8 The Role of Parsing in High-Level Motion Processing

Peter Tse, Patrick Cavanagh, and Ken Nakayama

In a typical apparent motion display, one shape disappears at one location and a second shape then appears elsewhere. Motion is seen between the two locations. In this chapter we will describe a different apparent motion display which reveals new, fundamental properties of motion processing. In these displays, the second shape overlaps the first in location and is seen as an extension or growth of the first. We will call this type of display *Transformational Apparent Motion* because it creates impressions of changing figural shape, not just changing position. Transformational Apparent Motion occurs when a stimulus changes discretely from one configuration to another that overlaps with the first (figure 8.1a); rather than an abrupt exchange of shapes, a smoothly animated transition is seen from the first shape to the next. The perception of a smooth shape change despite the discrete nature of the stimulus is very much like the smooth shape change seen in computer graphics MORPHing programs.

The study of Transformational Apparent Motion reveals that an advanced degree of figural parsing and matching is integral to the perception of this class of motion (Tse, Cavanagh, and Nakayama, 1998). This type of apparent motion reveals several novel properties of motion processing. In particular, the study of Transformational Apparent Motion reveals that a stage of figural parsing and matching in the high-level motion pathway precedes the perception of motion. Transformational Apparent Motion comprises a class of apparent motions that obey different properties than those obeyed by standard apparent motion. Although we believe that both Transformational Apparent Motion and standard (translational or rotational) forms of apparent motion follow from the same high-level motion processing mechanism, we shall show that Transformational Apparent Motion allows us to study properties of parsing and matching that are not accessible using standard apparent motion as a probe.

Determining that something moved requires that the "something" be identified in the first instant and then paired off with what is presumed to be the same thing in the next instant. The first component of this process is to identify candidates at both instants and the second is to match them. We can call the first component *a parsing step* and the second *a matching step*. By

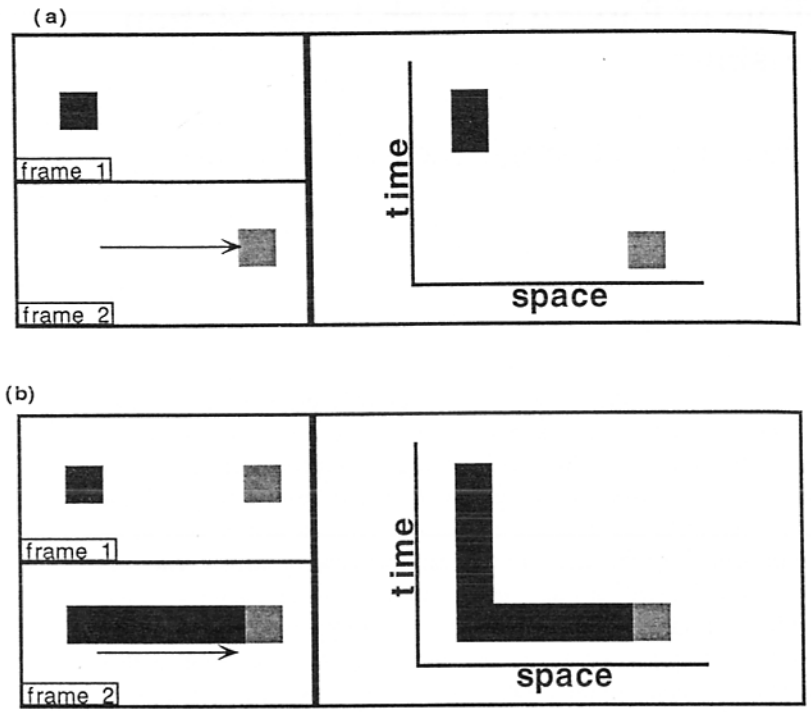

Figure 8.1 (a) In standard apparent motion, there is generally no space-time overlapping of frame 1 and frame 2 cues. There is therefore no problem of parsing the second frame to determine where the figures lie. The problem in standard apparent motion is simply one of matching given figures across frames. The arrow in frame 2 indicates the perceived direction of apparent movement. (b) In Transformational Apparent Motion, new image data can appear in the second frame that overlaps first-frame cues in space-time. Before this new image data can be matched to frame 1 figures, the figures in frame 2 must be parsed. There is no comparable parsing problem in most standard apparent motion displays. The arrow in frame 2 indicates the perceived direction of apparent movement.

parsing, we mean the spatial isolation and specification of individual objects, including any necessary segmentation away from overlapping objects or background elements of the image, as well as the discounting of noise. In this chapter, we will describe the nature of these two steps in high-level motion as they are revealed by Transformational Apparent Motion.

Low-level detectors such as, for example, directionally selective neurons, decompose the image pattern rather than segment it, allowing separate responses of many different units to any one location. A simplified low-level motion unit has two receptive fields with a particular spatial offset, and their activity is compared with a given temporal offset. If the two receptive regions become successively active in the preferred order, the unit responds, signaling that something moved in the direction that the unit is wired to detect. This property of registering the motion energy of multiple, superimposed components at a location, rather than treating the pattern at that location as a single entity, is for us the critical difference between low- and high-level processes. We will use this property to help separate any low-level response to a stimulus from the high-level analyses which are the subject of our chapter.

