# MOTION ANALYSIS BY COMPUTER

Dynamic vision is an area in computer vision that studies acquisition and processing of time-varying imagery for scene interpretation, and it obtains three-dimensional structure and motion of the environment in particular. There are a number of techniques that provide the information necessary to obtain the three-dimensional structure of the scene from a single static image, such as shape from shading, shape from texture, deformation of areas, and vanishing point analysis. However, these techniques are not always reliable. They may fail when the underlying assumptions regarding the shape of the world surface are invalid or under unfavorable illumination conditions. On the other hand, a computer vision system is not necessarily passive, but can be active. The perceptual activity of human vision system is exploratory, probing, and searching. Percepts do not simply fall onto sensors as rain falls onto ground. We do not just see, we look. Our pupils adjust to the level of illumination, our eyes bring the world into sharp focus, our eyes converge or diverge, we move our heads or change our position to get a better view of something, and sometimes we even put on spectacles. In fact, if there is relative movement between the camera and the object, the viewer is automatically provided with several distinctive views of the object. Therefore they can be combined to produce reliable three-dimensional information about the object.

In general, use of the dynamic properties of the objects in the images can provide information useful for the segmentation of the image into distinct objects, and it can determine the three-dimensional structure and motion. A variety of real-world problems have motivated current dynamic vision research. These include applications in industrial automation and inspection, robot assembly, autonomous vehicle navigation, biomedical engineering, remote sensing, and general three-dimensional scene interpretation.

## MOTION ANALYSIS

Time-varying motion images can be obtained by either (a) using a stationary camera to acquire a sequence of images containing one or more moving objects in the scene or (b) moving the camera in the environment to acquire a sequence of images. The later method is also known as active perception. In either case, the sequences of images contain information about the relative movement between the camera, the objects, and the environment.

We first describe the fundamental concepts and techniques of motion analysis with image sequences acquired by a stationary camera. In this situation, there is no relative movement between the camera and the surrounding background environment. However, one or more objects in the scene may move.

### Motion Detection

The first step of motion analysis is the motion (or change) detection. Motion detection is to find where are moving objects. The simplest approach is to find the difference between two images from a motion sequence. A straightforward pixel-wise subtraction of the two images will find regions with nonzero difference. These dynamic regions correspond to objects moving in the scene. However, the images acquired from the real world could be very noisy. Therefore, the motion (or change) detected by using the simple image subtraction may not be reliable. Some preprocessings of the images are necessary to reduce the noise in the images before motion detection is performed. Motion detection may also be performed in feature spaces derived from the images, such as edge space or multiresolution decomposed hierarchical space, in order to achieve better reliability and improved performance. In addition, motion detection results obtained from feature spaces can usually facilitate motion analysis in the next step. For example, edge space is often used for motion detection because the edge information usually corresponds to boundaries of objects or textures on object surface. Because the information in edge space is at a higher level and in a more compact form than the original image pixels, the computational cost can be reduced. Perhaps most importantly, the motion information detected in edge space will readily be available for interpretation of three-dimensional structures of objects in later stages.

### Motion Estimation

Once we know the dynamic regions in the images, we want to find out how the image pixels in the dynamic regions move from one to another in the image sequence. To do this, we must first find the corresponding points between the two images. This problem is known as the correspondence problem. The correspondence problem in motion analysis is the same as that defined in stereo vision. In general, the correspondence problem is to identify image "events" that correspond to each other in the image sequence. The term "event" should be interpreted in a broad sense, to mean any identifiable structure in the image—for example, image intensity in a local region, edges, lines, texture marks, and so on.

