The field of 3D reconstruction and modeling spans the complete process of obtaining a geometric or semantic model of an imaged scene or object. This encompasses not only the methods and algorithms for retrieving the depth of the image but also the deep understanding of the imaging process in its mathematical and geometrical terms. We described in Chapter 1 intuitively how a moving camera can be used to retrieve the 3 dimensional structure of a recorded scene. This procedure is also called Structure from Motion (SfM) \[1\].

The problem has been studied in great detail in the literature. In this chapter we follow the work of Hartley and Zisserman to formalize the geometry of the camera and the imaging process. We also describe the relation between the simplistic camera described in the previous chapter and modern digital cameras by means of a calibration matrix. Once we understand how the camera is defined mathematically we then move on to the actual methods to perform the reconstruction. Furthermore, there are a large amount of different conventions with respect to the geometrical frames. This makes it difficult sometimes to integrate different approaches or steps in one single reconstruction procedure. In this chapter we establish our point of view, defining the referential frames to describe geometrically: cameras, images and objects in space. Finally, we formalize the geometrical projection and describe the relation between corresponding sets of points in images by means of the fundamental and essential matrices. These two contain the essence of the camera motion which can then be used for the actual reconstruction of
the scene. The work of Hartley and Zisserman \cite{HartleyZisserman} is among the most well known on visual geometry and is used as a general guide, though deep details are left out. The standards and definitions presented in this chapter are used in the rest of this Thesis.

2.1 Formalizing the Geometry of the Imaging Process

Understanding the geometry of the camera implies understanding how the images are formed and how we can backtrack the process for 3D reconstruction. In this section we formalize the pinhole camera model by describing the mathematical projection, the parameters that define the camera and the referential frames used to describe it all.

2.1.1 The Pinhole Camera Model

As we described in Chapter 1, the pinhole camera is the most basic imaging device where rays of light pass through a small hole in one side of a dark chamber. These rays impact the image sensor behind the hole where the image is recorded. The hole is called the \textit{focal point}. The smaller the hole, the sharper the image, but also the more dim it becomes (save for degeneracies with very small holes due to diffraction). The size of the hole is commonly referred as the \textit{aperture} of the camera. The rays cross in the camera hole, producing therefore an upside down image (see Figure \ref{fig:pinhole} LEFT). In the Pinhole Camera Model, the aperture is assumed to be an infinitely small hole and all cameras treated from here on are assumed to be compliant with that model.

Cameras can be geometrically described as a central projection where the real life object is projected onto a recording device through a center of projection, which is of course the focal point. In such a representation, it is common to represent the recording device \textit{in front} of the center of projection \cite{HartleyZisserman} (see Figure \ref{fig:pinhole} RIGHT). The focal point is also called \textit{camera center} or \textit{optical center}.

The distance of the sensor to the focal point is called the \textit{focal distance} and usually referred as $f$. This parameter describes the geometry of the imaging process which can be encoded in a \textit{camera calibration matrix} called $K$. The camera matrix represents a mapping between 3D points in the real world to points on the image sensor.

The projection of objects into the imaging sensor is a linear mapping between 3D points and image points in homogeneous coordinates:
2.1 Formalizing the Geometry of the Imaging Process

Figure 2.1: LEFT: Pinhole camera model. Rays of light reflected from the real object pass through a small opening in the camera and intersect the imaging sensor. The image is therefore recorded flipped both vertically and horizontally. RIGHT: Camera as a central projection. Rays of light reflected from the real object are projected through the center of projection onto the imaging sensor.

\[
\begin{pmatrix}
U \\
V \\
W
\end{pmatrix} \rightarrow \begin{pmatrix}
f U \\
f V \\
W
\end{pmatrix} = \begin{bmatrix}
f & 0 & 0 & 0 \\
0 & f & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix} \begin{pmatrix}
U \\
V \\
W \\
1
\end{pmatrix},
\]  

(2.1)

where \((U,V,W,1)\) represents the homogeneous coordinates of a point of the imaged object in 3D space, and \((fU,fV,W)\) represents the projected homogeneous coordinates of a point in the image space.

At this point, we can already see that there are at least 2 different coordinate systems: the coordinate system of the objects in the real world, and the coordinate system of the projected objects in the 2D image plane. We will describe in more detail the coordinate systems in the following section, but let us for now assume that there are two generic reference frames for both the object in 3D space as seen from the camera and the imaged object in image space, and let us call them the Camera reference frame \((C_{rf})\) and the Image reference frame \((I_{rf})\).

