Computer Vision
A Modern Approach

David A. Forsyth
University of California at Berkeley

Jean Ponce
University of Illinois at Urbana-Champaign

An Alan R. Apt Book

Prentice Hall
Upper Saddle River, New Jersey 07458

12

Affine Structure from Motion

This chapter revisits the problem of estimating the three-dimensional shape of a scene from multiple pictures. In the context of stereopsis, the cameras used to acquire the input images are normally calibrated so their intrinsic parameters are known and their extrinsic ones have been determined relative to some fixed world coordinate system. This greatly simplifies the reconstruction process and explains the emphasis put on the binocular (or trinocular) fusion problem in conventional stereo vision systems. We consider in this chapter a more difficult setting where the cameras' positions and possibly their intrinsic parameters are a priori unknown and may change over time. This is typical of *image-based rendering* applications, where a video clip recorded by a hand-held camcorder, possibly zooming during the shoot, is used to capture the shape of an object and render it under new viewing conditions (chapter 26). This is also relevant for active vision systems whose calibration parameters vary dynamically and planetary robot probes for which these parameters may change due to the large accelerations at take-off and landing. Recovering the cameras' positions is of course just as important as estimating the scene shape in the context of mobile robot navigation.

We ignore the correspondence problem in the rest of this chapter, assuming that the projections of n points have been matched across m pictures.[1] We focus instead on the purely geometric *structure-from-motion* problem of using image matches to estimate both the three-dimensional positions of the corresponding scene points in some fixed coordinate system (i.e., the scene *structure*) and the projection matrices associated with the cameras observing them (or, equivalently, the *motion* of the points relative to the cameras). This chapter is concerned with scenes whose relief is small compared with their overall depth relative to the cameras observing them, so perspective projection can be approximated by the simpler *affine* models of the imaging process

[1]Methods for establishing such correspondences across both continuous image sequences and scattered views of a scene are discussed in chapters 17 and 23.

introduced in chapters 1 and 2. The full perspective structure-from-motion problem is discussed in the next chapter. Concretely, given n *fixed* points P_j ($j = 1, \ldots, n$) observed by m affine cameras and the corresponding mn (nonhomogeneous) coordinate vectors $\boldsymbol{p}_{ij}$ of their images, we rewrite the affine projection Eq. (2.19) as

$$\boldsymbol{p}_{ij} = \mathcal{M}_i \begin{pmatrix} \boldsymbol{P}_j \\ 1 \end{pmatrix} = \mathcal{A}_i \boldsymbol{P}_j + \boldsymbol{b}_i \quad \text{for} \quad i = 1, \ldots, m \quad \text{and} \quad j = 1, \ldots, n, \tag{12.1}$$

and define *affine structure from motion* as the problem of estimating the m 2×4 matrices $\mathcal{M}_i = (\mathcal{A}_i \quad \boldsymbol{b}_i)$ and the n positions $\boldsymbol{P}_j$ of the points P_j in some fixed coordinate system from the mn image correspondences $\boldsymbol{p}_{ij}$.

When the projection matrices $\mathcal{M}_i$ are allowed to take an arbitrary form (i.e., when the intrinsic and extrinsic parameters of the cameras are unknown, see chapter 2), Eq. (12.1) provides $2mn$ constraints on the $8m + 3n$ unknown coefficients defining the matrices $\mathcal{M}_i$ and the point positions $\boldsymbol{P}_j$. Since $2mn$ is greater than $8m + 3n$ for large enough values of m and n, it is thus clear that a sufficient number of views of a sufficient number of points allows the recovery of the corresponding structure and motion parameters via, say, the least-squares techniques presented in chapter 3. However, it is important to understand that, if $\mathcal{M}_i$ and $\boldsymbol{P}_j$ are solutions of Eq. (12.1), so are $\mathcal{M}'_i$ and $\boldsymbol{P}'_j$, where

$$\mathcal{M}'_i = \mathcal{M}_i \mathcal{Q} \quad \text{and} \quad \begin{pmatrix} \boldsymbol{P}'_j \\ 1 \end{pmatrix} = \mathcal{Q}^{-1} \begin{pmatrix} \boldsymbol{P}_j \\ 1 \end{pmatrix} \tag{12.2}$$

and $\mathcal{Q}$ is an arbitrary *affine transformation* matrix—that is, it can be written (see chapter 2 and next section) as

$$\mathcal{Q} = \begin{pmatrix} \mathcal{C} & \boldsymbol{d} \\ \boldsymbol{0}^T & 1 \end{pmatrix} \quad \text{with} \quad \mathcal{Q}^{-1} = \begin{pmatrix} \mathcal{C}^{-1} & -\mathcal{C}^{-1}\boldsymbol{d} \\ \boldsymbol{0}^T & 1 \end{pmatrix}, \tag{12.3}$$

where $\mathcal{C}$ is a nonsingular 3×3 matrix and $\boldsymbol{d}$ is a vector in $\mathbb{R}^3$. In other words, any solution of the affine structure-from-motion problem can *only be defined up to an affine transformation ambiguity*. Taking into account the 12 parameters defining a general affine transformation, we should thus expect a finite number of solutions as soon as $2mn \geq 8m + 3n - 12$. For $m = 2$, this suggests that four point correspondences should be sufficient to determine (up to an affine transformation) the two projection matrices and the three-dimensional position of any other point. This is confirmed formally in Section 12.2.

When the intrinsic parameters of the cameras are known so the corresponding calibration matrices can be taken equal to the identity, the parameters of the projection matrices $\mathcal{M}_i = (\mathcal{A}_i \quad \boldsymbol{b}_i)$ must obey additional constraints. For example, according to Eq. (2.20), the matrix $\mathcal{A}_i$ associated with a (calibrated) weak-perspective camera is formed by the first two rows of a rotation matrix, scaled by the inverse of the depth of the corresponding reference point. As shown in Section 12.4, constraints such as these can be used to eliminate the affine ambiguity when enough images are available. This suggests decomposing the solution of the affine structure-from-motion problem into two steps: (a) first use at least two views of the scene to construct a unique (up to an arbitrary affine transformation) three-dimensional representation of the scene, called its *affine shape*; then (b) use additional views and the constraints associated with known camera calibration parameters and specific affine models to uniquely determine the rigid Euclidean structure of the scene. The first stage of this approach yields the essential part of the solution: The affine shape is a full-fledged three-dimensional representation of the scene, which, as shown in chapter 26, can be used in its own right to synthesize new views of the scene. The second step simply amounts to finding a *Euclidean upgrade* of the scene (i.e., to computing a single affine transformation that account for its rigidity and map its affine shape onto a Euclidean one).

Using three or more images overconstrains the structure-from-motion problem and leads to more robust least-squares solutions. Accordingly, a significant portion of this chapter is devoted to the problem of recovering the affine shape of a scene from several (possibly many) pictures. We conclude with techniques for segmenting a set of data points into objects undergoing different motions.

12.1 ELEMENTS OF AFFINE GEOMETRY

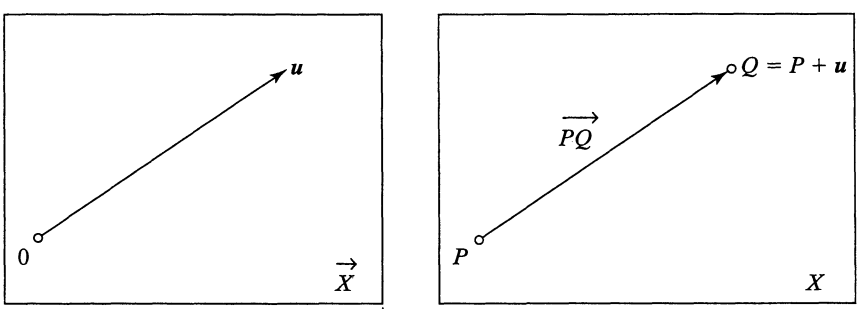
Let us start by introducing some elementary notions of affine geometry. The corresponding geometric and algebraic tools allow us to state and prove the fundamental properties of affine projection models. They also serve as building blocks for the structure-from-motion algorithms introduced in the rest of this chapter.

As noted in Snapper and Troyer (1989), affine geometry is, roughly speaking, what is left after all ability to measure lengths, areas, and angles has been removed from Euclidean geometry. The concept of parallelism remains, however, as well as the ability to measure the ratio of distances between collinear points. Giving a rigorous axiomatic introduction to affine geometry would be out of place here. Instead, we remain quite informal and recall the basic facts about real affine spaces that are necessary to understand the rest of this chapter. The reader familiar with notions such as barycentric combinations, affine coordinate systems, and affine transformations may safely proceed to the next section.

12.1.1 Affine Spaces and Barycentric Combinations

A *real affine space* is a set X of *points*, together with a real vector space $\vec{X}$, and an *action* ϕ of the additive group of $\vec{X}$ on X. The vector space $\vec{X}$ is said to *underlie* the affine space X. Informally, the action of a group on a set maps the elements of this group onto bijections of the set. Here, the action ϕ associates with every vector $\boldsymbol{u} \in \vec{X}$ a bijection $\phi_{\boldsymbol{u}} : X \to X$ such that, for any $\boldsymbol{u}, \boldsymbol{v}$ in $\vec{X}$ and any point P in X, $\phi_{\boldsymbol{u}+\boldsymbol{v}}(P) = \phi_{\boldsymbol{u}} \circ \phi_{\boldsymbol{v}}(P)$, $\phi_{\boldsymbol{0}}(P) = P$, and for any pair of points P, Q in X, there exists a unique vector $\boldsymbol{u}$ in $\vec{X}$ such that $\phi_{\boldsymbol{u}}(P) = Q$. These definitions may sound a bit abstract, so let us give some concrete examples. A familiar affine space is, of course, $\mathbb{E}^3$, where X is the set of physical points and $\vec{X}$ is the set of translations of X onto itself. Another affine space can be constructed by choosing both X and $\vec{X}$ to be equal to $\mathbb{R}^n$, with the action ϕ defined by $\phi_{\boldsymbol{u}}(P) = P + \boldsymbol{u}$, where P and $\boldsymbol{u}$ are both elements of $\mathbb{R}^n$ and "+" denotes the addition in that vector space.

Example 12.1 **$\mathbb{R}^2$ as an affine plane.**



The vector space $\mathbb{R}^2$ can be considered as an affine space by choosing $X = \vec{X} = \mathbb{R}^2$. Given $P = (x, y)^T$ and $\boldsymbol{u} = (a, b)^T$, we define $\phi_{\boldsymbol{u}}(P) \stackrel{\text{def}}{=} P + \boldsymbol{u} = (x + a, y + b)^T$. Given $P = (x, y)^T$ and

$Q = (x', y')^T$, the unique vector $\boldsymbol{u}$ such that $P + \boldsymbol{u} = Q$ is of course $\boldsymbol{u} = Q - P = \overrightarrow{PQ} \stackrel{\text{def}}{=} (x' - x, y' - y)^T$.

Following this example, we denote from now on the point $\phi_{\boldsymbol{u}}(P)$ by $P + \boldsymbol{u}$ and the vector $\boldsymbol{u}$ such that $\phi_{\boldsymbol{u}}(P) = Q$ by $\overrightarrow{PQ}$ or, equivalently, by $Q - P$. This is justified by the fact that choosing a point O as the *origin* of X allows us to identify every other point P with the vector $\boldsymbol{u} = \overrightarrow{OP}$ such that $\phi_{\boldsymbol{u}}(O) = P$. Indeed,

$$Q = P + \overrightarrow{PQ} \Longleftrightarrow \overrightarrow{OQ} = \overrightarrow{OP} + \overrightarrow{PQ} \quad \text{and} \quad Q - P = \overrightarrow{PQ} \Longleftrightarrow \overrightarrow{OQ} - \overrightarrow{OP} = \overrightarrow{PQ}.$$

The introduction of an origin is often useful for beginners who want to keep their affine notation straight. It should be absolutely clear, however, that the point $P + \boldsymbol{u}$ and the vector $\overrightarrow{PQ} = Q - P$ are *totally independent* of the choice of any origin whatsoever. Likewise, the symbols "+" and "−" in these expressions are used for notational convenience and *do not* convey their usual meaning of addition and subtraction in the additive group of a vector space.