Given a point $P$ in one image, we want to find the corresponding point $P'$ in the other image. The most direct approach is to match the light intensity reflected from a point in the environment and recorded in two images. Discrete correlation can be used to find the corresponding points. Discrete correlation is the process in which an area surrounding a point of interest in one image is "correlated" with areas of similar shape in a target region in the second image (Fig. 1), and the "best-match" area in the target region is discovered. The center of the best-match area in the second image is then regarded as the point $P'$. Among many suitable match-measures, two simple and commonly used formulas are *direct correlation:*

$$\text{Correlation} = \frac{1}{M^2}\sum_{i=0}^{M-1}\sum_{j=0}^{M-1} f_t\left(x-\frac{M}{2}+i,\, y-\frac{M}{2}+j\right) f_{t+1}\left(x'-\frac{M}{2}+i,\, y'-\frac{M}{2}+j\right) \quad (1)$$

and *least-mean-square error* (LMSE):

$$\text{LMSE} = \frac{1}{M^2}\left[\sum_{i=0}^{M-1}\sum_{j=0}^{M-1}\left(f_t\left(x-\frac{M}{2}+i,\, y-\frac{M}{2}+j\right) - f_{t+1}\left(x'-\frac{M}{2}+i,\, y'-\frac{M}{2}+j\right)\right)^2\right]^{1/2} \quad (2)$$

where $f_t(\ )$ and $f_{t+1}(\ )$ are two images, $M$ is the dimension of the window, $P = (x, y)$ is a given point in the first image, $P' = (x', y')$ is a matching point in the second image.

If the direct correlation is used to compute the match measure, then the best match is the candidate window that has the maximum value for the match measure. If the LMSE is used, then the best match is the candidate window that minimizes the match measure. The point at the center of the best-match candidate window in the second image is regarded as the corresponding point for the point of interest in the first image.

The above technique solves the local correspondence problem, because it provides constraints on the displacement of a point in the image based on image information in the immediate neighborhood of that point, and they are solved independently at all points of interest in the image.

It is also possible to apply nonlocal constraints to motion estimation. For example, we usually make an assumption of the spatial smoothness of the motion for all the points on rigid bodies in motion. It is also possible to impose on top of this framework a multifrequency, multiresolution approach. In this approach the images are preprocessed with a set of bandpass filters that are spatially local and which decompose the spatial frequency spectrum in the image in a convenient way. The matching can be performed at different frequency channels and different resolution levels. The matching results are then combined using a consistency constraint. Some researchers use the results obtained at a lower resolution level as the initial guess for a higher resolution level. The motion estimation results can be adaptively refined in this manner.

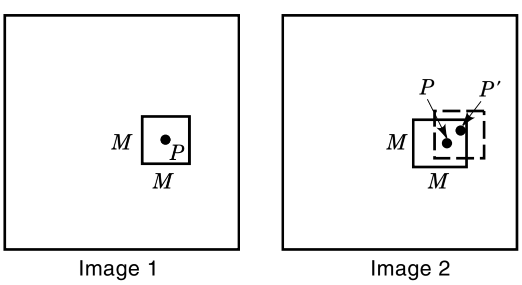


**Figure 1.** Given an image point $P$ in image 1 and an $M \times M$ window, a correspondence point $P'$ is searched in the neighborhood of $P$ in the second image.

### Optical Flow

Given the two corresponding points $P$ and $P'$ from two images, the vector $v = P - P'$ gives the direction and the distance of the point $P$ traveled from one image frame to the next. If the time interval between the two frames is considered to be unit time, the vector $v$ characterizes the velocity of point $P$ in motion. If such a motion vector is computed for every image point, it is called optical flow fields (see Fig. 2). We should keep in mind that the primary objective of dynamic vision is to recover the three-dimensional structure of objects in the scene and/or the motion of these objects in the three-dimensional world space. We discuss in the following the relationship between the motion of a three-dimensional

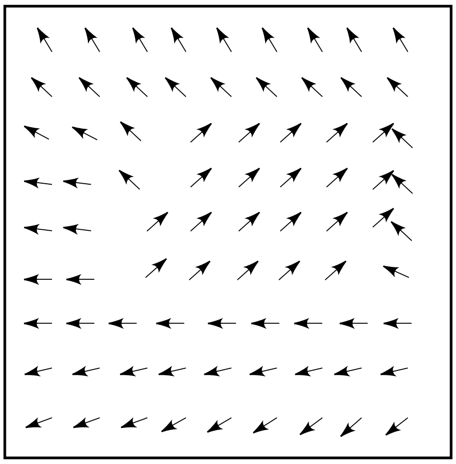

**Figure 2.** An optic flow fields in which each vector represents the velocity of the corresponding image point.

point and the corresponding motion of that point on the perspective projection image.

