High-Level Motion Processing
Computational, Neurobiological, and Psychophysical Perspectives

edited by Takeo Watanabe
A Theoretical Framework for Visual Motion

Alan L. Yuille and Norberto M. Grzywacz

There are many theories of motion measurement in both the biological (for reviews, see Smith and Snowden, 1994) and artificial vision (Blake and Yuille, 1992; Faugeras, 1993) literatures. Typically, each theory is only suitable for describing limited types of motion. However, the growing success of such theories leads us to ask the big question: Can we design a biologically plausible theory that will describe all visual motion phenomena and work on real images? In other words, can we design a theory for biological measurement of visual motion that embodies the "real-world" constraints of computer vision? Because computer vision theories are designed to perform useful visual tasks on realistic scenes, they tend to enforce the ecological constraints advocated by Gibson and others (Gibson, 1979; Marr, 1982). In this chapter, we first attempt to classify the most important visual motion phenomena. We then describe a theoretical framework that appears to be sufficiently rich to account for most existing psychophysical and physiological experiments. Specific models (parts of the framework which have been developed in detail) are shown to be in detailed agreement with experiments. Finally, the chapter discusses experimental and computational problems, which must be solved as a theory of motion measurement is developed.

MOTION PHENOMENA

Workers in psychophysics, physiology, and computer vision have investigated many different motion phenomena. Psychophysicists and physiologists have tended to work on simple but well-controlled stimuli in an attempt to tease out the fundamental mechanisms used by biological systems. By contrast, workers in computer vision have been more concerned with designing theories to work on real images while paying special attention to detecting specific types of visual motion that are important to the system. These different agendas mean that both groups have tended to stress somewhat different visual motion phenomena. Artificial vision systems have tended to work in a world of rigid or semirigid objects. These systems compute motion to determine optical flow, track objects, locate motion boundaries, and provide input to structure-from-motion algorithms. Many of
Figure 6.1 Lip and hand tracking. The algorithms make use of Kalman filter techniques to predict the positions of the tracked objects and to update the estimates of the positions using input intensity data. Parameters of the model are trained for specific types of motion. (A) The white line shows the estimated position of the lips. (B) The white line is inaccurate because the parameters of the model were trained for one type of motion, and tested on a different one, at one time instant. In (C) and (D), the solid lines show the estimated position of the hand, while the dashed lines show the uncertainty of the position estimates. Reprinted with permission from Blake, Curwen, and Zisserman (1993) and Blake, Isard, and Reynaud (1995).

these systems are not intended to model biology. For example, because of pragmatic considerations, current tracking and structure from motion systems often work on sparse representations of images (for example, Tomasi and Kanade, 1992) obtained by using edge or sparse-feature detectors. In particular, there has been considerable work on object tracking (see figure 6.1), but although the state of the art is rapidly improving (for a review, see Blake and Yuille, 1992), the algorithms work best in controlled environments and when asked to perform limited tasks.

Psychophysicists and physiologists have investigated a far wider class of motions. Indeed, some of these motions seem unlikely to arise in real-world scenes. See figure 6.2. Examples include: (i) motion capture (Ramachandran and Anstis, 1983a), (ii) transparency detection (Adelson and Movshon, 1982; Welch, 1989; Qian and Andersen, 1994; Bravo and Watamiani, 1995), (iii) motion-boundary detection (Anstis, 1970; Baker and Braddick, 1982; Julesz, 1971), (iv) motion-outlier detection (Watamiani, McKee, and Grzywacz, 1994; Grzywacz, Watamiani, and McKee, 1995), (v) second-order motion boundaries (Zanker, 1990, 1993), and (vi) motion-occlusion perception (Watamiani and McKee, 1995).

As we will argue, these phenomena are all consistent with a theory that uses local image velocity measurements followed by a global process which groups the measurements into regions with similar motion properties.
Figure 6.2 Psychophysical paradigms. The six subfigures illustrate the phenomena of capture, transparency, motion boundaries, temporal outlier detection, second-order boundaries, and motion occlusion. In each subfigure, the schematic input stimulus is labeled true. It consists of local velocity estimates of features represented by filled arrows, moving gratings denoted by bars, and motion boundaries represented by dotted lines. Open arrows indicate the motion of gratings or second-order defined objects. The perceived motion, labeled perceived, is typically more regular than the motion measurements.

thereby segmenting the image. These properties include spatial position and texture, and the regions occur in high-dimensional parameter space, called Beta-Space, which will be described later.

