Camera Calibration - ELL793

Shauryasikt Jena, 2018EE10500
Amokh Varma, 2018MT60527

Indian Institute of Technology, Delhi

Abstract

Cameras are a highly common occurrence in the modern world. Cameras allow us to retain a 2D image of our 3D world. However, this transformation is imperfect, i.e. we are bound to lose information. There is extensive research conducted in order to improve the quality of images captured by a camera. To deliver a worthwhile contribution to the field, it is highly important that we familiarise ourselves with the process of generation of images through a camera. This report contains our findings that we observed by attempting to create our own implementation for camera calibration.

Introduction

The introductory study of camera is done through a simplified model called the *Pinhole Camera* model. In this model, we analyse the image forming characteristic of a pinhole which only allows a single ray to go through it. Obviously, such an ideal case is not possible to replicate. However, given small enough apertures, the results derived for a simple pinhole camera model have been observed to be applicable. However, too small an aperture can lead to problems due to *diffraction*, which takes away the sharpness of the image. The effect of pinhole size can be observed Figure 1 from [1].

Clearly, the camera is responsible for translation 3D world information into a 2D image. This is bound to result in loss of information. Certain instances of loss of information include :

- Inability to capture size of different objects.
- Inability to capture angle between the objects.
- Inability to capture the actual distance between objects.

In order to be able to extract real world information from camera images, we need to know about the internal mechanism of image formation and estimate the parameters that affect said formation. Since, different parts of camera might be produced and assembled in different places, it becomes very important to be able to have a top down approach towards finding the parameters of a camera. This process is called **Camera Calibration** . Clearly, due to loss of information, the camera image itself is not enough to capture the properties of camera. So, we make use of a real world coordinate system and look at its translation in the image coordinates, to calculate the necessary details.

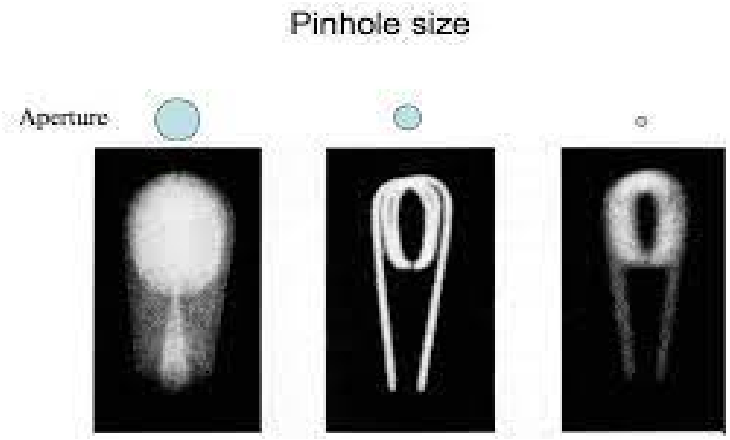


Figure 1: Effect of varying pinhole size to the sharpness of the image

Prerequisites and Notations

The calibration process requires basic knowledge of pinhole camera model and a fair bit of linear algebra. The pinhole camera model can be seen in Figure 2. As we can, the image of the candle is formed inverted and in front of the pinhole. The virtual image shown is for illustrative purposes, as it is sometimes convenient to only focus on one side of the camera.

Consider Figure 3 . O is the pinhole, which also coincides with the origin of the camera frame. The normalised image plane is a plane at a distance of 1m from the pinhole. The physical retina refers to the final image forming screen of camera. So, we have the following coordinate systems. $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are coordinate basis vectors for the camera coordinates. $(\hat{x}, \hat{y})$ and (x, y) are the coordinate systems of normalised and physical planes. The angle between x and y axes is θ , which occurs due to manufacturing error, and is close to 90. Note, that the centre C, which is the centre of camera coordinate system (CCS) , is not aligned with the centre of image plane. The world frame is denoted as W and image frame is denotes as C . A point P in frame A is