Although it is possible to "add" a vector to a point and to "subtract" two points, it is not possible to "add" two points or to "multiply" a point by a scalar (see Exercises). However, a restricted kind of "linear combination" of points can be defined: Consider $m+1$ points $A_0, A_1 \ldots, A_m$ and $m + 1$ weights $\alpha_0, \alpha_1, \ldots, \alpha_m$ such that $\alpha_0 + \alpha_1 + \cdots + \alpha_m = 1$; the corresponding *barycentric combination* of the points A_0 to A_m is the point

$$\sum_{i=0}^{m} \alpha_i A_i \stackrel{\text{def}}{=} A_j + \sum_{i=0, i \neq j}^{m} \alpha_i (A_i - A_j), \tag{12.4}$$

where j is an integer between 0 and m. The right-hand side of this equation defines a point by adding a vector (a linear combination of the vectors $A_i - A_j$) to a point (A_j). It is easily shown that this definition is *independent* of the value of j (see Exercises), which justifies the symmetrical role played by the points A_i ($i = 0, \ldots, m$) in the notation $\sum_{i=0}^{m} \alpha_i A_i$. This notation is further justified by introducing an origin O and noting that $\sum_{i=0}^{m} \alpha_i \overrightarrow{OA_i} = \overrightarrow{OA_j} + \sum_{i=0, i \neq j}^{m} \alpha_i (\overrightarrow{OA_i} - \overrightarrow{OA_j})$ when $\alpha_0 + \alpha_1 + \cdots + \alpha_m = 1$. However, the definition of barycentric combinations by Eq. (12.4) is preferable since it is obviously independent of any choice of origin.

A familiar example of barycentric combination is the *center of mass* of $m + 1$ points, corresponding to the case where all weights are equal to $1/(m + 1)$. Any other set of weight values adding to 1 yields a valid barycentric combination.

12.1.2 Affine Subspaces and Affine Coordinates

An *affine subspace* of X is defined by a point O and a vector subspace U of $\vec{X}$ as the set of points $O + U \stackrel{\text{def}}{=} \{O + \boldsymbol{u}, \boldsymbol{u} \in U\}$. Its *dimension* is the dimension of the associated vector subspace. Two affine subspaces $O' + U'$ and $O'' + U''$, such that U' is a subspace of U'', or U'' is a subspace of U' are said to be parallel. Affine subspaces of dimension 1 and 2 are, respectively, called *lines* and *planes*. When $\vec{X}$ is of finite dimension n, its affine subspaces of dimension $n - 1$ are called *hyperplanes*. Affine lines, planes, and hyperplanes take their usual meaning in the affine spaces associated with physical three-dimensional space and $\mathbb{R}^n$.

Example 12.2 The intersection of two affine subspaces is either empty or an affine subspace.

Consider two subspaces $Y' = O' + U'$ and $Y'' = O'' + U''$ of some affine space X, and denote by Z their intersection. Let P_0 denote some point in Z. We have by definition $P_0 = O' + \boldsymbol{u}'_0 = O'' + \boldsymbol{u}''_0$ for some vectors $\boldsymbol{u}'_0$ in U' and $\boldsymbol{u}''_0$ in U''. Likewise, given any other point P in Z, we can write $P = O' + \boldsymbol{u}' = O'' + \boldsymbol{u}''$ for some vectors $\boldsymbol{u}'$ in U' and $\boldsymbol{u}''$ in U''. In particular, we must have

$$P = P_0 + \boldsymbol{u}' - \boldsymbol{u}'_0 = P_0 + \boldsymbol{u}'' - \boldsymbol{u}''_0,$$

which implies that (a) $u' - u'_0 = u'' - u''_0$ is an element of $U' \cap U''$, and (b) P is an element of $P_0 + U' \cap U''$. Conversely, any point P in $P_0 + U' \cap U''$ can be written as $P = P_0 + u$ for some vector u in $U' \cap U''$; thus,

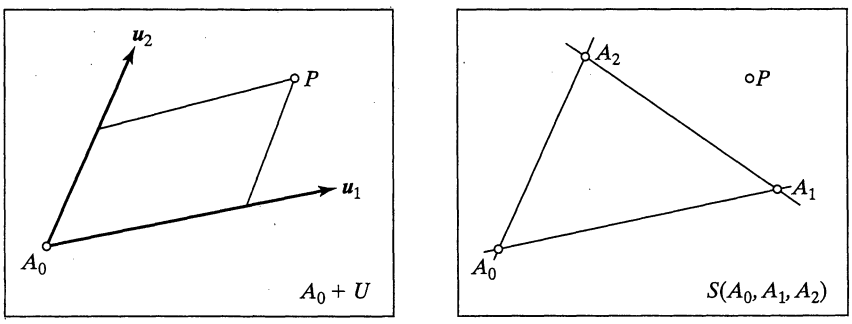
$$P = O' + (P_0 - O') + u = O' + u' + u = O'' + (P_0 - O'') + u = O'' + u'' + u,$$

which implies that P is an element of Z. We finally conclude that $Z = P_0 + U' \cap U''$. Note that the intersection of two affine spaces may be empty. For example two parallel lines do not intersect. Neither do two skew lines in space of course, although they are not parallel to each other.

Affine subspaces can also be defined purely in terms of points: Let $S(A_0, A_1 \ldots, A_m)$ denote the set of all barycentric combinations of $m+1$ points $A_0, A_1, \ldots, A_m$. It is easy to verify that $S(A_0, A_1, \ldots, A_m)$ is indeed an affine subspace (see Exercises), and that its dimension is at most m (e.g., two distinct points define a line, three points define [in general] a plane, etc.). We say that $m+1$ points are independent if they do not lie in a subspace of dimension at most $m-1$, so $m+1$ independent points define (or *span*) an m-dimensional subspace.

Example 12.3 Two complementary definitions of an affine plane.

Consider three noncollinear points A_0, A_1, and A_2 in $\mathbb{R}^3$ viewed as an affine space. These points define the plane $\Pi = A_0 + U$ of $\mathbb{R}^3$ associated with the point A_0 and the vector plane U spanned by the two vectors $u_1 = \overrightarrow{A_0A_1}$ and $u_2 = \overrightarrow{A_0A_2}$.



Equivalently, the plane Π can be viewed as the affine subspace $S(A_0, A_1, A_2)$ of $\mathbb{R}^3$, and any point P in Π can be represented as a barycentric combination of the points A_0, A_1, and A_2.

An *affine coordinate system* for $O + U$ consists of a point A_0 (called its *origin*) in $O + U$ and a coordinate system $(u_1, \ldots, u_m)$ for U. The *affine coordinates* of a point P in $O + U$ are defined as the coordinates of the vector $\overrightarrow{A_0P}$ in the coordinate system $(u_1, \ldots, u_m)$. It is crucial to understand that the (Euclidean) point coordinates used in chapter 2 and in conventional Euclidean geometry are just affine coordinates. The vectors u_i $(i = 1, \ldots, m)$ used to define the corresponding coordinate systems simply have the additional property of having unit length and being orthogonal to each other. This property is not required for general affine coordinate systems, and indeed the notions of lengths and angles may not be defined in general affine spaces.

Example 12.4 Affine coordinate changes.

Given some coordinate system $(F) = (O, u, v, w)$ for the affine space $\mathbb{E}^3$ and a point P of $\mathbb{E}^3$ such that $\overrightarrow{OP} = xu + yv + zw$, we can define, using the same notation as in chapter 2, the (affine) coordinate vector of P as ${}^F P = (x, y, z)^T$.

Given two affine coordinate systems $(A) = (O_A, u_A, v_A, w_A)$ and $(B) = (O_B, u_B, v_B, w_B)$ for the affine space $\mathbb{E}^3$, let us define the 3×3 matrix

$$\overset{B}{A}C = \begin{pmatrix} {}^B u_A & {}^B v_A & {}^B w_A \end{pmatrix},$$

where ${}^B\boldsymbol{a}$ denotes the coordinate vector of the vector $\boldsymbol{a}$ in the (vector) coordinate system $(\boldsymbol{u}_A, \boldsymbol{v}_A, \boldsymbol{w}_A)$. It is easy to show that ${}^BP = {}^B_AC\,{}^AP + {}^BO_A$, or, in (affine) homogeneous coordinates,

$$\begin{pmatrix} {}^BP \\ 1 \end{pmatrix} = {}^B_A\mathcal{T}\begin{pmatrix} {}^AP \\ 1 \end{pmatrix}, \quad \text{where} \quad {}^B_A\mathcal{T} = \begin{pmatrix} {}^B_AC & {}^BO_A \\ \mathbf{0}^T & 1 \end{pmatrix}.$$

Note the obvious similarity with the formula for a change of Euclidean coordinate system in chapter 2. Here, however, the basis vectors of the two coordinate frames do not form orthonormal bases, so ${}^B_AC$ is an ordinary nonsingular 3×3 matrix instead of a rotation matrix, and ${}^B_A\mathcal{T}$ is an affine transformation matrix.

An alternative way of defining a coordinate system for an n-dimensional affine space X is to pick $n+1$ independent points $A_0, A_1 \ldots, A_n$ in X. The *barycentric coordinates* α_i $(i = 0, 1, \ldots, n)$ of a point P in Y are uniquely defined by $P = \alpha_0 A_0 + \alpha_1 A_1 + \cdots + \alpha_n A_n$. They are related to affine coordinates in a simple way: Choosing $j = 0$ in Eq. (12.4) yields

$$P = \alpha_0 A_0 + \alpha_1 A_1 + \cdots + \alpha_n A_n = A_0 + \alpha_1(A_1 - A_0) + \cdots + \alpha_n(A_n - A_0),$$

showing that the affine coordinates of P in the basis formed by the points A_i $(i = 0, 1, \ldots, m)$ are $\alpha_1, \ldots, \alpha_m$.

When an n-dimensional affine space X has been equipped with an affine basis, a necessary and sufficient condition for $m+1$ points A_i to define a p-dimensional affine subspace of X (with $m \geq p$ and $n \geq p$) is for the $(n+1) \times (m+1)$ matrix

$$\mathcal{D} = \begin{pmatrix} x_{01} & x_{11} & \ldots & x_{m1} \\ \ldots & \ldots & \ldots & \ldots \\ x_{0n} & x_{1n} & \ldots & x_{mn} \\ 1 & 1 & \ldots & 1 \end{pmatrix}$$

formed by their coordinate vectors $(x_{i1}, \ldots, x_{in})^T$ $(i = 0, 1, \ldots, m)$ to have rank $p+1$. Indeed, a rank lower than $p+1$ means that any column of this matrix is a barycentric combination of at most p of its columns, and a rank higher than $p+1$ implies that at least $p+2$ of the points are independent.

Example 12.5 The equation of a line in the plane.

