9
Stereo Vision and Structure From Motion

The previous chapter developed a mathematical relationship between the position of a point P in a scene (expressed in world frame coordinates P_W), and the corresponding point p in pixel coordinates that gets projected onto the image plane of the camera. This relationship was derived based on the pinhole camera model, and required knowledge about the camera's intrinsic and extrinsic parameters. Nonetheless, even in the case where all of these camera parameters are known it is still impossible to reconstruct the depth of P with a single image (without additional information). However, in the context of robotics, recovering 3D information about the structure of the robot's environment through computer vision is often a very important task (e.g. for obstacle avoidance). Two approaches for using cameras to gather 3D information are therefore presented in this chapter, namely *stereo vision* and *structure from motion*[1,2].

Stereo Vision and Structure From Motion

Recovering scene structure from images is extremely important for mobile robots to safely operate in their environment and successfully perform tasks. While a number of other sensors can also be used to recover 3D scene information, such as ultrasonic sensors or laser rangefinders, cameras capture a broad range of information that goes beyond depth sensing. Additionally, cameras are a well developed technology and can be an attractive option for robotics based on cost or size.

Unfortunately, unlike sensors that are specifically designed to measure depth like laser rangefinders, the camera's projection of 3D data onto a 2D image makes it impossible to gather some information from a single image[3]. Techniques for extracting 3D scene information from 2D images have therefore been developed that leverage *multiple* images of a scene. Examples of such techniques include *depth-from-focus* (uses images with different focuses), *stereo vision* (uses images from different viewpoints), or *structure from motion* (uses images captured by a moving camera).

[1] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza. *Introduction to Autonomous Mobile Robots*. MIT Press, 2011

[2] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2011

[3] Unless you are willing to make some strong assumptions, for example that you know the physical dimensions of the objects in the environment.

9.1 Stereo Vision

Stereopsis (from *stereo* meaning solidity, and *opsis* meaning vision or sight) is the process in visual perception leading to the sensation of depth from two slightly different projections of the world onto the retinas of the two eyes. The difference in the two retinal images is called horizontal *disparity*, retinal disparity, or binocular disparity, and arise from the eyes' different positions in the head. It is the disparity that makes our brain fuse (perceive as a single image) the two retinal images, making us perceive the object as one solid object. For example, if you hold your finger vertically in front of you and alternate closing each eye you will see that the finger jumps from left to right. The distance between the left and right appearance of the finger is the disparity.

Computational stereopsis, or *stereo vision*, is the process of obtaining depth information of a 3D scene via images from two cameras which look at the same scene from different perspectives. This process consists of two major steps: fusion and reconstruction. Fusion is a problem of correspondence, in other words how do you correlate each point in the 3D environment to their corresponding pixels in *each* camera. Reconstruction is then a problem of *triangulation*, which uses the pixel correspondences to determine the full position of the source point in the scene (including depth).

9.1.1 Epipolar Constraints

As previously mentioned, the first step in the stereo vision process is to fuse the two (or more) images and generate point correspondences[4]. This task can be quite challenging, and erroneously matching features can lead to large errors in the reconstruction step. Therefore, several techniques are leveraged to make this task simpler. The most important simplifying technique is to impose an *epipolar constraint*.

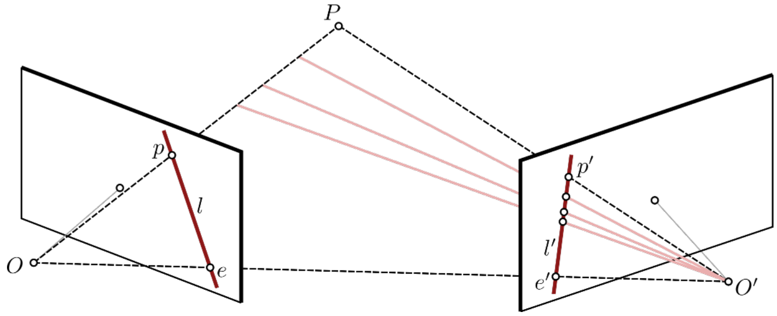

Figure 9.1: The point P in the scene, the optical centers O and O' of the two cameras, and the two images p and p' of P all lie in the same plane, referred to as the epipolar plane. The lines l and l' are the epipolar lines of the points p and p', respectively. Note that if the point p is observed in one image, the corresponding point in the second image must lie on the epipolar line l'!

