ANALYZING DYNAMIC SCENES
CONTAINING MULTIPLE MOVING OBJECTS[1]

J. K. Aggarwal[2,3]
W. N. Martin[2]

TR-125 January 1980

[1] This research was supported in part by the Air Force Office of Scientific Research under grant AFOSR 77-3190.

[2] Computer Sciences Department
The University of Texas at Austin
Austin, Texas 78712

[3] Department of Electrical Engineering
The University of Texas at Austin
Austin, Texas 78712

ANALYZING DYNAMIC SCENES
CONTAINING MULTIPLE MOVING OBJECTS

Table of Contents

ANALYZING DYNAMIC SCENES
CONTAINING MULTIPLE MOVING OBJECTS

This report provides an insight into the problems encountered in image sequence analysis when the viewed scene contains several objects all of which may be moving. A major emphasis is on the interaction of object images resulting from viewpoint placement and object movement. The interaction that concerns us here is the occlusion of some part of an object by other objects or even by other parts of the same object. Our discussion begins with the problems of interpreting single images containing occluded objects and then turns to the implications of occlusion on image sequence analysis. Further problems encountered in image sequence analysis will then be emphasized through the description of those problems as they have arisen in specific systems. Foremost among these problems is the identification, throughout the image sequence, of objects in spite of their continually changing appearance. Also important are the computational implications for attempting to process the flux of information presented by an image sequence.

1. Occlusion in General

1.1 Arbitrary Images

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, i.e., paintings on a museum wall or a pair of birds flying in a clear sky, there is no problem. 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 of one object will not interfere with the analysis of the remaining foreground objects. However, if the background has a complex structure, i.e., the museum wall has a highly patterned covering, or if the foreground objects are closely arranged in some structure, i.e., a flock of birds flying in the same direction, then the classic "figure-ground" problem arises.

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 psycho-physical problem where the preconceptions and expectations of the viewer play an important part in perception. Figure 1 indicates some of the subtleties of this problem. The lowest level figure-ground

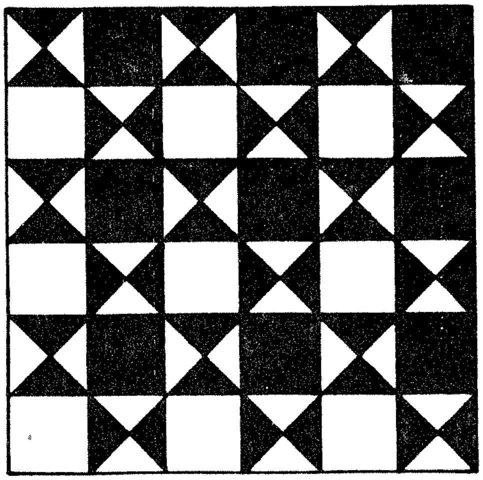

Figure 1. Geometric pattern having an ambiguous figure-ground interpretation.

ambiguity involved in Figure 1 is whether the figure depicts dark squares and triangles against a light background or light squares and triangles on a dark background. This ambiguity is inherent in the construction of the figure and when focusing on a single element, i.e., a triangle or square, it is fairly easy to form either interpretation. However, if one focuses on an area containing several elements it becomes more difficult to interpret the figure as being a simple collection of squares and triangles of either color.

The prevalence of pairs of single element edges that are colinear, parallel or perpendicular interfers with the perception of the individual elements. The lines through which the sides of the squares are colinear give rise to the perception of the figure as having three horizontal and three vertical bands with hourglass holes (or obstructions) along their length. Again bands of either color can be seen. The fact that many element edges can be linked together to form long straight lines reinforces this perception.

Another set of colinearities, the diagonal lines through the sides of the triangles, yields yet another interpretation. In this case the crossings of the diagonal lines emphasize the right angles in the triangles. The triangles are thus grouped into four element sets with each set forming a diamond whose corners correspond to the right angle corners of the constituent triangles. In this interpretation the figure can be seen as a checkerboard of black and white diamonds where each diamond is partially covered by a smaller square of the opposite color.

For the interpretations of Figure 1 discussed above, an initial decision is made as to what is the prevalent structure. The remaining parts of the figure must then be related to that structure. For the last of the above interpretations the checkerboard of light and dark diamonds is accepted as the primary structure. The smaller squares interfere with the perception of the checkerboard but can be accounted for by assuming that they are foreground objects occluding the diamonds. In this manner the concept of occlusion is brought into the perceptual process at a very low level in order to make sense of the ambiguous figure. The variety of relationships exhibited by the individual squares and triangles provides the opportunity to resolve the ambiguities into several different interpretations of the figure.

1.2 Scene Domain Imposed Constraints

In the abstract geometrical pattern of Figure 1, both the quantity and subtlety of the inter-element relationships are greater than that occurring in typical natural scenes. The constraints imposed by the three-dimensional structure and distribution of the objects appearing in typical scenes serve to limit the relationships possible in the images derived from such scenes. For instance, the boundary edges of non-contiguous objects are rarely colinear in natural scenes. This consideration makes reasonable the assumption that if the image of a scene contains disjoint edges which are colinear, 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 [1] 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 cues to apparent depth.

An elegant example of how scene domain constraints can be used in understanding occlusion is the system developed by Waltz [2]. 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 distribution of the objects in the scene, the location of the light source, and the orientation of the image plane are restricted only to eliminate certain coincidences which cause unresolvable ambiguities in the resulting images. The images analyzed are actually "perfect" line drawings formed by an orthogonal projection of the scene onto the image plane with a relative brightness associated with each region in the drawing.

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 [3] and Clowes [4]. Junctions are the line drawing representations of the vertices in the scene, thus the trihedral restriction on 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, thus when found in the drawings, provide a reliable indication of occlusion.

The so-called T junction is an example of a junction indicative of occlusion. The line forming the bar of the T must correspond to an obscuring edge in the scene. In fact, the edge which forms the bar line must be a convex edge. However, the isolated T junction remains ambiguous in that the post of the T may correspond to a partially obscured edge, or it may be a "crack" along which two objects coincide. In the latter case the bar line is actually the result of an "accidental alignment" of two convex edges and the obscuring object faces occur on the post side of the T junction. In the former case the bar line is generated by the convex edge bounding the obscuring object face opposite the T junction post.

This ambiguity can be resolved in two ways, each of which serves to provide non-local information to the labeling process. First, the system may be told that the region opposite the post of the T junction is the background or supporting surface. This insures that the post of the T junction is a "crack" and that the bar line corresponds to two aligned edges. Second, the labeling process may have already determined a legitimate labeling for one of the junctions which are connected to the given T junction. The labeled junction is, of course, connected to the given junction by a common line and thus the currently assigned label for the common line can be used to determine the labels for the other two lines of the T junction or at least to limit the number of possible labels for those lines. For example, the post of a given T junction may also be part of an L junction. This means that the post of the T junction could not possibly be a crack and therefore the bar line must be a single edge bounding the obscuring object face.

This latter method is fundamental to the overall labeling process and is based on another major constraint imposed by the scene domain. For the type of objects allowed in the scenes, a line must retain a given labeling throughout its extent. This constraint is assured by the fact that each edge in the scene is created by the intersection of exactly two planes. The constraint provides, in the actual structure of the input images, a mechanism by which a rather simple labeling process can obtain information that is not local to the junction currently under consideration. The non-local information is of prime importance because the ambiguities in the image caused by occlusion in the scene cannot usually be resolved on the basis of information contained at a single junction. Indeed, it may not even be possible to tell initially how far away (in terms of the number of interceding junctions) the necessary clarifying information is. For this reason the labeling process is embedded in a parallel-iterative procedure (since referred to as "relaxation", Rosenfeld, et al. [5]). This procedure allows the necessary information to propagate throughout the interconnected network of junctions, yet requires the labeling process to interrogate no more than two adjacent junctions at a time. Of course, the pairwise comparisons performed by the labeling process are not sufficient to resolve all ambiguous line drawings. Increasing the number of junctions that the labeling process may inspect at a given time will allow the system to disambiguate more scenes. However, for any fixed and finite limit there will always be scenes, albeit exceedingly complex scenes, which will not be resolvable.

From the previous discussion of the figure-ground problem for arbitrary pictures one can see that the removal of scene domain constraints only worsens the degree of ambiguity caused by occlusion. In particular, 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.