It is commonly assumed that one axis of the \(C_{rf}\) will intersect the image plane perpendicularly. This intersection point is called the principal point. This perpendicularity
assumption is related to the position of the image sensor. In most cameras the sensor is placed parallel to the focal plane. This allows the virtual plane in 3D space where imaged objects are in focus on the image to be also parallel to the sensor. Certain cameras or lenses do not follow this assumption, though their use is limited to creative or architectural photography, creating focus planes in the image that introduce visual effects.

If we could choose the location of the $I_{rf}$, we could set it such that it would coincide with the projection of the $W_{rf}$ in the image (see Figure 2.2 LEFT). However, in most cases, we have no choice in the position of the center of the $I_{rf}$ or it is not convenient to transform it. If the principal point and the origin of the $I_{rf}$ do not coincide, we need a mapping between them:

$$ (U, V, W)^T \rightarrow (fV/W + p_u, fV/W + p_v)^T, \quad (2.2) $$

where $(p_u, p_v)^T$ are the coordinates of the principal point in the $I_{rf}$.

Combining this with our previous definition of the projection, we obtain a more general description where the difference between the principal point and the center of $I_{rf}$ are accounted for:
2.1 Formalizing the Geometry of the Imaging Process

\[
\begin{pmatrix}
U \\
V \\
W \\
1
\end{pmatrix} \rightarrow \begin{pmatrix}
U \\
V \\
W \\
1
\end{pmatrix} = H \begin{pmatrix}
fU + Wp_u \\
fV + Wp_v \\
W
\end{pmatrix} = \begin{pmatrix}
f & p_u & 0 \\
f & p_v & 0 \\
1 & 0
\end{pmatrix} \begin{pmatrix}
U \\
V \\
W \\
1
\end{pmatrix}
\] (2.3)

A typical sensor in modern digital cameras is a finite sensor consisting of pixels that capture light rays. These pixels might not be arranged in a square array, which introduces a nonhomogeneous scale factor in each image axis. If we consider the number of pixels per unit of distance in image coordinates \((m_u, m_v)\), we can create a mapping that will account for a finite sensor with nonsquare pixel arrays:

\[
\begin{bmatrix}
fm_u & 0 & p_u & 0 \\
0 & fm_v & p_v & 0 \\
0 & 0 & 1 & 0
\end{bmatrix},
\] (2.4)

where \(fm_u\) and \(fm_v\) are usually written as \(\alpha_u\) and \(\alpha_v\).

There is one more parameter for a complete generalization of the camera model. The skew factor \(s\) can be interpreted as a skewing of the pixel elements in the sensor array so that the x-axis and y-axis are nonperpendicular. This could be the result of the image sensor not being parallel to the optical axis. This is encoded in \(K_{(1,2)} = s\). It is commonly assumed to be zero in most commercial and consumer cameras. Throughout our work, we assume the use of cameras and sensors with NO skew factor, therefore, we will always consider a camera calibration matrix on the form:

\[
K = \begin{bmatrix}
\alpha_u & 0 & p_u \\
0 & \alpha_v & p_v \\
0 & 0 & 1
\end{bmatrix},
\] (2.5)

where \(\alpha_u = fm_u\) and \(\alpha_v = fm_v\).

2.1.2 Real Life Imaging Devices and the Lens Problem

Modern cameras are different from the pinhole camera in that they have a lens placed at the hole. Lenses are used primarily to get sufficient amount of light in (achieve the desired image quality), to change the focal length (zooming in and out), to change the aperture and control sharpness, and to introduce certain visual effects.

The use of lenses can however introduce a distortion factor in the recorded image. This distortion can take a variety of shapes depending on the quality and construction.
2. ESSENTIAL CONCEPTS IN VISUAL GEOMETRY

materials of the lens. Changing the focal length by zooming in and out may also change the distortion factor. Some of these distortions are used for artistic reasons and some of them produce the effect of an image sensor that is not perpendicular to the camera coordinate system.

Most everyday consumer cameras only introduce two types of distortion: barrel and pincushion.

2.1.2.1 Barrel and Pincushion Distortion

Barrel distortion (see Figure 2.3 LEFT) produces and effect where points are imaged further from the center as the distance from the center increases. This distortion is typically found in wide angle lenses where the goal is to record a bigger portion of the scene, even if distorted. This is the most typical type of distortion found in consumer lenses in modern cameras.