Consider three points A_0, A_1, and A_2 in an affine plane, with coordinate vectors $(x_0, y_0)^T$, $(x_1, y_1)^T$, and $(x_2, y_2)^T$ in some basis of this plane. According to the previous paragraph, a necessary and sufficient condition for these points to lie in an affine subspace of dimension 1 (i.e., be collinear) is that the rank of the matrix

$$\mathcal{D} = \begin{pmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ 1 & 1 & 1 \end{pmatrix}$$

be equal to 2, or, equivalently, that its determinant be equal to zero. Note that

$$\text{Det}(\mathcal{D}) = x_1y_2 - x_2y_1 + x_2y_0 - x_0y_2 + x_0y_1 - x_1y_0 = \begin{pmatrix} x_1 - x_0 \\ y_1 - y_0 \end{pmatrix} \times \begin{pmatrix} x_2 - x_0 \\ y_2 - y_0 \end{pmatrix},$$

where "$\times$" denotes here the operator that associates with two vectors in $\mathbb{R}^2$ the determinants of their coordinates. Thus $\text{Det}(\mathcal{D}) = 0$ is indeed equivalent to $\overrightarrow{A_0A_1}$ and $\overrightarrow{A_0A_2}$ being parallel or to the three points being collinear. When the points A_0 and A_1 are fixed, $\text{Det}(\mathcal{D}) = 0$ can be seen as an equation defining the line passing through A_0 and A_1 in terms of the coordinates of A_2, and it has of course the form $ax_2 + by_2 + c = 0$. This method can be generalized to affine subspaces defined by arbitrary numbers of points: The corresponding equations are simply obtained by writing that the appropriate minors of the matrix $\mathcal{D}$ have zero determinants.

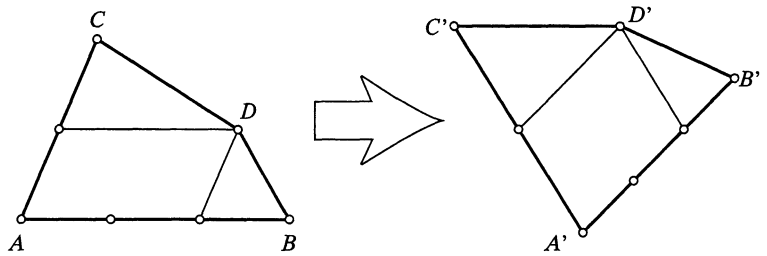


Figure 12.1 An affine transformation of the plane. The points A, B, C, and D are transformed into the points A', B', C', and D'. The affine coordinates of D in the basis of the plane formed by A, B, and C are the same as those of D' in the basis formed by A', B', and C'—namely 2/3 and 1/2.

12.1.3 Affine Transformations and Affine Projection Models

An *affine transformation* between two affine spaces X and Y is a bijection from X onto Y that maps m-dimensional subspaces of X onto m-dimensional subspaces of Y, maps parallel subspaces onto parallel subspaces, and preserves barycentric combinations (or, equivalently, affine coordinates; Figure 12.1). It can be shown that affine transformations can also be characterized by the (seemingly weaker) property of mapping lines onto lines and preserving the *ratio of the signed lengths of parallel line segments.*

An affine transformation between two affine spaces X and Y of dimension m is completely defined by the images $B_0, \ldots, B_m$ of $m+1$ independent points $A_0, \ldots, A_m$. Indeed, the image of any other point with affine coordinates α_i $(i = 0, \ldots, m)$ in the basis of X formed by the points A_i have the same coordinates in the basis of Y formed by the points B_i. Conversely, it can be shown that given any independent points $B_0, \ldots, B_m$ in Y, there is a unique affine transformation mapping the points A_i onto the points B_i. It is thus clear that affine transformations do not preserve angles or distances—a fact confirmed by Figure 12.1. In fact, it can also be shown that affine transformations of $\mathbb{R}^3$ can always be written as the combination of a translation, rotation, nonuniform scaling, and shear.

The relationship between vector and affine spaces induces a relationship between linear and affine transformations. In particular, it is easy to show (see Exercises) that an affine transformation $\psi : X \to Y$ between two affine subspaces X and Y associated with the vector spaces $\vec{X}$ and $\vec{Y}$ can be written as

$$\psi(P) = \psi(O) + \vec{\psi}(P - O),$$

where O is some arbitrarily chosen origin, and $\vec{\psi} : \vec{X} \to \vec{Y}$ is a linear mapping from $\vec{X}$ onto $\vec{Y}$ that is independent of the choice of O. When X and Y are of (finite) dimension m and an affine coordinate system with origin O is chosen, this yields the familiar expression

$$\psi(\boldsymbol{P}) = \boldsymbol{d} + \boldsymbol{C}\boldsymbol{P} = \boldsymbol{C}\boldsymbol{P} + \boldsymbol{d},$$

where $\boldsymbol{P}$ denotes the coordinate vector of P in the chosen basis, $\boldsymbol{d}$ denotes the coordinate vector of $\psi(O)$, and $\boldsymbol{C}$ is the $m \times m$ matrix representing $\vec{\psi}$ in the same coordinate system. Thus, affine transformations as defined in chapter 2 are indeed affine transformations as defined in this chapter.

A fundamental property of parallel projections is that they induce affine transformations from planes onto their images. Let us first show that they preserve the ratio of signed distances between collinear points: The triangles OAa, OBb, and OCc in Figure 12.2(left) are similar,

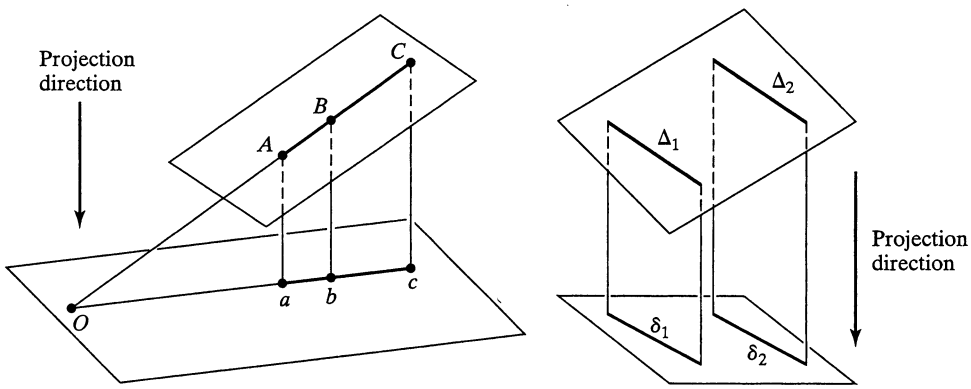


Figure 12.2 Parallel projection preserves: (left) the ratio of signed distances between collinear points and (right) the parallelism of lines.

and it follows that $\overline{AB}/\overline{BC} = \overline{ab}/\overline{bc}$ for any orientation of the lines OC and Oc. To show that parallel projections preserve the parallelism of lines, we use the fact that the intersection of a plane with two parallel planes consists of two parallel lines (see Exercises). Now consider the situation depicted in Figure 12.2(right), where two parallel lines Δ_1 and Δ_2 are projected onto a plane. The planes defined respectively by these two lines and the parallel projection direction are parallel to each other and therefore intersect the image plane along two parallel lines δ_1 and δ_2.

Weak- and paraperspective projections from one plane onto another are also affine transformations. This follows immediately from the fact that they can always be written as the composition of a parallel projection and an affine transformation of the image plane that compounds the effects of the inverse-depth scaling and intrinsic camera parameters. As shown by Theorem 2 in chapter 2, a general affine projection can always be written as a weak-perspective one, thus affine projections from one plane onto another are indeed affine transformations.

It follows immediately that affine projections preserve parallel lines and barycentric combinations. In particular, the center of mass of a set of scene points projects onto the center of mass of their images (which gives a simple method for selecting the reference point of a paraperspective camera; see chapter 2), and the ratio of signed distances between collinear points is an affine-projection invariant (which is useful in the object recognition context; see chapter 23, for example).

12.1.4 Affine Shape

We say that two (possibly infinite) point sets S and S' in some affine space X are *affinely equivalent* when there exists an affine transformation $\psi : X \rightarrow X$, such that S' is the image of S under ψ. It is easy to show that affine equivalence is an equivalence relation, and we define the *affine shape* of a point set S in X as the equivalence class of all affinely equivalent point sets. Affine structure from motion can thus be seen as the problem of recovering the affine shape of the observed scene (and/or the equivalence classes formed by the corresponding projection matrices) from features matched in an image sequence. We now have the right tools for solving this problem.

12.2 AFFINE STRUCTURE AND MOTION FROM TWO IMAGES

Let us start with the case where two affine images of the same scene are available (the case of multiple pictures is addressed in the following section). The two structure-from-motion tech-

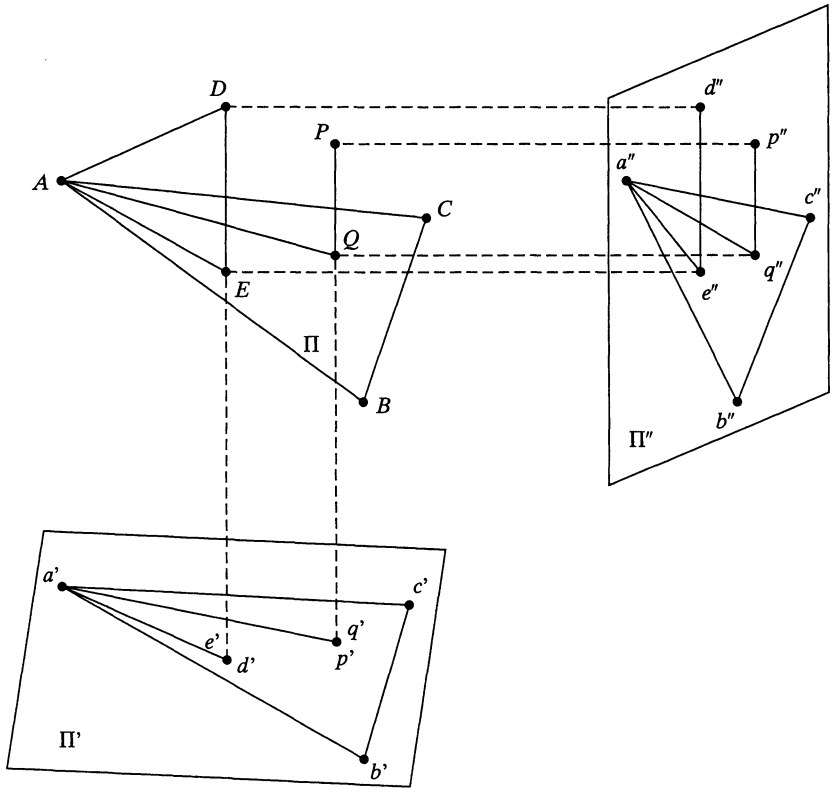


Figure 12.3 Geometric construction of the affine coordinates of a point P in the basis formed by the four points A, B, C, and D. This diagram illustrates the parallel projection case, but the reasoning used in this section is valid in the general affine setting.

niques discussed in this section are complementary: The first one uses geometric reasoning to uncover the affine shape of the scene (from which the projection matrices can be found if needed), whereas the second one uses simple algebraic manipulations to estimate the projection matrices (from which the positions of the scene points are easily calculated).

12.2.1 Geometric Scene Reconstruction

We already mentioned that two affine views of four points A, B, C, D should be sufficient to compute the affine coordinates of any other point P in the basis (A, B, C, D). This is indeed the case, and we now present the constructive proof from (Koenderink and Van Doorn, 1990). Remember that the affine projection of a plane onto another plane is an affine transformation. In particular, when the point P belongs to the plane Π that contains the triangle ABC, its affine coordinates in the basis of Π formed by these three points can be directly measured in either of the two images. Now let E (resp. Q) denote the intersection of the line passing through the points D and d' (resp. P and p') with the plane Π (Figure 12.3). The projections e'' and q'' of the points E and Q onto the plane Π'' have the same affine coordinates in the basis (a'', b'', c'') as the points d' and p' in the basis (a', b', c').