This discussion has brought to light several factors which 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. Third, the use of such cues may involve the complex integration of information taken from areas which are widely separated in the image. Fourth, occlusion necessarily results in the loss of information available about the obscured object thus causing uncertainties in the interpretation of the image. And finally, the resolution of some occlusion ambiguities depends on external expectations and goals.

1.3 Occlusion In Image Sequences

The discussion up to this point has dealt with the implications of occlusion on the analysis of single images. For the remainder of this chapter the focus will be on time varying images, that is, sequences of images representing systematic time order samplings of scenes. The question addressed is: How is the complexity of the occlusion analysis problem affected by the addition of time variation, i.e., sequences of images? 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, and are described at a detailed level in later sections where specific dynamic scene analysis systems are presented.

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, Mounts [6], 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 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, Underwood and Coates [7], which forms a three-dimensional description of a planar 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 complimentary 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 which will be discussed in more detail in later sections of this chapter.

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", Guzmán [8], 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, i.e., a gravel road bed, will vary as the sun angle changes throughout the day.

This problem of identifying "apparently different but semantically identical objects", Futrelle and Potel [9], 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 Figure 2. Here four identical disks are attached pairwise to the ends of two crossmembers 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 Figure 2a. However, when the position shown in Figure 2b is reached, the constancies no longer serve as identifying features and thus admit an ambiguity to the interpretation of the position displayed in Figure 2c.

Is the pair of velocities labeled A in Figure 2c the correct interpretation or is the pair labeled B correct? An assumption of 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 Figures 2b 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

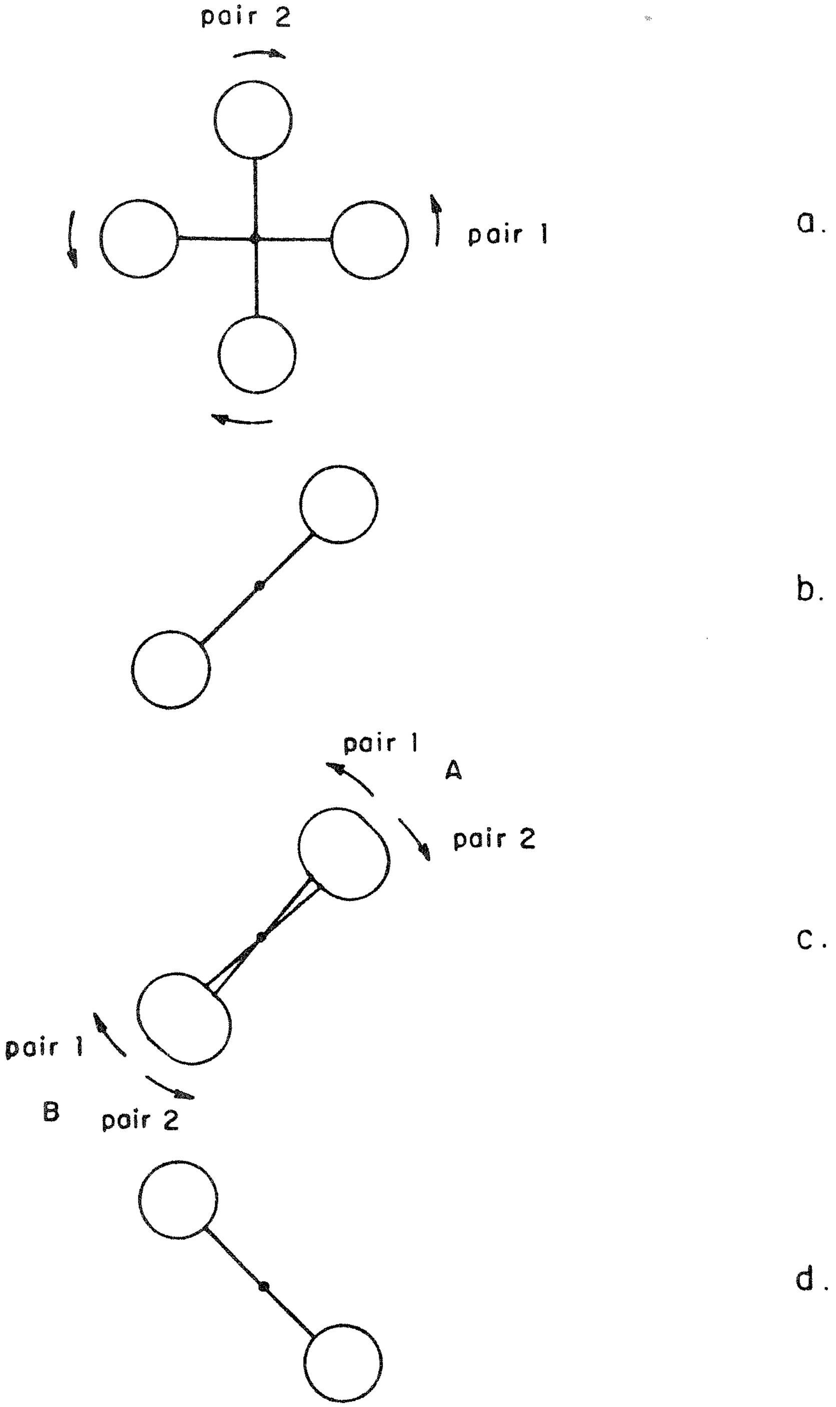

Figure 2. A motion illusion with four spinning disks.

sweeps both back and forth through a given quadrant as bounded by the positions shown in Figures 2b 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 Figure 2b 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 Figure 2b 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.

In this introductory section we have discussed several of the general problems in understanding occlusion as they relate to the analysis of time varying imagery. In the remaining sections we discuss the various problems encountered and the solutions derived in specific systems (for further references see Martin and Aggarwal [10], Nagel [11], and WCATVI [12]). We turn first to systems which reduce each image to point patterns and analyze the changes in successive patterns. Then we discuss several systems which operate on the object boundaries. Finally we conclude with a discussion of general representation problems and other areas for further research involving time varying imagery.

2. Dot Pattern Analysis

This section presents several specific dynamic scene analysis systems. These systems preprocess the image sequence so that each image is transformed into a dot pattern, where each dot constitutes a "token" of some significant feature in the image. In addition to spatial location, various attributes of the image features may be associated with their respective dots. In some cases the features represented by the dots are whole objects which, due to the scene domain, do not occlude or otherwise closely interact. For this type of scene, simple one-object tracking procedures (e.g., matching by shortest Euclidean distance) can be applied independently to each dot in a given image in order to identify the represented feature in the next image of the sequence. Identifying a particular feature in every image of the sequence by matching the proper dots in every pair of consecutive images allows the system to form a trace of the positions occupied by the feature throughout the sequence (as an example see Greaves [13]). Several motion measurements such as linear velocity, angular velocity, and acceleration can be computed from these traces.

2.1 Combined Motion and Correspondence Processes

If the represented features are objects which are likely to occlude one another, or are parts of a single object and interact closely, then the simple tracking procedures will not be applicable. In order to elaborate on this point we will discuss a system described in an early paper on cloud tracking, Endlich et al. [14]. This system represents the clouds by "brightness centers." The centers are obtained by first thresholding the satellite pictures to separate the clouds from the background, and then applying a clustering procedure, ISODATA, Ball and Hall [15], to the designated cloud pixels.

The brightness center representation is appropriate for images of clouds because the detailed description of the cloud region boundaries is quite complex and is continually changing. Through short time intervals these changes along the cloud boundaries have little effect on the location of the resulting brightness centers. Thus the amorphous nature of the clouds can be ignored while retaining a fairly reliable indication of the cloud locations and brightnesses. These attributes do not provide a sufficient basis for identifying the corresponding brightness centers between consecutive images of the sequence. However, the system is capable of successfully matching the centers under the assumption of a fairly uniform yet unknown common motion for most of the centers within arbitrarily selected subimages. The correspondence is determined by hypothesizing for each center in a selected subimage of a given image several possible matches to the centers in the appropriate subimage of the next image of the sequence. Each hypothesis specifies a motion vector so that the procedure initially yields a set of vectors. The most common vector in this set is chosen as the representative motion for the subimage.

Hypothesized matches which generate motion vectors that vary significantly from the representative motion are discarded, thus reducing the number of possible matches. An iterative application of this procedure is used to obtain unique assignments, i.e., no center is matched more than once. Having established the acceptable matches the system can use the vectors specified by those matches to form a vector map of the wind velocities over the area viewed in the image.

