AA274A: Principles of Robot Autonomy I
Course Notes

Oct 18, 2019

10 Stereo Vision and Structure from Motion

Range sensing is extremely important in mobile robotics, since it is a basic input for successful obstacle avoidance. As we have seen earlier in this chapter, a number of sensors are popular in robotics explicitly for their ability to recover depth estimates: ultrasonic, laser rangefinder, time-of-flight cameras. It is natural to attempt to implement ranging functionality using vision chips as well.

However, a fundamental problem with visual images makes rangefinding relatively difficult. Any vision chip collapses the 3D world into a 2D image plane, thereby losing depth information. If one can make strong assumptions regarding the size of objects in the world, or their particular color and reflectance, then one can directly interpret the appearance of the 2D image to recover depth. But such assumptions are rarely possible in real-world mobile robot applications. Without such assumptions, a single picture does not provide enough information to recover spatial information.

The general solution is to recover depth by looking at *several* images of the scene to gain more information, hopefully enough to at least partially recover depth. The images used must be different, so that taken together they provide additional information. They could differ in camera geometry—such as the focus position or lens iris—yielding depth from focus (or defocus) techniques that we have described in the past couple lectures. An alternative is to create different images, not by changing the camera geometry, but by changing the camera viewpoint to a different camera position. This is the fundamental idea behind *structure from stereo* (i.e., stereo vision) and *structure from motion* that we will present in the next sections[1]. As we will see, stereo vision processes two distinct images taken at the same time and assumes that the relative pose between the two cameras is known. Structure-from-motion conversely processes two images taken with the same or a different camera at different times and from different unknown positions; the problem consists in recovering both the relative motion between the views and the depth. The 3D scene that we want to reconstruct is usually called *structure*.

[1]Much of this lecture is a direct excerpt from [SNS11] Chapter 4.2.5 and 4.2.6 unless noted otherwise.

10.1 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 the 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. The differences 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 a one and solid. To have a clearer understanding of what disparity is, as a simple test, hold your finger vertically in front of you and close each eye alternately. 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 from a pair of images coming from two cameras which look at the same scene from different positions. In stereo vision we can identify two major problems:

1. Correspondence

2. 3D reconstruction

The first consists in matching (pairing) points of the two images which are the projection of the same point in the scene. These matching points are called *corresponding points* or *correspondences* as described in the top part of Figure 1. Determining the corresponding points is made possible based on the assumption that the two images differ only slightly and therefore a feature in the scene appears similar in both images. Based only of this assumption, however, there might be many possible false matches. As we will see, this problem can be overcome by introducing an additional constraint which makes the correspondence matching feasible. This constraint is called *epipolar constraint* and states that the correspondent of a point in an image lies on a line (called *epipolar line*) in the other image (bottom image of Figure 1. Because of this constraint, we will see that the correspondence search becomes one-dimensional instead of two-dimensional.

10.1.1 Epipolar geometry

We start our discussion with epipolar geometry[2] before discussing how to solve the correspondence problem.

Consider the images p and p' of a point P observed by two cameras with optical centers O and O'. These five points all belong to the *epipolar plane* defined by the two intersecting rays OP and $O'P$ (Figure 2). In particular, the point p' lies on the line l' where this plane and the retina Π' of the second camera intersect. The line l' is the epipolar line associated with the point p, and it passes through the point e' where the baseline joining the optical centers O and O' intersects Π'. Likewise, the point p lies on the epipolar line l associated

[2]The majority of this section is a direct excerpt from [DAF11].