High-level motion has been studied most extensively in standard apparent motion displays where a configuration of items in the first time frame is replaced with a new configuration in the second, and perceived motion is reported (figure 8.1a). High-level motion has been described as a process of identifying forms and then matching those forms across time intervals (Anstis, 1980; Braddick, 1980). On the one hand, the stage of form identification is potentially the same as that available with unlimited viewing of static images. It could conceivably call upon all the processes of parsing (grouping, segmentation, amodal and modal completion, three-dimensional analysis, and object recognition) which are invoked for standard scene analysis. No author has found experimental evidence that would specify constraints on the processes available here, although many of the elements of scene analysis have been studied in the context of apparent motion. On the other hand, what these studies have shown is that few of the elements of scene or form analysis have much impact on matching in apparent motion (see, e.g., Kolers and Pomerantz, 1971; Burt and Sperling, 1981; Navon, 1976; Victor and Conte, 1990; Dawson, 1991; Watanabe and Cole, 1995). The most potent factor is simply spatiotemporal proximity: items in frame 1 will match most readily to their nearest neighbors in frame 2, irrespective of, say, shape or color changes.

This result is somewhat unsettling because it is not much different from what would be predicted by low-level mechanisms. That is, if large-scale, directionally selective units were available to respond to the isolated items of the two frames, they would signal motion of the nearest neighbor pairs and more or less ignore the form and color details of the items. There are examples of apparent motion which deviate from the expected properties of low-level detectors (such as matching opposite-contrast items or matching across eyes), but these examples do not rule out contributions from low-level units in most apparent motion displays. They only suggest that low-level mechanisms cannot be solely responsible for all motion phenomena. There are also examples of object properties—shape, color, depth—playing a role in matches (e.g., Green, 1986a, 1986b, 1989). But the effect of these factors is only revealed when the much stronger factor of proximity is carefully controlled.

Our studies with Transformational Apparent Motion displays suggest that these ambiguities concerning the insignificant contribution of form and feature information to high-level motion mechanisms are largely attributable to the spatial configurations used in standard apparent motion studies. When we use spatially overlapping elements we reveal a predominance of figural effects over proximity effects which cannot be seen in the standard displays.

The nature of these effects convinces us that high-level motion parsing and matching are sophisticated figural processes unrelated to the passive decomposition of low-level mechanisms.

We shall show that in the high-level motion processing pathway, it is parsed figures that are matched to parsed figures from scene to scene, in some cases ignoring nearest-neighbor principles. By *parsed figures*, we mean the attended portions of the completed segmentation which would occur with unlimited viewing of each individual frame. The unattended portions comprise the "background" to such figures. In Transformational Apparent Motion, new image data can appear closer to one figure than another, and still get matched as comprising a shape change in the more distant figure. We find that a set of parsing and matching principles that aids in determining figural identity within and between scenes holds for both Transformational Apparent Motion and standard apparent motion, and that this set approximately reduces to the nearest-neighbor principle for cases of standard apparent motion.

The importance of parsing could not have been revealed by research into standard apparent motion, because in standard apparent motion displays, the parsing of each successive scene is generally given unambiguously (figure 8.1a). In both scenes of a two-frame standard apparent motion display, there may be one or more figures, but what counts as a figure is unambiguous because these figures are usually spatially distinct. That is, they generally do not overlap, either within a scene or between scenes. In typical standard apparent motion, a figure seems to disappear at one location and reappear at a different, nonoverlapping location some time later. The problem in standard apparent motion experiments has generally been the match between figures, not the parsing of figures.

However, in Transformational Apparent Motion displays, there is usually ambiguity in determining which figure in one scene has become which figure in the following scene because of the spatiotemporal overlap of succeeding figures (figure 8.1b). That is, in the case of the apparent shape transformations of figures, new image data generally appear without the disappearance of the figure(s) that existed in the previous scene, although brief "figureless" intervals are tolerated, as in standard apparent motion. In the Transformational Apparent Motion displays that we employ, this problem of parsing must be solved before the problem of determining figural identity across scenes can be attempted. Since the image itself is not parsed, the visual system faces a problem of underdetermination in its efforts to correctly parse the image so as to coincide with the actual segmentation of the world into independent, but abutting or overlapping, objects. Since many possible parsings are consistent with a single image, the visual system, we argue, has evolved default "rules" for solving the parsing problem. These rules might be realized, for example, in the contour formation and segmentation processes of neural networks in visual cortex. Note that the parsing problem cannot be solved *after* the solution of the matching problem (i.e., of matching corresponding figures between successive scenes), because the parsing process is what defines the figures that are to be matched. Figural matching must therefore either follow or coincide with figural parsing. The visual system thus faces a deeper and logically prior problem than the correspondence (matching) problem in cases of apparent motion. This deeper problem is the problem of figural parsing.

Whereas standard apparent motion is generally insensitive to shape and color constraints so long as the two stimuli presented remain within the optimal range of spatiotemporal offsets (see, e.g., Cavanagh, Arguin, and von Grünau, 1989; or Kolers and von Grünau, 1976), Transformational Apparent Motion *is* sensitive to such constraints, because these are used by the parser to disambiguate figures in scenes that can only be ambiguously parsed. In Transformational Apparent Motion, we maintain, shape and feature (texture, luminance, color) constraints are used to parse the scene and match which extant figure, if any, has changed to subsume new image data. Motion perception follows this critical stage of figural parsing and matching. In particular, Transformational Apparent Motion is sensitive to geometrical constraints of contiguity, smooth contour continuity, and occlusion, and is influenced by more ecological constraints as well, such as an assumption that figures tend to persist across scenes, transforming and translating, rather than vanishing or appearing out of nowhere. These constraints, we argue, fall out of the nature of the parsing mechanism and its relationship with attentional mechanisms involved in the tracking of figures.

We also address the question of whether the parser tackles each successive scene in isolation. A few examples, in fact, show the contrary. The parser acts on the spatiotemporal flow of the image to segment figures from scenes, and takes the pattern of figures extant in the previous scene into account in parsing the present scene. We argue that the parser does not act on each successive scene regardless of the parsing of the previous scene. Rather, we conceive of a parser that is spatiotemporal, acting to segment figures from scenes in spite of spatial noise, such as within-scene occlusion, and temporal noise, such as a changing flux of partial information about a figure across scenes. Thus, the traditional distinction between within-scene parsing and between-scene correspondence matching is found to be too limited. Within-scene parsing and between-scene matching may be aspects of a single spatiotemporal parsing mechanism that segments figures over a certain optimal spatiotemporal extent.