What exactly do we mean by "similar"? This is a critical question because similarity should at least capture the idea of standard motion-flow patterns or motion fields. What are these standard patterns? Presumably, they correspond mostly to the commonly occurring motion patterns in the world. Thus, for instance, we might expect that the visual system groups local measurements that are consistent with translation in the image plane, rigid motion in three dimensions, and pure expansion motion (figure 6.3) (see Hildreth and Royden, chapter 9; Tanaka, chapter 10; Warren, chapter 11). What other types of motion might the visual system group into regions? There is a vast range of more complex motion flows, which humans have little difficulty detecting—for example, swarms of bees, leaves blowing in the wind, and specularities on a river. Moreover, Zanker (1990, 1993) describes stimuli for which the motion of an object's boundary is different from the motion of the object's interior. Such a motion flow can occur in natural scenes. Consider, for example, a person walking through a forest with complicated shading patterns induced because the sunlight is partially obscured by tree branches. The motion of these shading patterns will be quite different from the motion of the person. Another example for which the boundary and interior motions differ is a mirror on a carousel where the reflected
motion flow is the opposite of the true motion of the mirror. Thus, the set of possible motion flows is fairly rich.

Most of the motion fields we have mentioned, such as expansion and rotation (figure 6.3), are regular spatially. To tap these regularities the human visual system may use a number of different motion models, each corresponding to a different type of motion field. These models could compete among themselves to explain the motion measurements (to be addressed later). Thus, if certain motion measurements were consistent with an expansion motion model, they would be grouped together as an expansion region. Similarly, other motion measurements might be grouped by the rotation model to form a rotation region. This notion of competitive models seems capable of explaining some psychophysical results (Yuille and Bulthoff, 1996). The visual system is also flexible enough to deal with unexpected motion patterns, which it has never seen before, and with motion fields that are fairly random. For example, what happens when a set of features is undergoing Brownian motion (figure 6.4)? Would a biological vision system group them into one region (corresponding to “an object”), or would they be described as many individual features, or small groups of features, with little spatiotemporal coherence? If the former is correct, then we should observe motion boundaries between two sets of features undergoing different Brownian motions (figure 6.4). Informal observations suggest that boundaries can be detected in such circumstances.

This raises an important question: Supposing that biological visual systems are sensitive to different types of Brownian motion, does this imply that these systems have, in any sense, a model for such motion? Or do they have a more general purpose mechanism, which determines, on the basis of local tests, that the motion of local features is consistent and that they should therefore be grouped together? In mathematical terms, the difference is between having a parametric model for motion flow (for example, a parametric model for expansion motion would have a focus of expansion and an expansion rate as parameters to be determined) and using nonparametric techniques (such
as the Wilcoxon rank-sum test or the Kolmogorov-Smirnov tests) to determine motion similarity. Such nonparametric tests have been used ( Spoerri and Ullman, 1987; Bülthoff, Little, and Poggio, 1989) to locate motion boundaries. Another advantage of a general purpose, nonparametric model is that it would work for motion fields that have never been seen before. Of course, it might be possible to learn (over a period of time) models for novel motion fields, particularly if they are important for the visual system. We speculate that certain motion flow models can be learned for specific applications; for example, a fighter pilot might have a different class of motion models from those of a deep-sea diver.

OVERVIEW OF THE FRAMEWORK

This section gives a schematic overview of the framework. Technical details will be given in a later section. The input to our system will be a set of local motion measurements. At each point in the visual field, we assume that there are cells, implementing nonlinear filters, which are tuned to different spatiotemporal frequencies of the stimuli (Grzywacz and Yuille, 1990). As the system observes a moving stimulus, different cells will be activated with differing strengths. These local set of estimates provide the input to a high-dimensional parameter space, called Beta-Space, which is schematically illustrated in figure 6.5. At each spatial point \( x \), this space corresponds to a set of velocity-sensitive filters parametrized by \( \beta \).