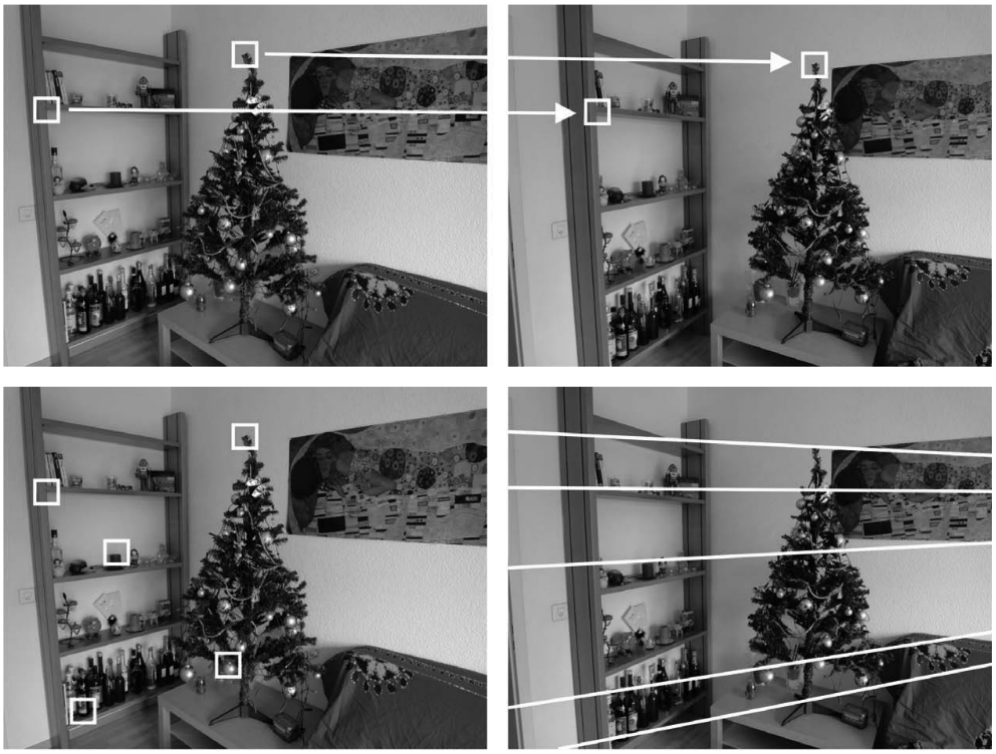

Figure 1: A stereo pair. Corresponding points are projections of the same scene point. Because of the epipolar constraint, conjugate points can be searched along the epipolar lines. This heavily reduces the computational cost of the correspondence search: from a two-dimensional search it becomes a one-dimensional search problem. [SNS11]

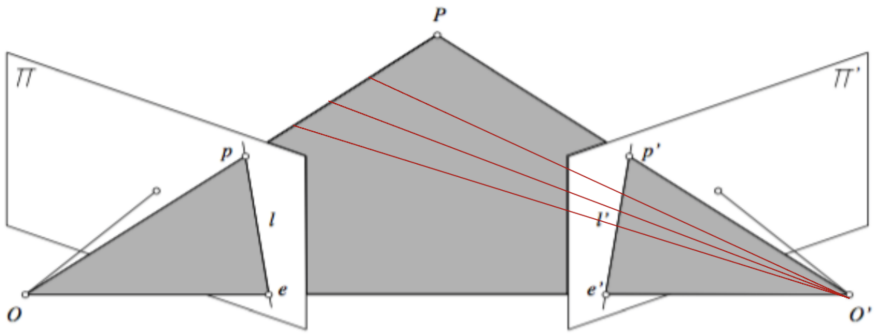

Figure 2: Epipolar geometry: the point P, the optical centers O and O' of the two cameras, and the two images p and p' of P all lie in the same plane.

with the point p', and this line passes through the intersection e of the baseline with the

plane Π.

The points e and e' are called the *epipoles* of the two cameras. The epipole e' is the (virtual) image of the optical center O of the first camera in the image observed by the second camera, and vice versa. As noted before, if p and p' are images of the same point, then p' must lie on the epipolar line associated with p. This *epipolar constraint* plays a fundamental role in stereo vision and motion analysis.

