## CHAPTER 1

## THE BASIC-ELEMENTS PROBLEM

### 1.1 Theoretical Considerations

At the core of the capacity to visually perceive motion lies the ability to identify distinct elements in the incoming visual array as representing the same physical object. For a given element $\mathbf{X}$ in $A(t)$ (the visual array at time $t$ ) its counterpart $X^{\prime}$ in $A\left(t^{\prime}\right)$ (the visual array at a later time $t^{\prime}$ ) must be located. $X$ ' need not be identical to X in appearance, in fact the very difference between them might serve for the subsequent analysis of the motion, or change, attributed to the object that both $X$ and $X^{\prime}$ represent.

Before investigating the fundamental correspondence function which matches elements in successive views, we face the more elementary problem of finding the domain and range of this function. That is, what is the set of elements that are mapped in the process of motion perception. I shall refer to the basic elements comprising this set as the correspondence tokens. When formulated in terms of the domain and range of a function, this basic elements problem seems misleadingly innocuous. Yet some of the more profound controversies in theories of motion perception stem from a disagreement as to what the basic elements are. Different approaches suggested for the visual analysis of motion have differed in the assumptions they make about the nature of the basic elements. As the particular choice is usually embedded in the theory rather than explicitly stated, it is rarely justified and its implications are seldom discussed. The extreme empiricist view, for instance, suggests that humans learn to associate objects with their different views and are thereby able to recognize these objects in motion. The problem of accepting or refuting this theory is tantamount to the question of whether learned views are indeed the correspondence tokens. Similarly, the controversy of whether or not motion analysis is
based upon object recognition is a dispute concerning whether or not recognized objects (or object descriptions) constitute the basic elements of the correspondence function.

The basic elements problem should be the first one considered because of its fundamental importance and the resulting impact it must have on the course of the research. If, for example, object recognition indeed precedes motion analysis, then the correspondence procedure becomes rather simple: the particular element in $A\left(t^{\prime}\right)$ corresponding to, say, the white rabbit under the oak tree in $A(t)$, is probably uniquely determined without great difficulty. On the other hand, in such theories of perception, both the correspondence and the threedimensional interpretation of unfamiliar objects in motion become enigmatic, especially in the case of objects whose twodimensional views are unrecognizable, as will be discussed in Chapter 4. If, however, the basic elements are more primitive, such as lines, dots, and edge fragments, then familiarity and recognizability cease to play an important role. The correspondence problem then becomes difficult, as there might be many candidates in $A\left(l^{\prime}\right)$ for a possible match to, say, a grey line segment in $A(t)$.

The goal of the current section is to suggest, on theoretical grounds, a plausible domain of basic elements. It will be argued that the proper level at which to carry out the correspondence process is that immediately following the initial organization of the raw intensity data into meaningful units. (Meaningful units are, as discussed in the introduction, symbols in the representation whose meanings are founded in the environment, not in the intensity array.) The next two sections provide evidence supporting this view. Section 1.2 shows that organization of the raw intensity data into meaningful units precedes the correspondence operation, while Section 1.3 shows that this antecedent organization is limited to simple, primitive units.

## A plausible domain of basic elements

A common approach to picture-matching problems, found in the psychological literature [Kabrisky, 1966; Anstis, 1970; Bell \& Lappin, 1973; Pantle \& Picciano, 1976] as well as in applications [Leese, Novak \& Taylor, 1970; Smith \& Phillips, 1972; Wolferts, 1974] is to suggest that the correspondence process takes place prior to any organization of the raw intensity data.

This general view can be further divided according to the particular operation used to establish the correspondence. In one approach [Anstis, 1970] individual points are paired on the basis of their intensity similarity. In this view the role of correspondence tokens is assumed by single intensity points. In more global approaches, a sub-region $A$ of a given intensity array is considered a basic element, and its counterpart $A^{\prime}$ in a second intensity array is sought on the basis of similarity between their intensity distributions. The similarity of the sub-arrays is usually measured either by a cross-correlation technique, in which maximum correlation is sought, or by subtraction [1.1], wherein the match is indicated by a minimal value.

In the above approaches the correspondence is determined on the basis of similarity between intensity distributions. There are two main arguments against establishing the correspondence by such grey level similarity comparisons.