We hypothesize that the measurements are clustered into regions in Beta-Space (using homogeneity criteria we will discuss later) whose boundaries will correspond to motion boundaries. These boundaries will typically occur at places where the measurement properties change greatly and will be represented by contours which, in turn, may have their own motions. The clustering is performed by a number of competitive processes corresponding to different motions. One generic form of clustering uses a nonparametric test
Figure 6.5  Schematic of Beta-space. The black dots in the left panel show the activated cells in Beta-Space as a function of position x (both β and x represent several variables). The right panel shows how the responses in Beta-Space may cluster into two regions, given by the black and white dots. In this case, the regions correspond to transparent motion because the clusters overlap in space.

to group regions that the test decides are similar. This process competes with more structured tests, which try to detect familiar motions such as the translations of objects, pure expansion motion, or pure rotation motion. Thus, grouping can be performed by either parametric models or a nonparametric test. The idea is that the parametric models correspond to important types of motion to which the visual system is sensitive (such as expansion or rotation), while the nonparametric test allows the visual system to deal with general types of motion that it may never have seen before. It is expected that the regions obtained after grouping correspond to different objects.

There are therefore three main parts in our framework: the motion measurement stage, the motion segmentation stage, and the motion model stage. These stages will interact with each other. For example, the motion segmentation will influence, and be influenced by, the motion models. We propose that they interact with each other as shown in figure 6.6. Experimental justification for having a separate motion boundary and motion model stage comes from double disassociation of segmentation from motion and spatial integration of motion cues due to damage in different brain areas (Vaina, Grzywacz, and Kikinis, 1994). Computational justification comes from the need of a model-independent process (nonparametric test) for segmentation.

Our framework requires that the motion system can make predictions over time. There are two possible ways to implement this. One uses the intrinsic lateral connections in the area of the cortex implementing the measurements (see Grzywacz, Watamaniuk, and McKee, 1995, for a model of this type). The second uses feedback from higher areas (the velocity flows computed by the motion models) to prime the cells that perform motion measurements (see Watanabe and Miyaiuchi, chapter 3; Grossberg, chapter

192  Motion Coherence and Grouping
1. Both lateral and feedback connections exist but, to our knowledge, there are no experiments that distinguish between these two possibilities. This is an interesting topic for future research but we will not deal with the possible role of feedback connections in this chapter.

Either approach could be considered as an implementation of standard prediction and estimation techniques, such as the Kalman filter used by Blake and his collaborators (figure 6.1). It predicts a response in the future and adjusts it based on the observed new input data. The motion measurements also develop a confidence estimate for the motion (by combining the current confidence estimate with the prediction and the new measurement); the confidence estimate is high if the motion is coherent in time. This confidence could be used as an additional feature in the Beta-Space. This would allow temporally coherent features to emerge out of incoherent Brownian motion and would allow objects to be represented even when they are temporarily occluded and hence do not receive direct input. If an object is occluded for too long it will receive little input so its confidence will decay and the region will cease to exist.

Sketch of How the Framework Describes Different Motion Phenomena

Before describing the framework more formally, we will sketch how it can be used to describe different motion phenomena. This section describes psychophysical phenomena, and the next addresses physiological and anatomical issues. Motion capture and motion coherence (the perception of a single motion in a very noisy flow pattern) are explained by a single region getting
Figure 6.7 Coherence and transparency in beta-space. (Left) The motion measured, represented by the arrows, is roughly coherent, so the filter responses in Beta-Space lie in one cluster. (Right) The motion is transparent (the arrows indicating the motion measurements point in two different directions), so the responses in Beta-Space cluster into two regions. Because these regions overlap in space, transparency results.

grouped (figure 6.7) using a coherence model (a model suitable for the translation of large objects). The motion coherence model, however, is only one of several motion models that we suggest try to account for the data. If motion coherence, at least as it was formulated by Yuille and Grzywacz (1988, 1989), was applied to pure expansion motion or pure rotation motion it would tend to degrade them. Thus, we propose that motion coherence is only one of several motion models that the visual system uses. For instance, Regan and Beverley (1979, 1983) obtained psychophysical evidence for internal models (channels) of expansion and rotation.

Certain types of transparent motion can also be explained naturally within our framework (figure 6.7). If we have two regions that overlap in space, but differ in other properties (for instance, texture and velocity), then the regions will! be separated in measurement space and so be perceived as different objects. The model of Smith and Grzywacz (1993) gives a biologically plausible explanation of this and also explains how differences in contrast can lead to region segmentation. Temporal motion coherence was described by Anstis and Ramachandran (1987) and Ramachandran and Anstis (1983b), who call this phenomenon motion inertia. We interpret it as being due to grouping by a temporal coherence model, which involves prediction and measurement (for example, implemented by a Kalman filter). The prediction encourages the system to move with temporally coherent velocity. Experimentally the prediction is more specific for direction than for speed. Anstis
Figure 6.8 Temporal integration in Beta-Space. (Left) A dot moving coherently in time (the long straight arrow) is perceptually separated from a randomly moving background (the short arrows). The long arrow indicates a long straight trajectory, not faster speed (the outlier dot has the same speed as the background dots). (Right) In beta-space, the outlier dot forms a temporal cluster (thick dark line) while the background dots give rise to jerky trajectories (thin jagged lines).