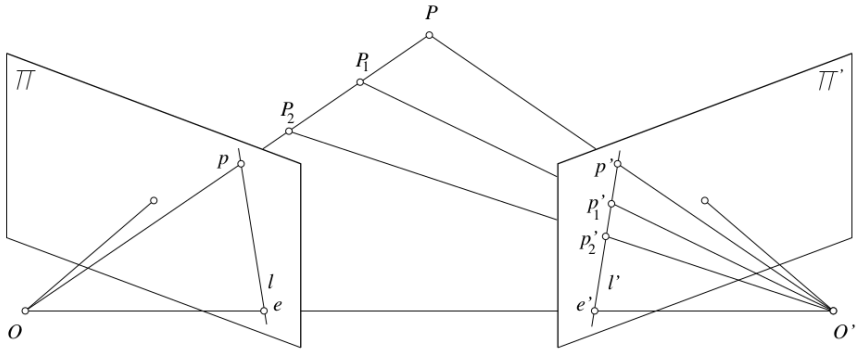

Figure 3: Epipolar constraint: given a calibrated stereo rig, the set of possible matches for the point p is constrained to lie on the associated epipolar line l'.

Let us assume for example that we know the intrinsic and extrinsic parameters of the two cameras of a stereo rig. The most difficult part of stereo data analysis is often establishing correspondences between the two images, i.e., deciding which points in the right picture match the points in the left one. The epipolar constraint greatly limits the search for these correspondences: indeed, since we assume that the rig is calibrated, the coordinates of the point p completely determine the ray joining O and p, and thus the associated epipolar plane $OO'p$ and epipolar line. The search for matches can be restricted to this line instead of the whole image (Figure 3). In two-frame motion analysis on the other hand, each camera may be internally calibrated, but the rigid transformation separating the two camera coordinate systems is unknown. In this case, the epipolar geometry obviously constrains the set of possible motions.

10.1.2 Epipolar Constraints in Uncalibrated Cameras and the Essential Matrix

Mathematically, epipolar constraints can be written as:

$$\overline{Op} \cdot [\overline{OO'} \times \overline{O'p'}] = 0, \tag{1}$$

since $\overline{Op}$, $\overline{O'p'}$, and $\overline{OO'}$ are coplanar. To see this is true, note that $\overline{OO'} \times \overline{O'p'}$ is a vector perpendicular to the epipolar plane, and consequently perpendicular to $\overline{Op}$. The dot product of these perpendicular vectors is then defined to be 0.

Note that a cross product can be expressed as the product of skew-symmetric matrix and a vector.

$$a \times b = [a]_x b$$
$$\text{where } [a]_x = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}.$$

For internally calibrated cameras, intrinsic parameters are known so that image positions can be expressed in normalized coordinates. When these parameters are unknown, we can write $\boldsymbol{p} = \mathcal{K}\hat{\boldsymbol{p}}$ and $\boldsymbol{p}' = \mathcal{K}'\hat{\boldsymbol{p}}'$ where $\mathcal{K}$ and $\mathcal{K}'$ are 3×3 calibartion matrices, and $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{p}}'$ are normalized image coordinate vectors. It can be shown[3] that the following relationship

$$\boldsymbol{p}^T \mathcal{F} \boldsymbol{p}' = 0 \tag{2}$$

is true, where $\mathcal{F} = \mathcal{K}^{-T}\mathcal{E}\mathcal{K}'^{-1}$ is called the *fundamental matrix* and $\mathcal{E}$, the *essential matrix*. $\mathcal{F}$ has rank two, and the eigenvector of $\mathcal{F}$ corresponding to its zero eigenvalue is as before the position $\boldsymbol{e}'$ of the epipole. Note that $\mathcal{F}\boldsymbol{p}'$ represents the epipolar line corresponding to the point $\boldsymbol{p}'$ in the first $\boldsymbol{p}'$ in the first image.

We now have a single matrix that defines the epipolar constraint for any normalized point correspondences between two calibrated cameras. If we note that a line in homogeneous coordinates is defined by the points that satisfy $\boldsymbol{p} \cdot l = 0$ then we can see right away that the epipolar lines can be calculated as:

$$l = \mathcal{F}\hat{\boldsymbol{p}} f' \tag{3}$$
$$l' = \mathcal{F}^T \boldsymbol{p}. \tag{4}$$

$\mathcal{F}$ is singular when $\mathcal{F}^T e = \mathcal{F}e' = 0$. $\mathcal{F}$ is defined by 9 elements but is constrained by $det(\mathcal{F}) = 0$ and a common scaling so in the end $\mathcal{F}$ has only 7 degrees of freedom.