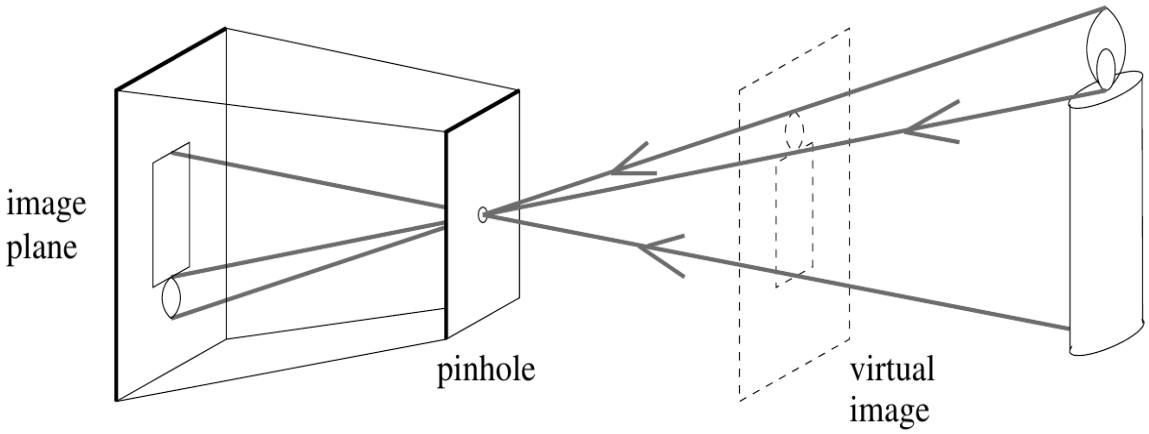


Figure 2: Image formation in pinhole camera model

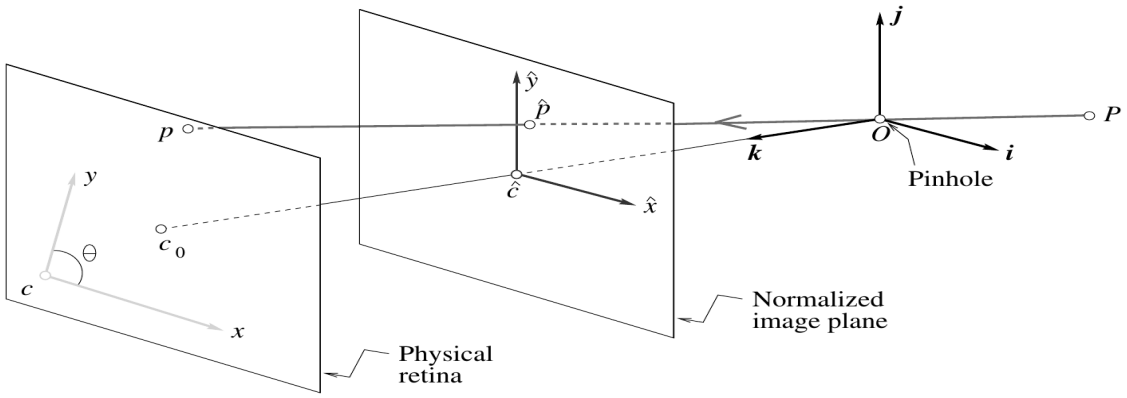


Figure 3: Different coordinate systems involved in image formation

written as ${}^{A}P$. As we can see, the image formation depends on the location and angel of the camera w.r.t the point P, as well as the scale, translation and angle of image plane inside the camera. The former details are collectively known as *extrinsic parameters* and latter ones are known as *intrinsic parameters*. Further, it is worth noting that the camera has to translate the length information into pixel information. This introduces scaling factors α and β explained by $\alpha = kf$ where, f is the focal length (where the image is formed if object is at infinity, we will use this simplification throughout this assignment) . Overall α, β will have the unit *pixels*.

Method Used

Using some geometry, we can find out that, under negligible radial distortion, the point ${}^{W}P$ in homogeneous world coordinates can be converted into image coordinates using a single transformation matrix $\mathcal{M}$. The relation is

$$ {}^{C}P = \mathcal{M}\,{}^{W}P \tag{1} $$

$\mathcal{M}$ will be a 3x4 matrix, which depends on both the intrinsic and extrinsic properties of the camera. Further, we can decompose $\mathcal{M}$ as

$$ \mathcal{M} = \mathcal{K}[\mathcal{R}|\mathbf{t}] = [\mathcal{K}\mathcal{R}|\mathcal{K}t] \tag{2} $$

Here, the $\mathcal{K}$ is the intrinsic parameter matrix. It can be written as :

$$ \mathcal{K} = \begin{bmatrix} \alpha & -\alpha\cot\theta & x_0 \\ 0 & \beta/\sin\theta & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3} $$

Here, x_0 and y_0 are the coordinates of C_0 in the retinal frame. Further the R matrix is a rotation matrix. Clearly, R will be **orthogonal** and $\mathcal{K}$ is **upper triangular** . As explained in [2], the RQ decomposition of this $\mathcal{K}R$ matrix will be unique if diagonal entries of $\mathcal{K}$ are positive. Since $\theta \approx 90$, we can ensure that $\sin\theta > 0$. To estimate $\mathcal{M}$, we generate a dataset of points whose real world and pixel coordinates are known. Using these points, we try to approximate the matrix which results in lowest error in satisfying the Equation (1) .