In addition, since the two segments ED and QP are parallel to the first projection direction, the two line segments $e''d''$ and $q''p''$ are also parallel, and we can measure the ratio

$$\lambda = \frac{\overline{q''p''}}{\overline{e''d''}} = \frac{\overline{QP}}{\overline{ED}},$$

where $\overline{AB}$ denotes the signed distance between the two points A and B for some arbitrary (but fixed) orientation of the line joining these points.

If we now denote by $(\alpha_{d'}, \beta_{d'})$ and $(\alpha_{p'}, \beta_{p'})$ the coordinates of the points $d' = e'$ and $p' = q'$ in the basis (a', b', c'), we can write

$$\begin{aligned}\overrightarrow{AP} &= \overrightarrow{AQ} + \overrightarrow{QP} = \alpha_{p'}\overrightarrow{AB} + \beta_{p'}\overrightarrow{AC} + \lambda\overrightarrow{ED} \\ &= (\alpha_{p'} - \lambda\alpha_{d'})\overrightarrow{AB} + (\beta_{p'} - \lambda\beta_{d'})\overrightarrow{AC} + \lambda\overrightarrow{AD}.\end{aligned}$$

In other words, the affine coordinates of P in the (A, B, C, D) basis are $(\alpha_{p'} - \lambda\alpha_{d'}, \beta_{p'} - \lambda\beta_{d'}, \lambda)$. This is the *affine structure-from-motion theorem*: Given two affine views of four noncoplanar points, the affine shape of the scene is uniquely determined (Koenderink and Van Doorn, 1990). Figure 12.4 shows three projections of the synthetic face used in Koenderink and Van Doorn's experiments, along with an affine profile view computed from two of the images.

12.2.2 Algebraic Motion Estimation

Let us now explore a completely different approach, where geometric insight is somewhat neglected in favor of simple algebraic manipulations that exploit the affine ambiguity of structure from motion to simplify the form of the projection matrices. The outcome is an extremely simple technique for recovering these matrices and the corresponding affine shape.

Let us start by introducing the affine equivalent of the epipolar constraint. We consider two affine images and rewrite the corresponding projection equations

$$\begin{cases} \boldsymbol{p} = \mathcal{A}\boldsymbol{P} + \boldsymbol{b} \\ \boldsymbol{p}' = \mathcal{A}'\boldsymbol{P} + \boldsymbol{b}' \end{cases} \quad \text{as} \quad \begin{pmatrix} \mathcal{A} & \boldsymbol{p} - \boldsymbol{b} \\ \mathcal{A}' & \boldsymbol{p}' - \boldsymbol{b}' \end{pmatrix} \begin{pmatrix} \boldsymbol{P} \\ -1 \end{pmatrix} = \boldsymbol{0},$$

and a necessary and sufficient condition for these equations to admit a nontrivial solution is that

$$\text{Det} \begin{pmatrix} \mathcal{A} & \boldsymbol{p} - \boldsymbol{b} \\ \mathcal{A}' & \boldsymbol{p}' - \boldsymbol{b}' \end{pmatrix} = 0,$$

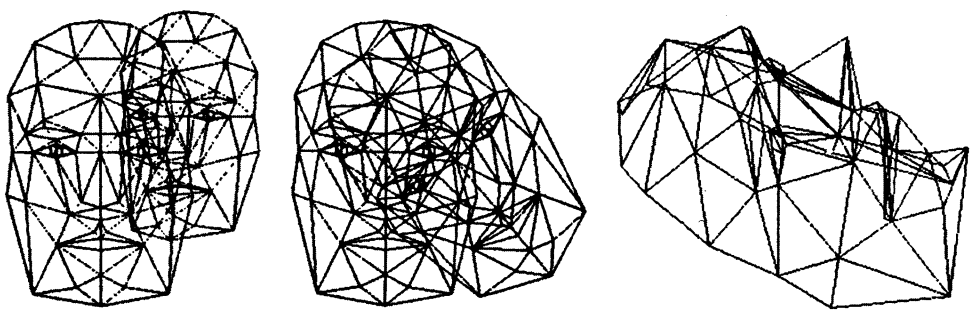


Figure 12.4 Affine reconstruction from two views. Left and middle: three views of a face; Images 0 and 1 are overlaid on the left, and Images 1 and 2 are overlaid in the middle part of the figure. Right: A profile view of the affine face computed from Images 0 and 1 (the third picture is used in Section 12.4 to turn this affine reconstruction into a Euclidean one). *Reprinted with permission from "Affine Structure from Motion," by J.J. Koenderink and A.J. Van Doorn, Journal of the Optical Society of America A, 8:377–385, (1990). © 1990 Optical Society of America.*

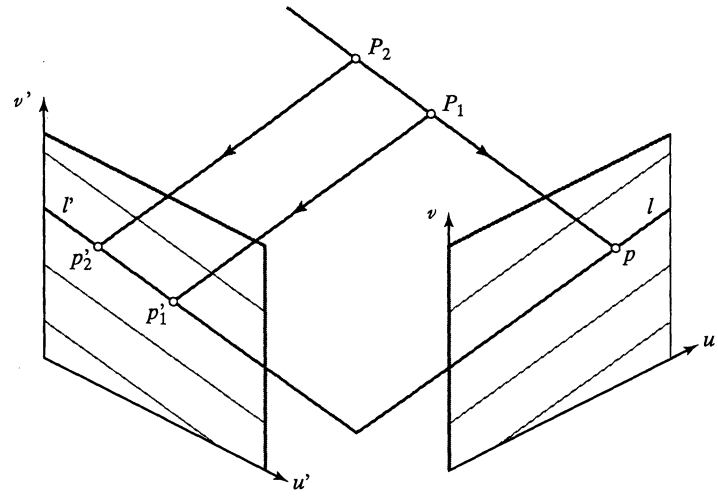


Figure 12.5 Affine epipolar geometry: Given two parallel-projection images, a point p in the first image and the two projection directions define an epipolar plane that intersects the second image along the epipolar line l'. As in the perspective case, any match p' for p is constrained to belong to this line.

or

$$\alpha u + \beta v + \alpha' u' + \beta' v' + \delta = 0, \tag{12.5}$$

where α, β, α', β', and δ are constants depending on $\mathcal{A}$, $\boldsymbol{b}$, $\mathcal{A}'$, and $\boldsymbol{b}'$. This is the *affine epipolar constraint*. Indeed, given a point p in the first image, the position of the matching point p' is constrained by Eq. (12.5) to lie on the line l' defined by $\alpha' u' + \beta' v' + \gamma' = 0$, where $\gamma' = \alpha u + \beta v + \delta$ and vice versa (Figure 12.5).

Note that the epipolar lines associated with each image are parallel to each other: For example, moving p changes γ' or, equivalently, the distance from the origin to the epipolar line l', but does not modify the direction of l'.

The affine epipolar constraint can be rewritten in the familiar form

$$(u, v, 1)\mathcal{F}\begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} = 0, \quad \text{where} \quad \mathcal{F} \stackrel{\text{def}}{=} \begin{pmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \\ \alpha' & \beta' & \delta \end{pmatrix}$$

is the *affine fundamental matrix*. This suggests that the affine epipolar geometry can be seen as the limit of the perspective one. Indeed, it can be shown that an affine picture is the limit of a sequence of images taken by a perspective camera that zooms in on the scene as it backs away from it (see Exercises for details).

Let us now show that the projection matrices can be estimated from the epipolar constraint. The inherent affine ambiguity of affine structure from motion actually allows us to simplify the calculations: According to Eqs. (12.2) and (12.3), if $\mathcal{M} = (\mathcal{A} \quad \boldsymbol{b})$ and $\mathcal{M}' = (\mathcal{A}' \quad \boldsymbol{b}')$ are solutions of our problem, so are $\tilde{\mathcal{M}} = \mathcal{M}\mathcal{Q}$ and $\tilde{\mathcal{M}}' = \mathcal{M}'\mathcal{Q}$, where

$$\mathcal{Q} = \begin{pmatrix} \mathcal{C} & \boldsymbol{d} \\ \boldsymbol{0}^T & 1 \end{pmatrix}$$

is an arbitrary affine transformation. The new projection matrices can be written as $\tilde{\mathcal{M}} = (\mathcal{A}\mathcal{C} \quad \mathcal{A}\boldsymbol{d} + \boldsymbol{b})$ and $\tilde{\mathcal{M}}' = (\mathcal{A}'\mathcal{C} \quad \mathcal{A}'\boldsymbol{d} + \boldsymbol{b}')$. Note that, according to Eq. (12.3), applying this

transformation to the projection matrices amounts to applying the inverse transformation to every scene point P, whose position $\boldsymbol{P}$ is replaced by $\tilde{\boldsymbol{P}} = C^{-1}(\boldsymbol{P} - \boldsymbol{d})$.

Now let us denote by $\boldsymbol{a}_1^T$ and $\boldsymbol{a}_2^T$ (resp. $\boldsymbol{a}_1'^T$ and $\boldsymbol{a}_2'^T$) the two rows of $\mathcal{A}$ (resp. $\mathcal{A}'$) and introduce the vectors $\boldsymbol{b} = (b_1, b_2)^T$ and $\boldsymbol{b}' = (b_1', b_2')^T$. We can rewrite the epipolar constraint as

$$0 = \text{Det}\begin{pmatrix} \mathcal{A}C & \boldsymbol{p} - \mathcal{A}\boldsymbol{d} - \boldsymbol{b} \\ \mathcal{A}'C & \boldsymbol{p}' - \mathcal{A}'\boldsymbol{d} - \boldsymbol{b}' \end{pmatrix} = \text{Det}\left(\begin{array}{cc|c} \multicolumn{2}{c|}{\boldsymbol{a}_1^T C} & u - \boldsymbol{a}_1^T\boldsymbol{d} - b_1 \\ \multicolumn{2}{c|}{\boldsymbol{a}_2^T C} & v - \boldsymbol{a}_2^T\boldsymbol{d} - b_2 \\ \hline \multicolumn{2}{c|}{\boldsymbol{a}_1'^T C} & u' - \boldsymbol{a}_1'^T\boldsymbol{d} - b_1' \\ \multicolumn{2}{c|}{\boldsymbol{a}_2'^T C} & v' - \boldsymbol{a}_2'^T\boldsymbol{d} - b_2' \end{array}\right)$$

$$= \text{Det}\left(\begin{array}{cc|c} \multicolumn{2}{c|}{\boldsymbol{a}_1^T C} & u - \boldsymbol{a}_1^T\boldsymbol{d} - b_1 \\ \multicolumn{2}{c|}{\boldsymbol{a}_2^T C} & v - \boldsymbol{a}_2^T\boldsymbol{d} - b_2 \\ \multicolumn{2}{c|}{\boldsymbol{a}_1'^T C} & u' - \boldsymbol{a}_1'^T\boldsymbol{d} - b_1' \\ \hline \multicolumn{2}{c|}{\boldsymbol{a}_2'^T C} & v' - \boldsymbol{a}_2'^T\boldsymbol{d} - b_2' \end{array}\right) = \text{Det}\begin{pmatrix} \mathcal{S}C & \boldsymbol{q} - \mathcal{S}\boldsymbol{d} - \boldsymbol{r} \\ \boldsymbol{c}^T & v' - d \end{pmatrix},$$

where

$$\mathcal{S} = \begin{pmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \boldsymbol{a}_1'^T \end{pmatrix}, \quad \boldsymbol{q} = \begin{pmatrix} u \\ v \\ u' \end{pmatrix}, \quad \boldsymbol{r} = \begin{pmatrix} b_1 \\ b_2 \\ b_1' \end{pmatrix}, \quad \boldsymbol{c} = C^T\boldsymbol{a}_2' \quad \text{and} \quad d = \boldsymbol{a}_2'^T\boldsymbol{d} + b_2'.$$

When $\mathcal{S}$ is nonsingular, we can choose $C = \mathcal{S}^{-1}$ and $\boldsymbol{d} = -\mathcal{S}^{-1}\boldsymbol{r}$. If $\boldsymbol{c} = (a, b, c)^T$, this reduces the two projection matrices to the canonical forms

$$\tilde{\mathcal{M}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \tilde{\mathcal{M}}' = \begin{pmatrix} 0 & 0 & 1 & 0 \\ a & b & c & d \end{pmatrix}, \tag{12.6}$$

and allows us to rewrite the epipolar constraint as

$$\text{Det}\begin{pmatrix} 1 & 0 & 0 & u \\ 0 & 1 & 0 & v \\ 0 & 0 & 1 & u' \\ a & b & c & v' - d \end{pmatrix} = -au - bv - cu' + v' - d = 0,$$

where the coefficients a, b, c, and d are related to the parameters α, β, α', β', and δ by $a : \alpha = b : \beta = c : \alpha' = -1 : \beta' = d : \delta$.

