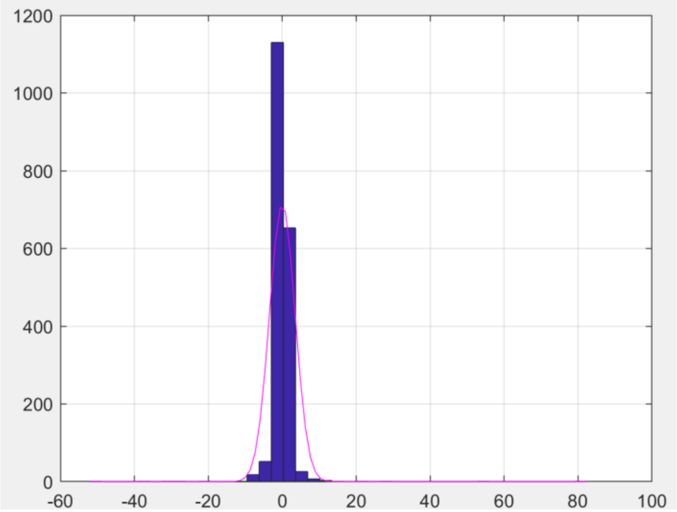
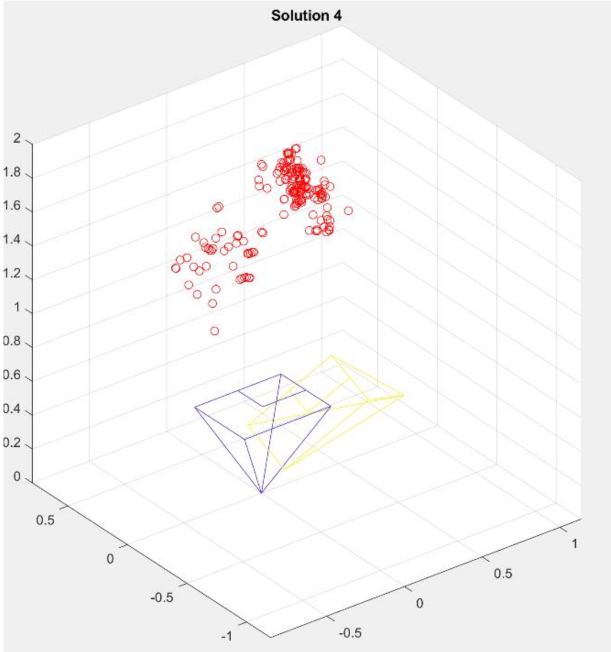
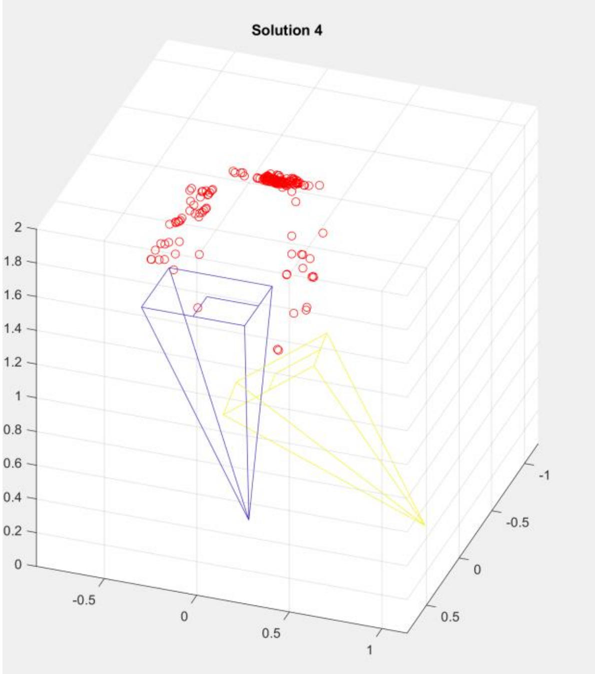


Figure 13Error histogram after bundle adjustment

After bundle adjustment the mean error is almost zero. When using only two cameras, the first and the last, the matched points will not be 100% accurate (will have some noise) in addition the 3D point estimate and reprojection will have errors as well due to the projective ambiguity and the estimated projective matrices not being exact. In resectioning, we are trying to 3D points estimate we have from two cameras to sever intermediate others. The projective matrix estimate for those cameras will have a certain degree of error, similar to the two cameras, and this combined with 3D point estimate noise results in a higher reprojection error. In other words, we are trying to find the point reprojections using noisy 3D point estimates and projective matrix by linearly transforming the 3D points (matrix multiplication) however this can't be satisfied as there's noisy and there is no exact linear fit. As a result, we use the bundle adjustment step to find the reprojected points that fit the best but this isn't guaranteed to always yield the best estimate as there is randomness and only limited number of iterations can be done.