First, grey level correlations can be expected to yield the correct match only in the very simple case of translation in the image plane. In the general case, in which the two pictures to be compared represent an object in general motion, there is no reason why any of the above grey level comparisons should yield veridical results [1.2]. One of the problems that arises is the "window size" problem. The intensity comparison cannot be performed on single points, nor can it be performed on the image as a whole. The correlation ought to be established between patches of the "right size", but there is no satisfactory way of predetermining that size. The situation becomes even
more complex and less amenable to grey level comparisons in the case of several objects which are simultaneously engaged in different motions.

The second objection stems from the fact that grey level distributions and their changes do not correspond directly to physical entities and their motion, while it is the latter that should be established. As illustrated, for instance, by the Cornsweet illusion [Cornsweet, 1970; Ratliff, 1972], a visible edge can have radically different underlying intensity distributions that are perceptually indistinguishable from one another. The Cornsweet illusion should not seem surprising; a given physical edge can, under different illumination and orientation conditions, induce different intensity distributions which by themselves are of no interest to the perceiver who is to recover the physical structure of the environment (c.f. [Marr, 1974; Marr \& Poggio, 1976] for the same argument concerning the computation of stereo disparity). Once the edge has been detected, the underlying intensity distribution can be replaced by an edge representation, and a correspondence may then be established between two edge representations. In Section 1.2 the above reasoning will serve to construct a counter-example to the grey level correlation hypothesis.

The foregoing discussion suggests that the discernment of motion should be performed only after the raw intensity data have been organized into units. Such units are probably detected and organized hierarchically, in the sense that units such as edge fragments, bars and small blobs are detected first, then organized into more structured forms, and finally into distinct objects.

If such an organization scheme holds, the appropriate candidates for correspondence tokens are the units situated near the lower end of the hierarchy, (which we shall therefore call low-level units). The argument supporting this claim depends in part on the way the correspondence process is carried out and can therefore be fully appreciated only in combination with the discussions in later sections. The gist of the argument, however,
is the following. In teleological terms, the problem faced by the visual system in establishing a correspondence is one of guessing the probability that elements $X$ and $X$ ' are the same object in motion. It therefore needs some measure of the likelihood that $X$ and $X$ ' represent the same object after a slight movement. It is inconceivable that all possible figures are stored in memory together with their likelihood measures, hence this measure must be computed. Only for a certain class of units, namely the members of the basic elements domain, is there indeed a "stored" likelihood measure which we shall call affinity. For more complex figures the correspondence is computed from the affinities of their constituents via interactions such as those specified in Section 2.1. The basic elements are therefore expected to be the building blocks out of which complex figures can be structured. The next two sections will support this view by providing evidence that the correspondence is indeed established by matching basic units such as edges, line segments and small blobs. It is of interest to introduce in this context the notion of the primal sketch termed by Marr in his theory of early visual processing [Marr, 1976]. The primal sketch is a set of basic units that are the first to be formed in the course of visual analysis, and serve as building blocks for higher-order constructs. From the above discussion it is expected that the domain of correspondence tokens will be roughly equivalent to the elements comprising the primal sketch. This appears indeed to be the case: the two searches for basic units do seem to converge to a similar set of elements.

### 1.2 The Correspondence is not a Grey Level Operation

The preceding section argued that grey level operations are inadequate for the determination of motion and that organization of the raw data into elementary meaningful units must precede the correspondence process.

In the current section, a demonstration supporting this claim is presented. The demonstration is based on the apparent
motion between two pictures. These are designed in such a way that grey level operations imply one kind of motion, whereas a scheme based on pre-organizing the data into meaningful units predicts a different motion.

The intensity profiles of the two pictures used in the demonstration are shown in Figure 1.1 as profiles $A$ and $B$. Both profiles are derived from graph $S$ in Figure 1.1. Profile $A$ is obtained from $S$ by "smoothing out" the right-hand step, profile $B$ by smoothing out the left-hand step. Perceptually, $S$ contains two sharp edges at positions $p$ and $q$, while $A$ has a single sharp edge at $p$, and $B$ a single edge at $q$. For the subsequent exposition a definition of $B$ s position relative to $A$ is needed. Let the position at which $A$ and $B$ overlap be the 0 position. A positive position will mean that $B$ was displaced to the right, and negative position -- to the left. The entire picture measures 250 units, as indicated in Figure 1.1. Both the position and the intensity units are intended to be on a relative scale only, the actual values can vary within a wide range.