Given enough point correspondences, the coefficients a, b, c, and d can be estimated via linear least squares, similar to the perspective case studied in chapter 10. Once these parameters have been found, the two projection matrices are known, and the position of any point can be estimated from its image coordinates by using once again linear least squares to solve the corresponding system of four equations,

$$\begin{pmatrix} 1 & 0 & 0 & u \\ 0 & 1 & 0 & v \\ 0 & 0 & 1 & u' \\ a & b & c & v' - d \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{P}} \\ -1 \end{pmatrix} = 0, \tag{12.7}$$

for the three unknown coordinates of $\tilde{\boldsymbol{P}}$.

Note that the first three equations in Eq. (12.7) are in principle sufficient to solve for $\tilde{\boldsymbol{P}}$ as $(u, v, u')^T$ without estimating the coefficients a, b, c, and d and without requiring a minimum number of matches. This is not as surprising as one may originally think: In the case of two cali-

brated orthographic cameras with perpendicular projection directions and parallel v axes, taking $x = u$, $y = v$, and $z = u'$ does yield the correct Euclidean reconstruction (have another look at Figure 12.5, assuming orthographic projection and imagining that the epipolar lines are parallel to the u and u' axes). In practice, of course, using all four equations may yield more accurate results. The proposed method reduces the first row of $\mathcal{A}'$ to $(0, 0, 1)$ via the affine transformation $\mathcal{Q}$. When $\mathcal{S}$ is (close to) singular, it is possible to apply instead the same reduction to the second row of $\mathcal{A}'$. When both $\mathcal{S}$ and the matrix constructed in that fashion are singular, the two image planes are parallel and the scene structure cannot be recovered.

12.3 AFFINE STRUCTURE AND MOTION FROM MULTIPLE IMAGES

The methods presented in the previous section are aimed at recovering the affine scene structure and/or the corresponding projection matrices from a minimum number of images. We now address the problem of estimating the same information from a potentially large number of pictures. We first show that any *fixed* set of affine images of a scene exhibits an affine structure; then we use this property to derive the factorization method of Tomasi and Kanade (1992) for estimating the affine structure and motion of a scene from an image sequence.

12.3.1 The Affine Structure of Affine Image Sequences

We suppose in this section and the next that we observe a static scene with a fixed set of m affine cameras and denote by $p_1, \ldots, p_m$ the m projections of the scene point P. Stacking the corresponding m instances of Eq. (12.1) yields

$$\boldsymbol{q} = \boldsymbol{r} + \mathcal{A}\boldsymbol{P},$$

where

$$\boldsymbol{q} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{p}_1 \\ \ldots \\ \boldsymbol{p}_m \end{pmatrix}, \quad \boldsymbol{r} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{b}_1 \\ \ldots \\ \boldsymbol{b}_m \end{pmatrix} \quad \text{and} \quad \mathcal{A} \stackrel{\text{def}}{=} \begin{pmatrix} \mathcal{A}_1 \\ \ldots \\ \mathcal{A}_m \end{pmatrix}.$$

If I denotes the set of all images taken by the m cameras, we have

$$I = \{\boldsymbol{r} + \mathcal{A}\boldsymbol{P} | \boldsymbol{P} \in \mathbb{R}^3\} = \boldsymbol{r} + V_{\mathcal{A}},$$

where $V_{\mathcal{A}}$ denotes the *range* of the $2m \times 3$ matrix $\mathcal{A}$ (i.e., the three-dimensional vector subspace of $\mathbb{R}^{2m}$ spanned by its column vectors. In other words, I is a three-dimensional subspace of the affine space $\mathbb{R}^{2m}$). In particular, if we consider as before n points $P_1, \ldots, P_n$ observed by m cameras, we can define the $(2m + 1) \times n$ data matrix

$$\mathcal{D} = \begin{pmatrix} \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_n \\ 1 & \ldots & 1 \end{pmatrix},$$

and it follows from Section 12.1 that this matrix has (at most) rank 4.

12.3.2 A Factorization Approach to Affine Structure from Motion

Tomasi and Kanade (1992) exploited the affine structure of affine images in a robust factorization method for estimating the structure of a scene and the corresponding camera motion through singular value decomposition (see insert).

Technique: Singular Value Decomposition

Let $\mathcal{A}$ be an $m \times n$ matrix, with $m \geq n$, then $\mathcal{A}$ can always be written as

$$\mathcal{A} = \mathcal{U}\mathcal{W}\mathcal{V}^T,$$

where

- $\mathcal{U}$ is an $m \times n$ column-orthogonal matrix (i.e., $\mathcal{U}^T\mathcal{U} = \mathrm{Id}_n$),
- $\mathcal{W}$ is a diagonal matrix whose diagonal entries w_i $(i = 1, \ldots, n)$ are the singular values of $\mathcal{A}$ with $w_1 \geq w_2 \geq \cdots \geq w_n \geq 0$,
- and $\mathcal{V}$ is an $n \times n$ orthogonal matrix, i.e., $\mathcal{V}^T\mathcal{V} = \mathcal{V}\mathcal{V}^T = \mathrm{Id}_n$.

This is the *singular value decomposition* (*SVD*) of the matrix $\mathcal{A}$, and it can be computed using the algorithm described in Wilkinson and Reich (1971).

As shown by the following theorem, the singular value decomposition of a matrix is related to the eigenvalues and eigenvectors of its square.

Theorem 3. *The singular values of the matrix $\mathcal{A}$ are the eigenvalues of the matrix $\mathcal{A}^T\mathcal{A}$ and the columns of the matrix $\mathcal{V}$ are the corresponding eigenvectors.*

This theorem can be used to solve overconstrained homogeneous linear equations of the form $\mathcal{A}\boldsymbol{x} = \boldsymbol{0}$ as defined in chapter 3 without explicitly computing the corresponding matrix $\mathcal{A}^T\mathcal{A}$. The solution is simply the column vector of the matrix $\mathcal{V}$ in the singular value decomposition of $\mathcal{A}$ that is associated with the smallest singular value.

The SVD of a matrix can also be used to characterize matrices that are rank-deficient. Suppose that $\mathcal{A}$ has rank $p < n$. Then the matrices $\mathcal{U}$, $\mathcal{W}$, and $\mathcal{V}$ can be written as

$$\mathcal{U} = \left[\begin{array}{c|c} \mathcal{U}_p & \mathcal{U}_{n-p} \end{array}\right] \quad \mathcal{W} = \left[\begin{array}{c|c} \mathcal{W}_p & 0 \\ \hline 0 & 0 \end{array}\right] \quad \text{and} \quad \mathcal{V}^T = \left[\begin{array}{c} \mathcal{V}_p^T \\ \hline \mathcal{V}_{n-p}^T \end{array}\right],$$

and

- the columns of $\mathcal{U}_p$ form an orthonormal basis of the space spanned by the columns of $\mathcal{A}$ (i.e., its *range*),
- and the columns of $\mathcal{V}_{n-p}$ for a basis of the space spanned by the solutions of $\mathcal{A}\boldsymbol{x} = 0$ (i.e., the *null space* of this matrix).

The $m \times p$ and $n \times p$ matrices $\mathcal{U}_p$ and $\mathcal{V}_p$ are both column-orthogonal, and we have of course $\mathcal{A} = \mathcal{U}_p\mathcal{W}_p\mathcal{V}_p^T$.

The following theorem shows that singular value decomposition also provides a valuable *approximation* procedure. In both cases, $\mathcal{U}_p$ and $\mathcal{V}_p$ denote as before the matrices formed by the p leftmost columns of the matrices $\mathcal{U}$ and $\mathcal{V}$, and $\mathcal{W}_p$ is the $p \times p$ diagonal matrix formed by the p largest singular values. This time, however, $\mathcal{A}$ may have maximal rank n, and the remaining singular values may be nonzero.

Theorem 4. *When $\mathcal{A}$ has a rank greater than p, $\mathcal{U}_p\mathcal{W}_p\mathcal{V}_p^T$ is the best possible rank-p approximation of $\mathcal{A}$ in the sense of the Frobenius norm.*

This theorem plays a fundamental role in the factorization approach to structure from motion presented in this chapter.

Assuming that the origin of the object coordinate system is one of the observed points or their center of mass, say P_0, we can translate the origin of the image coordinate system to the corresponding image point, say p_0. The transformation $\boldsymbol{p} \rightarrow \boldsymbol{p} - \boldsymbol{p}_0$ freezes the origin of the set of images I, which becomes the three-dimensional *vector space* $V_{\mathcal{A}}$. In other words, we can write, for any point P, and for $i = 1, \ldots, m$, that $\boldsymbol{p}_i = \mathcal{A}_i \boldsymbol{P}$. Equivalently, $\boldsymbol{q} = \mathcal{A}\boldsymbol{P}$, and

$$I = \{\mathcal{A}\boldsymbol{P} | \boldsymbol{P} \in \mathbb{R}^3\} = V_{\mathcal{A}}.$$

Given m images of n points $P_1, \ldots, P_n$, we can now define the $2m \times n$ data matrix

$$\mathcal{D} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{q}_1 & \ldots & \boldsymbol{q}_n \end{pmatrix} = \mathcal{A}\mathcal{P}, \quad \text{with} \quad \mathcal{P} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{P}_1 & \ldots & \boldsymbol{P}_n \end{pmatrix}.$$

As the product of a $2m \times 3$ matrix and a $3 \times n$ matrix, $\mathcal{D}$ has, in general, rank 3. If $\mathcal{U}\mathcal{W}\mathcal{V}^T$ is its singular value decomposition, this means that only three of the singular values are nonzero, thus $\mathcal{D} = \mathcal{U}_3\mathcal{W}_3\mathcal{V}_3^T$, where $\mathcal{U}_3$ and $\mathcal{V}_3$ denote the $2m \times 3$ and $3 \times n$ matrices formed by the three leftmost columns of the matrices $\mathcal{U}$ and $\mathcal{V}$, and $\mathcal{W}_3$ is the 3×3 diagonal matrix formed by the corresponding nonzero singular values.