Consider the images p and p' of a point P observed by two cameras with optical centers O and O' (see Figure 9.1). These five points all belong to the *epipolar plane* defined by the two intersecting rays OP and $O'P$. In particular, the point p lies on the line l where the epipolar plane and the image plane intersect. The line l is referred to as the *epipolar line* associated with the point p, and it

[4] This generally assumes that the perspective of each image is only a slight variation from the other, such that the features appear similarly in each.

passes through the point e (referred to as the *epipole*). Based on this geometry, if p and p' are images of the same point P, then p must lie on the epipolar line l and p' must lie on the epipolar line l'.

Therefore, when searching for correspondences between p and p' for a particular point P in the scene it makes sense to restrict the search to the corresponding epipolar line. This is referred to as an *epipolar constraint*, and greatly simplifies the correspondence problem by restricting the possible candidate points to a line rather than the entire image (i.e. a one dimensional search rather than a two dimensional search). Mathematically, the epipolar constraints can be written as:

$$\overline{Op} \cdot [\overline{OO'} \times \overline{O'p'}] = 0, \tag{9.1}$$

since $\overline{Op}$, $\overline{O'p'}$, and $\overline{OO'}$ are coplanar. Assuming the world reference frame is co-located with camera 1 (with an origin at point O) this constraint can be written as:

$$p^\top F p' = 0, \tag{9.2}$$

where F, referred to as the *fundamental matrix*, has seven degrees of freedom and is singular. For a derivation of the epipolar constraint see Section 7.1 from Forsyth et al.[5]. Additionally, the matrix F is only dependent on the intrinsic camera parameters for each camera and the geometry that defines their relative positioning, and can be assumed to be constant. The expression for the fundamental matrix in terms of the camera intrinsic parameters is:

$$F = K^{-\top} E K'^{-1}, \quad E = \begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{bmatrix} R, \tag{9.3}$$

where K and K' are the intrinsic parameter matrices for cameras 1 and 2 respectively, and R and $t = [t_1, t_2, t_3]^\top$ are the rotation matrix and translation vector that map camera 2 frame coordinates into camera 1 frame coordinates. Note that with the epipolar constraint defined by the fundamental matrix (9.2), the epipolar lines l and l' can be expressed by $l = Fp'$ and $l' = F^\top p$. Additionally, it can be shown that $F^\top e = Fe' = 0$ where e and e' are the epipoles in the image frames of cameras 1 and 2, since by definition the translation vector t is parallel to the coordinate vectors of the epipoles in the camera frames. This in turn guarantees that the fundamental matrix F is singular.

If the parameters K, K', R, and t are not already known, the fundamental matrix F can be determined in a manner similar to the intrinsic parameter matrix K in the previous chapter. Suppose a number of corresponding points $p^h = [u, v, 1]^\top$ and $(p^h)' = [u', v', 1]^\top$ are known and are expressed as homogeneous coordinates. Each pair of points has to satisfy the epipolar constraint (9.2), which can be written as:

$$\begin{bmatrix} u & v & 1 \end{bmatrix} \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = 0$$

[5] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2011

This expression can then be equivalently expressed by reparameterizing the matrix F in vector form f as:

$$\begin{bmatrix} uu' & uv' & u & vu' & vv' & v & u' & v' & 1 \end{bmatrix} f = 0 \tag{9.4}$$

where $f = [F_{11}, F_{12}, F_{13}, F_{21}, F_{22}, F_{23}, F_{31}, F_{32}, F_{33}]^\top$. For n known correspondences (p, p') these constraints can be stacked to give:

$$Wf = 0, \tag{9.5}$$

where $W \in \mathbb{R}^{n \times 9}$. Given $n \geq 8$ correspondences, an estimate $\tilde{F}$ of the fundamental matrix estimate is given by:

$$\begin{aligned} \min_{f} \ & \|Wf\|^2, \\ \text{s.t.} \ & \|f\|^2 = 1. \end{aligned} \tag{9.6}$$

Note that the estimate $\tilde{F}$ computed by (9.6) is not guaranteed to be singular. A second step is therefore taken to enforce this additional condition. In particular it is desirable to find the matrix F that is closest to the estimate $\tilde{F}$ that has a rank of two:

$$\begin{aligned} \min_{F} \ & \|F - \tilde{F}\|^2, \\ \text{s.t.} \ & \det(F) = 0, \end{aligned} \tag{9.7}$$

which can be accomplished by computing a singular value decomposition of the matrix $\tilde{F}$.