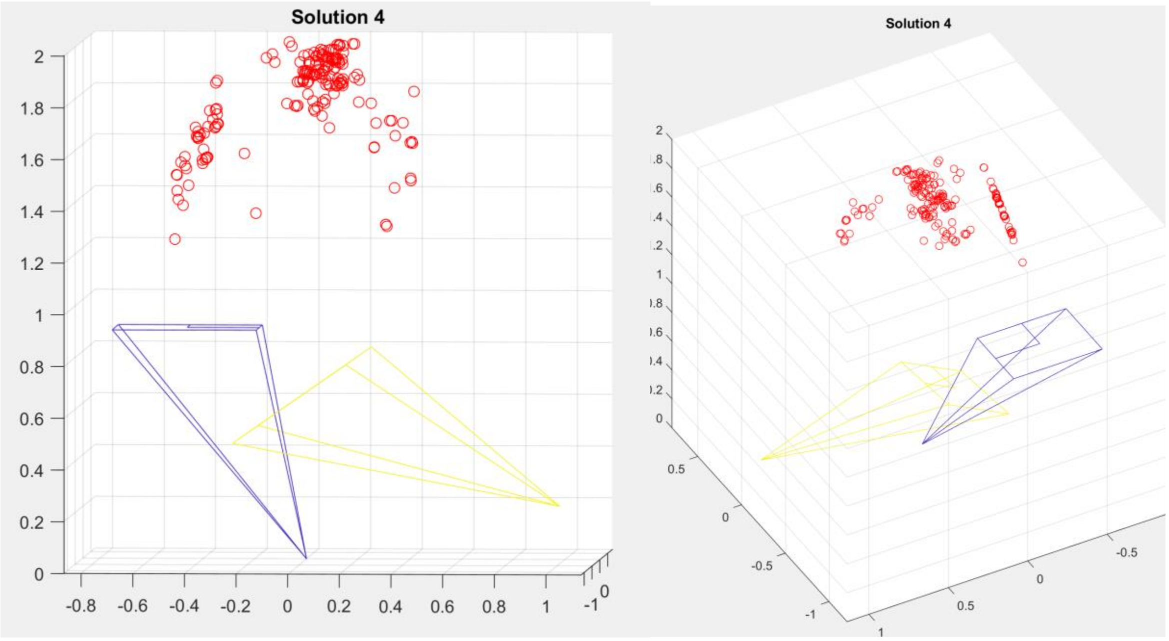


Figure 14 multiple angles of scene 3D reconstruction

Extra

Drawing cloud points with corresponding image color and adding small image region around point to cloud point were added. They can be accessed by setting **color** and **crop** attributes of **params** in section_3_n_view.m. Size of the cropped region can be set with **params.crop_size** attribute.

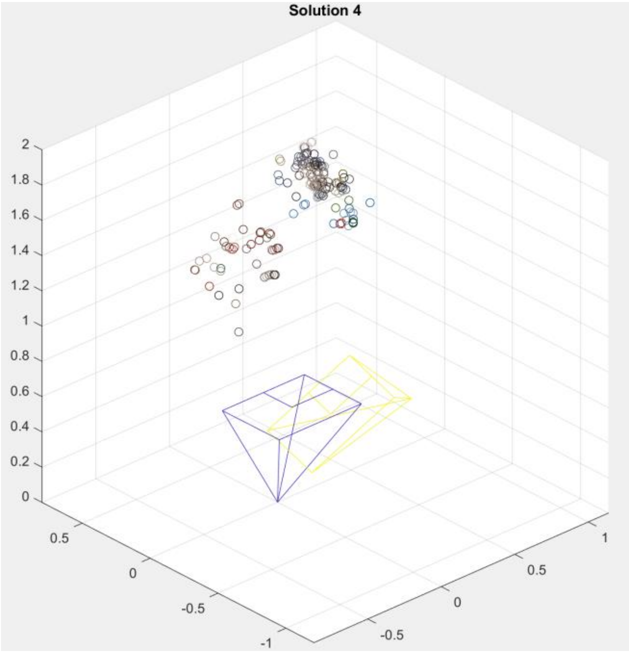


Figure 15cloud point with image colors

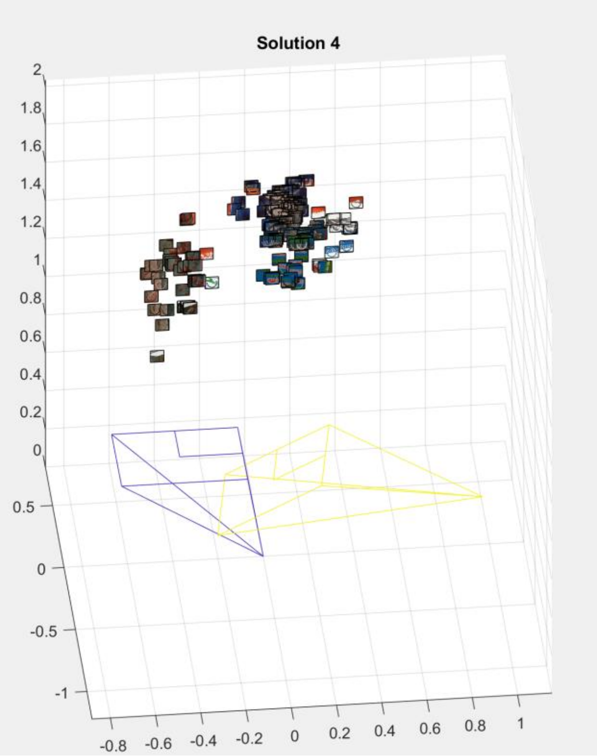


Figure 16cloud point with corresponding region crops