The pincushion distortion (see Figure 2.3 RIGHT) has the opposite effect of the barrel distortion. As the distance with the center of distortion decreases, the points are imaged closer to the center.

Figure 2.3: LEFT: Barrel distortion. RIGHT: pincushion distortion.
2.1 Formalizing the Geometry of the Imaging Process

2.1.2.2 Radial distortion

Both distortions can be seen as a radial distortion, where the location of the true pixel is altered based on some function of the distance with respect to the center of the distortion. It is important to understand that in order to obtain a linear projection (where a ray can be traced from the imaged point to the real object) into the imaging device, these distortions need to be accounted for. Otherwise, the backtracking and the point matching procedures required for the reconstruction step would be incorrect. In computer vision, the distortion in a lens is measured and the corresponding inverse transformation is applied so that the camera becomes effectively linear projection device. Radial distortion can be described as a function of the distance to the center of distortion:

\[
\begin{pmatrix}
u_d \\
v_d
\end{pmatrix} = L(\sqrt{u_l^2 + v_l^2}) \begin{pmatrix} u_l \\
v_l
\end{pmatrix},
\]

(2.6)

where \((u_d, v_d)\) are the distorted image pixels, \((u_l, v_l)\) are the linearized, or corrected, image pixels.

\(L(r)\) is the distortion function dependent on the distance \(r\) to the center of distortion and is usually approximated by a Taylor series or order \(n\).

\[L(r) \approx 1 + k_1 r + k_2 r^2 + k_3 r^3 + ... + k_n r^n \]

(2.7)

Note however that the distortion function \(L(r)\) differs when modeling the transformation from distorted pixels to linearized pixels, or the other way around. The function is not easily inverted, though an approximation can be found when the distortion factor is sufficiently small. It is then common to model the distortion some way or other depending on the reconstruction procedure and the way the images are further processed.

2.1.3 Geometry of Projection and the Camera Motion

We have precisely defined the camera model, which allows us to map points in space to points in an image. However, we have not defined how the position of the camera (sometimes called camera motion) can be described nor have we integrated it in the projection process, that is, we have assumed that the camera center is at the origin of a World Reference Frame \(W_{rf}\).
2. ESSENTIAL CONCEPTS IN VISUAL GEOMETRY

The projection from camera coordinates $X_{\text{cam}}$ to image coordinates $x$, defined earlier in Equation 2.3, can also be described as:

$$x = K[I|0]X_{\text{cam}}, \quad (2.8)$$

with $K$ as in 2.3, $I$ a $3 \times 3$ identity matrix and $0$ a $3 \times 1$ vector of zeros.

This projection operation however only works for cameras that are located at the center of the world. If the camera is located elsewhere, we need to incorporate the transformation between the world center and the camera center in the projection process. Hartley and Zissermann chose to describe that transformation as the rotation and translation $(R', t')$ that is required to bring the center of the world to the camera center. We chose however the inverse transformation, that is the rotation and translation $(R, t)$ that is required to bring the camera to the center of the world. The reason to chose this representation is that it depicts more accurately the fact that we have a center of the world that is universal and nonmovable, while the camera is the one moving around the environment.

Since we will be dealing with sequences of images it is also sensible to reference the camera centers in a relative fashion. We can then distinguish between the absolute and relative camera poses. The absolute camera pose is described as the rotation and translation that will be required to bring the camera to the center of the world. The relative camera pose is described as the rotation and translation that is required to bring each camera to the previous camera pose.

This relative and absolute description of motion is very convenient for us, as we will see in the following sections. Additionally, we can elegantly accumulate the relative camera motion by a simple matrix multiplication.

The projection operation for cameras not located at the center is then described with respect to the absolute rotation and translation required to bring the camera to the origin of the $W_{r.f}$:

$$x = K[R|t]X_{\text{cam}} = PX_{\text{cam}}, \quad (2.9)$$

where we define the camera position, or projection matrix, as $P = K[R|t]$. $P$ is also sometimes called camera motion since in monocular systems it represents the different positions of a camera moving through the environment.
2.1 Formalizing the Geometry of the Imaging Process

Both absolute and relative motions are described with a $3 \times 4$ matrix containing a $3 \times 3$ rotation matrix $R$ and a $3 \times 1$ translation vector $t$. When expressed in homogeneous coordinates using a $4 \times 4$ matrix with last row $[0^T|1]$ the relative motion of cameras can be accumulated by a simple multiplication.