BACKGROUND

Our initial point of departure in studying Transformational Apparent Motion phenomena was the examination of illusory line motion, originally described by Hikosaka, Miyauchi, and Shimojo (1993a, 1993b). This phenomenon arises most simply when a spot is briefly flashed, followed by a line adjacent to the location of the spot's offset. This is a perceptual illusion

because, contrary to the strong impression that there is motion along the length of the line away from the point of the spot's offset, the physical presentation of the line occurs all at once. Hikosaka, Miyauchi, and Shimojo had a novel explanation for illusory line motion, supported by several observations. They hypothesized that attention is drawn to the spot in the first frame and that this allocation of attention sets up an attentional gradient such that regions of the line closest to the attended spot are processed faster, thereby reaching a higher-level motion detector before more distant regions of the line. We have shown elsewhere (Tse and Cavanagh, 1995; Tse, Cavanagh, and Nakayama, 1998) that this speeded-processing account is inadequate, and that line motion is in fact an example of the more general class of phenomena that we call *Transformational Apparent Motion.*

In following up the line motion phenomenon, von Grünau and Faubert (1994) found that low-level features such as color, luminance, texture, and motion can drive the illusion of line motion, even when the initial spot and subsequently presented line possess different features. Faubert and von Grünau (1995) presented two initial spots followed by a line that appeared all at once, spanning the distance between them. The initial spots did not disappear when the spanning line appeared. Observers typically noted motion away from both spots toward the center of the line when the spots and bridging line were all the same color and luminance. In a critical variant of this two-spot case, Faubert and von Grünau (1992) differentiated the two spots using different colors such as, for example, red and green. They then presented a bridging line of the same color as one of the spots. According to the attentional speeding hypothesis, attention should still be attracted to both spots, and motion should be seen commencing from both spots, meeting in the center, just as in the case where the two spots are the same color. However, inward motion from both spots does not occur when the spots are different colors. Rather, motion seems invariably to commence only from the spot that is the same color as the line and proceed along the line all the way to the edge of the other spot, which seems to just stand idly by. Faubert and von Grünau (1995, experiment 2) extended this result by presenting one spot earlier than the other by various SOAs, from 0 to 1,200 milliseconds. The line always appeared to move away from the spot that was the same color or luminance as the line, regardless of SOA, even though attention presumably was attracted to the spot that had appeared more recently. Our own research has involved extending the competing-cue paradigm to study the spatial and temporal relationships between cues that determine the direction of Transformational Apparent Motion. Although Faubert and von Grünau did not interpret their results as evidence for parsing, we find that their competing-cue paradigm is a powerful probe of the factors that underlie parsing and matching. We have found that nonattentional properties, such as color, contiguity, and contour continuity, can bias the direction of illusory motion and can override the allocation of attention in certain circumstances.

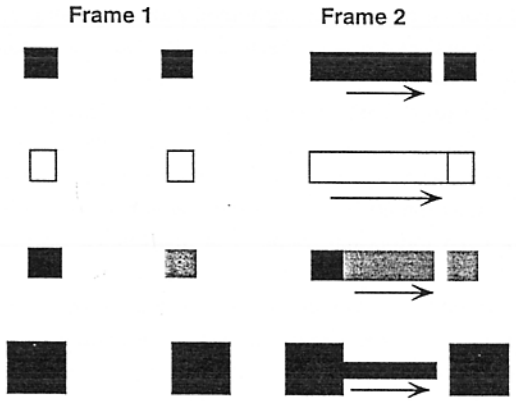

Figure 8.2 The arrow in frame 2 indicates the perceived direction of motion in this and all other figures. Here, motion always proceeds away from the contiguous cue.

THE ROLE OF CONTIGUITY

When there are two "competing" cues, as shown in figure 8.2, observers report that motion commences away from the contiguous cue upon the replacement of frame 1 by frame 2. When new image data, such as the line in frame 2, figure 8.2a or d, are presented that are the same color and luminance as the two cues shown in frame 1, the line appears to move away from the cue with which it is contiguous. More surprisingly, even if the image is defined by outlines, as in figure 8.2b, or if the contiguous line of frame 2 is the same color as the noncontiguous cue, as in figure 8.2c, the new image region appears to move away from the contiguous cue.

Although we will later reject the distinction between parsing and matching, let us assume for now that successive images are first parsed into figures, and that figures in each image are then matched with figures from the previous image. Consider the present case of contiguity of new image data to extant cues. When the new image data enter the visual system they are parsed. Since the new image data (the new regions in frame 2 of figure 8.2) are contiguous with only the lefthand cue of frame 1, the parsing at frame 2 includes the unchanged righthand cue and the longer lefthand line. Since this line overlaps the lefthand cue of frame 1 in space-time, it is matched to this cue. As with standard apparent motion, this discrete change between two steps is not seen as an abrupt exchange but rather as a smooth transition between the lefthand cue of frame 1 to its new state, the lefthand elongated cue of frame 2. In general, then, it appears that the matching operator matches successive parsed figures that overlap in space-time.