What motion should arise when the two pictures are presented in alternation? If the correspondence is established between perceivable edges the prediction is simple: edges $p$ and $q$ should be seen in motion. If the match is governed by grey level correlation the prediction is different. Graph 1.2a depicts the cross-correlation function between profiles $A$ and $B$. As seen from the graph, the cross-correlation reaches its maximum at position 0 . It is therefore predicted that:
(i) If $A$ and $B$ are shown in registration (position 0 ) no motion should arise.
(ii) If $A$ is shown first, followed by $B$ displaced by, say, -20 units, then a movement of $A$ by 20 units to the left is expected, since a displacement by this amount will maximize the crosscorrelation.

The two methods are thus brought into a critical test, predicting opposite results. When the two pictures were presented tachistoscopically, the observed motion was between the visible edges, contrary to the cross-correlation prediction.




Figure 1.1 Intensity profiles. Profile $S$ gives rise to two distinct edges. Profiles $A$ has a single sharp edge at $p$, and $B$ a single edge at $q$.

Presentation times were between 100 and 200 msec ., with an inter-stimulus interval (ISI) of between 30 and 70 msec . The angular extension of each picture was 3.5 degrees of visual angle, and the separation of the two edges was 1.5 degrees.

Various other grey level operations besides crosscorrelation have been suggested for picture comparisons. Graph 1.2 b shows the results of applying a second method, called the "subtraction operation" to the profiles in question [1.3]. In this operation, the match is indicated by the minimum value of Graph 1.2 b . As with cross-correlation, the match reaches its optimum at position 0 . The predictions of the subtraction method are therefore equivalent to those based on the crosscorrelation technique, and are likewise rejected by the experimental results. Other grey level operations, such as local correlation (Graph 1.2 c ) and local subtraction (Graph 1.2d), were examined as well and refuted in a similar manner for the same underlying reason: the changes in the raw intensity distributions do not directly reflect changes in the visible environment. Hence, organization of the visual input into units corresponding to physical entities is a prerequisite for the recovery of physical motion from the changing optical array.

The conclusion that motion correspondence is based on the matching of tokens, not intensity distributions, must be qualified in the case of small displacements by the following comment. There are indications (e.g. the "reversed phi" motion discovered by Anstis [1970], and the short-range process in [Braddick, 1974]) for the existence of a motion detection process that responds to changes in intensity distributions. This process is characterized by its short range ( $15-20$ minutes of an arc), and is more effective in peripheral vision, in contrast with the correspondence of tokens that is long range, and effective primarily in central vision. If the two processes do exist, it seems plausible that they might serve different functions. The intensity based peripheral process is adequate for an "early warning system", detecting changes and directing attention. It might also be useful in detecting discontinuity boundaries where


Figure 1.2 Measuring the similarity between profiles $A$ and $B$ of Figure 1.1 (vertical axis) as a function of their position (horizontal axis). In 1.2 a global cross-correlation is used, in 1.2 b global subtraction process, in 1.2 c local crosscorrelation, and in 1.2 d local subtraction process.
velocities in the visual field change abruptly. The correspondence process, on the other hand, by identifying and tracing tokens, is instrumental in maintaining the perceptual identity of moving objects, and in the 3-D interpretation stage, as will be elaborated in Chapter 4. For the purpose of the present study the possible intensity-based detection process is thus of secondary importance, and will not be discussed further.

In conclusion, the current section places a lower bound on the amount of processing required prior to the establishment of correspondence between image elements. In the next section the upper bound problem is considered. I have argued on theoretical grounds that in the hierarchy of units organized by the visual system, the correspondence tokens are expected to be found at or near the lowest level. The following section provides evidence in support of this view.

### 1.3 The Correspondence Tokens are not Structured Forms

In this section, five demonstrations will be described in support of the view that the correspondence process does not rely on elaborate form perception. It will be argued that the correspondence perceived between structured forms in motion is not established between the complete forms on the basis of their similarity. Rather, it is the result of a match established between simple constituents of the structured forms. Each demonstration will be described, followed by a brief discussion of the results.

## Demonstration 1: The "broken wheel"

The broken wheel display is a modification of a wellknown motion picture effect, sometimes called the "wagon wheel" phenomenon, in which a spoked wagon wheel seems to rotate in the direction opposite to its real sense of rotation. This phenomenon indicates the visual system's disposition to choose, from two possible matches, the one which involves minimal change (angular change in this case). As far as the basic
elements problem is concerned the wagon wheel phenomenon admits two different interpretations:

1. The organization of small units into the complete form (the wheel) comes first, and then the form in the first image $A(t)$ is matched against the one found in the later image $A\left(t^{\prime}\right)$. There is more then one way of matching them perfectly, so the one which involves minimum change is selected.
2. Correspondence is established between small sub-units of the wheels, and the motion of the entire form is constructed at a later stage from the motions of the constituents.