This definition of the camera position based on the projection process is very intuitive since it encompasses the motion as a rotation and translation in 3D space. However, since we are interested in the motion between images, it is important to define the relation between points in two different images. The epipolar geometry "is the intrinsic projective geometry between two views" [1]. It does not depend on the image itself but rather on the internal parameters and the relative position of the cameras. This projective geometry does not say anything about the 3D structure, contrary to equation 2.9

The relation with the previous camera motion definition is straightforward: if a point $X$ in the real world is projected as $x$ in a first image, and then projected as $x'$ in a second image, then those points must satisfy the following relation defined through the fundamental matrix $F$:

$$x'^T F x = 0$$

(2.10)

The proposed relation however does not state anything about the camera calibration. A more specific relation that accounts for $K$ is the essential matrix $E$, which relates the calibrated points in two images.

Say we consider the calibrated imaged points $\hat{x}$, which are the projection $x$ in the image plane of a point $X$ in 3D space, but accounting for the calibration parameters:

$$\hat{x} = K^{-1} x$$

(2.11)

Then introduce those calibrated image points in equation 2.10 obtaining:

$$\hat{x}'^T K^T F K \hat{x} = 0$$

(2.12)

So the relation between the essential matrix and the fundamental matrix is defined as:

$$\hat{x}'^T E \hat{x} = \hat{x}'^T K'^T F K \hat{x} = 0$$

(2.13)
2. ESSENTIAL CONCEPTS IN VISUAL GEOMETRY

The difference between the fundamental matrix $F$ and the essential matrix $E$ is that the latter already encodes the camera calibration. This implies a total of 5 degrees of freedom: 3 for the translation, 3 from the rotation minus 1 for the overall scale. Several solutions for the camera matrices can be retrieved up to a scale ambiguity \cite{1} from the essential matrix using the method from Horn \cite{2}. Horn’s method first extracts the rotation $R$ and then the translation $t$ from $E$ by observing that $E = BR$ where $B$ is a skew symmetric matrix that satisfies $Bv = t \times v$ for all vectors $v$. The method provides four possible solutions ($[R_+, t], [R_-, t], [R_+, -t], [R_-, -t]$). Choosing the correct one involves a voting approach, accepting the solution where most of the structure can be recovered in front of both cameras.

2.2 About Referencing Cameras, Images and Objects

So far we have assumed some coordinate system to be present to reference both the real life object and the image pixels. It is important to define clearly those coordinate systems since all transformations that need to be applied for the reconstruction will depend on them. Furthermore, we have introduced the need for a coordinate system to reference real life objects, and one to reference pixel locations in the image. Additionally, when we introduce multiple images we also need to introduce a universal reference frame so we can reference those camera positions with respect to the real world.

2.2.1 Camera Reference Frame

The first reference frame we need is the camera coordinate system $C_{rf}$. It is typically placed to have its origin at the focal point (the center of projection) and we must decide the direction of two of the axes since one is commonly perpendicular to the image sensor. Hartley and Zisserman chose a right hand coordinate system where the z-axis points towards the scene, and the y-axis and x-axis point upwards and to the left respectively (when looking from the center towards the z-axis). We chose however an x-axis that points to the right while retaining the upward y, effectively adopting a left-hand side coordinate system. We make this choice given that image features are typically represented from left to right. Furthermore, we chose the z-axis to be perpendicular to the image sensor. From this frame, we can reference the
2.2 About Referencing Cameras, Images and Objects

points in space that belong to the object, the imaging sensor position at a distance \( f \) from the center and the size of the sensor.

### 2.2.2 Image Reference Frame

We also need to define a 2 dimensional local frame \( I_{rf} \) to reference the points in the image. These points will be the pixels in the images that are obtained through the projection mechanism.

It is important to decide if we chose to put the sensor behind the center of projection, or in front. This, along with considerations on the software used to read the images, will determine the selection of our reference frame. Since we chose a left-hand camera reference frame, and we want to avoid an inverted image behind the projection center, we chose to place the sensor in front of the projection center and use the same orientation on the x- and y-axis as in the camera reference frame.

Therefore, the x-axis points to the right of the image as seen on a screen or paper, while the y-axis points towards the top of the image. On the other hand, given that Matlab references the images from the top-left corner, we place the center of the image reference frame in that corner.

### 2.2.3 World Reference Frame

Finally, having defined the image and camera reference systems, we need one more coordinate system \( W_{rf} \) to be able to reference multiple cameras in space. This is the world reference coordinate system and it is used as an absolute coordinate system from which cameras or objects can be referenced in space.