Figure 4: Experimental Setup

Sno	Real World	Pixel Coordinates
1	(4,0,4)	(1220,2221)
2	(4,0,6)	(1237,1875)
3	(3,0,7)	(1386,1716)

Table 1: Example dataset with 3 elements. We take a total $n = 30$ such points

Dataset Generation

To create the dataset, we use the camera from a OnePlus 8T mobile phone. We use 3 images of a checkerboard and paste them in perpendicular planes. From this, we take out mark points in the world coordinate and take the photo of the set up. The image has a size of 3516 x 3516. Afterwards, we manually find the pixel values of each point. The entire set up has been shown in Figure 4. A small portion of the dataset can be seen in Table 1.

Results

We divide our results section into different subsections. Firstly, we talk about the normalisation of the data. Afterwards, we analyse the $\mathcal{M}$ matrix found by the calibration routine. Afterwards, we discuss about the quality of our predictions. Finally, we discuss properties of dataset.

Normalisation Transformations

Normalisation is a very common technique to make sure that mean of the data points is 0 and average distance is $\sqrt{dim}$. This ensures that all processing done will affect each of the points equally. Further, it makes sure that, in the new coordinates, the centres of the real world data, and the centre of their images, align properly. Consider our data set to be $X = [x_1, x_2 \dots x_n]^T$. It will have a dimension of $(m+1) \times (m+1)$ where n is the number of points and m is the dimension (2 or 3, in our case). This is because we do this transformation in homogeneous coordinates.

$$X_{norm} = \sqrt{n} * (X - \frac{\sum_1^n x_i}{n}) / \sqrt{\sum_1^n (x_i - \bar{x})^2}$$

We can pre-calculate this information and create a matrix $\mathcal{T}$ such that

$$X_{norm} = \mathcal{T}X$$

For world coordinates , we get :

$$\mathcal{T}_{4\times4} = \begin{bmatrix} 0.3892 & 0 & 0 & -0.9732 \\ 0.0 & 0.3892 & 0 & -0.9974 \\ 0.0 & 0 & 0.3892 & -2.45 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

For, image coordinates, we get

$$\mathcal{T}_{3\times3} = \begin{bmatrix} 0.00186 & 0 & -3.27 \\ 0.0 & 0.00186 & -3.38 \\ 0 & 0 & 1 \end{bmatrix}$$

We can notice how these matrices have different scales of values, because the real world and pixel coordinates have a different scale of values. Now, we move on to the second set of results.

Calibration Matrices

We take 30 data points to create the dataset. Using this data, we get the following matrix as our projection matrix $\mathcal{M}$

$$\mathcal{M}_{3\times4} = \begin{bmatrix} -121.658 & 42.449 & 2.59 & 1271.92 \\ -49.200 & -3.9709 & -1.1480 & 2093.792 \\ -0.02181 & -0.0219 & -0.000868 & 0.7378 \end{bmatrix} \tag{4}$$

We perform RQ-decomposition of the first 3×3 square of the projection matrix. As discussed in Section , this will give us the rotation and intrinsic parameter matrices.

We get the following values for $\mathcal{K}$ and $\mathcal{R}$

$$\mathcal{K}_{3\times3} = \begin{bmatrix} 3756.671 & 92.986 & 1799.535 \\ 0 & 3659.477 & 2135.335 \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

Clearly, since the diagonal elements are positive, the decomposition is unique.

$$\mathcal{R}_{3\times3} = \begin{bmatrix} -0.7085 & 0.7030 & 0.0605 \\ 0.0231 & -0.0625 & 0.9977 \\ 0.7053 & 0.7083 & 0.0281 \end{bmatrix} \tag{6}$$

It can be verified that $\mathcal{R}^T\mathcal{R} = I$. Let the last column of $\mathcal{M}$ be equal to $\mathbf{b}$. Finally, we get $\mathbf{t} = \mathcal{K}^{-1}\mathbf{b}$. So, we get

$$\mathbf{t} = [0.5932, -4.579, -23.851]^T \tag{7}$$

Finally, using Equation (3) and Equation (5), we can find the actual values of the parameters, which are