In some cases, however, centers will be left unmatched indicating that there is not a center in the second image which when hypothesized as a match for the given center yields a motion vector consistent with the chosen representative motion. It is reasonable to have unmatched centers because clouds tend to both dissipate and gather through time. However, it is also possible that the selected subimage contains clouds from several altitudes which are moving with radically different velocities.

Figure 3 illustrates this situation (for a similar example, as processed by the system, see Figure 5 of [14]). Two consecutive frames are displayed in Figure 3 with the majority of the dots moving down and to the left. There is also a small group of dots moving to the left and slightly up. Due to the statistical weight of the downward moving group of dots, the chosen representative velocity vector is similar to the vector labeled A in Figure 3c. This choice precludes the correct matchings for the dots in the smaller group, which would result in vectors such as the one labeled B in Figure 3c. Since there are no matches for the dots of the small group that yield vectors similar to A, these dots are left unmatched in both frames.

A more extreme case is displayed in Figure 4. In this example two equally numbered sets of dots are moving in perpendicular directions with approximately equal speeds, as indicated by the dashed vectors of Figure 4c. These velocities in conjunction with the original locations of the dots allow an alternate interpretation in which a single vector, such as the solid vector of Figure 4c, can be chosen as representative of

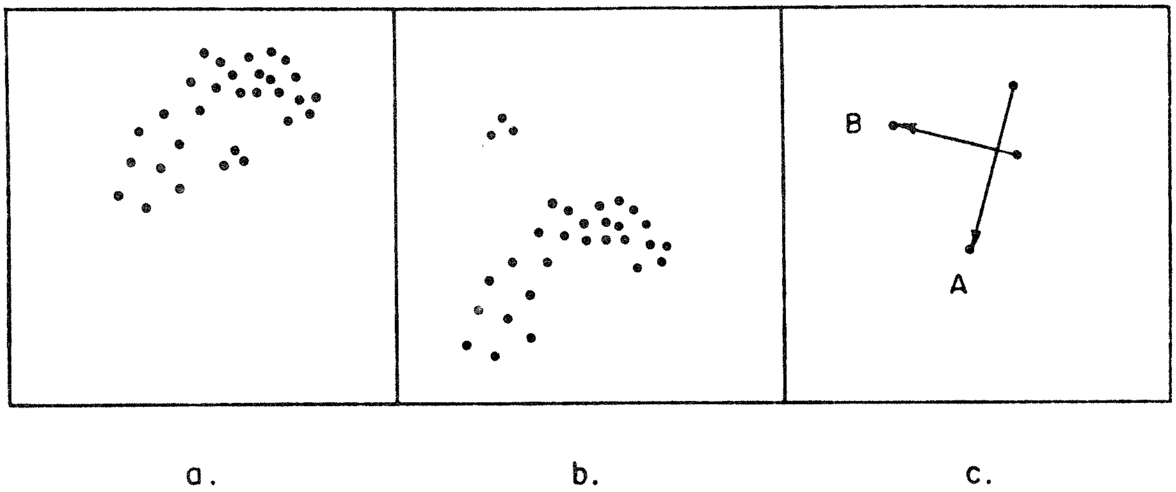

Figure 3. Dot patterns with interfering motion.

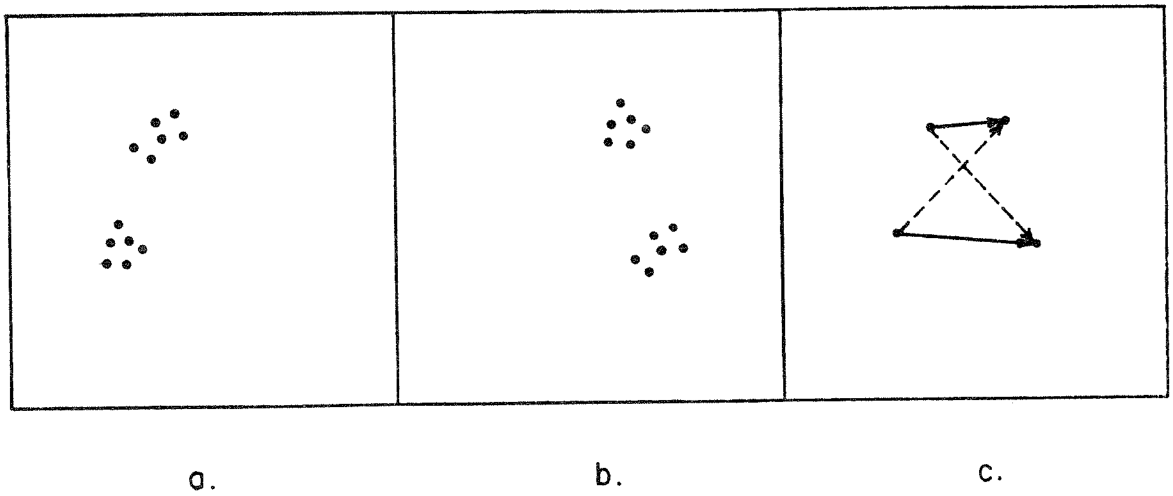

Figure 4. Dot patterns with dual motion interpretations.

the motions of all the dots. No dots will be left unmatched, but the chosen vector will specify an incorrect match for every dot. The problem arises from the essentially circular use of hypothesized motions in determining the dot correspondence. The motion measurements are necessarily based on the derived matches, thus the correspondence must itself be based on some other constancy feature of the moving objects. Otherwise, limitations must be imposed on the types of movement allowed, making the motion constant in some respect. In this system the motion is assumed to be nearly constant over the subimages.

2.2 Separate Correspondence Determination

A similar approach to the matching problem is taken in a recent book on motion, Ullman [16], where it is argued that the "correspondence process is a low level operation which precedes, or is independent of, the [3D] interpretation scheme." The correspondence process receives each image in the form of a set of tokens and constructs a "minimal" mapping of those tokens to the ones of the previous image in the sequence. It is referred to as a minimal mapping because the computational mechanism suggested creates the mapping by minimizing a cost function which reflects the most likely movement of the tokens. "Most likely" is in terms of an assumed probability distribution function for the velocities exhibited by the tokens in the image. In a discussion of human visual behavior, evidence is given for an affinity function upon which the low level correspondence process could be based. It can reasonably be argued that a typical human response to the example of Figure 4 would be that there is only one group of dots and the group is moving horizontally to the right.

This interpretation is consistent with a spatial proximity affinity function. However, the correct interpretation can only be made if information other than the dot displacements between the given two frames is used in the analysis. This information might be in the form of projected movements derived from previous frames in the sequence. Such projections would provide a probability distribution for the motion of each dot in a given frame which had been successfully matched to a dot in the immediately preceding frame. The most likely match for each dot would then depend on its associated distribution rather than a globally assumed distribution. The minimizing function could begin with a much closer initial estimate containing fewer candidate matches to choose from. The problem that would arise is in trying to allow for objects which change their direction or speed of travel. Such movement variations invalidate the previously compiled projections leaving the analysis system with the original problem again. This complication is also encountered in higher level predictive schemes, for example see Chow and Aggarwal [17].

Other information useful to this analysis but not necessarily dependent on previous frames of the sequence is the spatial relationship of the dots in a given frame. This would involve grouping the dots into another level of tokens with each new token having as an attribute the spatial distribution of the dots grouped under that token. The new tokens could be matched between consecutive frames by this additional attribute, thus removing hypothesized motion measurements from the correspondence process. The problem encountered in trying to use this type of information is that of properly grouping the dots under the higher level tokens.

One obvious grouping is to consider all the dots of a given frame to be under a single token. This grouping is easily made, however, it will clearly be of no utility for scenes containing several moving objects. At the other extreme, assigning one dot to each token would not be a legitimate grouping because the spatial relationship attribute of the resulting tokens would be undefined. The desired grouping would be along the spectrum between these two possibilities, but exactly where would depend on the actual scene being analyzed. This grouping process is analogous to classical segmentation in scene analysis. The process could be quite complicated and depend on information from previous frames in the sequence or from system maintained models of the scene. It is worth observing that Ullman contends that the correspondence process operates at a much lower level than could support this grouping process. In fact a computational method [16] has been developed that uses the correspondence information from a minimal matching process to derive groupings which have three-dimensional interpretations. A discussion of that system follows.

