DYNAMIC SCENE ANALYSIS

J. K. Aggarwal[1,2]
W. N. Martin[2]

TR-82-3 September 1982

Department of Electrical Engineering[1]
Laboratory for Image and Signal Analysis[2]
The University of Texas at Austin
Austin, Texas 78712

This research was supported by the Air Force Office of Scientific Research under grant AFOSR 82-0064.

ABSTRACT

The three major components of dynamic scene analysis, namely segmentation, occlusion and the computation of three-dimensional information from images are discussed in depth. Segmentation refers to the process of determining features of interest, occlusion analysis includes the deriving of changes due to projection perspective, and computation of three-dimensional information entails the constructing of structural models and describing motions from image information. The purpose of the review is to give the reader a coherent view of the issues and the manner in which researchers are currently addressing these issues. Detailed descriptions of the systems developed at The University of Texas are presented.

1.0 Introduction

Dynamic Scene Analysis, also referred to as the Analysis of Time-Varying Imagery, is concerned with the processing of a sequence or a collection of images. The ultimate goal of the analysis is to assimilate information from the sequence as a whole that cannot be obtained from any one image by itself. The sequence of images usually represents a scene as sampled by a sensor at instants close in time and may arise from a variety of scenarios. Examples include motions of objects in a scene where the sensor is fixed, motion of the sensor relative to the scene or a combination of the two motions. A variety of applications have motivated the present research in Dynamic Scene Analysis. These include industrial automation and inspection, robotics, navigation, automatic surveillance and biomedical engineering. The research area of Dynamic Scene Analysis is rather new, however, it is receiving considerable attention as is evident from the Advanced Study Institute at Braunlage, other conferences and the recent literature [1-8].

The present paper addresses the important issues and ingredients of Dynamic Scene Analysis. Specifically, the three issues discussed in this review paper are segmentation, occlusion and the computation of three-dimensional information from images. Here segmentation refers to the process of determining features of interest together with distinguishing interesting changes from uninteresting changes and relating the features and components in one image to those of the succeeding images. Occlusion analysis includes deriving structural changes due to projection perspective and the appearance or disappearance of objects. Finally, the computation of three-dimensional information entails constructing structural models and describing three-dimensional motions through the analysis of two-dimensional image information. The next three sections discuss Dynamic Scene Segmentation, Occlusion in Image Sequences and Three-Dimensional Information from Images.

2.0 Dynamic Scene Segmentation

In almost every static scene analysis task, the first step is segmentation, i.e., to locate the significant scene components, to extract features from the image, or to separate the image into meaningful regions. Dynamic scene analysis is no different but must also consider dividing the images into parts that are changing and parts that are constant, or finding the the moving parts in each element of the sequence of images. In order to account for change and movement, information must be combined from consecutive frames or subsequences of images.

There are two distinct approaches to segmentation in dynamic scenes: Feature Based Segmentation and Pixel Based Segmentation. Feature based segmentation consists of finding edges, corners, boundaries, or surfaces in each of two images and then establishing a correspondence between various features in the two images. The process of establishing correspondence is at times difficult, especially if one has noisy images. Thus, the analysis proceeds with the static scene segmentation of each of the two images, and then establishes a feature correspondence between consecutive images to determine the changes in the images. Pixel based segmentation compares the two images at the pixel level by methods such as differencing, correlation or temporal-spatial gradient. In each case, pixel level comparisons are made and velocity estimates are assigned to various pixel positions. The velocity estimates become the basis for segmentation.

Neither type of process yields unique answers but generally the end product is a description of the moving parts of a scene. Each of the approaches makes the assumption that image components which move together are parts of the same underlying object in the scene. The above procedures are illustrated in the following discussion.

2.1 Feature Based Analysis

A variety of features have been used in segmenting each of the images and in establishing correspondence for the moving parts in a sequence of images. The list of features includes corners, straight edges, curvilinear edge segments, centroid, area, major and minor axes for moment of inertia, and others. The choice of features depends upon the problem domain and the assumptions that may be made about the moving parts. Two systems are briefly discussed to illustrate feature based analysis.

Corners and Straight Edges: Aggarwal and Duda [9] consider the motion of polygonal figures which are arbitrarily complex in shape, and possibly contain holes. In this case the polygons are software generated to appear as planar objects moving in planes parallel to the image plane. The parallel projection essentially creates the silhouette of the objects.

The overlapping of the actual polygons creates new vertices while removing occluded vertices and edges. The new vertices are referred to as "false" vertices and the visible vertices of the actual polygons are called "real" vertices. One of the main functions of the system is to classify the vertices of the input image into the appropriate one of these two categories. This classification process is facilitated by two characteristics of the input domain. First, no "false" vertex can have an interior angle which measures less than 180 degrees. Second, any vertex which changes its angular measure between two frames must be a "false" vertex. The first characteristic is due to the polygonal nature of the objects, while the restriction to rigid polygons assures the second. However, these two characteristics do not provide enough information to directly classify every vertex. There are vertices with obtuse interior angles which are not "false" vertices and there are "false" vertices which do not change their angular measure. One further restriction is necessary, and it is that no more than one "real" vertex can appear or become occluded between any two consecutive frames. The imoortance of this restriction is that it allows the system to determine the type of change that has occurred between two consecutive frames. This determination is based on the difference in the number of vertices having acute interior angles along with the difference in the number of vertices having obtuse interior angles.

The correspondence is established based on the nature of the vertex (i.e., acute or obtuse), the lengths of the polygons' sides, etc. In this fashion, the moving parts are isolated for further processing. Figure 2.1 shows an example of a sequence of images to which this process has been applied.

Curvilinear Boundaries: The system of Martin and Aggarwal [10] analyzes scenes containing figures with curvilinear boundaries in a manner similar to the system just described. The input is again restricted so that the objects move independently in planes parallel to the image plane. However, instead of software-generated images, homogeneously shaded, opaque, planar figures are moved in front of a TV camera to produce a sequence of images. The camera approximates an orthogonal projection into the digital images which are preprocessed to extract the boundaries of the figures [11]. The figure shading and the camera setup give rise to images in which overlapping figures are merged into single apparent objects. The task of the system is thus to derive descriptions of the

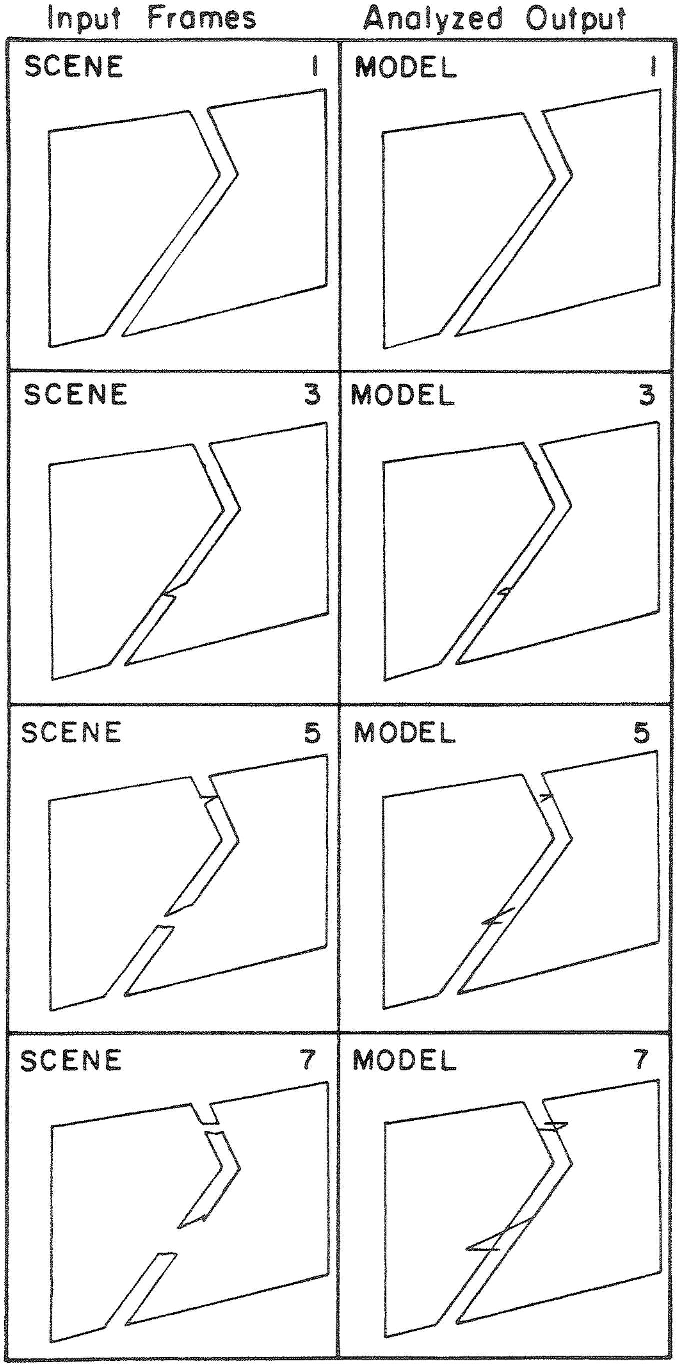


Figure 2.1 A sequence of polygonal objects illustrating feature based segmentation.

constituent actual figures and their motions by analyzing the apparent objects of the sequence of images. The analysis of the sequence is performed on pairs of consecutive images from the sequence and is based upon identifying shapes which are common to both images of any given pair. The matched shapes are interpreted as two views of the same object. In this way the moving objects can be tracked throughout the sequence while motion measurements are made from the displacements between the matched views.