10.1.3 Computing the Fundamental Matrix

To compute a fundamental matrix, we use the same approach as we did for projection matrix. We find a number of correspondences using clearly distinguishable points such as corners.

Assume we are given a number of corresponding points $\boldsymbol{p} = [u, v, 1]^T$ and $\boldsymbol{p}' = [u', v', 1]^T$ expressed in a homogeneous coordinate. Each pair of points has to satisfy the epipolar constraint $\mathcal{F}$.

$$[u, v, 1] \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = 0 \tag{5}$$

[3]Refer to Longuet-Higgins relation.

We can rewrite this into a scalar product of two one-dimensional vectors. The first vector contains the known coefficients from the given points and the second vector contains each element of the fundamental matrix. Now we have a scalar euqation, therefore one constraint, for each pair of given points.

$$[uu', uv', u, vu', vv', v, u', v', 1] \begin{bmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \\ F_{33} \end{bmatrix} = Wf = 0 \tag{6}$$

As in the projection matrix, the fundamental matrix can be defined up to scale, e.g. normalized by the last component. Therefore, we need 8 parameters to estimate 9 entries of the fundamental matrix. Given $n >= 8$ correspondences, we can solve $\tilde{\mathcal{F}}$

$$\min_{f \in R^9} \|Wf\|^2 \tag{7}$$

$$\text{subject to } \|f\|^2 = 1 \tag{8}$$

Here, while $\tilde{\mathcal{F}}$ satisfies the epipolar constraints, it is not necessarily singular. we can enforce rank-2 constraint via SVD decomposition to get the proper fundamental matrix from the candidates.

$$\operatorname{argmin}_{\mathcal{F}} \|\mathcal{F} - \tilde{\mathcal{F}}\|^2 \tag{9}$$

$$\text{subject to } \det(\mathcal{F}) = 0 \tag{10}$$

10.1.4 Image Rectification

Given a pair of stereo images, epipolar rectification is a transformation of each image plane such that all corresponding epipolar lines become collinear 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 4. Observe that after the rectification, all the epipolar lines in the left and right image are collinear and horizontal (Figure 4). The equations for the

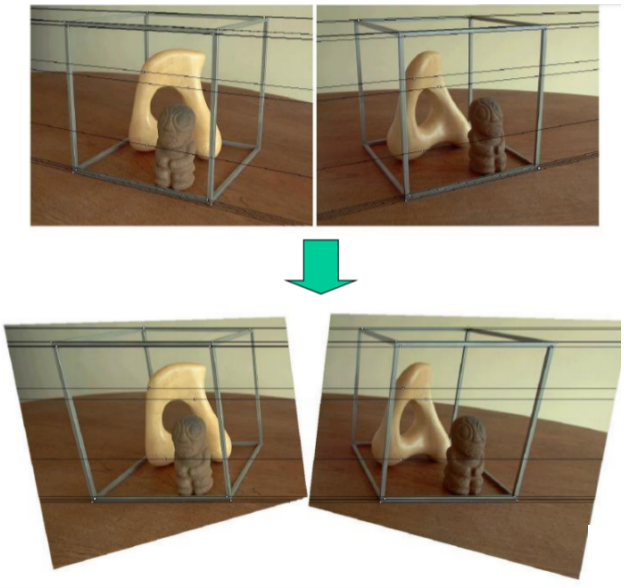

Figure 4: Image Rectification

epipolar rectification algorithm go beyond the scope of this class, but the interested reader can find an easy-to-implement algorithm in [FTV00]. For the remaining lecture, we will assume rectified image pairs.

10.2 Correspondence Problem

We digressed from our previous discussion on stereo vision to understand how stereopic images created by exploiting epipolar geometric properties. Now we understand how to obtain recified image pairs by imposing epipolar constraints, we can return to our discussion on how to create stereo vision problem. Earlier we mentioned that stereo vision is a two step process: first, we find the corresponding features, and (2) we reconstruct image with depth. This section is about correspondence.

Earlier, we observed two stereo image pairs contain the conjugate pair p_l and p_r in the left and right camera images, which originates from the same scene point $\tilde{P}_w$ (Figure 1). This fundamental challenge is called the *correspondence problem*. Intuitively, the problem is: given two images of the same scene from different perspectives, how can we identify the same object points in both images? For every such identified object point, we will then be able to recover its 3D position in the scene.