In figure 8.2d, the bridging line presented in frame 2 is not the same width as either of the square cues present in frame 1. Thus, there are deep concavities with both cues. What would happen if there were deep concavities with only one cue? We find that contour relationships, especially the presence

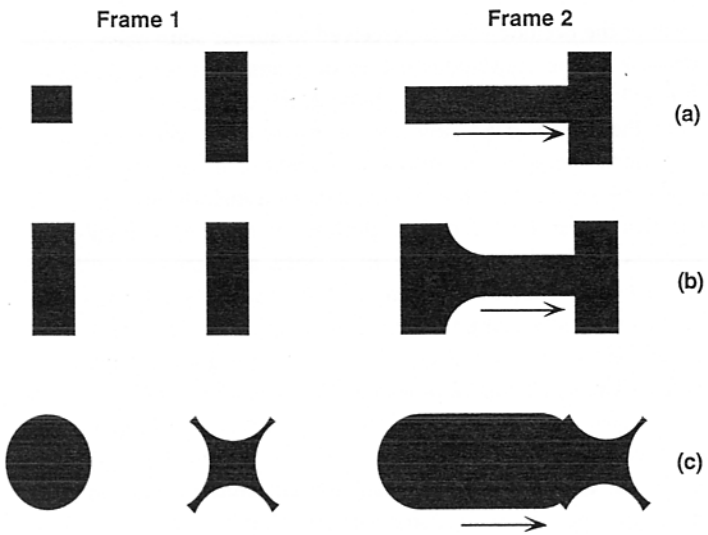

Figure 8.3 Motion proceeds away from the cue with which the bridging line shares smooth contours, and does not proceed away from the cue with which deep concavities are shared.

of deep concavities and smoothly continuous contours, play an important role in biasing the perceived direction of Transformational Apparent Motion between two competing cues, as described in the next section.

The Role of Contour Continuity and Deep Concavities

When the bridging line seen in frame 2 of figure 8.3, cases 8.a, 8.b, and 8.c, is presented contiguous with both competing cues, motion will appear to commence away from that cue whose contours smoothly continue into those of the new image data, as indicated by the arrow in each case.

Traditionally, it has been assumed that a parsing operator acts independently on each successive image, segmenting the scene into figures in a way not dependent on the parsing that occurred in previous scenes. Several researchers (Attneave, 1974; Marr and Nishihara, 1978; Hoffman and Richards, 1985; Biederman, 1987; Wilson and Richards, 1989; Kellman and Shipley, 1991) have described algorithms for figure segmentation, including separating figures at the locations of deep concavities. The deep concavities shared with just one of the cues in each case of frame 2 in figure 8.3, combined with the contour continuity of the new image data with the other cue in both cases, lead to a natural parsing of figures at the points of deep concavity. The new image data are parsed as belonging to the cue with which it shares smooth contours (i.e., where there are no deep concavities). The matching operation then matches new and previously existing figures such that only one cue is perceived to undergo a transformation in shape.

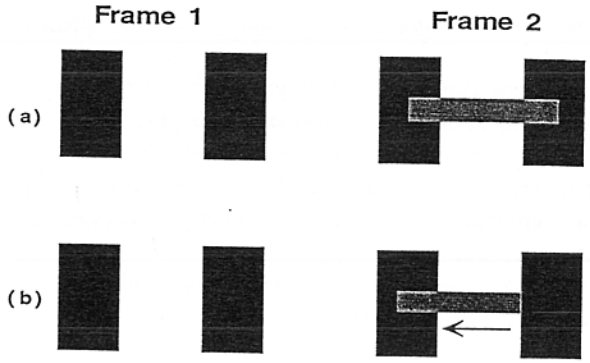

Figure 8.4 Only in case (b) there is overlap with one cue and contiguity with the other, does the line undergo transformational apparent motion from the contiguous cue. In (a) the line appears to emerge all at once and occlude the black rectangles.

From examples such as these, we conclude that a new portion of an image is parsed and matched to that cue with which it shares smooth contours. A bridging line will move away from this cue and not from the cue with which it shares deep concavities, when there are competing cues, as here. When there are no competing cues, mere contiguity is enough to parse new portions of an image as "belonging to" an already extant figure, even if there exist deep concavities, as in figure 8.2d. We assume that such parsing is an essential step in matching new portions of an image to extant figures, occurring either prior to or simultaneously with the matching operation.

Transformational Apparent Motion Is Figure-Centered

When a figure is progressively occluded, the occluded region continues to exist behind the occluder even though it is not visible in the image. Although the visible portions of the figure are changing in the image, the figure itself (visible plus invisible portions) does not change shape. Transformational Apparent Motion might operate on the visible image, such that any change in image shape would be perceived as motion. However, our tests demonstrate that Transformational Apparent Motion operates on figural changes of shape, not image-centered changes in shape.

When a line appears that seems to overlap both cues, as in frame 2 of figure 8.4a, the line seems to come on all at once. Although the occluded rectangles undergo a shape change in image coordinates, they do not appear to undergo Transformational Apparent Motion. They simply appear to become suddenly occluded, while maintaining their figural form. Moreover, the line does not appear to grow out from either of the rectangles of frame 1 because it occludes them, and therefore does not comprise a shape change in either one of them. In other words, the line is a new figure, and a new figure does not generate any shape change of an existing figure. We conclude from

this that changes in image shape do not underlie Transformational Apparent Motion.

When the line overlaps one cue but is contiguous with the other, as in figure 8.4b, the line seems to undergo Transformational Apparent Motion away from the abutting cue. We assume that it gets parsed as comprising a new state of the nonoverlapped cue, based on the role of contiguity described above. The line then appears to undergo Transformational Apparent Motion away from the contiguous cue, as a shape change of it. We conclude from this that changes in figural shape underlie Transformational Apparent Motion, whereas changes in image shape do not.

The Role of Occlusion in Figural Parsing

We have argued that parsing occurs prior to the perception of motion for displays such as we have shown. An integral part of parsing a scene is completing or inferring missing contour and figure information. Amodal completion (Kanizsa, 1979) involves completing a figure behind another occluding figure based on the presence of T-junctions and other border ownership cues (Nakayama and Shimojo, 1990). Modal completion involves inferring an occluding surface, complete with illusory contours (Kanizsa, 1979). If we are correct in our assertion that parsing takes place before Transformational Apparent Motion, then we must show that it demonstrates modal and amodal completion.