The tokens used by this system are circular arcs approximated by portions of the object boundaries. The arcs are derived by analyzing the subtended angle versus arc length, ψ-s, a function of the boundary as measured from an arbitrary starting point on that boundary. This function is useful because intervals of constant slope in the ψ-s function correspond to boundary sections of constant curvature, i.e., circular arcs. The appropriate intervals are determined by forming a piecewise straight line approximation of the pictorial graph of the ψ-s function. The set of straight lines in the ψ-s function approximation effectively decomposes the object boundary into a set of arcs. Figure 2.2 shows an object before and after being segmented into arcs by this process.

The shape representation, as entered in the data base which contains all the relevant information derived from the sequence of images, includes the coordinate list of the object boundary, the straight-line description of the ψ-s function, and pointers relating specific boundary sections to the appropriate elements of the straight-line set. This representation separates clearly the information needed for the shape matching from the information required in the movement measurement process. In fact the ψ-s function is invariant to translation and rotation (see [10] for minor qualifications) and is processed to eliminate the effects of arbitrarily choosing its starting point. This separation is in accordance with the system's use of the constancy in shape of the actual figures in order to interpret the movement of the apparent objects.

The initial correspondence is based on matching the tokens through their shape attributes, but is again aided by the higher level constraint imposed by the token ordering along object boundaries. Contiguous arcs from one image which match, in the same order, contiguous arcs from the second image are grouped into edge segments. This matching is performed by first choosing two arcs, one from each image of a consecutive pair, whose ψ-s function lines

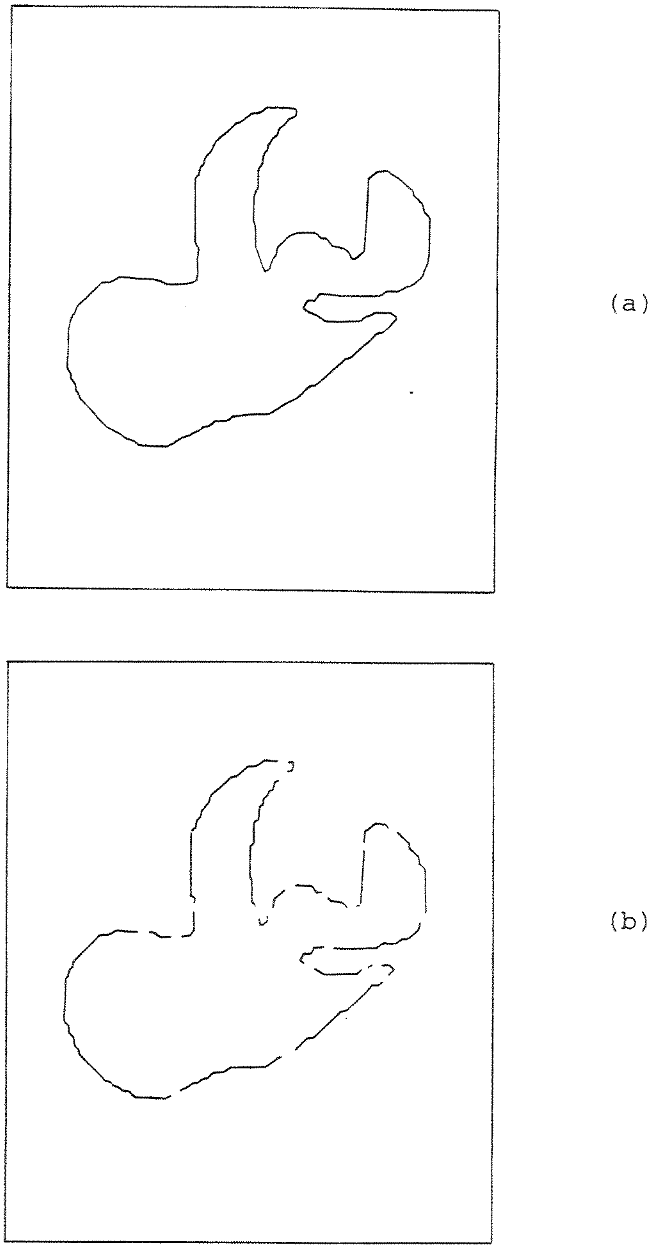


Figure 2.2 (a) A curvilinear object.
(b) Feature based segmentation into circular arcs and straight lines.

have similar slopes and lengths. From these "seed" arcs an edge segment can be "grown" by adding continuous arcs to either end of the already matched segments until a dissimilarity in the curves is found. The dissimilarity of two curves is measured by the area between the normalized pictorial graphs of their ψ-s functions. Two arcs are declared dissimilar when the measured value exceeds a preset threshold.

Edge segments grown in this way represent the portions of the object boundaries which have retained their shape through the sequence. Thus an edge segment relates two views of some part of an actual figure. The displacement between two such views provides motion measurements for the given edge segment. These measurements are then used to group the edge segments into object models under the assumption that edge segments which exhibit a common motion belong to the same object. Figure 2.3 shows two input frames and the derived models.

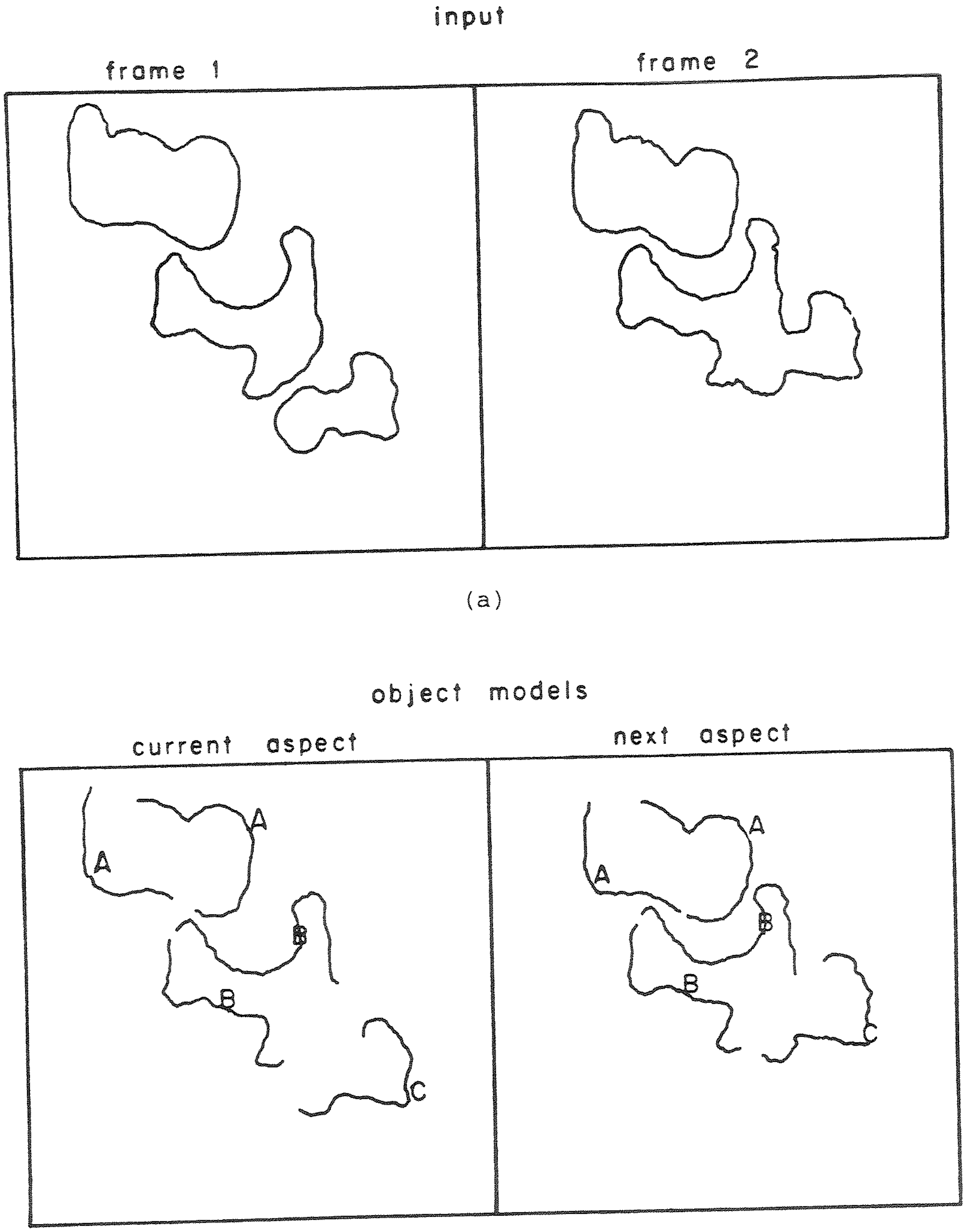
In contrast to the above research where detailed analysis is carried out on sections of the object boundaries, Chow and Aggarwal [12] compute measures over the complete boundaries, e.g., the centroid, major and minor axes, etc. These measures are used in conjunction with a predictive scheme to perform the dynamic scene segmentation for blob-like figures.

2.2. Pixel Based Analysis

The techniques of differencing and cross-correlation have been used extensively for comparing a pair of given images in several applications. In addition, in time-varying imagery, techniques based on temporal-spatial gradients have been developed. These techniques will be discussed in the following sections.

Differencing: A comparison of two images of a scene will indicate that the images differ in areas affected by the motions of objects. One method of comparison is to "subtract" or "difference" two images and record the results in another image referred to as the difference image. If the absolute value of the difference is above a preset threshold, the corresponding pixel is set to 1; otherwise it is set to 0. On analyzing a pair of images of a scene, it is assumed that all points in the two images and those of the difference image are referenced to a common grid, i.e. the three images are assumed to be registered.