9.1.2 *Image Rectification*

Given a pair of stereo images, epipolar rectification is a transformation of each image plane such that all corresponding epipolar lines become colinear and parallel to one of the image axes, for convenience usually the horizontal axis. The resulting rectified images can be thought of as acquired by a new stereo camera obtained by rotating the original cameras about their optical centers. The great advantage of the epipolar rectification is the correspondence search becomes simpler and computationally less expensive because the search is done along the horizontal lines of the rectified images. The steps of the epipolar rectification algorithm are illustrated in Figure 9.2. Observe that after the rectification, all the epipolar lines in the left and right image are colinear and horizontal. For an in-depth discussion on algorithms for image rectification see [6,7].

9.1.3 *Correspondence Problem*

Epipolar constraints and image rectification are commonly used in stereo vision to address the problem of correspondence, which is the problem of determining the pixels p and p' from two different cameras with different perspectives

[6] A. Fusiello, E. Trucco, and A. Verri. "A compact algorithm for rectification of stereo pairs". In: *Machine Vision and Applications* 12.1 (2000), pp. 16–22

[7] C. Loop and Z. Zhang. "Computing rectifying homographies for stereo vision". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 1. 1999, pp. 125–131

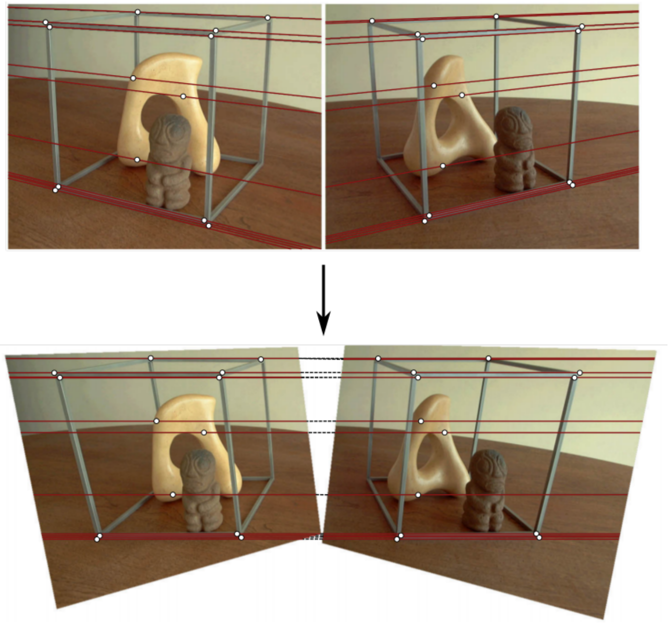

Figure 9.2: Epipolar rectification example from Loop et al. (1999).

that correspond to the same scene feature P. While these concepts make finding correspondences easier, there are still several challenges that must be overcome. These include challenges related to feature occlusions, repetitive patterns, distortions, and others.

9.1.4 *Reconstruction Problem*

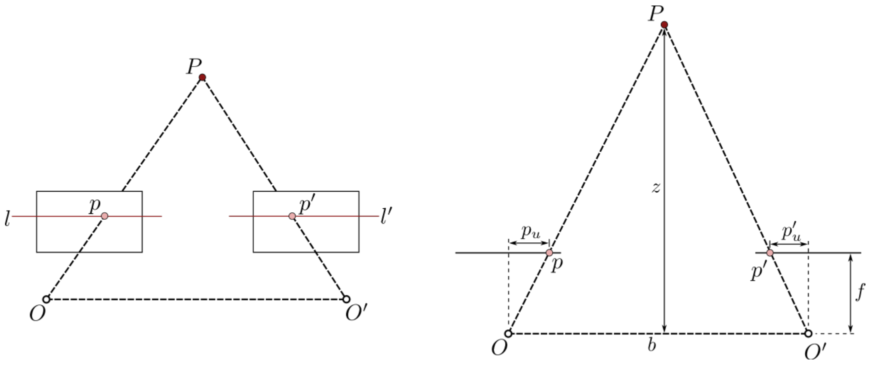

Figure 9.3: Triangulation with rectified images (horizontal view on the left, top-down view on the right).

In a stereo vision setup, once a correspondence between the two images is identified it is possible to reconstruct the 3D scene point based on *triangulation*. This process of triangulation has already been covered by the discussion on the epipolar geometry. However if the images have also be rectified such that the epipolar lines become parallel to the horizontal image axis the triangulation problem becomes simpler. This occurs, for example, when the two cameras have the same orientation, are placed with their optical axes parallel, and are separated by some distance b called the *baseline* (see Figure 9.3).