The correspondence search is based on the assumption that the two images of the same scene do not differ too much, that is, a feature in the scene is supposed to appear very similar in both images. Using an opportune image similarity metric, a given point in the first image can be paired with one point in the second image. The problem of false correspondences makes the correspondence search challenging. False correspondences occur when

a point is paired to another that is not its real conjugate. This is because the assumption of image similarity does not hold very well, for instance if the part of the scene to be paired appears under different illumination or geometric conditions. Other problems that make the correspondence search difficult are:

- *Occlusions*: the scene is seen by two cameras at different viewpoints and therefore there are parts of the scene that appear only in one of the images. This means, there exist points in one image which do not have a correspondent in the other image.
- *Photometric distortion*: there are surfaces in the scene which are nonperfectly lambertian, that is, surfaces whose behavior is partly specular. Therefore, the intensity observed by the two cameras is different for the same point in the scene as more as the cameras are farther apart.
- *Projective distortion*: because of the perspective distortion, an object in the scene is projected differently on the two images, as more as the cameras are farther apart.

10.3 Reconstruction Problem

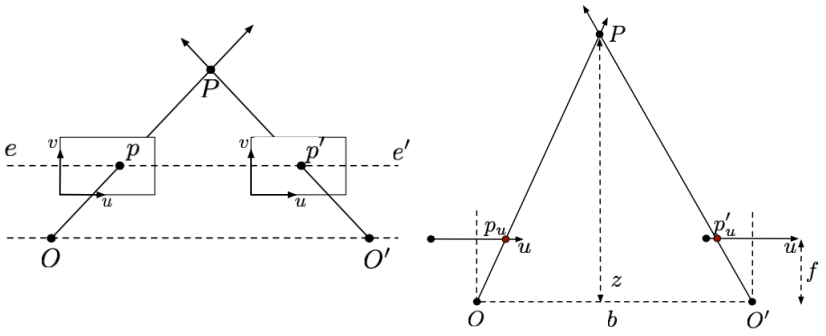

Figure 5: Triangulation under Rectified Images (horizontal view on the left, top-down view on the right)

Knowing the correspondences between the two images, knowing the relative orientation and position of the two cameras, and knowing the intrinsic parameters of the two cameras, it is possible to reconstruct the scene points (i.e., the *structure*). This process of *reconstruction* requires the prior calibration of the stereo camera; that is, we need to calibrate the two cameras separately for estimating their extrinsic parameters, but we also need to determine their extrinsic parameters, i.e. the camera relative position.

While deriving epipolar geometry in general (uncalibrated) case, we already saw how to triangulate correspondences in the general case. In other words, in rectified images, reconstruction problems are simply a triangulation problem, which is an idealized stereo vision problem in which two cameras have the same orientation and are placed with their optical axes parallel, at a separation of b (called baseline), shown in Figure 5.

In this figure, a point on the object is described as being at coordinate (x, y, z) with respect to the origin located in the left camera lens. The image coordinate in the left and right image are p_u and p'_u respectively. From Figure 5 and knowing the horizontal projection of p into two cameras, we can estimate the distance z of the point from the similar triangles

$$\frac{z}{b} = \frac{z - f}{b - p_u + p'_u} \tag{11}$$

from which we obtain

$$z = \frac{bf}{p_u - p'_u}. \tag{12}$$

Here b is the baseline and f is the focal length. If the baseline is very large, we have more resolution to estimate the distance z, some points may only be visible from one camera. If the baseline is small, we do not have problems in correspondences, we will have a larger relative error in the estimation of z.

Difference in the image coordinates, $p_u - p'_u$, is called *disparity*. This is an important term in stereo vision, because it is only by measuring disparity that we can recover depth information. Observations from this equation are as follows:

- Distance is inversely proportional to disparity. The distance to near objects can therefore be measured more accurately than that to distant objects, just as with depth from focus techniques. In general, this is acceptable for mobile robotics, because for navigation and obstacle avoidance closer objects are of greater importance.