In figure 8.5a, the smaller cue of frame 1 is reported by all observers to undergo a shape transformation into the occluded bar in frame 2. The "distant" visible half of the occluded bar is perceived to appear subsequent to the "near" half because of the continuous nature of Transformational Apparent Motion. (If frame 1 is shown again after frame 2, the occluded bar seems to shrink back into the smaller cue.) However, in figure 8.5b, where we eliminated the T-junctions necessary for amodal completion, an entirely different percept is usually reported. Here the cue appears to transform into only the overlapping cue of frame 2. The distant portion of the bar is usually described by subjects as simply turning on all at once, simultaneous with the occurrence of the Transformational Apparent Motion of the overlapping cue. We argue that this occurs because now the overlapping (lefthand) portion of the bar is segmented as a separate figure from the more distant (righthand) portion of the bar. The small square of frame 1 is matched to only the overlapping portion, allowing a shape change to occur between it and the overlapping portion. The nonoverlapping portion, because it is parsed as being a new figure, cannot undergo a shape change from anything else. In conclusion, then, the perception of Transformational Apparent Motion transmitted behind an occluder in figure 8.5a, but the failure of Transformational Apparent Motion to transmit behind the occluder in figure 8.5b, indicates that amodal surface completion occurs before the perception of Transformational Apparent Motion.

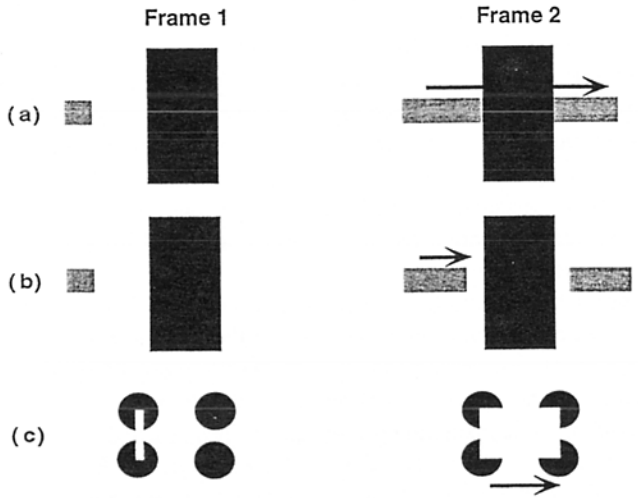

Figure 8.5 The parsing mechanism involves amodal and modal completion prior to the perception of motion. It does more than just segment the image. It is a generative process that completes inferred but missing contour and figural information.

Similarly, in the case of figure 8.5c, when presented with frame 1 followed by frame 2, observers attending to the frame 1 stimulus universally report seeing a white occluding surface change shape from a small rectangle into a large rectangle upon replacing frame 1 by frame 2. The illusory surface present in frame 2 must be completed as an occluding surface before the perception of its motion, because this illusory surface appears to change shape to cover all four inducers. This perception of an illusory surface undergoing a shape change indicates that modal surface completion occurs before, or as an integral part of, the perception of motion.

OTHER CONSTRAINTS ON TRANSFORMATIONAL APPARENT MOTION

We have seen that the problem of determining figural identity within and between scenes is intimately tied to the problem of parsing. We find that the geometry-based parsing principles discussed so far are not sufficient to fully account for the types of shape changes involved in certain displays where geometry supplies no basis for attributing new image data to one cue rather than another. Other constraints appear to operate on the parsed image prior to the perception of motion that help determine the trajectories underlying the best match.

Ullman (1979), following Attneave (1974), argued that correspondence matching between two multielement displays satisfies a minimal mapping, which, loosely speaking, can be thought of as a local nearest-neighbor correspondence mapping that satisfies certain additional constraints. One of these

constraints is that there is a preference for one-to-one mappings. According to this preference, each element in the frame 1 scene tends to map onto one and only one element in the following display, when the number of elements in the two frames is the same. Another constraint is the minimal cover property, according to which the number of motion trajectories is kept to a minimum in order to eliminate superfluous matches, while still supplying each element with a partner in the next frame. Although Ullman argued that correspondence matching occurs between low-level tokens, such as luminance blobs, and does not operate on high-level figures or object representations, we find that figures obey constraints similar to those outlined by Ullman. That set of motion trajectories that best conserves figure number and shape, and minimizes the number of figure trajectories, while satisfying the geometry-based constraints discussed above, will underlie the perception of motion. Such constraints were probably internalized by the motion system because they reflect the probabilities of motion events in the world. It seems that motion perception, like much else in vision, is an ex posteriori analysis, based on processing that has inherent in it inferential assumptions that constrain the multitude of interpretations compatible with the ambiguous input.

PARSING AND MATCHING COMPRISE A SINGLE OPERATION

Traditionally, parsing has been studied in terms of static images, and matching in terms of standard apparent motion displays. This has led to a certain implicit assumption that parsing, even if it precedes matching, acts on each successive scene independently. However, we find evidence that the parser takes the configuration of figures in a given scene into account in parsing the following scene. The parser tends to parse ambiguous scenes so that they include the figures that were present in the previous scene. But if the parser is concerned with between-scene figural correspondences, the distinction between parsing and matching is blurred.