The point $P = (x, y, z)$ on the moving rigid body has perspective project $P' = (u, v)$ on the image plane (Fig. 3), assuming the focal length $f$ (the distance from the center of projection to the image plane):

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{f}{z}\begin{pmatrix} x \\ y \end{pmatrix} \tag{3}$$

The motion of $(x, y, z)$ causes a motion of its projection $(u, v)$ on the image. By taking time derivatives on both sides of the above equation, we obtain the following relation:

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \frac{f}{z^2}\begin{pmatrix} z\dot{x} - x\dot{z} \\ z\dot{y} - y\dot{z} \end{pmatrix} \tag{4}$$

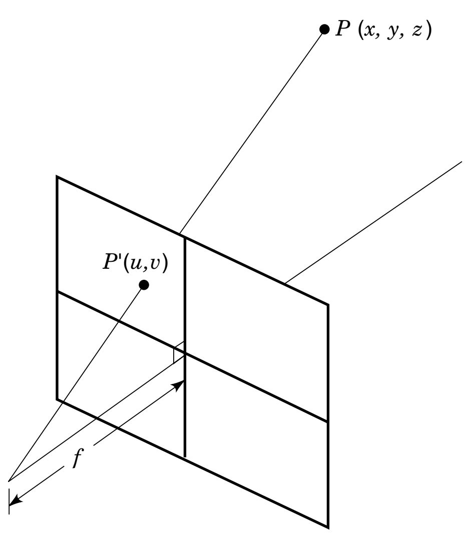


**Figure 3.** A point $P$ in the scene is perspectively projected onto the image plane at the point $P'$.

or in matrix form:

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \frac{1}{z}\begin{pmatrix} f & 0 & -u \\ 0 & f & -v \end{pmatrix}\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} \tag{5}$$

where $(\dot{u}, \dot{v})^T$ denotes the velocity of the point $(u, v)$ on the image plane and $(\dot{x}, \dot{y}, \dot{z})$ denotes the velocity of the point $(x, y, z)$ on the object.

The above equation is known as the fundamental optic flow equation.

The general solution of this equation is given by

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \frac{z}{f}\begin{pmatrix} \dot{u} \\ \dot{v} \\ 0 \end{pmatrix} + \lambda\begin{pmatrix} u \\ v \\ f \end{pmatrix} \tag{6}$$

where $\lambda$ is a free variable. The first term of the solution equation is the back-projected optical flow. It constitutes that particular solution to the optic flow equation in which all motion is in a plane parallel to the image plane. The second term is the general solution to the homogeneous equation. It indicates that any three-dimensional motion along the ray of sight is not captured in the optic flow.

## THREE-DIMENSIONAL MOTION AND STRUCTURE

The primary goal of motion analysis is to determine the three-dimensional structure of the objects in the environment and relative movement of the camera and the objects in the scene. The motion estimation and optical flow fields characterize the two-dimensional image displacements or velocities of the image points. This is only the first step in motion analysis. The interpretation of the displacement (or velocity) fields to determine the three-dimensional structure of the environment and the relative three-dimensional motion between the camera and the objects is another important step. In this section, we discuss the fundamental analysis related to rigid body motion.

### Rigid Body Motion

The geometrical nature of the optical flow fields can be understood through a serious of equations that relate the coordinates of the image points and the motion parameters to their velocity.

Let $X$–$Y$–$Z$ be a Cartesian coordinate system affixed to the camera and let $(u, v)$ represent the corresponding coordinate system on the image plane (Fig. 4). Without loss of generality, we assume the focal length of the camera to be 1.

Consider a point $P$ in the scene, located at $(X_p, Y_p, Z_p)$. The three-dimensional velocity $V = (\dot{X}_p, \dot{Y}_p, \dot{Z}_p)$ of the point is given by

$$V = \Omega \times P + T \tag{7}$$

where $\boldsymbol{\Omega} = (\Omega_X, \Omega_y, \Omega_z)$ is the rotation vector and $\boldsymbol{T} = (T_X, T_Y, T_Z)$ is the translation vector, whose direction and magnitude specify the direction of translation and the speed, respectively. The task of determining the three-dimensional motion of an object can be described as the task of recovering the parameters $\boldsymbol{\Omega}$ and $\boldsymbol{T}$. If $P' = (u, v)$ is the image position of