2.3 Motion Analysis Given Dot Correspondence

The input to Ullman's "structure from motion" system is a sequence of images, each formed by taking the orthographic projection of a three-dimensional scene containing identifiable feature points on moving objects. Thus every feature point or token is represented as a dot in each image with an assumed correspondence process identifying the feature points throughout the sequence of images. The transformations exhibited in the resulting two-dimensional dot patterns will, under certain conditions, give rise to a unique interpretation of both the three-dimensional spatial relationship of the feature points and their motions. The conditions are stated in the following theorem [16] whose proof constitutes a specification of a computational method to derive the correct interpretation.

THEOREM: Given three distinct orthographic views of four non-coplanar points in a rigid configuration, the structure and motion compatible with the three views are uniquely determined.

The most important of the constraints mentioned in the theorem is that the points be in a rigid non-coplanar configuration. The remaining details specify only minimum requirements for the computational method in the sense that more points or views can be used and that an analogous method can be applied to perspective projections.

The rigidity constraint reflects a more fundamental assumption of this system. The assumption is stated as follows: "Any set of elements undergoing a two-dimensional transformation which has a unique interpretation as a rigid body moving in space should be interpreted as such a body in motion." The overall strategy of the system is to find the sets of tokens which can be interpreted as configurations representing rigid bodies. The minimum requirements then prescribe a starting point for the system. The set of tokens for the first view are partitioned into four element subsets. The given correspondences specify the appropriate tokens for the subsets in the second and third views. Thus each subset meets the minimum requirements and all the system need do is determine if the rigidity constraint holds for every subset independently. The test performed, as detailed in [16], either determines that the subset can be interpreted as a rigid object by actually computing that configuration, or declares the relationship of the feature points in the subset to be non-rigid. The set of points from rejected subsets, the latter case, can be repartitioned and the test performed again until an interpretation is established for all the feature points.

In a similar manner the system can iteratively test for rigid configurations between previously defined subsets until a minimal, in terms of the number of subsets, covering the feature points is obtained. The resulting subsets correspond to the rigid objects in the scene and specify, through the successful rigidity tests, the structure and orientation of those objects. It should be noted that due to the orthographic projection the structure is determined up to a reflection about the image plane, while the orientation is resolved to within a translation in depth.

The approach taken by this system is to break the given data, i.e., the tokens in correspondence, into "nuclei" of identifiable features, i.e., four element subsets, which can be independently analyzed by a general knowledge based process. It is important that the analyzing process not depend on specific knowledge about the components of the scene or on disjoint sources of information indicative of the three-dimensional characteristics of the scene. The process should be such that "segmentation, structure, rotation and translation are uniquely determined, although none of them is determined by any 'cue' in isolation," Ullman [16]. Equally important is that the data is broken into small "nuclei" which having been properly analyzed can be grouped into larger structures. These structures can then be elaborated upon to derive full descriptions of the objects in the scene and their motions.

These considerations return us to the fundamental concept of exploiting the constancies in a changing environment in order to understand the components of that environment and their transformations. At the lower level the extracted tokens have associated features which the correspondence process can treat as locally constant. The correspondence process uses this constancy to identify the tokens throughout the sequence of images. At the higher level are the constancies which reflect what the system considers to be a significant component or object in the scene. The set of identified tokens is decomposed into groupings exhibiting this constancy. The groupings are thus taken to represent the objects of interest in the scene.

This strategy, however, does not depend upon reducing the input images to dot patterns. The tokens can be more complex structures, as we will see in the next section. In that section several systems which use edge and boundary descriptions in the initial analysis are discussed.

3. Edge and Boundary Analysis

3.1 Straight Edge Domain

We begin this section by discussing a system, Aggarwal and Duda [18] and Petermann [19], which analyzes software generated scenes containing rigid polygons. These polygons can be arbitrarily complex in shape and can possibly contain holes. Each polygon is restricted to move in a plane parallel to the image plane so that an input frame is formed by taking the orthographic projection of all the object planes onto the image plane. This projection essentially creates the silhouette of the object planes. In this way apparent objects in the image are formed from two or more overlapping actual polygons. The system must be able to derive descriptions of the actual polygons from the sequence of images of the apparent objects in motion.

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, i.e., is acute. 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 importance 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.

With the problem domain set in this manner the system begins processing by forming a correspondence between the objects in a given frame and the objects in the immediately preceding frame. This is accomplished by finding in both frames known "real" vertices, i.e., vertices with acute interior angles, which have similar angular measures. The matching of more than one vertex between two objects is restricted by the order in which the candidate vertices appear on each object in question. So the tokens, in this case the vertex points with the attribute of angular measure, are not matched independently on their attribute values. Rather, higher level constraints are imposed on the inter-token relationships.

The ordering constraint combines with the matching of known "real" vertices to provide a cue to which vertices having obtuse interior angles might be matching "real" vertices and should be compared for similarity in angular measure. If the vertices do have matching angular measures then the system assumes that they correspond and are "real" vertices. Of course, if the vertices actually represent a false vertex which is not exhibiting any change in its angular measure then the "real" vertex assumption may be incorrect. In this case, however, the system can either use information from previous frames to determine that the vertex is "false" or it can accomodate the error until there is a clear indication that the vertex is "false".

The matching of the "real" vertices between two frames provides the system with the correspondence (possibly many-to-one or one-to-many) of objects in the scene. With this correspondence the system can then determine what objects are in which one of the basic types of change and call special procedures developed to analyze each of those types. The function of these procedures is to create and continually update models of the actual objects in the scene. The vertex classification enables these procedures to decompose the apparent objects into sections comprising the visible portions of the constituent actual objects. The visible portions may simply indicate that the location and velocity components of the models should be updated. Or the models may require modifications such as adding a newly visible vertex or marking as currently not visible a just occluded vertex. In addition, two models might have to be merged or a given model split as more frames are analyzed. In this way the system uses these models not only to track the actual objects, but also to compile complete descriptions of the objects even though the objects may not have been totally visible in any of the given frames. Thus if an appropriate sequence of partial views of an object at some point presents each vertex and edge of that object, then the system will be able to derive a description for the object as a whole.