- $\alpha =$ $\boxed{3756.6698 \text{ pixels}}$
- $\beta =$ $\boxed{3658.3752 \text{ pixels}}$
- Aspect ration = $a = \frac{\alpha}{\beta} =$ $\boxed{1.02}$

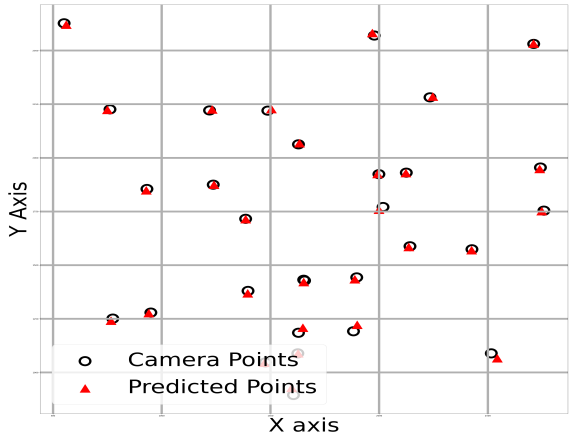


Figure 5: Projections using the predicted matrices $n = 30$

- $x_0 =$ $\boxed{1799.535 \text{ pixels}}$
- $y_0 =$ $\boxed{2135.335 \text{ pixels}}$
- $\theta = 1.59552$ radians $=$ $\boxed{91.42^\circ}$
- $f =$ $\boxed{26 \text{ mm}}$

Visualisation

We plot our projected positions in the pixel coordinates using the world coordinates and compare it with the ones generated by the camera. The best results were obtained when $n = 30$ and had a rmse of 5.85 . Our predictions and the actual points are plotted in Figure 5.

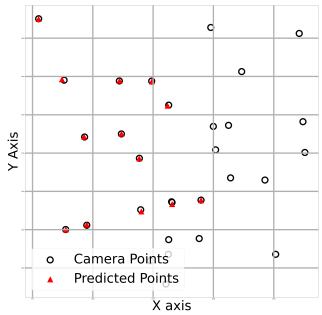
We can notice that we are able to form a rough prediction of the positions of the points in pixel coordinates.

Dataset Discussion

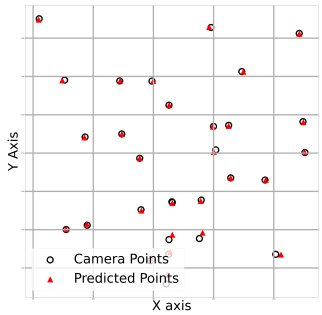
We can notice that increasing the dataset size increases the quality of our predictions. Further, it also improves our performance in the prediction on a holdout set that we kept. Figure 6 depicts the importance of having sufficient data points. We can see how increasing the dataset leads to better predictions. We can also observe the same from Table 2.

Conclusion

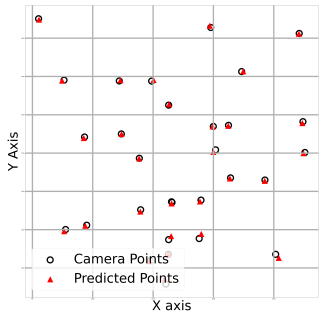
In this document, we have learned how to formulate the camera calibration problem and how to solve it using computational methods. Through this assignment, we are able to find the intrinsic and extrinsic parameters of our camera. In addition, we also show the importance of having sufficient number of labelled real world images. Though we have worked under the assumptions of a pinhole camera model, we are able to extract characteristics of the camera which seem to have a good predictive power. We have also avoided the discussion on distortion due to the lens. Considering these factors can further strengthen our model of the camera and allow us to make stronger predictions on the real world coordinates of different objects, based on pixel information and vice-versa. This finds numerous applications in tasks like robot coordination, medical imaging etc. Future works related to this can include taking multiple images of the checkerboard and averaging the parameters across different images.



(a) Projections using the predicted matrices $n = 7$



(b) Projections using the predicted matrices $n = 12$



(c) Projections using the predicted matrices $n = 25$

Figure 6: Performance with varying training data sizes

References

[1] 2011. *Computer vision*. Pearson Education (US).

Data Points	Root Mean Square Error
8	57.32
10	10.12
12	8.43
14	8.41
16	8.53
18	8.04
20	8.19
22	8.02
25	7.88
30	5.85

Table 2: Performance with varying training data sizes

[2] Belkhouf. 2009. Machine Vision.