We claim that we can take $\mathcal{A}_0 = \mathcal{U}_3$ and $\mathcal{P}_0 = \mathcal{W}_3\mathcal{V}_3^T$ as representative of the true (affine) camera motion and scene shape. Indeed, the columns of $\mathcal{A}$ form by definition a basis for the range $V_{\mathcal{A}}$ of $\mathcal{D}$, whereas the columns of $\mathcal{A}_0$ form by construction another basis for this vector space. This implies that there exists a 3×3 matrix $\mathcal{Q}$ such that $\mathcal{A} = \mathcal{A}_0\mathcal{Q}$ and, thus, $\mathcal{P} = \mathcal{Q}^{-1}\mathcal{P}_0$. Conversely, $\mathcal{D} = (\mathcal{A}_0\mathcal{Q})(\mathcal{Q}^{-1}\mathcal{P}_0)$ for any invertible 3×3 matrix $\mathcal{Q}$. Adding to this linear ambiguity the degrees of freedom corresponding to the position of the origin of the world coordinate system confirms once again the affine ambiguity of the structure-from-motion problem, and the fact that singular value decomposition provides representative estimates of the affine motion and scene structure.

Our reasoning so far is only valid in an idealized, noiseless situation. In practice, due to image noise, errors in localization of feature points, and to the mere fact that actual cameras are not affine, the equation $\mathcal{D} = \mathcal{A}\mathcal{P}$ does not hold exactly, and the matrix $\mathcal{D}$ has (in general) full rank. Let us show that singular value decomposition still yields a reasonable estimate of the affine structure and motion in this case: the best we can hope for is to minimize

$$E \stackrel{\text{def}}{=} \sum_{i,j} |\boldsymbol{p}_{ij} - \mathcal{A}_i \boldsymbol{P}_j|^2 = \sum_j |\boldsymbol{q}_j - \mathcal{A}\boldsymbol{P}_j|^2 = |\mathcal{D} - \mathcal{A}\mathcal{P}|^2$$

Algorithm 12.1: The Tomasi–Kanade factorization algorithm for affine shape from motion. Note that the original algorithm, proposed in Tomasi and Kanade (1992) uses $\mathcal{A}_0 = \mathcal{U}_3\sqrt{\mathcal{W}_3}$ and $\mathcal{P}_0 = \sqrt{\mathcal{W}_3}\mathcal{V}_3^T$. Both solutions are mathematically and numerically equivalent.

1. Compute the singular value decomposition $\mathcal{D} = \mathcal{U}\mathcal{W}\mathcal{V}^T$.
2. Construct the matrices $\mathcal{U}_3$, $\mathcal{V}_3$, and $\mathcal{W}_3$ formed by the three leftmost columns of the matrices $\mathcal{U}$ and $\mathcal{V}$, and the corresponding 3×3 submatrix of $\mathcal{W}$.
3. Define

$$\mathcal{A}_0 = \mathcal{U}_3 \quad \text{and} \quad \mathcal{P}_0 = \mathcal{W}_3\mathcal{V}_3^T;$$

the $2m \times 3$ matrix $\mathcal{A}_0$ is an estimate of the camera motion, and the $3 \times n$ matrix $\mathcal{P}_0$ is an estimate of the scene structure.

with respect to the matrices $\mathcal{A}_i$ $(i = 1, \ldots, m)$ and vectors $\boldsymbol{P}_j$ $(j = 1, \ldots, m)$ or, equivalently, with respect to the matrices $\mathcal{A}$ and $\mathcal{P}$.

According to Theorem 4, the matrix $\mathcal{A}_0\mathcal{P}_0$ is the closest rank-3 approximation to $\mathcal{D}$. Since the rank of $\mathcal{A}\mathcal{P}$ is 3 for any rank-3 $2m \times 3$ matrix $\mathcal{A}$ and rank-3 $3 \times n$ matrix $\mathcal{P}$, the minimum value of E is thus reached for $\mathcal{A} = \mathcal{A}_0$ and $\mathcal{P} = \mathcal{P}_0$, which confirms that $\mathcal{A}_0$ and $\mathcal{P}_0$ are the optimal estimates of the true camera motion and scene structure. This does not contradict the inherent ambiguity of affine structure from motion: All affinely equivalent solutions yield the same value for E. In particular, singular value decomposition can be used to estimate the affine structure and motion from the data matrix $\mathcal{D}$ as shown in Algorithm 12.12.1.

12.4 FROM AFFINE TO EUCLIDEAN IMAGES

Let us assume that a rigid scene is observed by two calibrated orthographic cameras so the image points are represented by their normalized coordinate vectors. In this case, the transformation between the coordinate systems attached to the cameras goes from affine to Euclidean (i.e., it can be written as the composition of a rotation and a translation). Under orthographic projection, a translation in depth has no effect, and a translation in the image plane (frontoparallel translation) is easily eliminated by aligning the two projections of some scene point A. Any rotation about the viewing direction is also easily identified and discarded. At this stage, the two views differ by a rotation about some axis in a frontoparallel plane passing through the projection of A. Koenderink and Van Doorn (1990) showed that there exists a one-parameter family of such rotations, determining the shape up to a depth scaling and a shear, and that the addition of a third view finally restricts the solution to one or two pairs related through a reflection in the frontoparallel plane (Figure 12.4). The details of this construction are a bit too involved to be included here. Instead, we introduce in the rest of this section a simple method for going from affine to Euclidean structure when the cameras' affine projection matrices have been estimated.

12.4.1 Euclidean Constraints and Calibrated Affine Cameras

Let us first have another look at the orthographic, weak-perspective, and paraperspective models of the imaging process (we do not detail the parallel projection case since it is rarely used in practice), assuming that the cameras have been calibrated. Obviously, the affine projection Eq. (12.1) still holds in this case, but this time there are some constraints on the components of the projection matrix $\mathcal{M} = (\mathcal{A} \quad \boldsymbol{b})$.

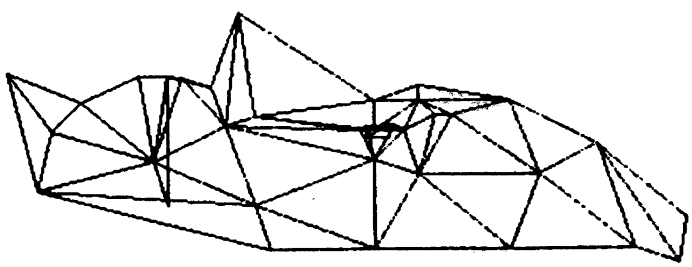


Figure 12.6 Euclidean reconstruction from the three views of a face shown in Figure 12.4. *Reprinted with permission from "Affine Structure from Motion," by J.J. Koenderink and A.J. Van Doorn, Journal of the Optical Society of America A, 8:377–385, (1990). © 1990 Optical Society of America.*

Recall from Eq. (2.20) in chapter 2 that a weak-perspective projection matrix can be written as

$$\mathcal{M} = \frac{1}{z_r}\begin{pmatrix} k & s \\ 0 & 1 \end{pmatrix}(\mathcal{R}_2 \quad \boldsymbol{t}_2),$$

where $\mathcal{R}_2$ is the 2×3 matrix formed by the first two rows of a rotation matrix and $\boldsymbol{t}_2$ is a vector in $\mathbb{R}^2$. When the camera is calibrated, we can use normalized image coordinates and take $k = 1$ and $s = 0$. The projection matrix becomes

$$\hat{\mathcal{M}} = (\hat{\mathcal{A}} \quad \hat{\boldsymbol{b}}) = \frac{1}{z_r}(\mathcal{R}_2 \quad \boldsymbol{t}_2). \tag{12.8}$$

An orthographic camera is a weak-perspective camera with $z_r = 1$, and it follows from Eq. (12.8) that the matrix $\hat{\mathcal{A}}$ is part of a rotation matrix, with unit row vectors $\hat{\boldsymbol{a}}_1^T$ and $\hat{\boldsymbol{a}}_2^T$ orthogonal to each other. In other words, an orthographic camera is an affine camera with the additional constraints

$$\hat{\boldsymbol{a}}_1 \cdot \hat{\boldsymbol{a}}_2 = 0 \quad \text{and} \quad |\hat{\boldsymbol{a}}_1|^2 = |\hat{\boldsymbol{a}}_2|^2 = 1. \tag{12.9}$$

The general weak-perspective case is similar, but the rows of the matrix $\hat{\mathcal{A}}$ are not unit vectors anymore. It follows that a weak-perspective camera is an affine camera with the two constraints

$$\hat{\boldsymbol{a}}_1 \cdot \hat{\boldsymbol{a}}_2 = 0 \quad \text{and} \quad |\hat{\boldsymbol{a}}_1|^2 = |\hat{\boldsymbol{a}}_2|^2. \tag{12.10}$$

Finally, it is easy to use the parameterization of paraperspective cameras given by Eq. (2.22) in chapter 2 to show (see Exercises) that a paraperspective camera is an affine camera that satisfies the constraints

$$\hat{\boldsymbol{a}}_1 \cdot \hat{\boldsymbol{a}}_2 = \frac{u_r v_r}{2(1+u_r^2)}|\hat{\boldsymbol{a}}_1|^2 + \frac{u_r v_r}{2(1+v_r^2)}|\hat{\boldsymbol{a}}_2|^2 \quad \text{and} \quad \frac{|\hat{\boldsymbol{a}}_1|^2}{(1+u_r^2)} = \frac{|\hat{\boldsymbol{a}}_2|^2}{(1+v_r^2)}, \tag{12.11}$$

where (u_r, v_r) denote the coordinates of the perspective projection of the reference point R associated with the paraperspective projection model.

12.4.2 Computing Euclidean Upgrades from Multiple Views

Let us focus on orthographic projection and assume that we have recovered the affine shape of a scene and the projection matrix $\mathcal{M}$ associated with each view. We already know that all solutions of the structure-from-motion problem are the same up to an affine ambiguity. In particular, if the position of a scene point in a *Euclidean* coordinate system is $\hat{\boldsymbol{P}}$ and the corresponding projection matrix is $\hat{\mathcal{M}} = (\hat{\mathcal{A}} \quad \hat{\boldsymbol{b}})$, there must exist some affine transformation

$$\mathcal{Q} = \begin{pmatrix} \mathcal{C} & \boldsymbol{d} \\ \boldsymbol{0}^T & 1 \end{pmatrix}$$

such that $\hat{\mathcal{M}} = \mathcal{M}\mathcal{Q}$ and $\hat{\boldsymbol{P}} = \mathcal{C}^{-1}(\tilde{\boldsymbol{P}} - \boldsymbol{d})$. Such a transformation is called a *Euclidean upgrade* because it maps the affine shape of a scene onto its Euclidean one.