- Disparity is proportional to b. For a given disparity error, the accuracy of the depth estimate increases with increasing baseline b.

- As b is increased, because the physical separation between the cameras is increased, some objects may appear in one camera but not in the other. This is due to the field of view of the cameras. Such objects by definition will not have a disparity and therefore will not be ranged.

After the calibration of the stereo-rig, the epipolar rectification, and the correspondence search, we can finally reconstruct the scene points in 3D by solving the system of equations by simple transformations. In particular, consider two calibrated cameras with intrinsic parameter matrices A_l and A_r, rotations R_l and R_r, and translations t_l, t_r, we can follow the common practice to assume the origin of the world coordinate system in the left camera, i.e., $R_l = I, R_r = R$. Then we can reconstruct perspective projection for the two cameras with:

$$s_l \tilde{\boldsymbol{p}}_l = A_l [R_l | t_l] \tilde{P}_w \tag{13}$$

$$s_r \tilde{\boldsymbol{p}}_r = A_r [R_r | t_r] \tilde{P}_w \tag{14}$$

where $\tilde{\boldsymbol{p}}_l = [u_l, v_l, 1]^T$ and $\tilde{\boldsymbol{p}}_r = [u_r, v_r, 1]^T$ are the image points (in homogeneous coordinates) corresponding to the world point $\tilde{P}_w = [x, y, z, 1]^T$ (in homogeneous coordinates) in the left

and right camera respectively. s_l and s_r are depth factors. After this series of operations, we can also output a disparity map. A *disparity map* appear as a grayscale image where the intensity of every pixel point is proportional to the disparity of that pixel in the left and right image: objects that are closer to the camera appear lighter, while farther objects appear darker. An example disparity map is shown in Figure 6. Disparity maps are very useful for obstacle avoidance

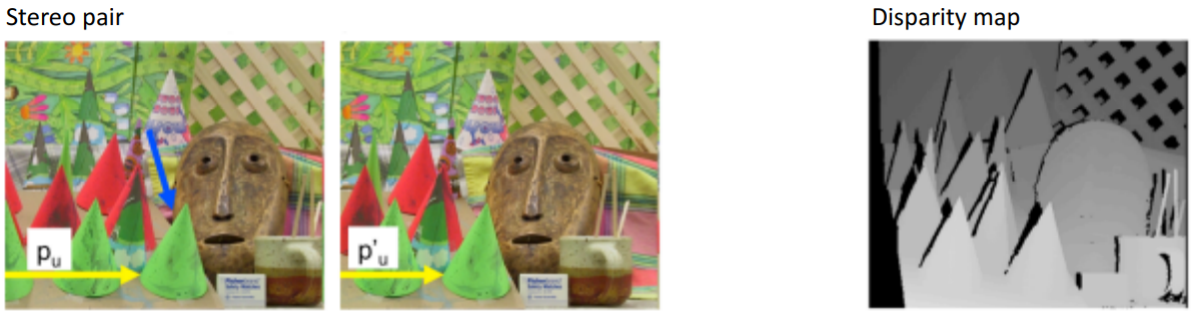

Figure 6: Disparity Map from a pair of Stereo Images

10.4 Stereo Fusion Techniques

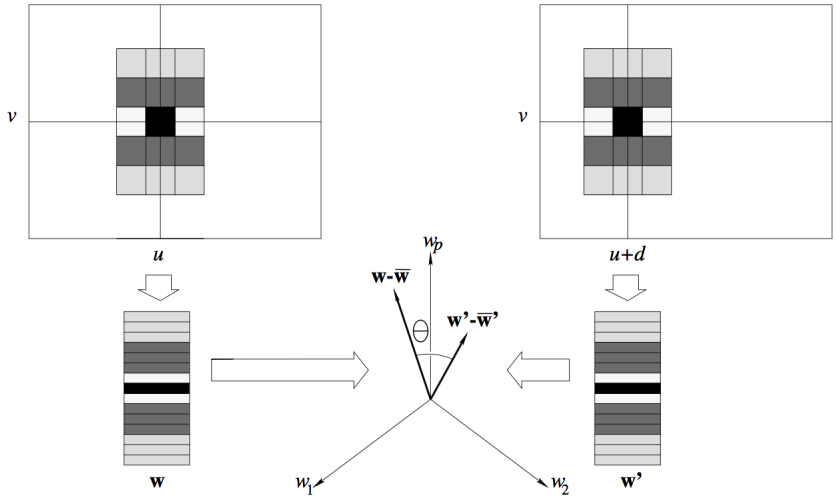

Figure 7: Correlation of two windows along an epipolar line (FP)

Given epipolar lines, several methods exist to solve for the corresponding coordinates along these lines.