Several researchers have used the difference images to characterize objects and their motions. For example, references [13-16]



(a)

(b)

Figure 2.3 (a) Input images.
(b) Object models based on feature based segmentation and velocity measurements.

describe a variety of results obtained. The differencing technique has been applied to synthetic scenes, laboratory scenes, and real world scenes. In general the technique is applicable to both polygonal and curvilinear objects. However, theoretical analysis is available only for the case of polygonal objects. For best results, the objects are assumed to be of a homogeneous gray level. If an object of interest were to comprise areas having distinct but uniform gray levels, each homogeneous portion could be treated as an individual object. It would then remain for some higher level process, possibly using a common motion constraint to determine that all the individually identified segments were parts of the same object.

By examining the difference image, object samples can be determined and then expanded, e.g. "grown" [13,14], to effectively include all image points that correspond to the object. An example illustrating the use of differencing is given in Figure 2.4.

Cross-Correlation Analysis: Given two images and a small window in the first image, the purpose of cross-correlation is to find the region in the second image that matches the windowed region of the first image. This situation is illustrated in the Figure 2.5a,b with the matching accomplished as in template matching.

Let the second image be denoted by f(x,y), the window of the first image by w(x,y), and the shifted version of the window by w(x-m,y-n), then their cross correlation function may be defined as

$$R(m,n) = \sum_x \sum_y f(x,y) w(x-m,y-n)$$

where the summation is taken over the region where w(x,y) is defined, and (m,n) vary over the entire image f(x,y). The range of the summation and the positions of the window and figure are illustrated in Figure 2.5c. As (m,n) vary, R(m,n) changes and reaches a maximum at the place where w(x,y) best matches f(x,y). A more complicated correlation function is given by

$$R'(m,n) = R(m,n)/D(f)$$

$$\text{with } D(f) = \sqrt{\sum_x \sum_y f^2(x,y)}$$

where both summations are defined over the region where w(x-m,y-n) is non-zero. The denominator varies with the position of

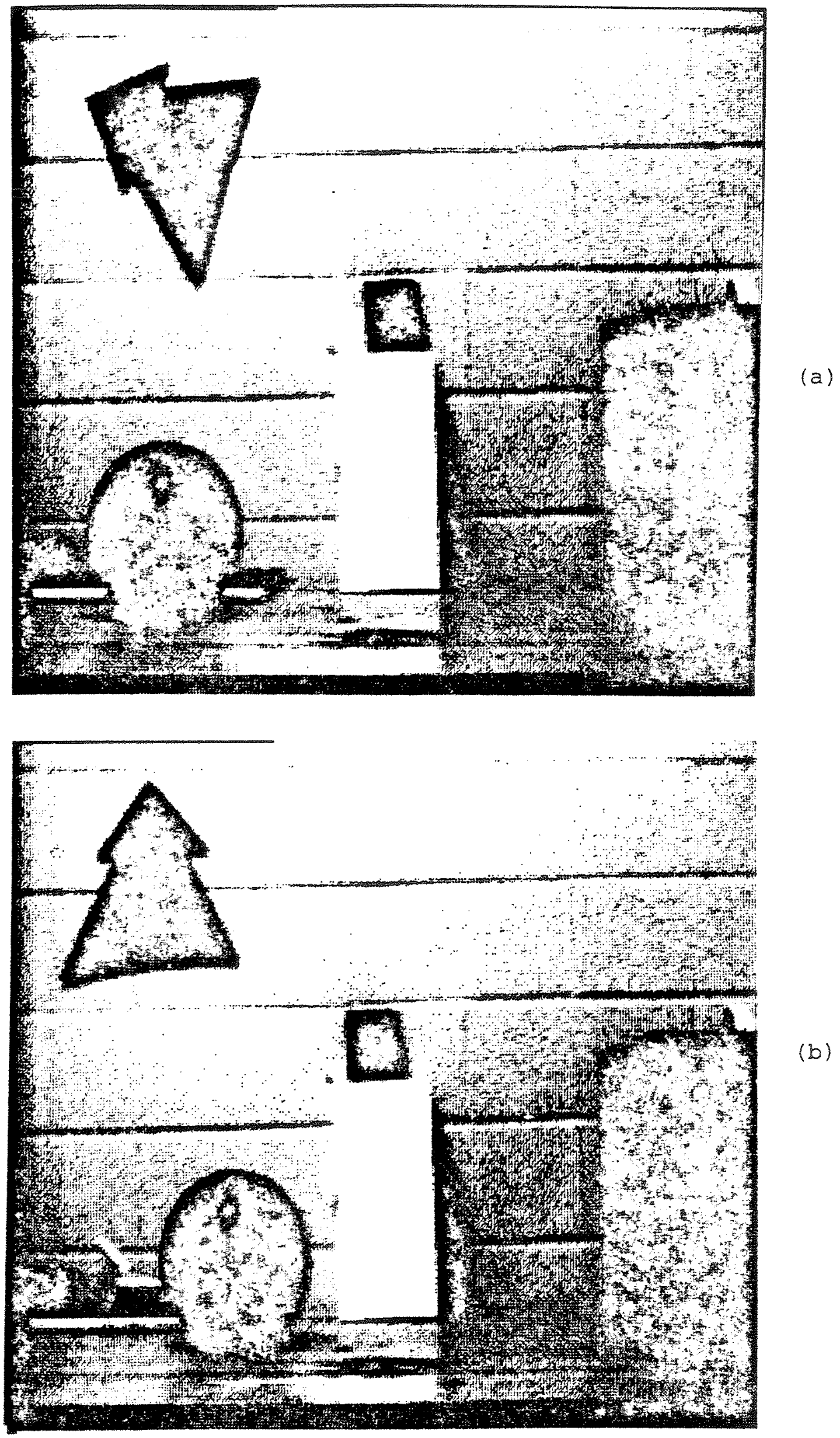


Figure 2.4 (a),(b) Two images of a laboratory scene.

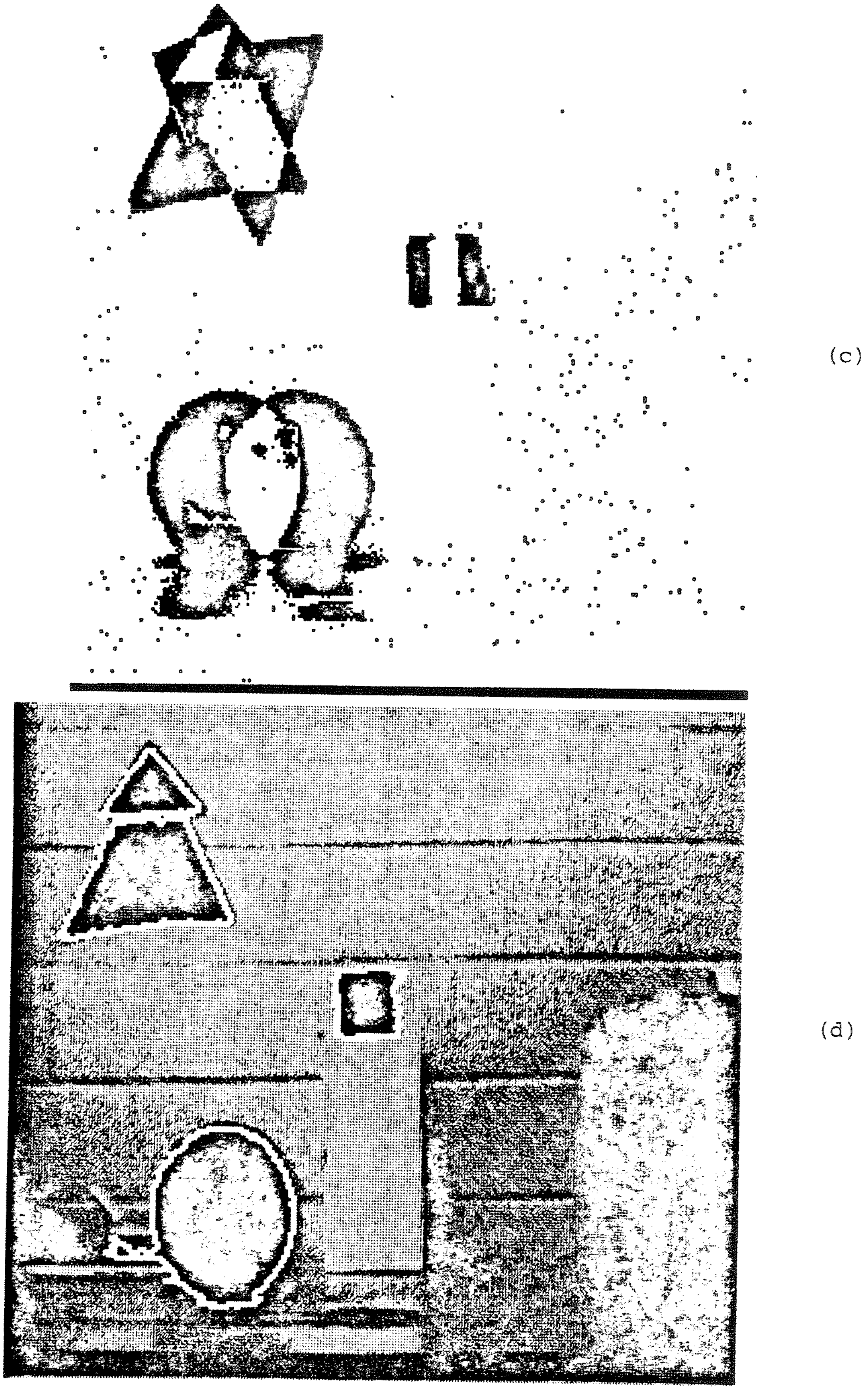


Figure 2.4 cont. (c) The difference picture of the two images.
(d) The results of the region growing process using the derived object samples.