When frame 2 of figure 8.6 is presented in isolation, observers do not report the presence of illusory contours or illusory occluding figures. Rather, they report an array of eight squares. Only when subjects see the transition between frames 1 and 2, do a majority of them report seeing an illusory "occluding" white bar, which is seen as the transformation of the small leftmost rectangle of frame 1 into the black-tipped white bar in frame 2 that "occludes" the large vertical rectangles from frame 1 as it changes its shape. If the parsing mechanism operated on each successive scene independently of the configuration of figures that was present in the previous scene, then we would expect subjects to report the presence of several unconnected squares in frame 2 even when presented with the transition between frames 1 and 2. This is not the case, however, as originally suggested by Kellman and Cohen (1984), who showed that an illusory occluding triangle could be generated by a succession of frames in none of which, considered as an isolated frame, an illusory figure could be generated. What we add to their

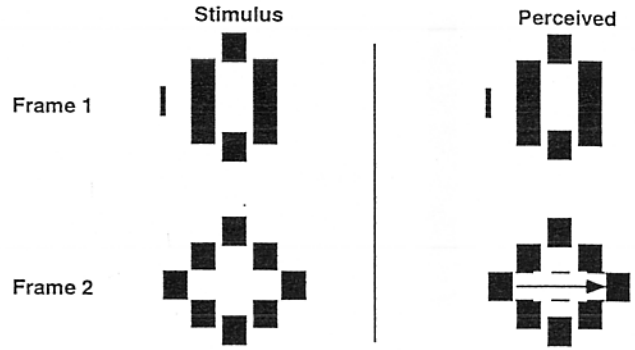

Figure 8.6 If frame 2 is presented in isolation, observers do not report seeing an occluding white bar. Rather, subjects generally report seeing a set of eight squares. However, if frame 2 is shown right after frame 1, observers report seeing a black-tipped occluding white bar undergoing Transformational Apparent Motion to the right.

work here is the ambiguity of parsing in frame 2. None of their frames was ambiguous for parsing. In our case, parsing is ambiguous, and the configuration of figures in frame 1 biases subjects to parse the scene in frame 2 to include an occluding illusory bar, as opposed to parsing the scene into several separate squares, as in the frame-2-only control case. This implies that the parsing mechanism must take into account the configuration of parsed figures in the previous scene in determining the parsing of the present scene. But if parsing is not a within-scene mechanism, but rather a between-scene mechanism, the distinction between parsing and matching becomes merely a heuristic distinction. Since the parsing mechanism seems to take into account the structure of the previous scene in segmenting the present scene, it becomes difficult to talk of independent or serially arranged parsing and matching mechanisms. In all likelihood, the parsing and matching operations are nonseparable aspects of a single spatiotemporal parsing operation that involves segmenting figures out from spatiotemporally noisy image data.

PARSING PRINCIPLES OVERRIDE MOTION ENERGY ACCOUNTS

Note that parsing in space-time cannot be accounted for in terms of low-level motion energy filters, because such filters are sensitive only to motion energy. Thus, a criticism we must meet is that Transformational Apparent Motion is simply a product of motion energy. As we have already maintained, many of our displays involve motion energy in several often opposing directions. Since the signal due to motion energy is often ambiguous, we have argued that while motion energy is an important low-level constraint on motion perception, there must be other higher-level constraints operating prior to motion perception, such as the simple parsing-related principles described here.

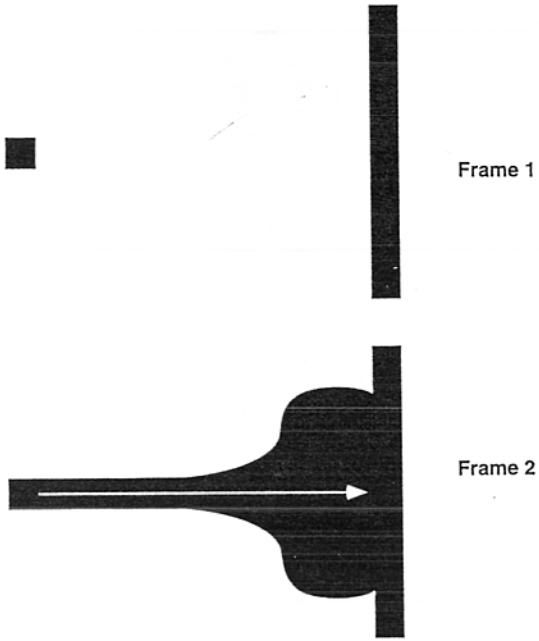

Figure 8.7 Even though the center of the new image data in frame 2 is closer to the righthand cue, motion proceeds away from the lefthand cue, suggesting that more is involved than the matching of luminance blobs, as is commonly assumed in motion energy models. In particular, the motion proceeds from left to right in this example because the new image data are matched as comprising a shape change in the lefthand cue, based on the location of smooth contours and deep concavities.

In order to directly answer the charge that Transformational Apparent Motion is simply a product of motion energy, we can do experiments of the sort shown in figure 8.7. Although there are many different motion energy models, all have in common that the centroid of a luminance blob remains the centroid regardless of spatial frequency. Moreover, motion energy is generally held to move from one centroid to the nearest centroid, as in the nearest-neighbor principle. For example, in figure 8.7, the centroid of the new luminance blob appearing in frame 2 is closer to the righthand cue. Nonetheless, motion is perceived to commence away from the cue that does not share deep concavities with the new image data.

In many of our displays, the motion perceived is only one of many possible motions that might have occurred between the endpoints defined by the initial and final shapes. If one were to conceive of the situation between successive frames in our displays from the perspective of motion-detecting filters, one would find that motion energy exists in multiple directions. However, subjects tend to perceive that the single-motion scenario that is most consistent with certain conservative assumptions that the visual system seems to make about how things do in fact translate and transform when they or we move about the world. Thus, only some higher-order principle or built-in inferential mechanism, such as might be realized by a parser that matches a figure onto a spatiotemporally overlapping figure in a subsequent frame, can account for the motion percept indicated in figure 8.7.