10.4.1 Local Methods for Stereo Fusion

One subset of stereo fusion methods search for corresponding points using only local information around the candidate points. One such method is the Correlation Method which is applied to rectified images and compares the intensities of pixels in the neighborhood of potential coordinates.

Consider a window of size $p = (2m+1)\cdot(2n+1)$ centered around (u,v) in image 1, as shown in Figure 7. We can associate with this window the vector $w(u, v) \in \mathbb{R}^p$ which is calculated based on the intensity of the pixels in each row of the window. Let the second window be centered around (u',v'). Because image 1 and image 2 are parallel, the second window is translated from the first window by an offset d. We can calculate the vector $w'(u + d, v)$ associated with this second window and define the normalized correlation function between the two vectors as

$$C(d) = \frac{1}{||w - \bar{w}||} \frac{1}{||w' - \bar{w'}||} [(w - \bar{w}) \cdot (w' - \bar{w'})] \tag{15}$$

where the coordinates of $\bar{w}$ are all equal to the mean of the coordinates of w. The output of this normalized function C ranges from -1 to +1. Matches between corresponding coordinates can be found by finding the maximum of the C function over a range of d.

Note that finding the maximum of the C function corresponds to calculating the minimum of

$$\left| \frac{1}{||w - \bar{w}||}(w - \bar{w}) - \frac{1}{||w' - \bar{w'}||}(w' - \bar{w'}) \right|^2 \tag{16}$$

A significant problem with this correlation method is that it assumes that the observed surface is parallel to the image planes. Figure 8 demonstrates how angled surfaces make the neighborhoods around corresponding points different between the two images.

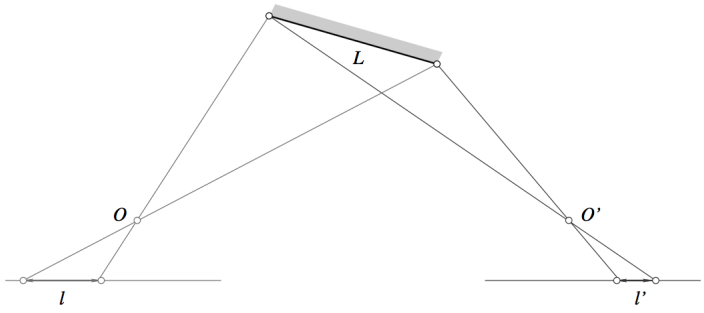

Figure 8: The foreshortening of an angled plane demonstrates that a distortion exists between the points on image 1 vs image 2.

Another method called the Multi-Scale Edge Matching method addresses the shortcomings of the correlation method by matching features, such as edges, instead of pixel intensities (see Algorithm 7.1 in [DAF11]).

10.4.2 Global Methods for Stereo Vision

A second set of methods for stereo fusion involve using non-local constraints to find correspondences.

One such constraint is the ordering constraint which assumes that the order of image points mapped from an object onto an epipolar line will always be the same, as shown in the left image of Figure 9. However, note that this constraint does not hold if a small object is in front of a larger object. In this case, the order of image points is not the same between the first and second images, as shown in the right image of Figure 9.

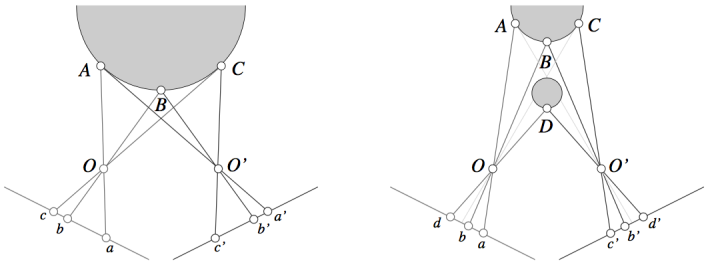

Figure 9: (left) Ordering Constraints (right) The effect of a small object in front of a larger one on the order of image points on epipolar lines.