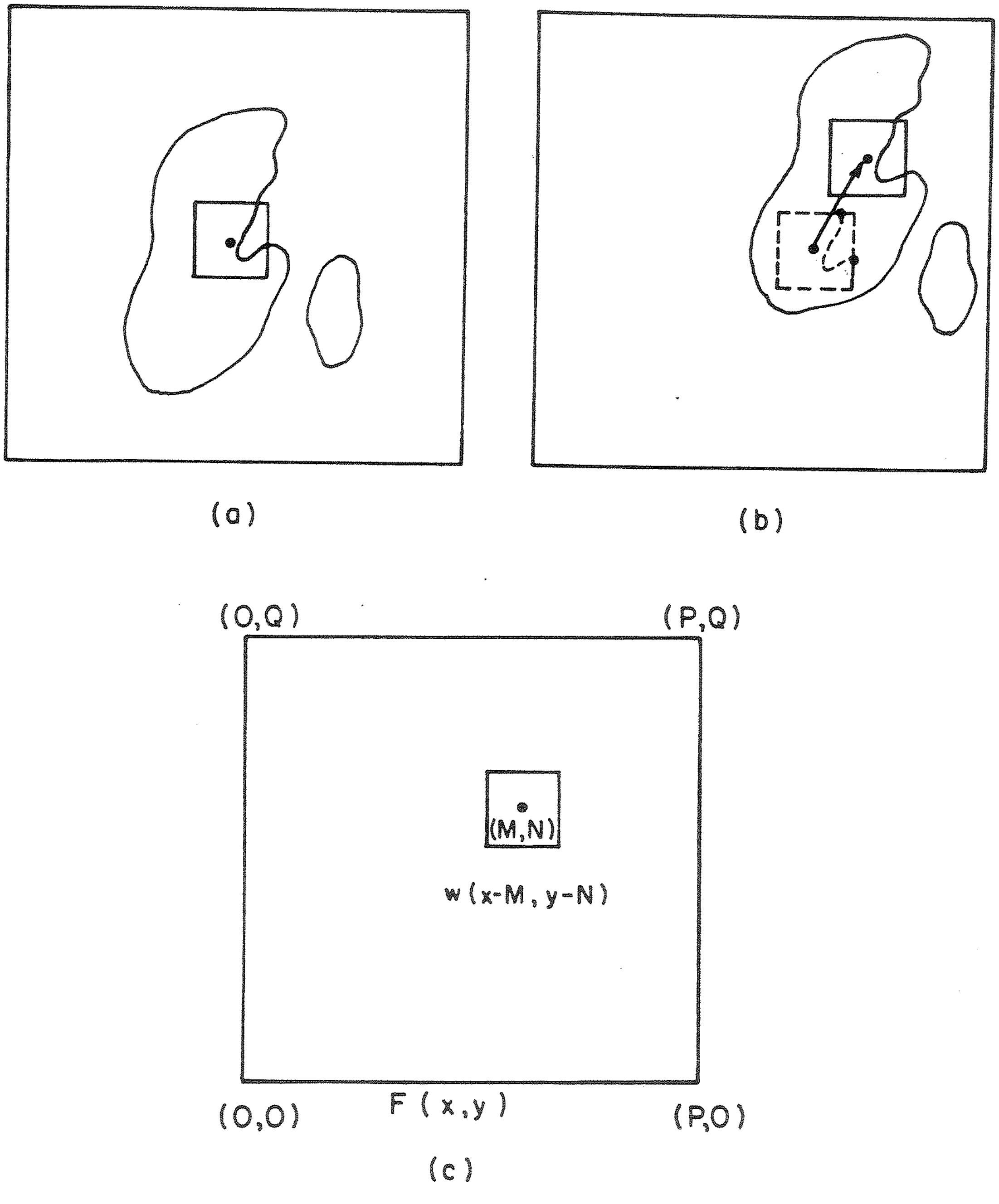


Figure 2.5 (a),(b) Two images of curvilinear objects with the windows indicating the matched regions, and the arrow denoting the movement of the object.
(c) The range of summation and the position of the window.

w(x-m,y-n) and it tends to sharpen the peaks of R(m,n). Details on cross-correlation are found in the book [17], whereas the application of cross-correlation to cloud motion analysis is found in [18,19].

Temporal-spatial Gradient: Let $f_1(x,y)$ and $f_2(x,y)$ denote the image intensities at the two instants, t_1 and t_2. During the intervening time interval, the image has moved by the amount Δx and Δy in the x- and y- directions, respectively. Now

$$\Delta f(x,y) = f_2(x,y)-f_1(x,y)$$

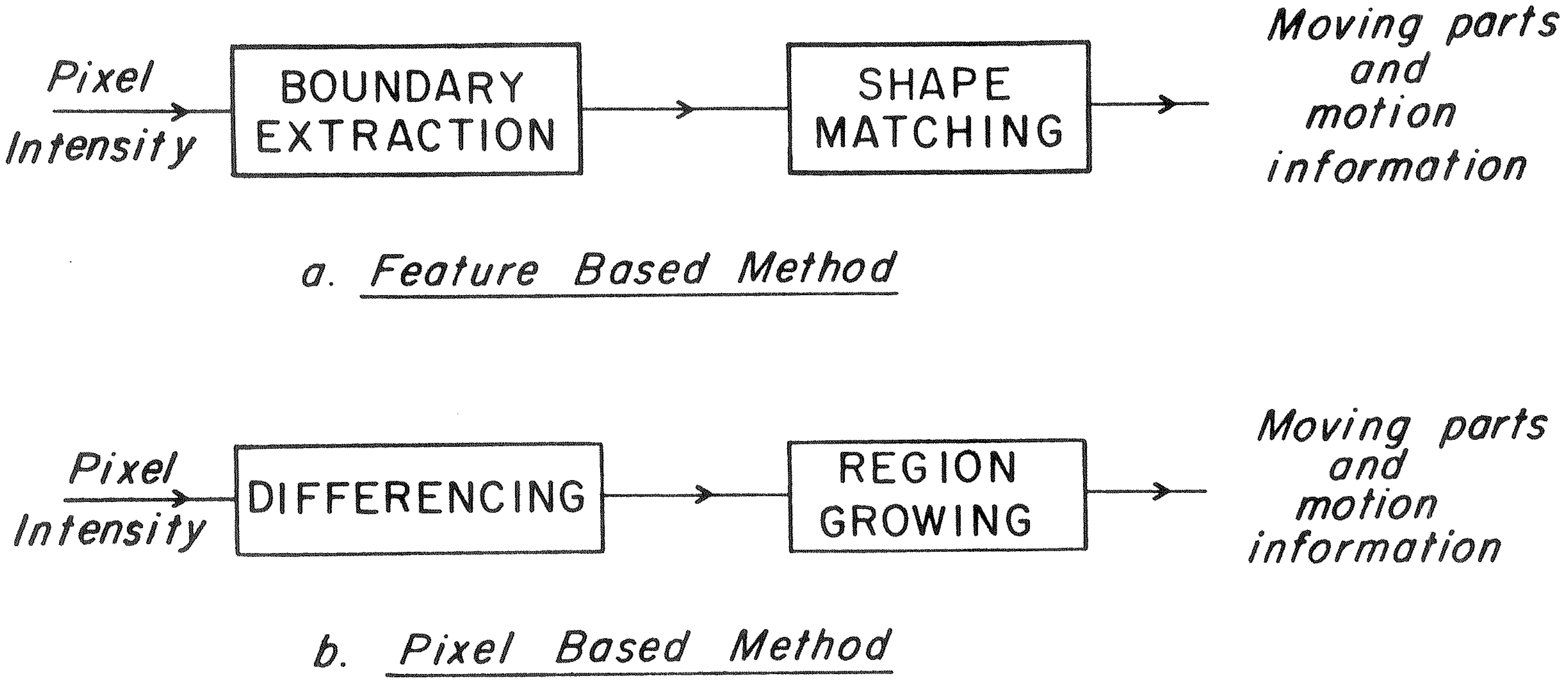
also $$f_1(x,y) = f_2(x+\Delta x,y+\Delta y)$$

therefore $$\Delta f(x,y) = f_2(x,y)-f_2(x+\Delta x,y+\Delta y)$$

Thus, at each point one can calculate Δf, $\partial f/\partial x$ and $\partial f/\partial y$; and obtain a linear equation for Δx and Δy . If one has two points, one can obtain two linear equations, and their solution would lead to the determination of Δx and Δy provided the equations were independent. In practice, one uses considerable redundancy in the number of points and obtains a least mean square solution. An excellent set of examples illustrating segmentation using temporal-spatial gradient and contrast is given by Thompson [20].

2.3 General Observations on Dynamic Scene Segmentation

Two broad approaches toward dynamic scene segmentation based on pixel and feature have been briefly presented and illustrated by example systems. The end product of the two approaches are similar, i.e., the location and the motion of moving parts in the sequence of scenes. It may be observed that the feature based methods must be supplemented by a correspondence procedure, whereas correspondence is a by-product of the pixel based analysis. This apparent superiority of pixel based techniques is quickly defeated by the simple observation that in the presence of occlusion, pixel based methods often break down. The same is true in case of structural changes in the objects. A graphical comparison of two types of methods is presented in Figure 2.6.



a. Feature Based Method

b. Pixel Based Method

Figure 2.6 Examples of methods for the extraction of moving parts.