In Figure 9.3, a point P on the object is described as being at coordinate (x, y, z) with respect to the origin located in the left camera at point O. The horizontal pixel coordinate in the left and right image are denoted by p_u and p'_u respectively. Based on the geometry the depth of the point P can be computed from the properties of similar triangles:

$$\frac{z}{b} = \frac{z - f}{b - p_u + p'_u}, \tag{9.8}$$

which can be algebraically simplified to:

$$z = \frac{bf}{p_u - p'_u}, \tag{9.9}$$

where f is the focal length. Generally a small baseline b will lead to larger depth errors, but a large baseline b may cause features to be visible from one camera but not the other. The difference in the image coordinates, $p_u - p'_u$, is referred to as *disparity*. This is an important term in stereo vision, because it is only by measuring disparity that depth information can be recovered. The disparity can also be visually represented in a *disparity map* (for example see Figure 9.4), which is simply a map of the disparity values for each pixel in an image. The largest disparities occur from nearby objects (i.e. since disparity is inversely proportional to z).

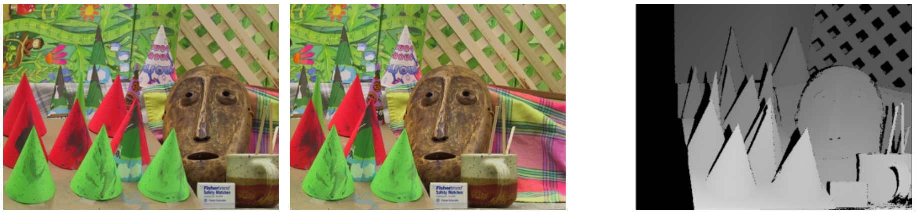

Figure 9.4: Disparity map from a pair of stereo images. Notice that the lighter values of the disparity map represent larger disparity, and correspond to the point in the scene that are closer to the cameras. The black points represent points that were occluded from one of the images and therefore no correspondence could be made. Images from Scharstein et al. (2003) .

9.2 *Structure From Motion (SFM)*

The structure from motion (SFM) method uses a similar principle as stereo vision, but uses *one* camera to capture multiple images from different perspectives while moving within the scene. In this case, the intrinsic camera parameter matrix K will be constant, but the extrinsic parameters (i.e. the rotation matrix R and relative position vector t) will be different for each image. Consider a case where m images of n fixed 3D points are taken from different perspectives. This would involve m homography matrices M_k and n 3D points P_j that would need to be determined by leveraging the relationships:

$$p^h_{j,k} = M_k P^h_j, \quad j = 1, \dots, n, \quad k = 1, \dots, m.$$

However, SFM also has some unique disadvantages, such as an ambiguity in the absolute scale of the scene that cannot be determined. For example a bigger

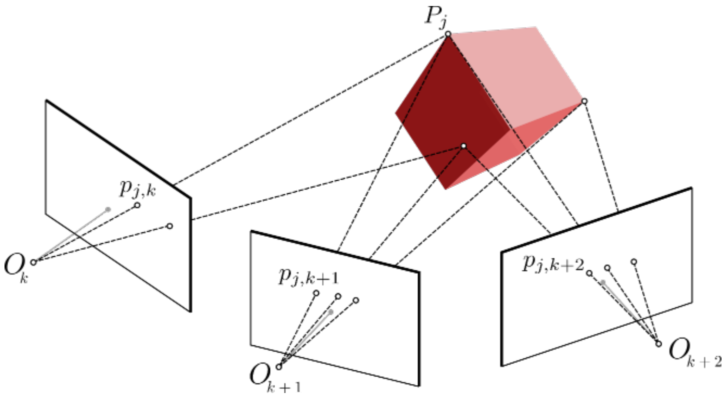

Figure 9.5: A depiction of the structure from motion (SFM) method. A single camera is used to take multiple images from different perspectives, which provides enough information to reconstruct the 3D scene.

object at a longer distance and a smaller object at a closer distance may yield the same projections.

One application of the SFM concept is known as *visual odometry*. Visual odometry estimates the motion of a robot by using visual inputs (and possible additional information). This approach is commonly used in practice, for example by rovers on Mars, and is useful because it not only allows for 3D scene reconstruction but also to recover the motion of the camera.