Let us now show how compute such an upgrade when $m \geq 3$ orthographic images are available. Let $\mathcal{M}_i = (\mathcal{A}_i \quad \boldsymbol{b}_i)$ denote the corresponding projection matrices, estimated using the factorization method of Section 12.3.2, for example. If $\hat{\mathcal{M}}_i = \mathcal{M}_i\mathcal{Q}$, we can rewrite the

orthographic constraints of Eq. (12.9) as

$$\begin{cases} \hat{\boldsymbol{a}}_{i1} \cdot \hat{\boldsymbol{a}}_{i2} = 0, \\ |\hat{\boldsymbol{a}}_{i1}|^2 = 1, \\ |\hat{\boldsymbol{a}}_{i2}|^2 = 1, \end{cases} \iff \begin{cases} \boldsymbol{a}_{i1}^T \mathcal{C}\mathcal{C}^T \boldsymbol{a}_{i2} = 0, \\ \boldsymbol{a}_{i1}^T \mathcal{C}\mathcal{C}^T \boldsymbol{a}_{i1} = 1, \\ \boldsymbol{a}_{i2}^T \mathcal{C}\mathcal{C}^T \boldsymbol{a}_{i2} = 1, \end{cases} \quad \text{for} \quad i = 1, \ldots, m, \tag{12.12}$$

where $\boldsymbol{a}_{i1}^T$ and $\boldsymbol{a}_{i2}^T$ denote the rows of the matrix $\mathcal{A}_i$. This overconstrained system of $3m$ quadratic equations in the coefficients of $\mathcal{C}$ can be solved via nonlinear least squares. An alternative is to consider Eq. (12.12) as a set of *linear* constraints on the matrix $\mathcal{D} \stackrel{\text{def}}{=} \mathcal{C}\mathcal{C}^T$. The coefficients of $\mathcal{D}$ can be found in this case via linear least squares, and $\mathcal{C}$ can then be computed as $\sqrt{\mathcal{D}}$ using Cholesky decomposition. It should be noted that this requires that the recovered matrix $\mathcal{D}$ be positive definite, which is not guaranteed in the presence of noise. Note also that the solution of Eq. (12.12) is only defined up to an arbitrary rotation. To determine $\mathcal{Q}$ uniquely and simplify the calculations, it is possible to map $\mathcal{M}_1$ (and possibly $\mathcal{M}_2$) to its canonical form and essentially follow the procedure given in the previous section.

Figure 12.7 shows an example, including four pictures in a video sequence of a house, a view of the recovered scene structure, and a real picture taken from a similar viewpoint for comparison.

The computation of a Euclidean upgrade for weak- and paraperspective projections follows a similar path, except for the fact that the two constraints of Eq. (12.10) or Eq. (12.11) written for m images replace the $3m$ constraints of Eq. (12.12). Note that in these cases it is not possible to determine the absolute scale of the scene since the Euclidean constraints of Eqs. (12.10) and

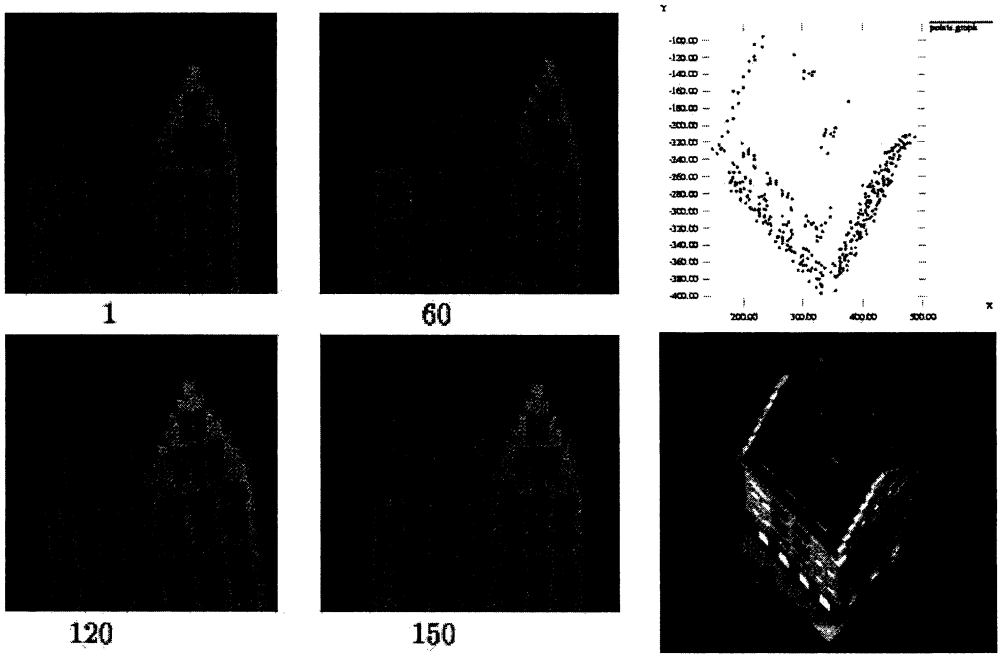


Figure 12.7 Euclidean structure from motion—experimental results. Left: Sample images of a house in a 150-frame sequence. Right: A view of the reconstructed structure (top) and a real picture of the house (bottom) taken from a similar viewpoint. *Reprinted from "Factoring Image Sequences into Shape and Motion," by C. Tomasi and T. Kanade, Proc. IEEE Workshop on Visual Motion, (1991).* © *1991 IEEE.*

(12.11) are homogeneous. In other words, the structure of the scene can only be recovered up to an arbitrary *similarity* (e.g., a rigid transformation followed by an isotropic scaling). Accordingly, we now take *Euclidean shape* to mean the equivalence class formed by point sets related by similarities (some authors use instead the term *metric shape* to emphasize the scale ambiguity).

12.5 AFFINE MOTION SEGMENTATION

We have assumed so far that the n points observed all undergo the same motion. What happens if these points belong instead to k objects undergoing different motions? This section presents two methods for segmenting the data points into such independently moving objects.

12.5.1 The Reduced Row-Echelon Form of the Data Matrix

Exactly as in Section 12.3.1, we can define the data matrix

$$\mathcal{D} = \begin{pmatrix} \boldsymbol{p}_{11} & \cdots & \boldsymbol{p}_{1n} \\ \cdots & \cdots & \cdots \\ \boldsymbol{p}_{m1} & \cdots & \boldsymbol{p}_{mn} \\ 1 & \cdots & 1 \end{pmatrix}.$$

This time, however, $\mathcal{D}$ does not have rank 4 anymore. Instead, the columns of the data matrix corresponding to each object define a four-dimensional subspace D_i ($i = 1, \ldots, k$) of its range, and the overall rank of $\mathcal{D}$ is (at most) $4k$. As remarked by Gear (1998), constructing the *reduced row-echelon form (RREF)* of $\mathcal{D}$ identifies the subspaces D_i and the column vectors that lie in them, providing a segmentation of the input points into rigid objects (or, more precisely, into objects that may undergo affine deformations).

The RREF of a matrix $\mathcal{U}$ is a matrix $\mathcal{V}$ whose rows are linear combinations of the rows of $\mathcal{U}$ and that satisfies the following conditions:

1. all rows consisting entirely of zeros are at its bottom;
2. the first nonzero entry in each row is a 1, called the *leading* 1;
3. the leading 1 in each row is to the right of all leading 1s in rows above it; and
4. each leading 1 is the only nonzero entry in its column.

A *base column* is a column that contains a leading 1. By construction, the only nonzero entries of any nonbase column are in rows in which exactly one base column has a 1. In addition, any nonbase column $\boldsymbol{v}$ lies in the subspace spanned by the base columns $\boldsymbol{v}_{j_1}, \ldots, \boldsymbol{v}_{j_k}$ associated with its nonzero entries $\alpha_1, \ldots, \alpha_k$, and $\boldsymbol{v}$ can be written as $\alpha_1 \boldsymbol{v}_{j_1} + \cdots + \alpha_k \boldsymbol{v}_{j_k}$. The number of base columns gives the rank r of the matrix.

Let us illustrate these properties with a sample 7×6 matrix $\mathcal{U}$ and its RREF $\mathcal{V}$ (the entries of $\mathcal{U}$ have been chosen to give $\mathcal{V}$ a simple form):

$$\mathcal{U} = \begin{pmatrix} 1 & 0 & 1 & -5 & 2 & -9 \\ 2 & 4 & 10 & 0 & 1 & 1 \\ -1 & 1 & 1 & 3 & 0 & 1 \\ 0 & 1 & 2 & -1 & 3 & -10 \\ 3 & -2 & -1 & 0 & 1 & 3 \\ 0 & 5 & 10 & 2 & -2 & 8 \\ -2 & 3 & 4 & 1 & 0 & -3 \end{pmatrix} \longrightarrow \mathcal{V} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 2 \\ 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Let us denote by $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ $(i = 1, \ldots, 6)$ the columns of the matrices $\mathcal{U}$ and $\mathcal{V}$. There are four base columns, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_4$, and $\boldsymbol{v}_5$, so the rank of $\mathcal{U}$ is 4. The nonzero entries of $\boldsymbol{v}_3$ are in the same rows as the leading 1s of base columns $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, indicating that $\boldsymbol{v}_3$ lies in the subspace of $\mathbb{R}^7$ spanned by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. The values of these entries are 1 and 2, implying that $\boldsymbol{v}_3 = \boldsymbol{v}_1 + 2\boldsymbol{v}_2$. Likewise, the nonzero entries of $\boldsymbol{v}_6$ are in the same rows as the leading 1s of $\boldsymbol{v}_1$, $\boldsymbol{v}_4$, and $\boldsymbol{v}_5$, indicating that $\boldsymbol{v}_6$ lies in the subspace of $\mathbb{R}^7$ spanned by $\boldsymbol{v}_1$, $\boldsymbol{v}_4$, and $\boldsymbol{v}_5$, and the values of these entries are 2, 1, and -3, showing that $\boldsymbol{v}_6 = 2\boldsymbol{v}_1 + \boldsymbol{v}_4 - 3\boldsymbol{v}_5$. In fact, these properties also hold for the original matrix $\mathcal{U}$ (i.e., $\boldsymbol{u}_3 = \boldsymbol{u}_1 + 2\boldsymbol{u}_2$ and $\boldsymbol{u}_6 = 2\boldsymbol{u}_1 + \boldsymbol{u}_4 - 3\boldsymbol{u}_5$, as immediately confirmed by inspection of $\mathcal{U}$). This is due to the fact that the rows of $\mathcal{V}$ are linear combinations of the rows of $\mathcal{U}$.

The same properties hold for arbitrary matrices and their RREFs, and this shows that the RREF of the data matrix $\mathcal{D}$ can, in theory, be used for affine motion segmentation. Indeed, it identifies a basis for the range of $\mathcal{D}$ (the columns of this matrix corresponding to base columns in its RREF) as well as all the column vectors that lie in the subspaces spanned by subsets of this basis. When the four-dimensional subspaces D_i associated with each object only intersect at the origin (which is expected to be true for large enough values of m), the corresponding groups of points form connected components of the graph whose nodes are the columns of the RREF and whose arcs link pairs of columns with nonzero entries in at least one common row.

Unfortunately, the situation is more complicated in practice due to noise and numerical errors. A plain implementation of the RREF using, say, Gauss–Jordan elimination with pivoting, normally results in a full-rank matrix, with none of the nonbase columns lying in a four-dimensional subspace of the range (see Exercises). Gear (1998) gives several "robustified" methods for computing the RREF of a matrix, including Gauss–Jordan elimination with a test for discarding small pivot values and QR reduction followed by Gauss–Jordan elimination applied to the corresponding triangular matrix $\mathcal{R}$, and presents successful segmentation experiments involving both synthetic and real image sequences.

12.5.2 The Shape Interaction Matrix

The approach presented in the previous section relies only on the affine structure of affine images. Costeira and Kanade (1998) have proposed a different method, based on a factorization of the data matrix. We present this technique in the case of two groups of points undergoing different motions. The generalization to an arbitrary number of independently moving objects is straightforward.