This center is effectively the center of our reconstruction universe. It can be placed arbitrarily anywhere. In a scenario where multiple cameras are used, it seems natural to choose the center of the world to be placed coincident with one of the camera centers. Since we are concerned with sequences of image, we chose to place the center of the world at the position of the first image of a given set. Therefore, image zero is always placed at the origin of the coordinate system of the reconstruction.
2. ESSENTIAL CONCEPTS IN VISUAL GEOMETRY

2.3 Some Computer Vision Concepts

The extension of the pinhole camera model with the calibration matrix accommodates modern cameras where the imaging sensor captures arrays of pixels. Those pixels are the ones referenced in the previous sections as image points. We need to define a procedure to establish sets of corresponding points or pixels that when backtracked intersect the real object in the same spot. In this section we describe how these set of points can be found in every image and then matched across different images.

2.3.1 Image Features

We consider an image feature as a pixel in the image for which we can find some unique description. The image feature is then composed of the pixel location (in $I_{r_f}$) and its descriptor. Traditionally (before 1996), these image features were defined by user input. This meant the user had to click on given remarkable pixels in the image such as corners or salient features in the structure.

As early as 1988, Harris et. al. \cite{harris1988combined} present a corner and edge detector extending the work from Moravec \cite{moravec1978integrated} eight years earlier. The Harris corner detector was used in one
2.3 Some Computer Vision Concepts

of the first automated reconstruction procedures in 1996 which we shall discuss in the next chapter. Even though this detector was popular, we fast forward to 1999 where Lowe proposed Scale Invariant Features (SIFT). Later in 2004 he refined the method. Since then, SIFT features have become one of the most popular choices for computer vision algorithms. They are robust against rotation and scaling. Lowe provided not only an algorithm to find the features but also a rich descriptor.

In 2006 Bay et. al. present a more efficient alternative to SIFT called Speeded Up Robust Features (SURF). Even though they report better performance that SIFT, both have been frequently used in computer vision.

In 2010 Tola et. al. present a local image descriptor called DAISY. As opposed to SURF or SIFT, DAISY is very efficiently computed densely throughout the whole image. This allows for a much larger set of image features that can be used for reconstruction. Tola’s method produces dense depth maps when using wide baseline stereo pairs of images.

Several implementations for extracting SIFT features are available. Lowe provides both a feature detector and descriptor. Vedaldi provides an open implementation of both detector and descriptor.

2.3.2 Image Feature Matching

In the previous section we gave an overview of the most widely used methods for finding points of interest in the images. Those points need to be matched with points found in different images so that the reconstruction process can be performed. Traditional work using user selected image features also required the user to provide the correspondence with other images. That is a laborious job that can hardly be done for more than a couple of images.

In 1996 Deriche et. al. use the Harris corner detector to propose a feature matching method that was later used to compute the epipolar geometry. Their matching method consist of exploring the second image using a window of a certain size centered on the location of the image feature to be matched with. This limits their approach to small camera motion where the features are not expected to move much from their original location.

SIFT, SURF and DAISY all provide the means to perform feature matching with no limitation with respect to where the target image feature might lie.
2. ESSENTIAL CONCEPTS IN VISUAL GEOMETRY

In our work we use the SIFT detector and descriptor given its ability to find large number of features in the structure and texture present in facades. For the matching procedure we use the algorithm proposed by Lowe [5] based on finding the nearest neighbor in the descriptor space and using a threshold to determine how close the matches need to be. This matching procedure is used to match image features across consecutive images, though a more advanced approach is used to improve robustness as we will discuss in Chapter 4.

2.4 Summary

In this chapter we have reviewed and discussed the essential concepts from visual geometry that are required to understand the 3D reconstruction process. We have introduced the pinhole camera model that fits most consumer cameras and used this definition to develop the fundamental theoretical groundwork. We have defined the camera calibration \( K \) and lens distortion, the projection operation that we introduced in Chapter 1 and the two primordial matrices \( E \) and \( F \) that define the relation between matching projecting points in two different images. Furthermore we have described how objects, cameras and image features can be referenced and our choices with respect to those coordinate systems. Finally we have introduced the concept of image feature and described some alternatives for finding and matching image points across multiple images. These concepts represent the basics over which in the following chapters we are going to develop our methods and techniques for large scale and semantic 3D reconstruction.