An example analyzed by the system is shown in Figure 5. The input

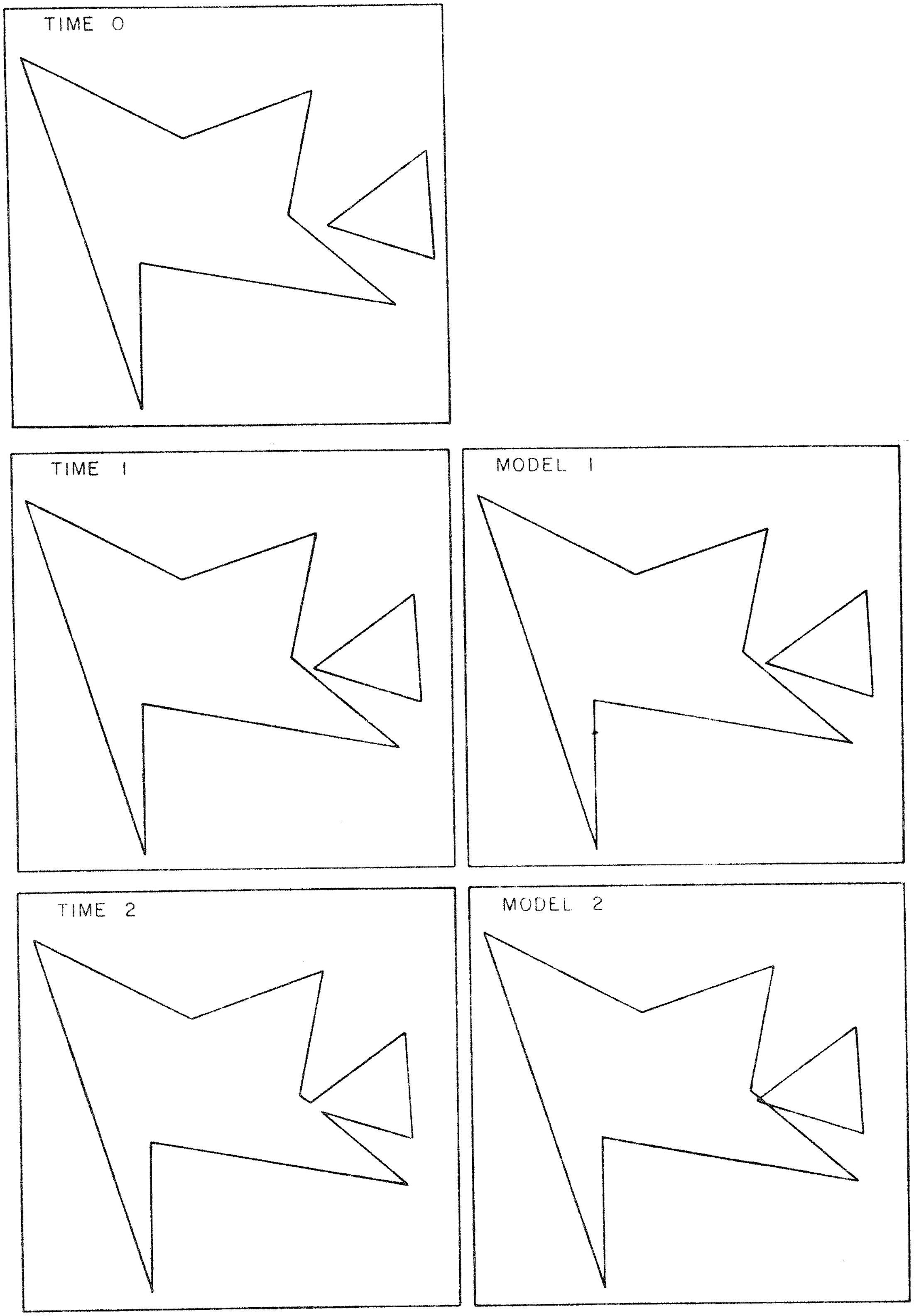

Figure 5. Example scene with derived models for a polygonal domain.

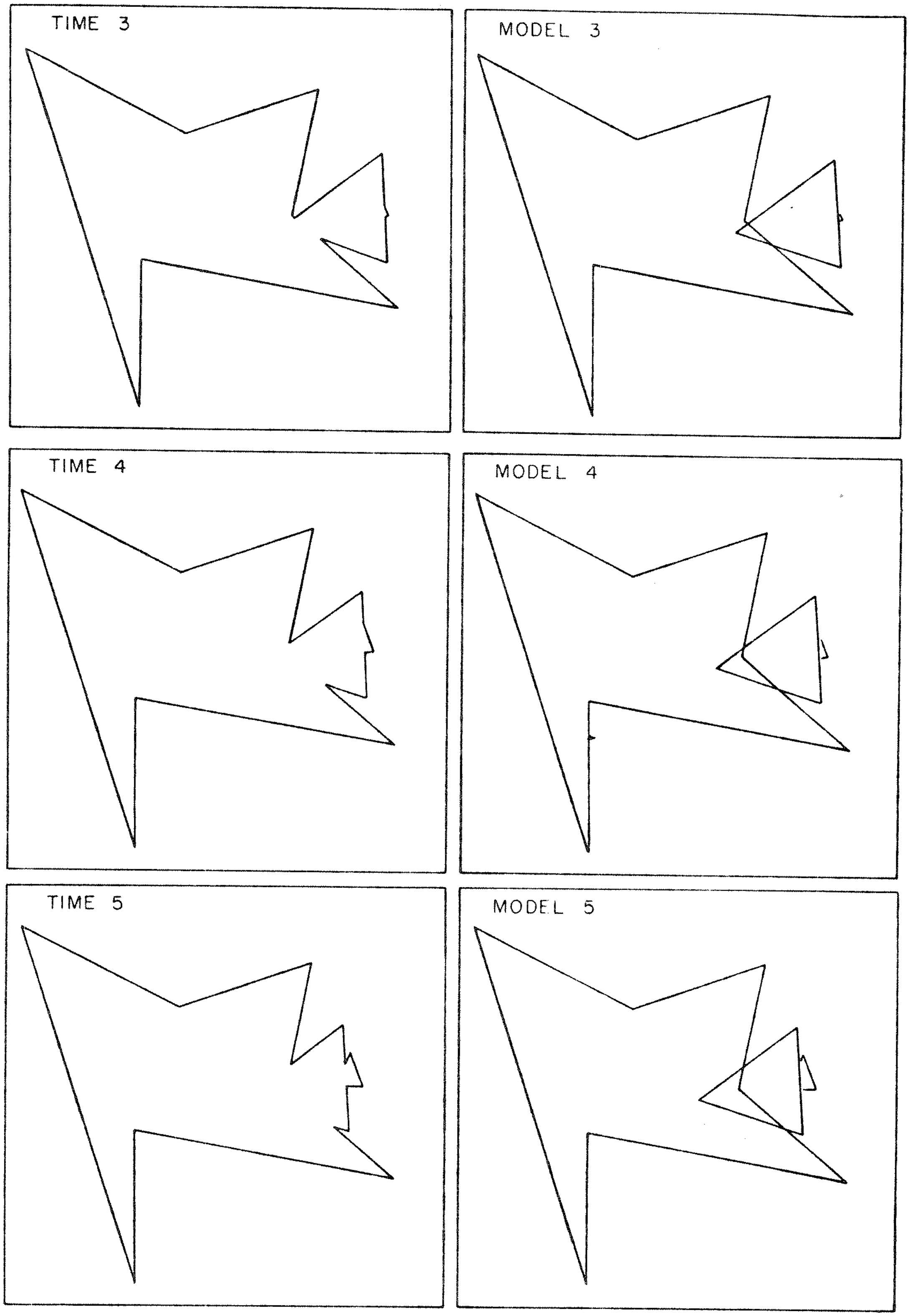

Figure 5. Continued.

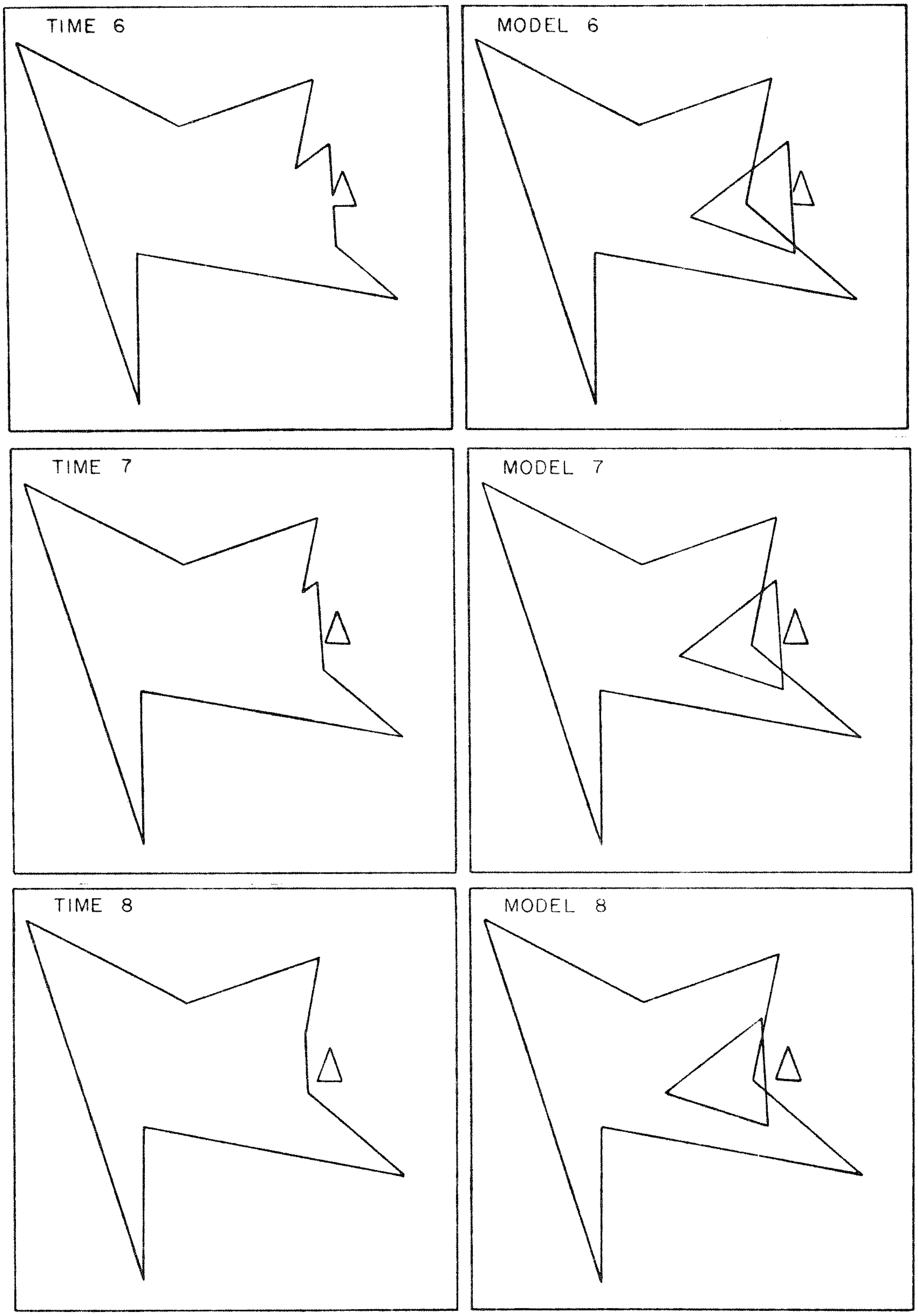

Figure 5. Continued.