In the case of the wagon wheel phenomenon, these two different methods of analysis yield the same result. The "broken wheel" display was constructed in such a way that the two hypotheses would have opposite implications. The rotating figure in this experiment consists of a wagon wheel in which every other spoke is broken, and its middle part is missing (Figure 1.3a). Let $a$ be the angle between two neighboring spokes, and suppose that between successive views the wheel is rotated $\beta$ degrees counterclockwise. Consider now what happens when $\alpha>$ $\beta>a / 2$ (Figure $1.3 \mathrm{~b} ; a$ was 12 degrees and $\beta$ was 8 degrees). Taking the figure as a whole, a perfect match is achieved by rotating the first wheel $\beta$ degrees counterclockwise. However, if a short line segment ( x in Figure 1.3b) were considered a basic element, and its closest counterpart were sought, then a line segment in the clockwise direction (y in Figure 1.3b) might be chosen. Such a choice is impossible according to the first view, but is highly plausible according to the second (though not necessary, for reasons to be discussed in Section 2.4.2). The outcome of the experiment is the following: when appropriately timed, (presentation time of 50 msec . and ISI of 30 msec . in a dark room were required to obtain good, coherent motion), the wheel breaks into three distinct rings. The innermost and outermost rotate clockwise while the middle ring rotates counterclockwise. This breakdown shows that the motion in this case was not established between the complete forms, but between the forms' constituents.

a

b

Figure 1.3 The broken wheel demonstration. Solid lines represent the first frame; dotted lines the second.


Figure 1.4 The block-train demonstration. Solid lines represent the first frame; dotted lines the second.

## Demonstration 2: The "block train"

The figure in this demonstration is a "block train" comprised of cars with windows as shown in Figure 1.4a. Neighboring vertical lines are separated by $x$ units, and the train moves $y$ units to the right between two successive views (Figure 1.4 b ). When $\mathrm{x} / 2<\mathrm{y}<\mathrm{x}$ (the actual values employed were $\mathrm{x}=$ 0.4 degrees of visual angle, $y=0.3$ degrees, with presentation time of 50 msec . and ISI of 30 msec .) there are two principal modes in which the "moving train" is perceived. First, the figure may split: the "windows" move to the left while the rest of the train moves to the right. Alternatively, the entire train may move to the right. The first of the above modes is similar to the broken wheel phenomenon with linear translation substituted for rotation. This mode is seen whenever the viewer fixates at a stationary point on the screen and does not allow his eyes to track the moving train. The second mode is probably the result of eye tracking movements. Suppose the viewer tracks the figure perfectly, that is, the "train" does not move relative to his eye. From the fact that the picture remained unchanged, but eye tracking motions were needed to keep the picture stable, the implication is that the whole figure underwent a translation.

## Demonstration 3: The "zebra star"

This experiment involves a rotating star in which every other line is widened (in another version: made longer). Again, let $\alpha$ be the angle between neighboring spokes and $\beta$ the angle of rotation (Figure 1.5 a ). If $\beta=\alpha / 2$ the zebra star rotates in the direction that will match each wide spoke with a neighboring wide spoke, and each narrow one with its narrow neighbor. If a $>\beta>\alpha / 2$ the star is perceived as rotating in the "wrong" direction and at the same time each narrow spoke becomes wider and vice versa.

In the "zebra star" display the figure did not split into sub-units, therefore one might question its significance to the issue in question on the grounds that it can still be explained in


Figure 1.5 The zebra-star demonstration. Solid lines represent the first frame; dashed lines the second.
terms of the figure as a single unit. The figure as a whole, it might be argued, has two possible matches: a perfect match $\beta$ degrees away, or another match, closer spatially ( $\alpha-\beta$ degrees), but which implies changes in the figure. One can thus propose the construction of a metric space based on spatial distance as well as on similarity, in which the second match will be "closer" to the original than the first.

This objection is unconvincing for the following reason. The primary advantage of doing form analysis prior to the matching operation is the ability to subsequently identify two figures as corresponding on the basis of figural similarity. One would expect therefore that two complex, identical, and proximate figures should inevitably be matched, a conclusion that
runs contrary to the described findings. If complete form analysis does precede the correspondence process one would further expect that a perfect match between complex figures would be a stronger indication of correspondence than a match between small and simple constituents thereof. Experiments with single spokes show, however, that the same ratio $\beta / \alpha$ is needed both for single spokes and for the whole star, in order to switch the direction of preferred motion.