CONCLUSION

We find that Transformational Apparent Motion reveals several properties of apparent motion that are not obvious in standard apparent motion displays. In particular, the matching stage appears to operate on parsed scene descriptions where contour continuity, color, texture, and shape play a major role, because they aid in the parsing process. Moreover, parsing seems to be an integral part of a combined parsing/matching operation which segments objects over time and space, not independently within each frame. The high level of representation used as a basis for these motion computations ensures that the trajectory calculated by the high-level motion processing stream is seldom at odds with the segmentations of objects which would arise in static scenes or which emerge once the motion is completed.

Although some researchers (Anstis, 1980; Braddick, 1980) argued for a stage of form extraction in the high-level motion processing stream, subsequent tests using standard apparent motion found little or no contribution of form or feature to correspondence matching. Thus, throughout the 1980s, many vision researchers sought to reduce the high-level motion processing stream to a variant of the motion-energy–detecting low-level stream (e.g., Marr and Ullman, 1981; Adelson, 1982; Adelson and Bergen, 1985; Burr, Ross, and Morrone, 1986; Watson and Ahumada, 1983; van Sandeen and Sperling, 1984; Chubb and Sperling, 1988; Cavanagh and Mather, 1989). We suggest that the reason researchers concluded that form and feature information plays little if any role in correspondence matching was because they studied matching using standard apparent motion displays where the solution to the deep problem of figural parsing is given, leaving only the problem of matching. By using Transformational Apparent Motion displays instead, the contribution of the form extraction process is made explicit (see also Grossberg, 1998, in chapter 1; Yuille and Grzywacz, 1998, in chapter 6; Watanabe and Miyauchi, 1998, in chapter 3). In particular, the contribution of features to parsing, and therefore to matching, can be studied in a way not accessible using standard apparent motion as a test. We hope that our research stimulates further experimental and theoretical investigations into the nature of form-motion interactions which have been heretofore neglected. Indeed, the work described in this chapter was recently modeled by Baloch and Grossberg. Francis and Grossberg (1996) recently modeled form-motion interactions to simulate Korte's law (see also Grossberg, chapter 1).

We have concentrated on the high-level, figural aspects of motion while trying to avoid discussion of the contributions of motion-energy detectors. Typically, objects do not change shape in discrete steps, and both low-level

and high-level motion signals will be produced in tandem and quite often be consistent. There are several inconsistencies, nevertheless. In particular the "aperture problem" leads to a variety of local motion vectors from the contours of a large moving object that may be inconsistent with its global motion vector. The high-level signal, on the other hand, is the global direction of the object, and so suffers no ambiguities. It serves as a solution to the aperture problem and may explain part of the reason for the evolution of this separate system.

The human visual system needs to track the movements of figures, such as those posed by predators and prey. Presumably, the low-level motion system contributes to the detection of figures and their motion, and thus must feed into the parsing/matching mechanism. Attention is also intimately involved in the tracking of figures. While we have largely avoided any in-depth discussion of the relationship between parsing and attention, we assume that the relationship is an intimate one. Roughly put, attention selects and defines a figure from the preattentive output of the parser, and the parser then aids in detecting shape and location changes in this attended figure in subsequent scenes. We argue that the parsing/matching mechanism and attention work in tandem in tracking figures as the attended figures move about.

In conclusion, we believe Transformational Apparent Motion to be a form of motion perception that follows a set of rules that are not applicable in the case of standard apparent motion. This is primarily because in translational apparent motion displays, parsing is given in each separate image, whereas in Transformational Apparent Motion displays, the problem of figural parsing within a scene is not separable from the problem of matching, which is really the problem of determining figural identity over successive scenes. As such, Transformational Apparent Motion, particularly in the context of the competing-cue paradigm that we have used, allows us to probe properties of the parsing/matching process more deeply than was the case using standard apparent motion displays.

REFERENCES

Adelson, E. H. (1982). Some new illusions and some old ones analyzed in terms of their Fourier components. *Investigative Ophthalmology and Visual Science, Suppl. 22,* 144.

Adelson, E. H., and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America, A 2,* 284–299.

Anstis, S. M. (1980). The perception of apparent movement. *Philosophical Transactions of the Royal Society of London, B 290,* 153–168.

Attneave, F. (1974). Apparent movement and the what-where connection. *Psychologia, 17,* 108–120.

Baloch, A., and Grossberg, S. (1996). Neural dynamics of morphing motion. *Investigative Ophthalmology and Visual Science, 37,* 3419.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115–117.

Braddick, O. J. (1980). Low-level and high-level processes in apparent motion. *Philosophical Transactions of the Royal Society of London, B 290,* 137–151.

Burr, D. C., Ross, J., and Morrone, M. C. (1986). Seeing objects in motion. *Proceedings of the Royal Society of London, B 227,* 249–265.

Burt, P., and Sperling, G. (1981). Time, distance, and feature tradeoffs in visual apparent motion. *Psychological Review, 88,* 171–195.

Cavanagh, P., and Mather, G. (1989). Motion: The long and short of it. *Spatial Vision, 4,* 103–129.

Cavanagh, P., Arguin, M., and von Grünau, M. (1989). Interattribute apparent motion. *Vision Research, 29,* 1197–1204.

Chubb, C., and Sperling, G. (1989). Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *Journal of the Optical Society of America, A 5,* 1986–2007.

Dawson, M. R. W. (1991). The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. *Psychological Review, 33,* 569–603.

Faubert, J., and von Grünau, M. W. (1992). Split attention and attribute priming in motion induction. *Investigative Ophthalmology and Visual Science, 33,* 1139.

Faubert, J., and von Grünau, M. W. (1995). The influence of two spatially distinct primers and attribute priming on motion induction. *Vision Research, 35,* 3119–3130.