3.0 Occlusion

Occlusion occurs whenever the image to be analyzed is a projection of some three-dimensional scene onto a two-dimensional plane. In this general case there is always a background obscured by the objects which are considered to be the foreground. For objects widely spaced over a homogeneous background, e.g., paintings on a museum wall or a pair of birds flying in a clear sky, there is no problem in understanding the image. The background is understood to be homogeneous so that the characteristics of the obscured portions are indicated by the visible sections. The foreground objects are assumed to have image characteristics which are distinct from the background making the foreground objects readily detectable in the image. In addition, the spacing of the objects assures that the presence of the features on one object will not interfere with the analysis of the remaining foreground objects. However, if the background has a complex structure, e.g., the museum wall has a highly patterned covering, or if the foreground objects are closely arranged in some structure, e.g., a flock of birds flying in the same direction, then the classic "figure-ground" problem arises, for example see [21]. In the figure-ground problem, the spatial relationships between disjoint elements of the viewed scene combine to interfere with the perception of the individual elements. In its full generality, this is a psychophysical problem where the preconceptions and expectations of the viewer play an important part in perception. For a fuller discussion of this topic, the reader is referred to [22].

3.1 Scene Domain Imposed Constraints

In abstract geometrical patterns, both the quantity and subtlety of the inter-element relationships are greater than those occurring in typical natural scenes. Similarly, the constraints imposed by the three-dimensional structure and distribution of the objects appearing in typical scenes are greater than those in abstract patterns. For instance, the boundary edges of noncontiguous objects are rarely collinear in natural scenes. This consideration makes reasonable the assumption that if the image of a scene contains disjoint edges which are collinear, then those edges correspond to a single boundary in the scene and the discontinuity is caused by the boundary being partially obscured in the given view. Barrow and Tenenbaum [23] argue that certain psychological phenomena, such as subjective contour, are the result of the human visual system attempting to use such evidence of occlusion as a cue to apparent depth.

An elegant example of how scene domain constraints can be used in understanding occlusion is the system developed by Waltz [24]. In this case, the domain is that of scenes having a single light source illuminating a set of planar-faced objects whose vertices are trihedral. The strong constraints imposed by this scene domain are primarily embedded in a junction classification and line labeling scheme generalized from the system first discussed by Huffman [25] and Clowes [26]. Junctions are the line drawing representations of the vertices in the scene, thus the trihedral restriction of the object vertices provides extensive constraints on both the types of junctions possible and the allowable labelings of the lines forming those junctions. In particular, certain of the labeled junction types can only arise through cases of occlusion, and thus, when found in the drawings, provide a reliable indication of occlusion.

Many pictures are inherently ambiguous, and no information derived from the image can resolve the uncertainties. Frequently, it is not that the image has no consistent interpretation, but rather that there are several mutually exclusive interpretations which are each independently consistent. The choice among such alternatives must be based on the expectations or goals of the viewer, not simply on features actually exhibited in the image.

Several factors are fundamental to the understanding of scenes containing occluding objects. First, the concept of occlusion is used at a very early stage in the human visual system in order to provide interpretations in terms of apparent depth. Second, effective cues to occlusion can be derived from scene-domain constraints. And third, occlusion necessarily results in the loss of information available about the obscured object, thus causing uncertainties in the interpretation of the image. Finally, it may be observed that the use of occlusion cues may involve the complex integration of information taken from areas which are widely separated in the image and that the resolution of some occlusion ambiguities depends on external expectations and goals.

3.2 Occlusion in Image Sequences

The discussion up to this point has dealt with the implications of occlusion on the analysis of simple images. For the remainder of this section the focus will be on time-varying images. The question addressed is: How is the complexity of the occlusion analysis problem affected by the addition of time variation? The broad answer to this question is that time variation simplifies some aspects of the problem, complicates other aspects, and introduces several new problems. These points are discussed on a general level in the following.

The time variation can simplify the initial feature extraction phase of processing through both the redundancy inherent in the dynamic scenes and the opportunities provided for acquiring new information. Typically the sampling rate along the time axis is such that the majority of the scene does not change through short sequences of images. This property has been exploited for data reduction by frame-to-frame encoding of video signals [27], but can also be used to attenuate noise and produce more reliable feature values. The new information can be obtained from the changing views of the objects in the scene. For instance, if one of the occluding objects in the foreground is moving, then additional portions of the objects that it is obscuring will become visible in each successive image. Similarly, since any three-dimensional object is self-obscuring, the object's motion will usually bring into view previously unseen portions of the object. This concept has been used in a system [28] which forms a description of a planar-faced object from a sequence of views taken while the object rotates. The description is in terms of the object faces and their interconnections, which are "learned" as previously hidden faces become visible. New views also result from changes in the orientation of the image plane caused by eye (camera) movement. In these situations, areas of ambiguity in a given image may be clarified by the additional information contained in the subsequent images.

The continual change in the information content of the images, which is an advantage when the change adds information, can be a disadvantage when the change results in a reduction of available information. In each of the information-adding cases discussed above, there can be a complementary aspect in which information is lost. For instance, the moving foreground object is probably proceeding to obscure some other objects or even other portions of the same object that it is elsewhere uncovering. This aspect raises the question as to what can be said about previously visible features once they are no longer visible. If a recently obscured feature is part of an object which is still partially visible, then the relationship of the feature to the currently visible portion, as determined in preceding images, can be used to infer the location and orientation of the feature in the present scene. This type of implication is based on the assumption that the object is rigid and thus that the spatial relationships of the various features of an object will remain constant through time. This is an extremely important scene-domain constraint.

The information flux in time-varying images also creates new problems at the image segmentation and object identification levels. The problems encountered here involve the additional "semantic noise" [29] exhibited in time-varying images. Typical systems for static image analysis must be capable of interfacing with preprocessors which occasionally fail to detect, erroneously produce, or incorrectly locate image feature descriptions. Systems for time-varying images will have similar preprocessing problems but must furthermore be prepared to interpret features which, through time, may take on different values yet signify the same scene component semantically. For example, the effects of shadows on a textured outdoor surface, e.g., a gravel road bed, will vary as the sun angle changes throughout the day.

This problem of identifying "apparently different but semantically identical objects" [30] indicates a fundamental concept in the analysis of time-varying images: in order to understand the changes that a given aspect of an entity in a scene may be undergoing, there must be some form of constancy in other aspects of that same entity to serve as the identifying features of the entity. This is particularly important when there are several objects moving about the scene, because the simple detection of change cannot attribute that change to the proper object.

As an illustration, consider the illusion depicted in Fig. 3.1. Here four identical disks are attached pairwise to the ends of two cross members which are slightly offset in depth and spin in opposite directions about the center point. They exhibit constancy in both shape and color. These features make it easy to track the disks while they are moving through positions such as that of Fig. 3.1a. However, when the position shown in Fig. 3.1b is reached, the constancies no longer serve as identifying features and thus admit an ambiguity to the interpretation of the position displayed in Fig. 3.1c.

Is the pair of velocities labeled A in Fig. 3.1c the correct interpretation or is the pair labeled B correct? An assumption of

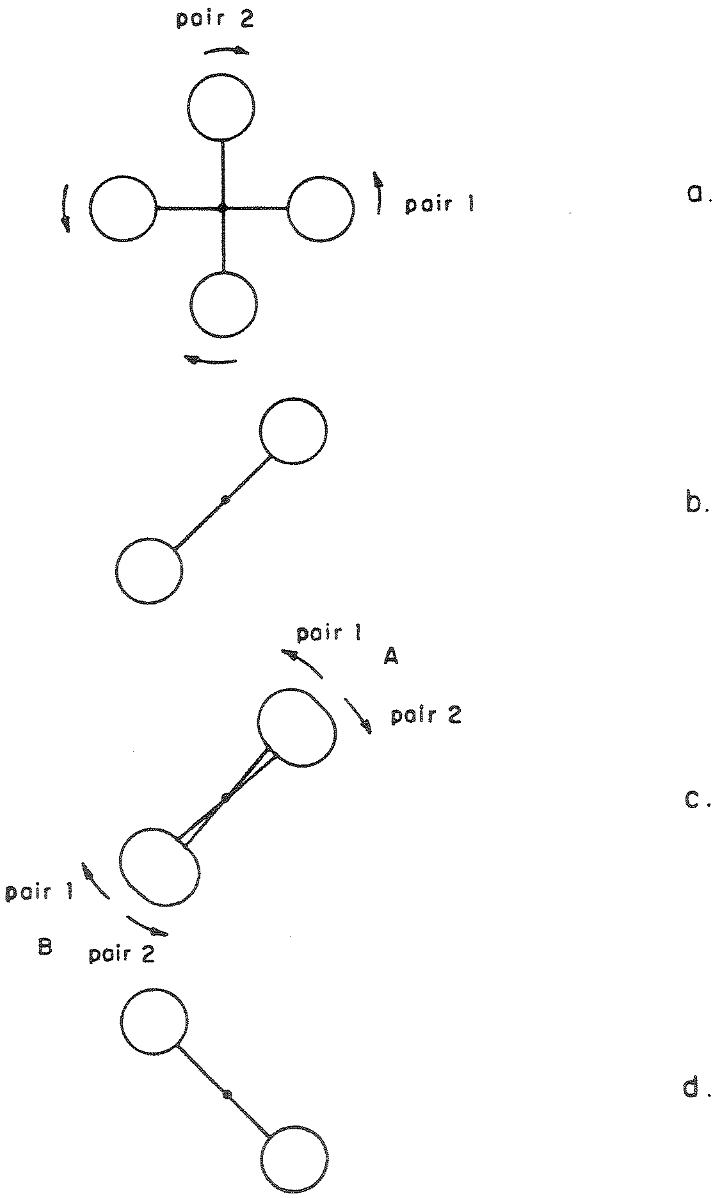


Figure 3.1 Four spinning disks forming a motion illusion.

minimal velocity change for each object results in a perception according to the velocities labeled A. In such a perception the disks appear to have circular paths and pass completely through one another at positions such as those shown in Fig. 3.1b and d. A rather more complicated proximity criterion, which holds that the disk last viewed completely in a given quadrant will return immediately to that quadrant, yields a perception according to the velocities labeled B. In this latter case, each disk sweeps both back and forth through a given quadrant as bounded by the positions shown in Fig. 3.1b and d. At these positions the disks appear to "bounce" off each other thus exactly reversing their velocities.