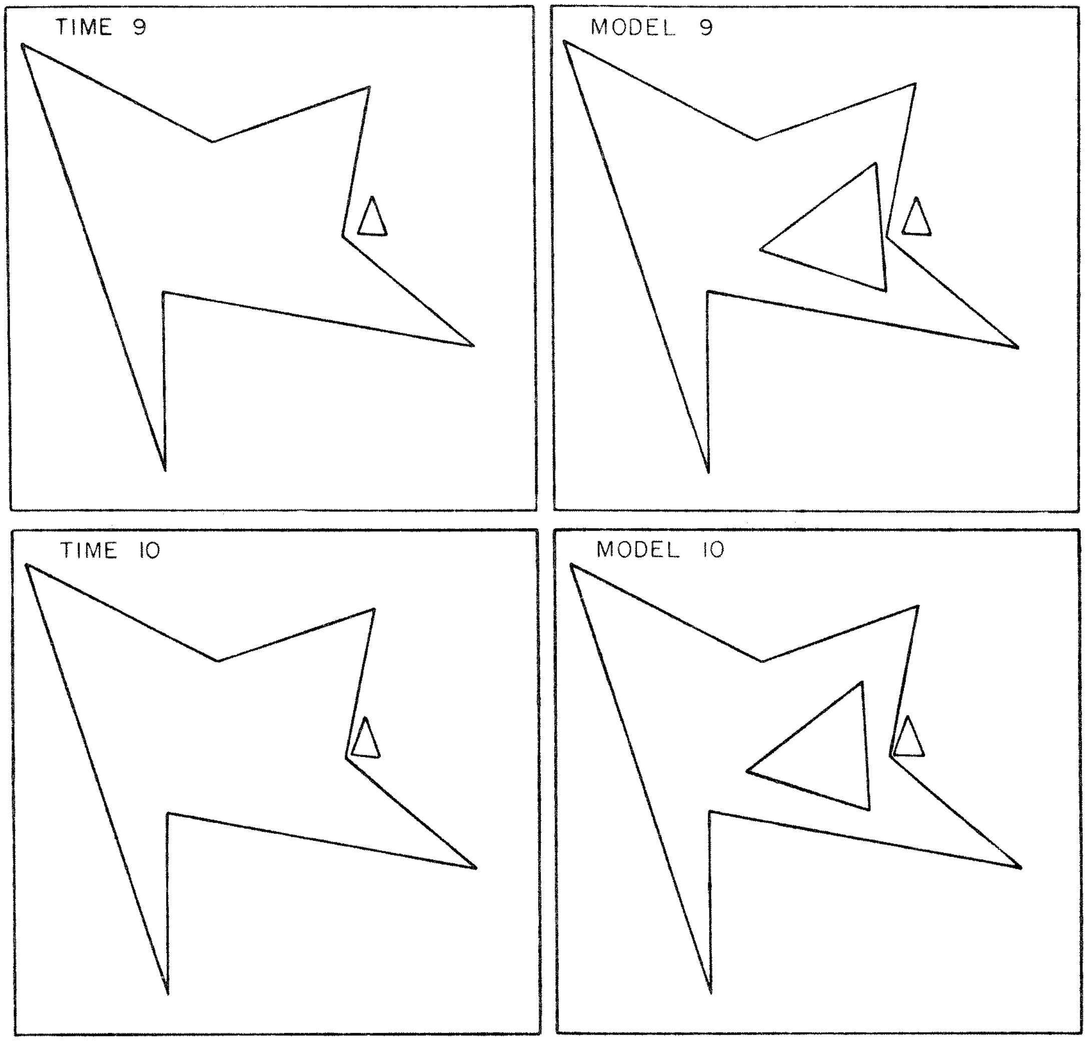

Figure 5. Continued.

sequence of frames is labeled TIME K for each time K, while the object models resulting from processing the sequence through time K are shown beside TIME K and labeled MODEL K. There are three actual objects in this scene: one large stationary object; a triangle moving to the left; and a smaller triangle, initially obscured, but later visible as it moves more slowly to the left. The portions of the objects which become occluded during the sequence are displayed at their projected positions. These projections are based on velocity estimates made while the particular section was visible. The single apparent object present in frames 2 through 6 is correctly decomposed into its constituent actual objects. Beginning with frame 3 this decomposition includes a partial object model for the smaller triangle which becomes visible one vertex at a time. This partial object model is updated when the next vertex is seen in frame 5 and is completed by the appearance of the final vertex in frame 7. Also of interest is that each of the basic types of change possible in the polygon domain of this system is instantiated in this example.

This system employs the constancy of angular measure at "real" vertices, as implied by the rigid object restriction, and several general observations about the polygon domain in the successful analysis of complex scenes of moving polygons. The "real" vertex constancy serves as the basis for matching individual vertices between consecutive frames. Higher level constraints are then used to form a correspondence for the objects in the frames. Finally, the correspondence provides cues, based on general knowledge, which enable the system to correctly interpret the changes as they occur and to use the interpretation to create and maintain detailed models of the actual objects in the scene and their movements.

3.2 Curvilinear Boundary Domain

The system, Martin and Aggarwal [20], discussed in the following, attempts to apply a similar type of analysis to scenes containing figures with curvilinear boundaries. The input is again restricted so that the objects independently move in planes parallel to the image plane. However, instead of software generated images, homogeneously shaded, opaque, planar figures are moved in front of an image dissector camera and a sequence of images is made. The camera approximates an orthogonal projection into the digital images which are preprocessed, McKee and Aggarwal [21], to extract the boundaries of the figures. The figure shading and the camera set-up form images in which overlapping figures are merged into single apparent objects. The task of the system is thus to derive descriptions of the 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.

For the system then the low level tokens are to be matched on the basis of a shape attribute. Thus the tokens must represent structures derived from the object boundaries having identifiable shapes, i.e., individual edge points are not adequate. 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, function of a 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 6 shows an object as 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 shape matching from the information required in the movement measurement process. In fact the ψ-s function is invariant to translation and rotation (see Martin and Aggarwal [20] 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 an object of 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