## Demonstration 4: The rotating spiral

In a well-known illusion, a rotating spiral (under either continuous or discrete presentations) seems to expand or contract, depending on its sense of rotation [Kolers 1966]. If the endpoints of the spiral are concealed, only the inward-outward motion is perceived, the rotation is not [Wallach, Weisz \& Adams 1956]. The spiral as a whole is involved only in a rotary motion. However, when considering small fragments of the spiral as basic elements, one plausible explanation of the illusion suggests itself. A correspondence between small sub-units of the spiral and their closest neighbors indeed implies a sense of motion perpendicular to the rotation. One can actually observe the outward and inward motion induced by the local correspondence by viewing the display through a narrow radial slit.

## Demonstration 5: Correspondence and form

Various attempts have been made in the past to examine the influence that similarity of form exerts on the perceived correspondence between figures. Following the assumption that the matching process should prefer to match similar figures, Kolers [1972] compared the "smoothness" of perceived motion between similar and dissimilar figures. Since the smoothness ratings were found to be the same, independent of figural similarity, Kolers concluded that for the visual system all twodimensional figures are equally similar. There were, on the other hand, some reports [Orlansky, 1940; Frisby, 1972] that for simple stimuli, especially line segments of different orientations,
there were some effects of similarity on the "optimality" of perceived motion.

The findings that similarity between complex forms does not affect their correspondence in any clear way seem to agree with the idea advanced in this section that the matching process occurs prior to the organization of the basic units into structured forms. However, they cannot be accepted as relevant to the problem in question. The main reason is that most of these findings (with the exception of [Navon, 1976] and some demonstrations in [Kolers, 1972]) were based on smoothness of motion judgements which, as shown in Section 2.4, are not a faithful measure of the figures' "tendency to fuse" which they were intended to measure.

A direct method for testing the effect of figural similarity on the matching process is a method which I shall call the competing motion technique. In this method two frames are presented in alternation. The first one contains a single element (or figure), while the second frame contains two elements. The question asked is whether the figure in the first frame is seen in motion with one or the other of the elements in the second frame. Figure 1.6 a shows an example of a competing motion display. The first frame presents the middle square $A$ alone, while the second frame presents both the outermost square $B$ and the innermost triangle $C$ (presentation time was 120 msec ., ISI 40 msec .). The perceived correspondence upon presentation of this display is $A \leftrightarrow B$. That is, the motion between the two squares is preferred. Unlike some past conclusions, these results suggest that figures do differ in their "tendency to fuse". But does this preference indicate an effect of figural similarity? Not necessarily. When the individual lines composing the display are tested in isolation, the preference remains the same. For instance, when $x$ in Figure 1.6a is shown in competing motion with $y$ and $z$, the motion towards $y$ is preferred. The tendency of square $A$ to fuse with square $B$ rather than with triangle $C$ can thus be explained on the basis of the motion of the constituent elements. There is no need to suppose that the
complete forms are the basic elements, nor that a similarity measure between forms determines the observed correspondence. While the correspondence in this example is compatible with both the similarity between the figures and the local match between components, in the next example similarity and local match have conflicting implications.

a

b

Figure 1.6 Correspondence and form. Solid lines represent the first frame; dashed lines the second. In 1.6 a the predominant correspondence is between similar forms, in 1.6 b between dissimilar ones.

In Figure 1.6b the observed correspondence is between dissimilar figures: the preferred match is between the rectangle $A$ and the triangle $B$ rather than the inner rectangle $C$. Once again, this preference is consistent with the correspondence among constituent elements, e.g. the $x \leftrightarrow y$ match is preferred over $x \leftrightarrow z$. Thus, it is the motion of the constituent elements rather than the similarity between the complete forms that governs the matching process. To sum up the discussion of demonstration 5:

1. Unlike past conclusions, different figures do differ in their "tendency to fuse", but the preference is consonant with the motion being
established between their components.
2. There are no indications that structured figures are part of the basic elements domain, or that the correspondence process is based on figural similarity.

All of the demonstrations described above support the claim that no elaborate form analysis must precede the correspondence operation, and that the motion of complex figures is constructed from the motion of their constituents.

Additional support for and elaborations of this view are indicated in two of the subsequent sections. First, the discussion of some relations between the matching process and structured figures is deferred to Section 2.4.2, as they are examined in light of the correspondence scheme advanced in Chapter 2. Second, Section 2.5 shows that the correspondence operation is independent of the three-dimensional interpretation of the scene, thus supporting the view that the correspondence is low level in nature [1.4].