The two cases discussed above can, however, be understood in terms of two different types of constancy, one involving velocity, and the other, occupancy. These constancies can only be used to resolve the ambiguity in an indirect way because the ambiguity occurs when one is trying to understand the progression from an image of a position such as that of Fig. 3.1b to the immediately succeeding image. In the constant velocity case, for example, the instantaneous velocity is measured as the displacement in disk location between two successive images. But the location of the disk in the image after that of Fig. 3.1b is precisely what is in question. Thus the analysis of these two images in isolation cannot resolve the ambiguity. Instead the velocity information must be derived from the preceding images in which the constancies of shape and color can be used to locate each disk, thereby allowing the calculation of its velocity. The velocity information can then be applied to the given pair of images as part of a predictive analysis or as the criterion for a hypothesis and test procedure.

4.0 Three-Dimensional Information from Images

Early studies of sequences of images were motivated by the desire to analyze two-dimensional motion, for example, the satellite imagery of clouds. Several researchers also considered abstract models of two-dimensional motion using polygonal as well as curvilinear figures. The use of planar figures and parallel projection allowed these systems to ignore considerations of the third dimension. In contrast to the above purely two-dimensional works (reviewed in [1]), certain researchers have considered scenes containing objects undergoing three-dimensional motion. The initial research, however, analyzed only the image plane motions taking the two-dimensional approximation to be adequate. For example, in the work of Jain and Nagel [31] and Yachida et al [32], there was no attempt to recover the three-dimensional structure of the objects or their three-dimensional motion. This emphasis on two-dimensional motion was a natural outgrowth of the research. Recovery of three-dimensional structure of objects and the parameters of motion is certainly more complex.

In the present discussion it is assumed that the low-level processing problems have already been solved, i.e., the feature points on the surface of the objects have already been extracted in each of the images and the correspondence between the feature points in various images have been established. As mentioned earlier, this is a non-trivial task. Later in this section, we shall consider another scenario where this establishing of correspondence is not necessary. The correspondence problem is further complicated by the disappearance of points on an object due to occlusion from other objects, self occlusion as points rotate out of view, and shadows. Also, the assumption of rigidity plays an important role.

Ullman [33] considers the problem of determining the three-dimensional structure of an object from its two-dimensional projections. Under the assumptions of object rigidity and parallel projection, Ullman proved that three distinct views of four non-coplanar points in a rigid configuration, enable one to determine uniquely the motion and structure compatible with the given three views. Roach and Aggarwal [34] give an alternate solution for the case of central projection. They showed that two views of five points leads to 18 nonlinear equations whose solution yields the three-dimensional structure of points under consideration. Bonde and Nagel [35] consider a restricted case of the above general three-dimensional motion. Badler [36] uses a spherical projection model and is able to predict the point positions in succeeding images of moving objects. More recently Tsai and Huang [Chapter 1 in ref. 7, 37] reformulate the problem in terms of five unknown motion parameters and show certain existence results. Also Nagel [38] has derived a compact vector equation to determine three-dimensional points from two-dimensional image points. In view of the several formulations and different results, it may be emphasized that the results obtained depend on the following assumptions: (i) the nature of projection: parallel, central or spherical; (ii) the number of points and the existence of any relationships among these points, e.g., object rigidity; and (iii) the number of available views. In all of the above works, rigidity of the object under consideration and pre-establishment of the correspondence of points are assumed.

Our group at The University of Texas at Austin has considered three distinct problems: the recovery of three-dimensional structure under the assumption of central projection [34]; the motion of articulated objects under parallel projection [39]; and the derivation of volumetric descriptions from occluding contours with viewpoint specifications [40]. These projects will be briefly discussed.

Feature Points from Rigid Objects: The image of a point under central projection is a function of the point's three-dimensional position, the focal length of the camera, and the location and orientation of the camera's lens relative to the global coordinate system. Information about the camera's position is needed to relate the position of points given in two-dimensional focal plane coordinates to the global three-dimensional coordinate system. The necessary camera information is the camera focal length F, the orientation angles Θ,ϕ, and κ of the camera to the global coordinate system, and the three-dimensional coordinates of the lens center (X_0,Y_0,Z_0). The three angles orient the camera to the global coordinate system as follows: (assume for simplicity that the camera lens center has been translated to (0,0,0) of the global coordinate system) Θ is a rotation about the X-axis that brings the optical axis into the X-Z plane, ϕ is a rotation about the Y-axis so that the optical axis is aligned with the Z-axis, and κ is a rotation about the Z-axis so that the x', y' axes of the focal plane are aligned with the global X,Y axes. The use of primes in this section in general denotes the focal plane coordinate system. It is, of course, impossible to determine the original (x,y,z) position of a point from a single image. The best we can do is to determine a line in space on which the point falls. Further explanations of the equations in this section may be found in [34], and [41].

However, we want to know how much of the original three-dimensional information can be recovered given only a sequence of images of a moving object. It is possible to show that any sequence of images is inherently ambiguous. That is, there are an infinite number of objects that produce the same sequence of images. The objects are all similar in structure and movement.

In the following we discuss how to find the movement and three-dimensional model of points on an object's surface from a sequence of noise-free images up to a scaling factor; that is, by setting the scaling factor to an arbitrary value we can find a particular movement and model for the points on the object.

In the description above it was assumed that the camera is stationary and the object is moving. It is convenient to reformulate the problem such that the object is stationary and the camera moves. Under this formulation the three-dimensional structure and motion of an object can be derived from two views each containing five feature points from the object surface such that not all five of the surface points are in the same plane.

The solution can be obtained from a system of non-linear equations specified in the following manner. The global coordinates of each point are variable, so five points produce 15 variables. The global coordinates and the Θ,ϕ,κ orientation angles for each camera position are also variable producing 12 more variables. Thus, there are a total of 27 variables in the problem. Each point produces two projection equations per camera position for a total of 20 nonlinear equations. To make the number of equations and unknowns come out the same, seven variables must be known including one variable that will determine the scaling factor.

Six of the variables are specified by assuming the first camera position is coincident with the global axis system, that is, set the X_0,Y_0,Z_0 position and Θ, ϕ,κ orientation angles of the first camera to zero. In addition, the z-component of any one of the five points is set to an arbitrary positive constant. We mentioned earlier that the best result possible in locating the three-dimensional position of a point on an object is to find (sx, sy, sz) where s is an arbitrary scaling factor. By setting the z-component of the position of a point to an arbitrary constant, we are fixing the scaling factor. Once the z-component of a point is known, the x and y components can also be found using the inverse of the projection equations (see [34] that is, two of the 20 equations can be solved directly using the given z-component). There are now 18 projection equations in 18 unknowns; the equations of projection, however, are nonlinear. The situation is shown in Figure 4.1.

The system of nonlinear projection equations explained above can be solved by using a modified finite difference Levenberg-Marquardt algorithm due to Brown [42-44] without strict descent

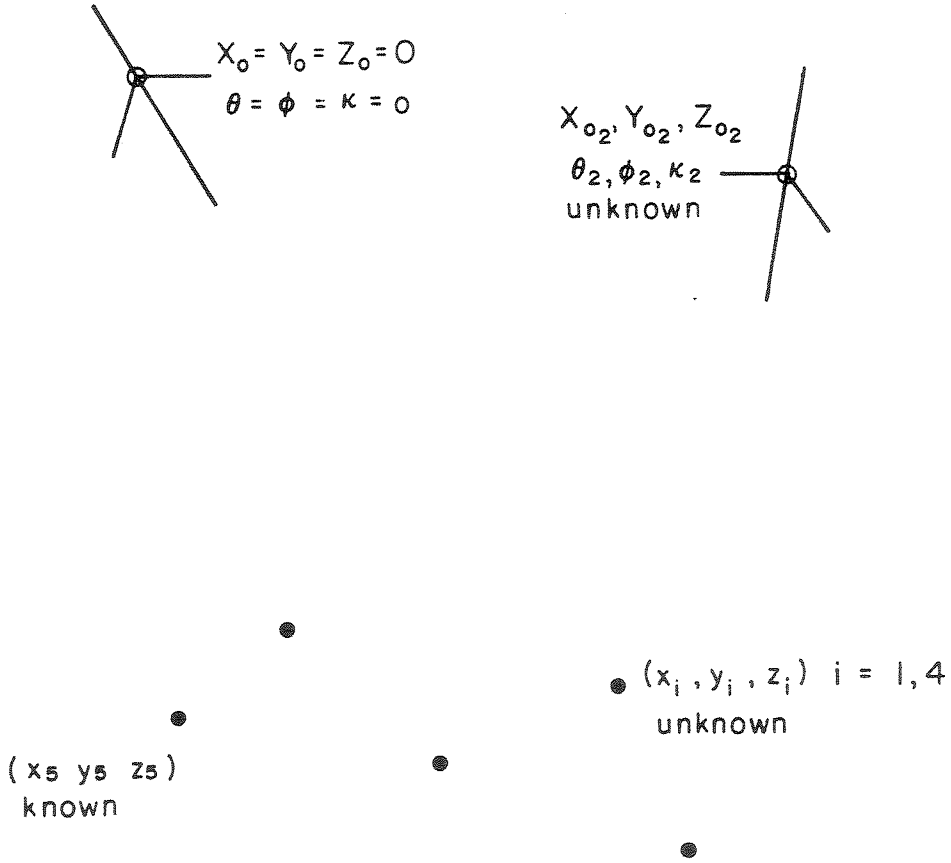

$X_o = Y_o = Z_o = 0$
$\theta = \phi = \kappa = 0$
$X_{o_2}, Y_{o_2}, Z_{o_2}$
$\theta_2, \phi_2, \kappa_2$
unknown
(x_i, y_i, z_i) $i = 1,4$
unknown
$(x_5\ y_5\ z_5)$
known

Figure 4.1 Camera configuration and point coordinates showing known and unknown parameters.

that minimizes the least-squared error of the 18 equations. The method employed is iterative and requires an initial guess for each unknown parameter.