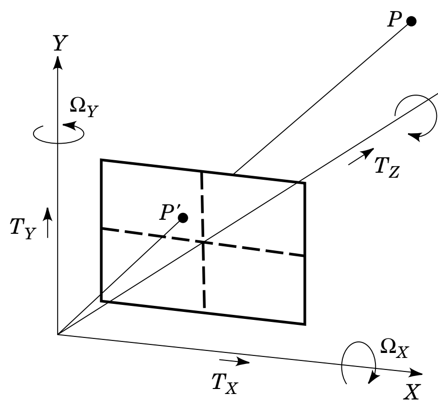


**Figure 4.** A Cartesian system illustrates the geometry of optical flow fields through rotation and translation.

the point $P$, and $U = (\dot{u}, \dot{v})$ is the velocity of the image point $P'$, then using the equations of perspective projection $x = X/Z$ and $y = Y/Z$, we derive from Eq. (7):

$$\dot{u} = -\Omega_X uv + \Omega_Y(1 + u^2) - \Omega_Z v + (T_X - T_Z u)/Z \quad (8)$$

$$\dot{v} = -\Omega_X(1 + u^2) + \Omega_Y uv + \Omega_Z u + (T_Y - T_Z v)/Z \quad (9)$$

Notice that the above equations are specific forms of the fundamental optical flow equation [Eq. (5)] for rigid bodies under rotation and translation. Six parameters describe the motion of an object and three parameters describe its three-dimensional structure. The three components each of $\boldsymbol{\Omega}$ and $\boldsymbol{T}$ specify the relative motion of the object and the camera. The $X$, $Y$, $Z$ coordinates of all the points on the object together specify the structure of the object. The parameters of motion typically do not vary from point to point in the image. All the points on a rigid object undergo the same motion and have the same motion parameters. Hence the number of motion parameters are few, with one set corresponding to each area of the image having an independent relative motion with respect to the camera. When only the camera moves, the whole image forms one coherently moving area, this situation is discussed in detail in the section entitled "Active Perception."

### Restricted Class of Motions

It should be noted in Eqs. (8) and (9) that, unless some assumptions are made regarding the structure of the environment, there is one unknown $Z$ value for each image point. There are three approaches for dealing with this problem.

The first type does not require prior computation of optical flow. Often, these techniques apply only to restricted camera motion. We illustrate these techniques by two examples: pure translation and pure rotation.

**Pure Translation.** When the camera motion is a pure translation toward the environment, all displacements of the image points appear to emanate radially from a single point in the image. This point, often known as the focus of expansion (FOE), is the point of intersection of the axis of translation with the image plane. This case is interesting because it is the practical situation for a pilot attempting to land. In this case, the problem of determining the motion of camera reduces to that of locating the FOE or, equivalently, the axis of translation. The number of parameters is two, thus greatly simplified the problem of general motion, which has six parameters. Additionally, knowing that all the displacements have to lie along the radial lines from FOE provides a powerful constraint that simplifies the correspondence problem.

The displacement $\Delta D$ of the image of the projection of a point in the three-dimensional environment is directly proportional to the distance $D$ of the projection from the FOE and inversely proportional to the distance $Z$ of the point from the camera:

$$\frac{\Delta D}{D} = \frac{\Delta Z}{Z} \quad (10)$$

where $\Delta Z$ is the displacement of the camera toward the environment along its optical axis. If the FOE is known, then $D$ is known and $\Delta D$ can be measured. Then, the ratio $\Delta Z$ and $Z$ can be obtained. If we assume that $\Delta Z$ is the unit length, then the depth $Z$ is recovered. Alternatively, some point in the image can be chosen as a reference point, and then the relative depth of the others can be obtained. Some simple algorithms are available for finding the location of FOE.

**Pure Rotation.** When the motion of the camera is pure rotation about an arbitrary axis, each image point follows a conic path. The exact curve along which the point travels is the intersection of the image plane with a cone passing through the image point (Fig. 5). Given a hypothesized axis of rotation, the path of each point can be determined.