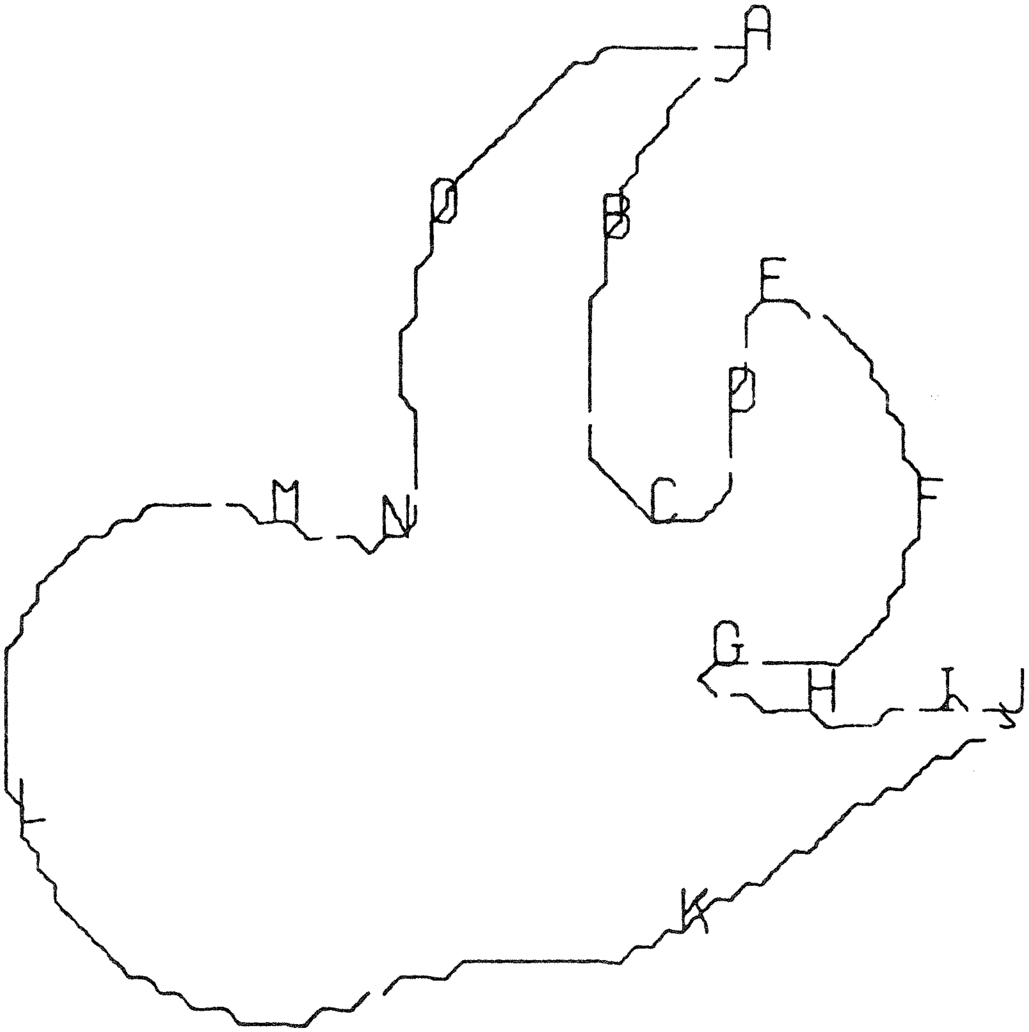

Figure 6. Image as decomposed into arcs.

consecutive pair, whose ψ-s function lines have similar slopes and lengths. From these "seed" arcs an edge segment can be "grown" by adding contiguous 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.

The example shown in Figure 7 is taken from a scene containing three actual objects: one central stationary object; one object on the left side moving from top to bottom; and one object on the right side moving toward the upper left corner. The first two images, however, contain only one apparent object. When comparing the shapes of the first image to those of the second, the system forms four edge segments. These edge segments are then grouped into three object models based on motion measurements. The object models formed in this way are inserted into the data base with arbitrarily chosen names. In Figure 7 each edge segment is labeled with the name of the appropriate object model. The observant reader will have noticed that no edge segments are formed from the center of the scene when the first two images are compared. This is due to the

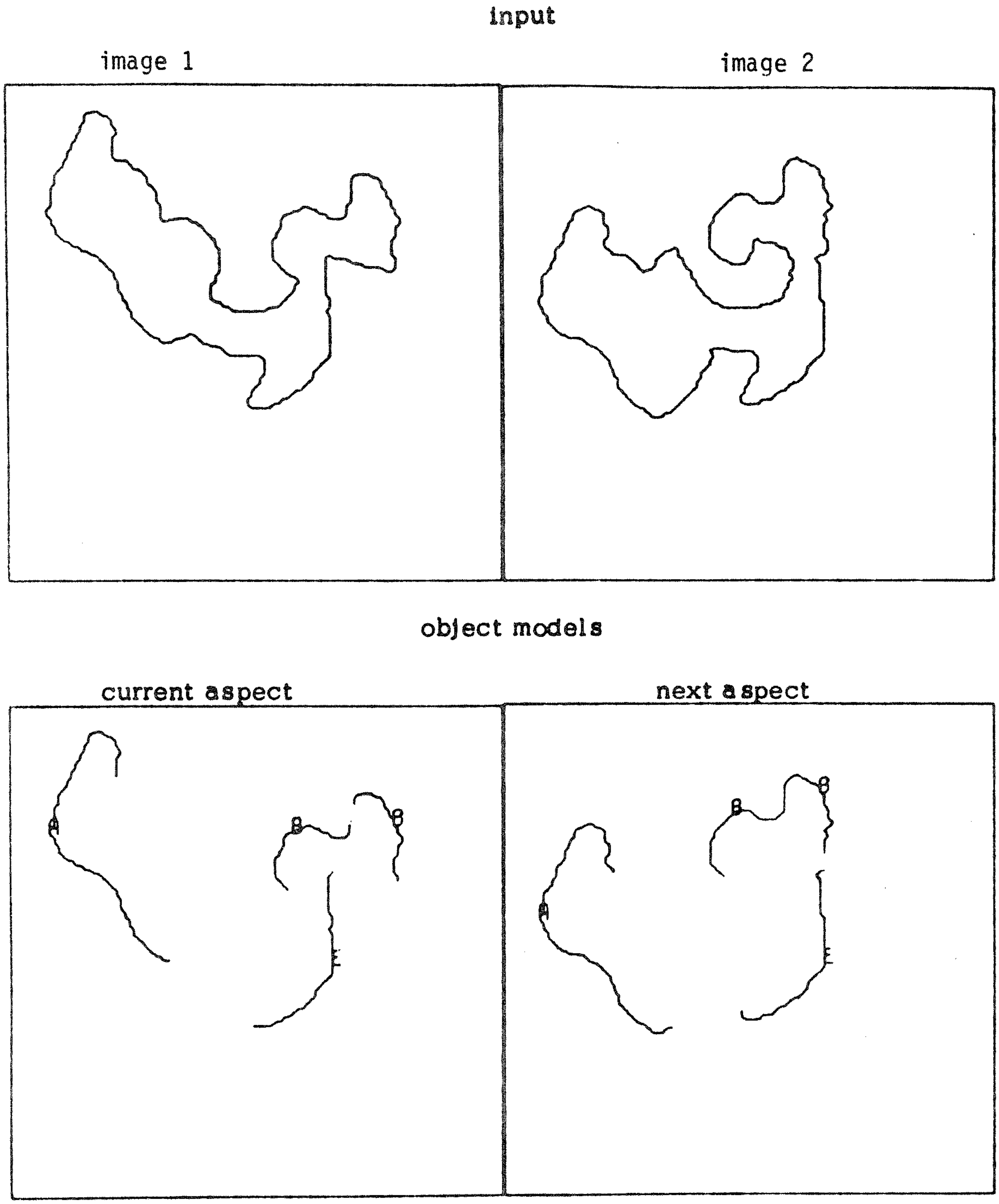

Figure 7. Example scene containing three occluding objects.