In the setting of motion segmentation, it is not possible to define a rank-3 data matrix for each object since the centroid of the corresponding points is unknown. Instead, let us assume noiseless data and define the data matrices $\mathcal{D}^{(i)}$ $(i = 1, 2)$ by

$$\mathcal{D}^{(i)} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{p}_{11}^{(i)} & \cdots & \boldsymbol{p}_{1n_i}^{(i)} \\ \cdots & \cdots & \cdots \\ \boldsymbol{p}_{m1}^{(i)} & \cdots & \boldsymbol{p}_{mn_i}^{(i)} \end{pmatrix},$$

where n_i is the number of points associated with object number i and $n_1 + n_2 = n$. Each data matrix has rank 4 since it can be rewritten as $\mathcal{D}^{(i)} = \mathcal{M}^{(i)}\mathcal{P}^{(i)}$ where, this time,

$$\mathcal{M}^{(i)} \stackrel{\text{def}}{=} \begin{pmatrix} \mathcal{M}_1^{(i)} \\ \cdots \\ \mathcal{M}_m^{(i)} \end{pmatrix} \quad \text{and} \quad \mathcal{P}^{(i)} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{P}_1^{(i)} & \cdots & \boldsymbol{P}_{n_i}^{(i)} \\ 1 & \cdots & 1 \end{pmatrix}.$$

Let us define the $2m \times n$ composite data matrix $\mathcal{D} \stackrel{\text{def}}{=} (\mathcal{D}^{(1)} \quad \mathcal{D}^{(2)})$ as well as the composite $2m \times 8$ (motion) and $8 \times n$ (structure) matrices

$$\mathcal{M} \stackrel{\text{def}}{=} \begin{pmatrix} \mathcal{M}^{(1)} & \mathcal{M}^{(2)} \end{pmatrix} \quad \text{and} \quad \mathcal{P} \stackrel{\text{def}}{=} \begin{pmatrix} \mathcal{P}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathcal{P}^{(2)} \end{pmatrix}.$$

With this notation, we have $\mathcal{D} = \mathcal{M}\mathcal{P}$, which shows that $\mathcal{D}$ has (at most) rank 8. Now the rows of the matrix $\mathcal{P}$ form a basis for the 8-dimensional subspace of $\mathbb{R}^{2m}$ spanned by the rows of the matrix $\mathcal{D}$. As shown in Strang (1980) for example, the operator that maps any vector onto its orthogonal projection into the space spanned by the columns of a matrix $\mathcal{A}$ can be represented by the matrix $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{A}(\mathcal{A}^T\mathcal{A})^{-1}\mathcal{A}^T$. In particular, the matrix $\mathcal{Z}$ associated with the rows of $\mathcal{D}$ (or equivalently the columns of $\mathcal{D}^T$) is by construction block diagonal since $\mathcal{P}^T$ is also block diagonal.

Of course, $\mathcal{P}$ is unknown in our case, but any other matrix whose rows form a basis for the row space of $\mathcal{D}$ can be used as well. For example, if the rank-8 SVD of $\mathcal{D}$ is $\mathcal{U}_8\mathcal{W}_8\mathcal{V}_8^T$, we can use the rows of $\mathcal{V}_8^T$ as a basis, and we obtain $\mathcal{Z} = \mathcal{V}_8(\mathcal{V}_8^T\mathcal{V}_8)^{-1}\mathcal{V}_8^T = \mathcal{V}_8\mathcal{V}_8^T$ since $\mathcal{V}_8$ is orthogonal. The matrix $\mathcal{Z}$ constructed in this fashion is called the *shape interaction matrix* by Costeira and Kanade (1998), and it is once again block diagonal.

The above construction assumes that the data points are ordered consistently with the object they belong to. In general, of course, this is not the case. It can be shown that the values of the entries of the matrix $\mathcal{Z}$ are independent of the order of the points. Changing this order just swaps the columns of $\mathcal{D}$ and swap the rows and columns of $\mathcal{Z}$ accordingly. Thus, recovering the correct point ordering (and the corresponding segmentation into objects) amounts to finding the row and column swaps of the matrix $\mathcal{Z}$ that reduces it to block-diagonal form.

Costeira and Kanade have proposed several methods for finding the correct swaps in the presence of noise. One possibility is to minimize the sum of the squares of the off-diagonal block entries over all rows and column permutations (see Costeira and Kanade, 1998 for details). Figure

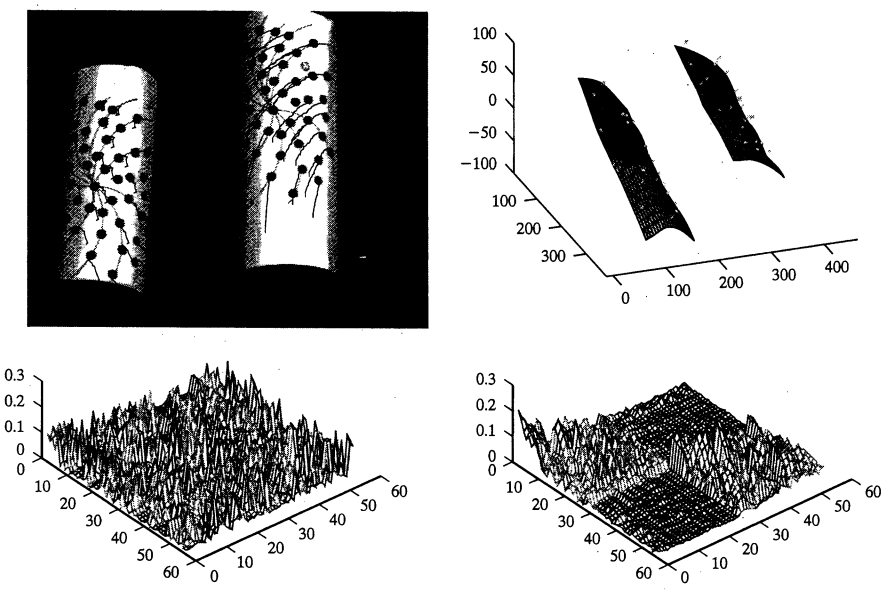


Figure 12.8 Motion segmentation—experimental results. Top-left: One frame from a sequence of pictures of two cylinders, including feature tracks. Top-right: The recovered shapes after motion segmentation. Bottom-left: The shape interaction matrix. Bottom-right: The matrix after sorting. *Reprinted from "A Multi-Body Factorization Method for Motion Analysis," by J. Costeira and T. Kanade, Proc. International Conference on Computer Vision, 1995.* © *1995 IEEE.*

12.8 shows experimental results, including the images of two objects and the corresponding feature tracks, a plot of the corresponding shape interaction matrix before and after sorting, and the corresponding segmentation results.

12.6 NOTES

The structure-from-motion problem was first studied in the calibrated orthographic setting by Ullman (1979). Its first solution in the affine setting is due to Koenderink and Van Doorn (1990). The factorization algorithm discussed in Section 12.3.2 is due to Tomasi and Kanade (1992). As shown in this chapter, the decomposition of structure from motion into an affine and a Euclidean stage affords simple and robust methods for shape reconstruction from image sequences. In essence, this *linearizes* the structure and/or motion estimation process, delaying the introduction of the nonlinear Euclidean constraints until the affine scene shape has been reconstructed. The affine stage is also valuable by itself since it is the basis for the motion-based segmentation methods introduced by Gear (1998) and Costeira and Kanade (1998) and discussed in Section 12.5; see Boult and Brown (1991) for another other approache to the same problem. As shown in chapter 26, other applications include interactive image synthesis in the augmented reality domain. Variations of the affine structure of affine images or, equivalently, of the rank 4 property of the data matrix associated with an affine motion sequence include the facts that an affine image is the linear combination of three model images (Ullman and Basri, 1991), and that the image trajectories of a scene point are linear combinations of the trajectories of three reference points (Weinshall and Tomasi, 1995). The nonlinear least-squares method for computing the Euclidean upgrade matrix $\mathcal{Q}$ is due to Tomasi and Kanade (1992). The Cholesky approach to the same problem is due to Poelman and Kanade (1997); see Weinshall and Tomasi (1995) for another variant. Various extensions of the approach presented in this chapter have been proposed recently, including the incremental recovery of structure and motion (Weinshall and Tomasi, 1995; Morita and Kanade, 1997), the extension of the affine/Euclidean decomposition to a projective/affine/Euclidean stratification (Faugeras, 1995), along with corresponding projective shape estimation algorithms (Faugeras, 1992; Hartley *et al.*, 1992; see also next chapter), and the generalization of the factorization approach of Tomasi and Kanade (1992) to the perspective case (Sturm and Triggs, 1996) and various other computer vision problems that have a natural bilinear structure (Koenderink and Van Doorn, 1997).

PROBLEMS

12.1. Explain why any definition of the "addition" of two points or of the "multiplication" of a point by a scalar is necessarily coordinate dependent.

12.2. Show that the definition of a barycentric combination as

$$\sum_{i=0}^{m} \alpha_i A_i \stackrel{\text{def}}{=} A_j + \sum_{i=0, i\neq j}^{m} \alpha_i (A_i - A_j),$$

is independent of the choice of j.

12.3. Prove that

$${}^B P = {}^B_A C\, {}^A P + {}^B O_A \iff \begin{pmatrix} {}^B P \\ 1 \end{pmatrix} = \begin{pmatrix} {}^B_A C & {}^B O_A \\ 0^T & 1 \end{pmatrix} \begin{pmatrix} {}^A P \\ 1 \end{pmatrix}.$$

12.4. Show that the set of barycentric combinations of $m+1$ points $A_0, \ldots, A_m$ in X is indeed an affine subspace of X, and show that its dimension is at most m.

12.5. Derive the equation of a line defined by two points in $\mathbb{R}^3$. (Hint: You actually need *two* equations.)

12.6. Show that the intersection of a plane with two parallel planes consists of two parallel lines.

12.7. Show that an affine transformation $\psi : X \to Y$ between two affine subspaces X and Y associated with the vector spaces $\vec{X}$ and $\vec{Y}$ can be written as $\psi(P) = \psi(O) + \vec{\psi}(P - O)$, where O is some arbitrarily chosen origin, and $\vec{\psi} : \vec{X} \to \vec{Y}$ is a linear mapping from $\vec{X}$ onto $\vec{Y}$ that is independent of the choice of O.

12.8. Show that affine cameras (and the corresponding epipolar geometry) can be viewed as the limit of a sequence of perspective images with increasing focal length receding away from the scene.

12.9. Generalize the notion of multilinearities introduced in chapter 10 to the affine case.

12.10. Prove Theorem 3.

12.11. Show that a calibrated paraperspective camera is an affine camera that satisfies the constraints

$$\hat{\mathbf{a}}_1 \cdot \hat{\mathbf{a}}_2 = \frac{u_r v_r}{2(1+u_r^2)}|\hat{\mathbf{a}}_1|^2 + \frac{u_r v_r}{2(1+v_r^2)}|\hat{\mathbf{a}}_2|^2 \quad \text{and} \quad \frac{|\hat{\mathbf{a}}_1|^2}{(1+u_r^2)} = \frac{|\hat{\mathbf{a}}_2|^2}{(1+v_r^2)},$$

where (u_r, v_r) denote the coordinates of the perspective projection of the point R.

12.12. What do you expect the RREF of an $m \times n$ matrix with random entries to be when $m \geq n$? What do you expect it to be when $m < n$? Why?

Programming Assignments

12.13. Implement the Koenderink–Van Doorn approach to affine shape from motion.

12.14. Implement the estimation of affine epipolar geometry from image correspondences and the estimation of scene structure from the corresponding projection matrices.

12.15. Implement the Tomasi–Kanade approach to affine shape from motion.

12.16. Add random numbers uniformly distributed in the [0, 0.0001] range to the entries of the matrix $\mathcal{U}$ used to illustrate the RREF and compute its RREF (using, e.g., the `rref` routine in MATLAB); then compute again the RREF using a "robustified" version of the reduction algorithm (using, e.g., `rref` with a nonzero tolerance). Comment on the results.