Francis, G., and Grossberg, S. (1996). Cortical dynamics of form and motion integration: Persistence, apparent motion, and illusory contours. *Vision Research, 36,* 149–173.

Green, M. (1986a). Correspondence in apparent motion: Defining the heuristics. *Proceedings of Vision Interface '86,* 337–242.

Green, M. (1986b). What determines correspondence strength in apparent motion? *Vision Research, 26,* 599–607.

Green, M. (1989). Color correspondence in apparent motion. *Perception and Psychophysics, 45,* 15–20.

Hikosaka, O., Miyauchi, S., and Shimojo, S. (1993a). Focal visual attention produces illusory temporal order and motion sensation. *Vision Research, 33,* 1219–1240.

Hikosaka, O., Miyauchi, S., and Shimojo, S. (1993b). Voluntary and stimulus-induced attention detected as motion sensation. *Perception, 22,* 517–526.

Hoffman, D. D., and Richards, W. (1985). Parts of recognition. *Cognition, 18,* 65–96.

Kanizsa, G. (1979). *Organization in Vision: Essays on Gestalt Perception,* 113–134. New York: Praeger Press.

Kellman, P. J., and Cohen, M. (1984). Kinetic subjective contours. *Perception and Psychophysics, 35,* 237–244.

Kellman, P. J., and Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology, 23,* 141–221.

Kolers, P. A., and Pomerantz, J. R. (1971). Figural change in apparent motion. *Journal of Experimental Psychology, 87,* 99–108.

Kolers, P. A., and von Grünau, M. (1976). Shape and color in apparent motion. *Vision Research, 16,* 329–335.

Marr, D., and Nishihara, H. K. (1978). Representation and recognition of three-dimensional shapes. *Proceedings of the Royal Society of London, B 200,* 269–294.

Marr, D., and Ullman, S. (1981). Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London, B 211,* 151–180.

Nakayama, K., and Shimojo, S. (1990). Towards a neural understanding of visual surface representation. In T. Sejnowski, E. R. Kandel, C. F. Stevens, and J. D. Watson, eds., *The Brain, 55,* 911–924. Cold Spring Harbor Laboratory, New York: Cold Spring Harbor Symposium on Quantitative Biology.

Navon, D. (1976). Irrelevance of figural identity for resolving ambiguities in apparent motion. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 130–138.

Shimojo, S., and Richards, W. (1986). "Seeing" shapes that are almost totally occluded: A new look at Parks's camel. *Perception and Psychophysics, 39,* 418–426.

Tse, P., and Cavanagh, P. (1995). Parsing occurs before line motion. *Investigative Ophthalmology and Visual Science, 36,* 4.

Tse, P., Cavanagh, P., and Nakayama, K. (1996). The roles of attention in shape change apparent motion. *Investigative Ophthalmology and Visual Science, 37,* 4.

Tse, P., Cavanagh, P., and Nakayama, K. (1998). *Transformational Apparent Motion and the Parsing Problem.* Unpublished manuscript.

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

van Sandeen, J. P. H., and Sperling, G. (1984). Temporal covariance model of human motion perception. *Journal of the Optical Society of America, A 1,* 451–473.

Victor, J. D., and Conte, M. M. (1990). Motion mechanisms have only limited access to form information. *Vision Research, 30,* 289–301.

von Grünau, M. W., and Faubert, J. (1994). Inter- and intra-attribute effects in motion induction. *Perception, 23,* 913–928.

Watanabe, T., and Cole, R. (1995). Constraint propagation of apparent motion. *Vision Research, 35,* 2853–2861.

Watson, A. B., and Ahumada, A. J., Jr. (1985). Model of human visual motion sensing. *Journal of the Optical Society of America, A 2,* 232–342.

Wertheimer, M. (1912, 1961). Experimental studies on the seeing of motion. In T. Shipley, ed., *Classics in Psychology,* 1032–1088. New York: Philosophical Library.

Wilson, H. R., and Richards, W. A. (1989). Mechanism of contour curvature discrimination. *Journal of the Optical Society of America, A 6,* 1.

III Heading and Structure from Motion

One of the most important functional aspects of motion is to recover visual information from an array of moving objects. Chapters 9 through 11 present research on the recovery of observer motion from optic flow from computational, physiological, and psychophysical perspectives. When an observer moves through an environment, images of objects flow across the surface of the retina. This pattern on the retina—optic flow—has been proposed as the basis by which an observer can recover both motion and the three-dimensional surface structure of the environment. As reported in chapters 9 and 11, recent psychophysical findings have shown that human mechanisms for judging heading work quite well under a variety of suboptimal conditions. In order to explain this fact, Hildreth and Royden in chapter 9, and Warren in chapter 11, have put forth different models.

Among the many computational models they review, Hildreth and Royden show that their models can use changes in image intensities directly to recover observer motion, without explicitly computing optical flow. Rotational components in the scene are eliminated by subtracting the image velocities from two points located at a depth discontinuity. Subsequently, the best point of intersection of the directions of these difference vectors is calculated. Warren, on the other hand, suggests that the robustness in human performance can be accounted for by spatial pooling. Accordingly he has suggested a two-layer neural network model in which each unit in the output layer pools all the signals from the input layer. The model becomes a kind of template model after training. While Hildreth's and Royden's model may be computationally more efficient than Warren's, Warren's model may be more consistent with the biological properties of the visual system. In chapter 10, Tanaka shows that the medial superior temporal area (MST) is tuned to expansion, contraction, clockwise and counterclockwise rotation, and translation over a large visual field, and suggests that these cells have properties that appear suited to detecting observer's motion. This implies that MSTd could act like a template for global optic flow fields.

In chapter 12, Todd reviews research on the recovery of surface structure from motion. He presents experimental results showing that observers