This work is somewhat like the camera calibration systems of Sobel [45] and Yakimovsky and Cunningham [46]. In their work multiple images of points together with a central projection model and numerical methods are used to determine camera parameters such as focal length, position, and orientation. These studies, however, have considerably more information about the three-dimensional positions of points than we are assuming. Thus, the problems being solved and the information given for the calibration systems are different from the work described in this section.

Implicit in this work are two very important assumptions: that the objects being observed are rigid and that the images of the object are noise free and thus completely accurate. To test the effect of the second assumption on the numerical method described above, from one to four pixels were randomly added to or subtracted from the exact photocoordinate data for a moving object. This perturbation of the data causes extreme instability in the numerical solutions. However, one of the main reasons for using a least-squared error technique to solve a problem is to make adjustments to observations that contain error (noise). Adjustment is only possible, however, when there are more equations than unknowns. Two views of five points are therefore inadequate for noisy data since there are the same number of equations as unknowns. Two views of six points or three views of four points produce 22 equations in 21 unknowns using the same problem model discussed above. Examination of experimental runs using overdetermined systems of equations shows that minimal overdetermination is not very accurate. It is only with considerable overdetermination (two views of 12 or even 15 points; three views of seven or eight points) that the results become accurate.

A Motion Constraint to Recover Structure: In this section we will consider a system [39] that uses a motion constraint to derive three-dimensional structures from sequences of feature point images. The constraint is that the motion of each rigid object in the scene consists of a translation and a rotation about an axis that is fixed in direction. Of course, given two distinct frames from a dynamic scene containing a moving rigid object, a transformation taking the first three-dimensional position of the object to the second three-dimensional position can always be defined that consists of only one translation and one rotation [47]. However, the restriction is that the direction of the axis of rotation is fixed throughout the sequence, or at least over subsequences of frames.

To understand the importance of this restriction consider first, a dynamic scene containing an object that is simply rotating about a stationary axis. The surface of the object is assumed to produce identifiable feature points in the images under an orthogonal projection without occlusion. Clearly, as the object rotates, each surface point will trace out a circle about the axis of rotation. The parallel projection assures that the feature point, i.e., image position, of each surface point will trace out an ellipse in the image plane. Now, the axes of the ellipse can be used to specify the orientation, in three-dimensions, of the plane containing the original circle and thereby specify the direction of the axis of rotation. Deriving this information for the given feature points of an object allows one to build a model of the three-dimensional structure of the object from the sequence of images. This simple case indicates the basic steps in the process.

In general, since the motion constraint does not demand zero translation nor a completely fixed axis of rotation, the surface points will not actually be tracing circles and their projections will not be ellipses. If, however, an arbitrary surface point is chosen as the reference point and the object motion is considered in a coordinate axis system having the reference point as origin, then the motion in that system follows the case described above. In other words, if the motion could be normalized to a coordinate axis system with a surface point as origin then the remaining surface points would trace circles about an axis through the reference point, the projected positions would trace ellipses and the three-dimensional structure could be derived.

Clearly, this normalization is not directly possible, but it has been shown [39] that the given image sequence can be normalized to yield the same effect. In particular, the feature point images are translated so that the image plane origin coincides in each case with the projected position of an arbitrarily selected surface point. The processing proceeds as follows: the image sequence is acquired; feature points are detected for various surface points of the object; the feature point images are normalized to have a selected feature point as the origin of each image; the parameters of the traced ellipses are determined; and finally, the three-dimensional structure is derived.

There are two major advantages of this approach. First, it can accomodate dynamic scenes containing several independently moving objects by assuming that a set of feature points mutually satisfying the motion constraint, i.e., each element of the set traces an ellipse relative to the same reference point, constitutes the feature points for a single object. In detail, a feature point is selected as the reference point, the images are normalized to this reference point, and all ellipse tracing points are included in an object description. These points are deleted from further consideration and the process is iterated until every feature point is part of some object model. Ullman [33] used his rigidity constraint in a similar manner.

The second advantage is that analytically only two feature points need be detectable in three consecutive frames to derive the three-dimensional model. In this way articulated objects with only two feature points on each part can be analyzed. By "articulated object" is meant an object comprising several rigid parts connected through various joints that allow distinct but not independent movements for each part. The experiments of Johansson [48] provide a good example of articulated objects with minimal numbers of feature points. In those experiments people with reflectors attached to their major joints moved about in front of a camera that was calibrated to record only the positions of the reflectors. The system discussed in this section adequately analyzed an example taken from a Johansson experiment as reported in [49].

Occluding Contours in Dynamic Scenes: In this section we will describe the results of work [40] done in the pursuit of two major goals. The first goal is the development of a dynamic scene analysis system that does not depend completely on feature point measurements. The second goal is the development of a scheme for representing three-dimensional objects that is descriptive of surface detail, yet remains functional in the context of structure from motion in dynamic scenes.

To lessen the dependency on feature point detection, occluding contours with viewpoint specifications are used. The term "occluding contour" means the boundary in the image plane of the silhouette generated by an orthogonal projection. Silhouettes can most often be formed by a simple thresholding of the intensity values. A connected component analysis [50, pp. 336-347] of the resulting binary valued image yields the boundary of the object silhouette. An ordered list of the image plane coordinates of the resulting boundary constitutes the initial representation of the occluding contour. Throughout the analysis of the dynamic image, however, another representation, referred to as the rasterized area [51], of the contour will also be used. For its use here, the most significant attribute of this representation is that given an area so represented and an arbitrary segment on a line parallel to the "raster direction" it is a simple process to determine what portions of the given segment intersect the area, i.e., to clip that segment to the represented area.

The three-dimensional structure to be derived from the sequence of occluding contours is a bounding volume approximation to the actual object. For this reason the representation incorporated in this system is based on volume specification through a "volume segment" data structure. The volume segment representation is a generalization to three-dimensions of the rasterized area description. For the rasterized area, each of the segments denoted a rectangular area. The generalization to three dimensions is to have each segment represent a volume, i.e., a rectilinear parallelepiped with edges parallel to the coordinate axes. In addition to grouping collinear segments into lists, the set of segment lists is partitioned so that the subsets contain lists having coplanar segments. The primary dimension of the parallelepiped specified by a segment is the length of the segment. The second dimension is given by the inter-segment spacing within the plane of the segment, while the third dimension is the inter-plane distance. The latter two dimensions are specified to be uniform throughout the volume segment representation.

The primary advantage of this structure in general situations is that the process of determining whether an arbitrary point is within the surface boundary consists of a simple search of three ordered lists: select a "plane" by z-coordinate; select a "line" by x-coordinate; and finally, check for inclusion of the y-coordinate in a segment.

This volume segment representation is created from a dynamic image by two processes. The first process combines information from frames 1 and 2 of the dynamic image to form an initial volume segment representation. The second process then accepts each succeeding frame in order to refine the approximation represented by the volume segment structure. Thus these processes analyze the occluding contours with their view orientations to initially construct and to continually refine the volume segment representation of the object generating the contours. Algorithm summaries of the two processes are given in more detail in [52,53].

5.0 Conclusions

The three major ingredients of Dynamic Scene Analysis, namely segmentation, occlusion and computation of three-dimensional information from images have been discussed in depth. The approaches to dynamic scene segmentation may be broadly divided into two classes - pixel and feature based techniques. Each has its advantages and disadvantages. It appears that a combination of the two approaches may be necessary to accomplish segmentation in a difficult dynamic scene segmentation task. Occlusion introduces some really difficult problems in dynamic scene analysis. Effective cues to occlusion may be derived from scene-domain constraints. The use of these cues depends upon the ability to integrate information derived from widely separated spatial areas in an image or widely separated temporal events in a sequence. The techniques for computation of three-dimensional information from images are still in their early stages of evolution. The solution of a large number of nonlinear equations and the sensitivity of solutions to noise pose serious hurdles in a satisfactory computation of three-dimensional information from images. The absence of an ideal description of three-dimensional structure of objects compounds the difficulties. Further, in each of the three areas discussed above, the correspondence problem plays an important role.

In view of the active participation at the Braunlage Advanced Study Institute, it appears that the next decade will be an exciting era in Dynamic Scene Analysis. In addition to the development of new techniques for the solution of problems discussed above, one will witness the application of dynamic scene analysis to many new areas.

References

1. W. N. Martin and J. K. Aggarwal, "Dynamic Scene Analysis: A Survey," Computer Graphics and Image Processing, Vol. 7, No. 3, pp. 356-374, June 1978.

2. H.-H. Nagel, "Analysis Techniques for Image Sequences," Proc. 4th International Joint Conference on Pattern Recognition, Kyoto, Japan, November 1978, pp. 186-211.