The second type of technique requires knowing the correspondences for sufficient number of points in the image to determine the three-dimensional structure and motion of the objects in the scene. The general idea is that displacement of each image point is a function of the motion parameters (six in number) and the depth of the point. Therefore, in order to be able to solve for these unknown parameters and depths, we need only to obtain a sufficient number of points and their displacements corresponding to the same rigid object in the scene. Several well-known algorithms are available for solving this problem.

The third type of technique requires an optical flow fields. There are two ways in which the optical flow fields can be used in this process. The local derivatives of the flow vectors can be used to provide information about the structure and motion of the object. Alternatively, some global measures of the flow vectors can be used by taking into account the fact

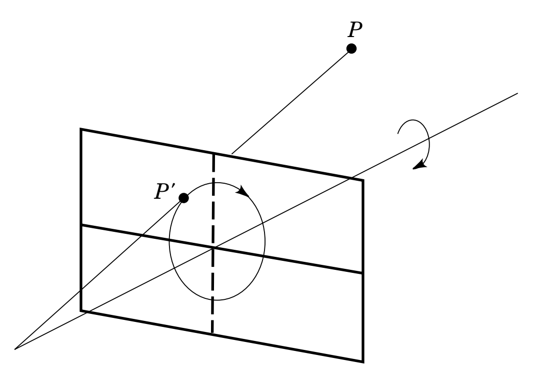


**Figure 5.** The motion of camera is a pure rotation about an arbitrary axis.

that the motion parameters are the same for an entire rigid object, and attempt to recover them. Several existing techniques for dealing with this problem are also available.

## ACTIVE PERCEPTION

Active perception leads naturally to exploration and mobility. Perception is a constructive but controlled process, and active perception can help us fill in missing information. Clearly, the whole sensing and perceptual process are actively driven by cognitive process incorporating *a priori* knowledge. The vision system receives feedback and actively follows up by seeking novel information. The exploratory behavior alluded to is one of the characteristics of dynamic vision. Furthermore, dynamic vision is also characterized by flexible perception, whereby hierarchical modeling can prime the operation and integration of modular processes toward actual recognition. This later aspect is also referred to as functional perception.

With a single perspective projection image, limited three-dimensional information may be derived about the objects by using techniques, such as shape from shading, shape from texture, and so on. If two cameras are placed apart to take a pair of perspective projection images, the depth information can be recovered for every image point by using the models developed in stereo vision. Furthermore, if we have only one camera, but we can move it around an environment, then two or more images can be acquired along the path at different positions during the camera movement. Motion analysis techniques using pairs of images are designed to process images that contain significant changes from one to another between views. These large changes force the techniques to tackle the difficult problem of stereo correspondence. On the other hand, if we take a sequence of images from positions that are very close together, this sampling frequency guarantees a continuity in the temporal domain that is similar to continuity in spatial domain. Thus, an edge of an object in one image appears temporarily adjacent to its occurrence in both preceding and following images. This temporary continuity makes it possible to construct a solid of data in which time is the third dimension and continuity is maintained over all three dimensions. By slicing the spatiotemporal data along a temporal dimension, locating features in these slices, we can compute three-dimensional positions of the objects.

We illustrate the above idea by moving a camera along a straight line. First consider two general arbitrary positions of the camera (Fig. 6). The camera is modeled by a center of projection and a projection plane in front of it. For each point $P$ in the scene, there is a plane, called an epipolar plane, which passes through the point $P$ and the line joining the centers of the two projections. The epipolar plane intersects with the two image planes along epipolar lines. All the points in the scene that are projected onto one epipolar line in the first image are also projected onto the epipolar line in the second image. The importance of these epipolar lines is that they reduce the search required to find matching points from two dimensions to one dimension. That is, to find a match for a given point along one epipolar line in an image, it is only necessary to search along the corresponding epipolar line in the other image. This is termed the epipolar constraint.