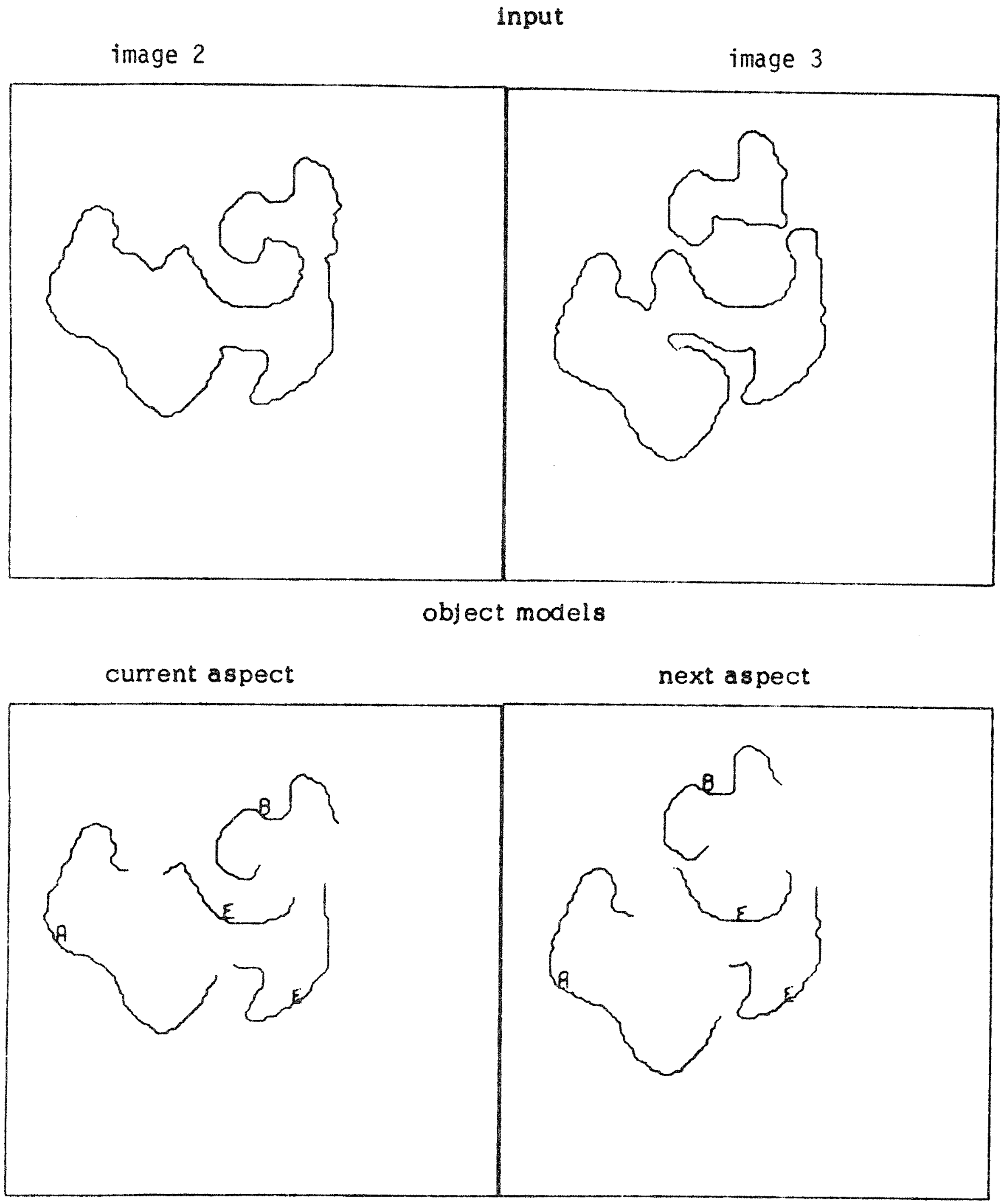

Figure 7. Continued.

extensive shape changes occurring in that part of the scene. It should also be noted that in the last image the upper object no longer overlaps the other object, causing the number of apparent objects to change. The shape matching procedures handle this case making the proper correspondence between the edges of the last image pair.

4. Conclusion

This chapter began with a discussion of the general problem of occlusion in scene analysis. Particular attention was brought to the analysis of occlusion in dynamic scenes, i.e., image sequences. Then several motion understanding systems were described in order to elucidate the interaction of change and constancy. Ostensively these systems were developed to analyze the changes which occur in sequences of images. However, prior to the change analysis the systems must determine the features of the scene which remain constant through at least short subsequences of the images. These constancy features are necessary to span the discontinuity inherent in the image sequence representation of the dynamic scene. In particular for each pair of consecutive images a correspondence must be formed to relate the appearance of a given token, i.e., a feature of interest, in one image to the occurrence of that token in the other image of the pair. The correspondence process essentially tracks the tokens throughout the sequence and provides the basis for the initial motion analysis. Convincing experimental results have been presented, Ullman [16], indicating that for certain "competing motion" stimuli the human visual system indeed accomplishes this task at a very early stage in the perceptual process. However, as argued in the introductory section of this chapter, the possibility of occlusion requires that some higher level information be used in forming the correspondence.

In three of the systems discussed, the movement measurements obtained from the tracked tokens played a crucial role in deriving object level interpretations of the dynamic scenes. In this way the changes detected at a low level are used in determining structures which provide constancy features for motion analysis at a higher level. For an example of top-down information flow in a hierarchical matching system see Roach and Aggarwal [22]. In any case, the inclusion of change information in the derivation of object descriptions has important implications on the data structures used by future scene analysis systems. Not only must the structures allow the descriptions to contain feature values which change in time and record traces of those changes, but also the structures must provide for features which appear, disappear, and change character. The first two changes are caused by occlusion of the object in question. The latter change, however, is due to the dynamic nature of the scenes being analyzed. For example, a person in a dynamic scene might be standing initially, then leaning slightly forward, then leaning forward and balancing on one foot but still standing, and then finally walking. The person has changed their status from "standing" to "walking" but has done so by gradually varying the value of their posture feature. These representation problems, along with the still important problems of token correspondence, object interpretation, and the interaction of low and high level information should form the core of much of the future research in dynamic scene analysis.

References

1. H. G. Barrow and J. M. Tenenbaum, "Recovering intrinsic scene characteristics from images," in *Computer Vision Systems,* A. Hanson and E. Riseman, eds., Academic Press, New York, 1978.

2. D. Waltz, "Understanding line drawings of scenes with shadows," in *Psychology of Computer Vision,* P. H. Winston, Editor, McGraw-Hill, 1975, pp. 19-92.

3. D. Huffman, "Impossible objects as nonsense sentences," in *Machine Intelligence 6,* B. Meltzer and D. Michie, Editors, Edinburgh University Press, Edinburgh, Scotland, 1971.

4. M. Clowes, "On seeing things," *Artificial Intelligence Journal,* vol. 2, no. 1, 1971, pp. 79-116.

5. A. Rosenfeld, R. Hummel and S. Zucker, "Scene labelling using relaxation operations," *IEEET-Systems, Man and Cybernetics, 6,* 1976, pp. 420-433.

6. F. W. Mounts, "A video encoding system with conditional picture-element replenishment," *Bell System Tech. J. 48,* 1969, pp. 2545-2554.

7. S. A. Underwood and C. L. Coates, "Visual learning from multiple views," *IEEET-Computers,* vol. C-24, no. 6, June 1975, pp. 651-661.

8. A. Guzmán, "Decomposition of a visual scene into three-dimensional bodies," in *Computer Methods in Image Analysis,* J. K. Aggarwal, R. O. Duda and A. Rosenfeld, Editors, IEEE Press, 1977, pp. 324-337.

9. 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, 1975, pp. 115-121.

10. W. N. Martin and J. K. Aggarwal, "Survey: dynamic scene analysis," *Computer Graphics and Image Processing,* vol. 7, no. 3, June 1978, pp. 356-374.

11. H.-H. Nagel, "Analysis techniques for image sequences," *International Joint Conference on Pattern Recognition,* Kyoto, Japan, Nov. 7-10, 1978.

12. Abstracts of the Workshop on Computer Analysis of Time-Varying Imagery, N. I. Badler and J. K. Aggarwal, eds., held in Philadelphia, PA., April 5-6, 1979.

13. J. O. B. Greaves, "The bugsystem: The software structure for the reduction of quantized video dots of moving organisms," *Proceedings of the IEEE,* vol. 63, 1975, pp. 1415-1425.

14. R. M. Endlich, D. E. Wolf, D. J. Hall, and A. E. Brain, "Use of a pattern recognition technique for determining cloud motions from sequences of satellite photographs," *Journal of Applied Meterology,* vol. 10, 1971, pp. 105-117.

15. G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behaviorial Science,* vol. 12, 1967, pp. 153-155.

16. S. Ullman, *The Interpretation of Visual Motion,* MIT Press, Cambridge, Massachusetts, 1979.

17. W. K. Chow and J. K. Aggarwal, "Computer analysis of planar curvilinear moving images," *IEEET-Computers,* vol. C-26, 1977, pp. 179-185.

18. J. K. Aggarwal and R. O. Duda, "Computer analysis of moving polygonal images," *IEEET-Computers,* vol. C-24, Oct. 1975, pp. 966-976.

19. R. Petermann, "Computer analysis of planar moving polygons," Masters Thesis, University of Texas at Austin, Jan. 1975.

20. W. N. Martin and J. K. Aggarwal, "Computer analysis of dynamic scenes containing curvilinear figures," *Pattern Recognition,* vol. 11, 1979, pp. 169-178.

21. J. W. McKee and J. K. Aggarwal, "Finding edges of the surfaces of three-dimensional curved objects by computer," *Pattern Recognition,* vol. 7, 1975, pp. 25-52.

22. J. W. Roach and J. K. Aggarwal, "Computer tracking of objects moving in space," *IEEET-Pattern Analysis and Machine Intelligence,* vol. PAMI-1, no. 2, 1979, pp. 127-135.