3. J. K. Aggarwal and N. I. Badler, eds., Abstracts for the Workshop on Computer Analysis of Time-Varying Imagery, Philadelphia, PA, April 1979.

4. J. K. Aggarwal and N. I. Badler, guest eds., Special Issue on Motion and Time-Varying Imagery, IEEE Trans. on Pattern Analysis and Machine Intelligence Vol. PAMI-2, No. 6, November 1980.

5. W. E. Snyder, guest ed., Computer Analysis of Time-Varying Images, Computer, Vol. 14, No. 8, August 1981.

6. J. K. Aggarwal, guest ed., Special Issue of Computer Graphics and Image Processing, (to appear).

7. T. S. Huang, Image Sequence Analysis, Springer-Verlag, New York, 1981.

8. NATO Advanced Study Institute, on Image Sequence Processing and Dynamic Scene Analysis, Advance Abstracts of Invited and Contributory Papers, June 21-July 2, 1982, Braunlage, W. Germany.

9. J. K. Aggarwal and R. O. Duda, "Computer Analysis of Moving Polygonal Images," IEEE Trans. on Computers, Vol. C-24, pp. 966-976, Oct. 1975.

10. W. N. Martin and J. K. Aggarwal, "Computer Analysis of Dynamic Scenes Containing Curvilinear Figures," Pattern Recognition, Vol. 11, pp. 169-178, 1979.

11. J. W. McKee and J. K. Aggarwal, "Finding the Edges of the Surfaces of Three-Dimensional Curved Objects by Computer," Pattern Recognition, Vol. 7, pp. 25-52, 1975.

12. C. K. Chow and J. K. Aggarwal, "Computer Analysis of Planar Curvilinear Moving Images," IEEE Trans. on Computers, Vol. C-26, pp. 179-185, 1977.

13. R. Jain, W. N. Martin and J. K. Aggarwal, "Segmentation Through the Detection of Changes Due to Motion," Computer Graphics and Image Processing, Vol. 11, pp. 13-34, 1979.

14. S. Yalamanchili, W. N. Martin and J. K. Aggarwal, "Extraction of Moving Object Descriptions Via Differencing," Computer Graphics and Image Processing, Vol. 18, pp. 188-201, 1982.

15. R. Jain and H.-H. Nagel, "On the Analysis of Accumulative Difference Pictures from Image Sequences of Real World Scenes," IEEE Trans. on Pattern Analysis and Machine Intelligence Vol. PAMI-1, No. 2, pp. 206-214, 1979.

16. H.-H. Nagel, "Formation of an Object Concept by Analysis of Systematic Time Variations in the Optically Perceptible Environment," Computer Graphics and Image Processing, Vol. 7, No. 2, pp. 149-194, 1978.

17. R. C. Gonzalez and P. Wintz, Digital Image Processing, Addison Wesley Publ. Co., Inc., Reading, MA, 1977, pp. 383-386.

18. J. A. Leese, C. S. Novak and V. R. Taylor, "The Determination of Cloud Pattern Motions from Geosynchronous Satellite Image Data," Pattern Recognition, Vol. 2, pp. 279-292, Dec. 1970.

19. J. A. Leese, C. S. Novak and B. B. Clark, "An Automated Technique for Obtaining Cloud Motion from Geosynchronous Satellite Data Using Cross-Correlation," J. Applied Meteorology, Vol. 10, pp. 118-132, Feb. 1971.

20. W. B. Thompson, "Combining Motion and Contrast for Segmentation," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 6, pp. 543-549, 1980.

21. D. R. Hofstadter, Godel, Escher, Bach: An Eternal Golden Braid, Vintage Books, New York, 1980.

22. J. K. Aggarwal and W. N. Martin, "Analyzing Dynamic Scenes Containing Multiple Moving Objects," in Image Sequence Analysis, T. S. Huang, ed., Springer-Verlag, New York, 1981, pp. 355-380.

23. H. G. Barrow and J. M. Tenenbaum, "Recovering Intrinsic Scene Characteristics from Images," in Computer Vision Systems, A. R. Hanson and E. M. Riseman, eds., Academic Press, New York, 1978.

24. D. Waltz, "Understanding Line Drawings of Scenes with Shadows," in Psychology of Computer Vision, P. H. Winston, ed., McGraw-Hill, New York, 1975, pp. 19-92.

25. D. Huffman, "Impossible Objects as Nonsense Sentences," in Machine Intelligence 6, B. Meltzer and D. Michie, eds., Edinburgh University Press, Edinburgh, Scotland, 1971.

26. M. Clowes, "On Seeing Things," Artificial Intelligence, Vol. 2, No. 1, pp. 79-116, 1971.

27. F. W. Mounts, "A Video Encoding System With Conditional Picture-Element Replenishment," Bell Syst. Tech. J., 48, pp. 2545-2554, 1969.

28. S. A. Underwood and C. L. Coates, "Visual Learning from Multiple Views," IEEE Trans. on Computers, Vol. C-24, pp. 651-661, 1975.

29. A. Guzman, "Decomposition of a Visual Scene into Three-Dimensional Bodies," in Computer Methods in Image Analysis, J. K. Aggarwal, R. O. Duda and A. Rosenfeld, eds., IEEE Press, New York, 1977, pp. 324-337.

30. R. P. Futrelle and M. J. Potel, "The System Design for GALATEA, an Interactive Real-Time Computer Graphics System for Movie and Video Analysis," Computer Graphics, Vol. 1, pp. 115-121, 1975.

31. R. Jain and H.-H. Nagel, "On a Motion Analysis Process for Image Sequences from Real World Scenes," IFI-HH-B-48/78, Institut fuer Informatik der Universitat Hamburg, Hamburg, Germany, April 1978.

32. M. Yachida, M. Asada and S. Tsuji, "Automatic Motion Analysis of Moving Objects from the Record of Natural Process," Proc. 4th International Joint Conference on Pattern Recognition, Kyoto, Japan, November 1978, pp. 726-730.

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

34. J. W. Roach and J. K. Aggarwal, "Determining the Movement of Objects from a Sequence of Images," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 6, pp. 554-562, 1980.

35. T. Bonde and H.-H. Nagel, "Deriving a 3-D Description of a Moving Rigid Object from Monocular TV-Frame Sequences," WCATVI-79, Philadelphia, PA, pp. 44-45.

36. N. Badler, "Temporal Scene Analysis: Conceptual Descriptions of Object Movements," Ph.D. dissertation, University of Toronto, Toronto, Ont., Canada, 1975.

37. R. Y. Tsai and T. S. Huang, "Estimating Three-Dimensional Motion Parameters of a Rigid Planar Patch," Proc. of the IEEE Conf. on Pattern Recognition and Image Processing, Dallas, TX, August 1981, pp. 94-97.

38. H.-H. Nagel, "On the Derivation of 3-D Rigid Point Configurations from Image Sequences," Proc. of the IEEE Conf. on Pattern Recognition and Image Processing, Dallas, TX, August 1979, pp. 103-108.

39. J. A. Webb and J. K. Aggarwal, "Structure from Motion of Rigid and Jointed Objects," to appear in Artificial Intelligence.

40. W. N. Martin and J. K. Aggarwal, "Occluding Contours in Dynamic Scenes," Proc. of the IEEE Conf. on Pattern Recognition and Image Processing, Dallas, TX, August 1981, pp. 189-192.

41. J. W. Roach, "Determining the Three-Dimensional Motion and Model of Objects from a Sequence of Images," Ph.D. dissertation, Department of Computer Sciences, The University of Texas, Austin, TX, May 1980.

42. K. M. Brown and J. E. Dennis, "Derivative Free Analogues of the Levenberg-Marquardt and Gauss Algorithms for Nonlinear Least Squares," Numer. Math., Vol. 18, pp. 289-297, 1972.

43. K. Levenberg, "A Method for the Solution of Certain Non-Linear Problems in Least Squares," Quart. Appl. Math., Vol. 2, pp. 164-168, 1944.

44. D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," J. SIAM, Vol. 11, No. 2, 1963.

45. I. Sobel, "On Calibrating Computer Controlled Cameras for Perceiving 3-D Scenes," Artificial Intelligence, Vol. 5, No. 2, pp. 185-198, 1974.

46. Y. Yakimovsky and R. Cunningham, "A System for Extracting Three-Dimensional Measurements from a Stereo Pair of TV Cameras," Computer Graphics and Image Processing, Vol. 7, No. 2, pp. 195-210, 1978.

47. J. Coffin, Vector Analysis, second edition, John Wiley & Sons, New York, 1911.

48. G. Johansson, "Visual Motion Perception,"Scientific American, Vol. 232, No. 6, pp. 76-89, June 1975.

49. J. A. Webb and J. K. Aggarwal, "Visually Interpreting the Motion of Objects in Space," Computer, Vol. 14, No. 8, pp. 40-46, August 1981.

50. A. Rosenfeld and A. C. Kak, Digital Picture Processing, Academic Press, New York, 1976.

51. W. M. Newman and R. F. Sproull, Principles of Interactive Computer Graphics, second edition, McGraw-Hill, New York, 1979.

52. W. N. Martin and J. K. Aggarwal, "Analyzing Dynamic Scenes," Laboratory for Image and Signal Analysis, The University of Texas at Austin, TR-81-5, December 1981.

53. W. N. Martin and J. K. Aggarwal, "Volumetric Descriptions of Objects from Multiple Views," to appear.

Acknowledgements

This research was supported by the Air Force Office of Scientific Research under grant AFOSR 82-0064.

The authors thank Chiun-Hong Chien for proofreading, Sudha Yalamanchili for the generation of the pictures and Linda Melnick for inputting the manuscript.