Now consider a simple motion in which the camera moves from right to left along a straight line, with its optical axis

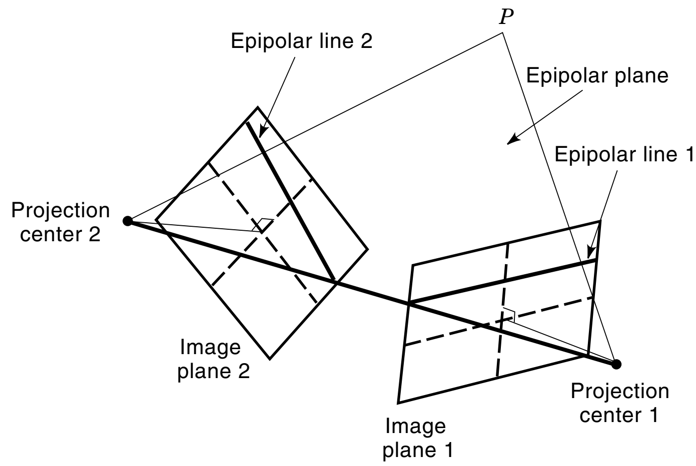


**Figure 6.** A configuration illustrates the definitions of epipolar plane and epipolar lines.

orthogonal to its direction of motion (Fig. 7). For this type of motion, the epipolar planes for a point $P$ in the scene is the same for all pairs of camera positions. Furthermore, the epipolar lines associated with one epipolar plane are horizontal scan lines in the images. The projection of $P$ onto these epipolar lines moves to the right as the camera moves to the left. The velocity of this movement along the epipolar line is a function of $P$'s distance from the line joining the projection centers. The closer the distance is, the faster the projection point moves. Therefore, a horizontal slice of the spatiotemporal data formed from this motion contains all the epipolar lines associated with one epipolar plane (Fig. 7). This type of slice is called epipolar plane image.

There are several things to notice about the epipolar image. First, it contains only linear structures. Given a point $P$ in the scene, let $P'$ be the projection point of $P$ on the image plane. Assume $P$ is not moving. When the camera translates from right to left along a straight line, the projection point $P'$ shifts from left to right. In the epipolar plane image, the shift of $P'$ as a function of time forms a sloped line. Second, the slopes of the lines determine the distances to the corresponding points in the scene. The greater the slope, the farther the point.

The $(x, y, z)$ location of a point $P$ in scene can be derived as the follows. Figure 8 is a diagram of a trajectory in an epipolar plane image derived from the right-to-left motion illustrated in Fig. 7. The scanline at $t_1$ in Fig. 8 corresponds to the epipolar line $l_1$ in Fig. 7. Similarly, the scanline at $t_2$ corresponds to the epipolar line $l_2$. The point $(u_1, t_1)$ in the epipolar plane image corresponds to the point $(u_1, v_1)$ in the image taken at time $t_1$ at position $c_1$. Thus, as the camera moves from $c_1$ to $c_2$ in the time interval $t_1$ to $t_2$, the point $P'$ moves from $(u_1, t_1)$ to $(u_2, t_2)$ in the epipolar plane image.

Given the speed of the camera, $s$, which is assumed to be constant, the distance from $c_1$ to $c_2$, $\Delta x$, can be computed as follows:

$$\Delta x = s\Delta t \tag{11}$$

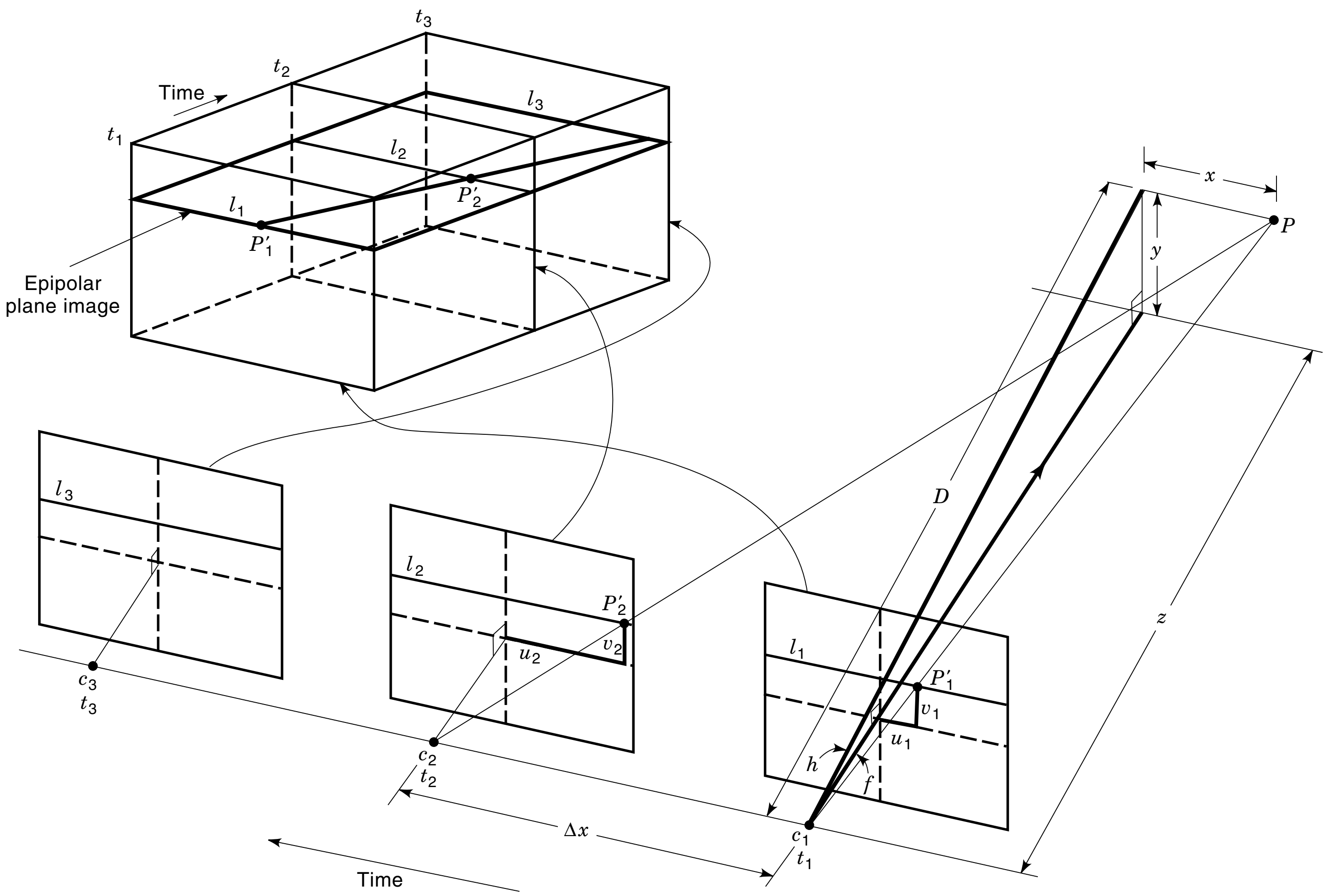


**Figure 7.** When the camera translates from right to left, the image point shifts from left to right.

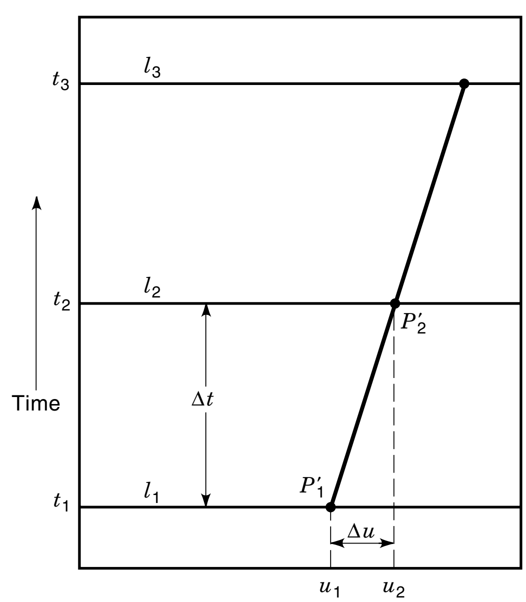


**Figure 8.** The trajectory an image point on an epipolar plane image.

where $\Delta t = (t_1 - t_2)$. By similar triangles (see Fig. 7) we obtain

$$\frac{u_1}{h} = \frac{x}{D} \tag{12}$$

$$\frac{u_2}{h} = \frac{\Delta x + x}{D} \tag{13}$$

From the above two equations, we can derive

$$\Delta u = (u_2 - u_1) = \frac{h}{D}\Delta x \tag{14}$$

Thus, $\Delta u$ is a linear function of $\Delta x$. Since $\Delta t$ is also a linear function of $\Delta x$, $\Delta t$ is linearly related to $\Delta u$, which means that trajectories in an epipolar plane image derived from a lateral motion are straight lines. The slope of line corresponding to a point $P$ in the scene is defined by