The ordering constraints can be used with dynamic programming to create efficient algorithms to find correspondences (see Algorithm 7.2 in FP).

Another non-local constraint is the smoothness constraint for the energy function $E : \mathbb{D}^n \to \mathbb{R}$ given by

$$E(d) = \sum_{p \in V} U_p(d_p) + \sum_{(p,q) \in E)} B_{pq}(d_p, d_q) \tag{17}$$

where d is a vector containing n integer disparities d_p corresponding to pixels p and $U_p(d_p)$ calculates the discrepancy between pixel p in the first image and pixel $p + d_p$ in the second one. A pair of corresponding points can be found by minimizing this error function.

10.5 Structure From Motion (SFM)

The Structure From Motion (SFM) method uses a similar principle than Stereopsis but in a different fashion: we do not take images from two cameras at the same time, but instead we take images from one camera at different points in time as we move the camera around the object. Here, the intrinsics parameters are the same because we are using one camera, but the extrinsic changes as we move the camera.

Let's consider we have different points P_j on an object and we take the image at time k. Given m images of n fixed 3D points, we get m projection matrices M_k (for the kth image): $p^h_{j,k} = M_k P^h_j$, where $p_{j,k}$ is the projection of the point P_j on the camera image k.

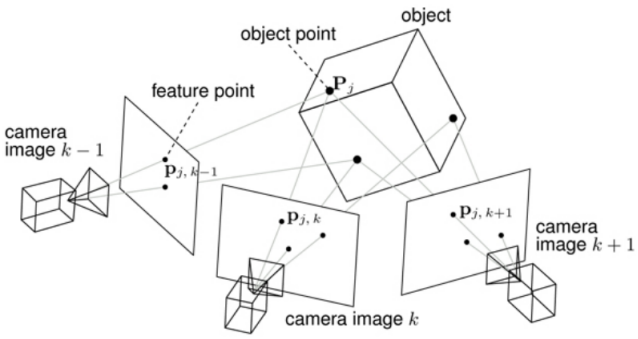

Figure 10: Structure From Motion (SFM)

However, SFM is difficult to use because it has ambiguities. That is, we cannot recover the absolute scale of the observed scene. If we have a bigger object in a longer distance in one situation and a smaller object in a closer distance, the projections will be the same (see Figure 11). To address this ambiguity, we can take several approaches such as algebraic approach (by fundamental matrix) and bundle adjustment.

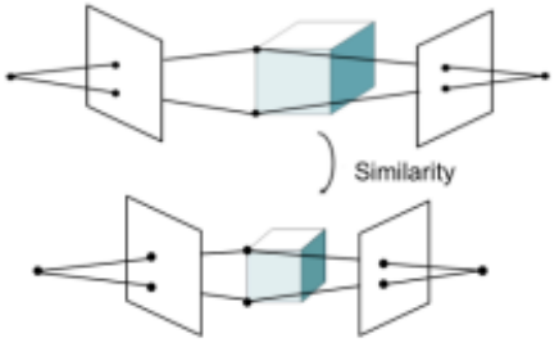

Figure 11: Ambiguity in the projection of an object using SFM

The concept of SFM is similar to visual odometry, which estimate the motion of a robot by using visual inputs in series. This is widely used in reality, including on Mars by rovers. SFM can be a very powerful method because it not only allows us to reconstruct the environment but also to recover the motion of the camera.

References

[DAF11] Jean Ponce David A. Forsyth. Geometric camera models. In *Computer Vision: A Modern Approach*. Prentice Hall, 2nd Edition, 2011.

[FTV00] Andrea Fusiello, Emanuele Trucco, and Alessandro Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, Jul 2000.

[SNS11] Roland Siegwart, Illah Reza Nourbakhsh, and Davide Scaramuzza. *Introduction to autonomous mobile robots*. MIT press, 2011.