$$m = \frac{D}{h} = \frac{\Delta x}{\Delta u} \tag{15}$$

From similar triangles, the $(x, y, z)$ position of $P$ can be obtained by

$$(x, y, z) = \left(\frac{D}{h}u_1, \frac{D}{h}v_1, \frac{D}{h}f\right), \text{ or} \tag{16}$$

$$(x, y, z) = (mu_1, mv_1, mf) \tag{17}$$

Similar analysis can be applied to other types of camera motions to derive the corresponding trajectories in the epipolar plane images and to find the formula for solving three-dimensional positions. However, the formula may be very complicated depending on the type of motion.

The epipolar plane analysis method described above has assumed that the object is not moving. Tracking of moving objects, though still in the developmental stages, is becoming increasingly recognized as important capabilities in vision systems. An active camera tracking system could operate as an automatic cameraperson. It is hoped that tracking combined with other technologies can produce effective visual serving for robotics in a changing work cell. Recent research indicates that tracking facilitates motion estimation.

***Reading List***


D. H. Ballard and C. M. Brown, *Computer Vision,* Englewood Cliffs, NJ: Prentice-Hall, 1982.

S. T. Barnard and W. B. Thompson, Disparity analysis of images, *IEEE Trans. Pattern Anal. Mach. Intell.,* **PAMI-2**: 333–340, 1980.

P. J. Burt, C. Yen, and X. Xu, Multi-resolution flow-through motion analysis, *Proc. IEEE CVRP Conf.,* 1983, pp. 246–252.

O. Faugeras, *Three-Dimensional Computer Vision,* Cambridge, MA: MIT Press, 1993.

Fisher and Firschein (ed.), *Reading in Computer Vision: Issues, Problems, Principles and Paradigms,* Los Altos, CA: Morgan Kaufmann, 1987.

R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision,* Vol. II, Reading, MA: Addison-Wesley, 1993.

R. A. Hummel and S. W. Zucker, On the foundations of relaxation labeling process, *IEEE Trans. Pattern Anal. Mach. Intell.,* **PAMI-5**: 267–286, 1983.

D. T. Lawton, Processing dynamic image sequences from a moving sensor, Ph.D. thesis, COINS Department, University of Massachusetts, TR 84-05, 1984.

D. Marr, *Vision,* San Francisco: Freeman, 1982.

D. Murray and A. Basu, Motion tracking with an active camera, *IEEE Trans. Pattern Anal. Mach. Intell.,* **PAMI-6**: 449–459, 1994.

J. W. Roach and J. K. Aggarwal, Determining the movement of objects from a sequence of images, *IEEE Trans. Pattern Anal. Mach. Intell.,* **PAMI-2**: 554–562, 1980.

R. Y. Tsai and T. S. Huang, Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces, *IEEE Trans. Pattern Anal. Mach. Intell.,* **PAMI-6**: 13–27, 1984.

C. Torras (ed.), *Computer Vision: Theory and Industrial Applications,* New York: Springer-Verlag, 1992.

S. Ullman, *The Interpretation of Visual Motion,* Cambridge, MA: MIT Press, 1979.

H. Wechsler, *Computational Vision,* San Diego: Academic Press, 1990.

X. Zhuang, T. S. Huang, and R. M. Haralick, A simplified linear optic flow-motion algorithm, *Comput. Vis., Graphics Image Process.,* **42**: 334–344, 1988.


XUE DONG YANG
University of Regina

**MOTION ANALYSIS, HUMAN.** See HUMAN MOTION ANALYSIS.

**MOTION CONTROL, ROBOT.** See ROBOT PATH PLANNING.

**MOTION SENSORS.** See TACHOMETERS.

**MOTOR DISABILITIES.** See ASSISTIVE DEVICES FOR MOTOR DISABILITIES.

**MOTOR DRIVES.** See SWITCHED RELUCTANCE MOTOR DRIVES.