and Ramachandran (1987) and Grzywacz, Smith, and Yuille (1989) conjectured that this is because if an object is rigidly moving in space then the direction of motion in the image plane is more likely to be constant than the speed. A further form of grouping by temporal coherence, Motion outlier detection (Watamaniuk, McKee, and Grzywacz, 1994; Grzywacz, Watamaniuk, and McKee, 1995), can occur when one dot, the outlier, moves coherently in a background of incoherent motion (figure 6.8). In this case, the outlier forms a narrow coherent region in space-time, which separates off from the background. This region is generated because the motion measurement constantly reinforces the motion prediction.

Temporal grouping may also account for recent results of Watamaniuk and McKee (1995) on motion occlusion. In their experiments, a dot moves so that part of its trajectory is obscured by a stationary object formed by dots moving randomly. There are two types of perception: The signal dot continues moving behind the stationary object and reappears on the far side (this is the motion occlusion case), or the signal dot disappears when it gets obscured. Watamaniuk and McKee showed that motion occlusion does not occur if the distribution of the motion properties of the dots composing the fixed object overlap with the motion properties of the signal dot. We explain these results by saying that occlusion can only occur if the responses to the two objects do not overlap in Beta-Space (figure 6.9). In this case, the region corresponding to the signal dot will continue to exist for a limited time, due to temporal coherence, even if there is no input stimulus.

In addition, we consider the theta motions by Zanker (1990, 1993) (figure 6.10). In these stimuli, which are a particular case of the so-called second-order motions (Cavanagh, Arguin, and von Grünau, 1989; Cavanagh and Mather, 1989; Zanker, 1990, 1993), the motions of the interior and the exterior of an object are different from the motion of the boundary. We suggest that the motions of the interior and the exterior are grouped by two
Figure 6.9  Motion occlusion in beta-space. Figure (A) shows a moving object, represented by the straight line, disappearing and reappearing behind a fixed object, shown by the square. The occlusion case is shown in (B), where the regions corresponding to the two objects (the shaded circle and the shaded ellipse) do not overlap in Beta-Space. The dotted circles in frames $t_1$ and $t_2$ show the position of the moving object region in the previous time frames $t_1$ and $t_2$ respectively. Temporal coherence causes the moving object region to exist at $t_3$, even when it receives no stimulus input (the rectangle indicates the clustering of the shaded and dotted circles). Figure (C) shows when occlusion does not occur because the regions for the two objects overlap in beta-space at time and the moving object region gets lost.

Second-Order Motion  

Second-Order Motion in β Space

Figure 6.10  Second-order motion in beta-space. (Left) The motion stimuli is represented by arrows, where the exterior and interior regions consist of dots moving horizontally and vertically (respectively) but the boundary moves diagonally, shown by the white arrow. (Right) The responses in beta-space show that the interior, the exterior, and the boundary all form coherent regions over time.

196  Motion Coherence and Grouping
different motion models, thereby forming two clustered regions in Beta-Space. In turn, because the boundary of the object moves with temporal coherence, it also forms a region in Beta-Space.

Finally, we note that Burr, Ross, and Morrone (1986) and Watamaniuk (1992) used motion prediction to explain why humans do not get motion blur. Moreover, performance in direction (Watamaniuk, Sekuler, and Williams, 1989) and speed discrimination (McKee and Welch, 1985; Turano and Pantle, 1989; Werkhoven, Snippe, and Toet, 1992) increases in time consistent with models such as Kalman filters. Motion inertia and the temporal buildup of the upper displacement limit (Nakayama and Silverman, 1984; Snowden and Braddick, 1989) show that the correspondence problem is solved along directions of past motion. All these results seem consistent with our framework.

PHYSIOLOGICAL AND ANATOMICAL INTERPRETATIONS

It is worthwhile to try to cast our framework in neural cortical mechanisms. This would provide a test of whether this framework, besides being in agreement with psychophysics, is consistent with the known physiology and anatomy of cortical pathways. Furthermore, casting this framework in neural language could lead to interesting interpretations of how the brain encodes complex motions. The starting point in interpreting cortical motion processing should be Hubel and Wiesel’s hierarchical view (Hubel and Wiesel, 1962, 1965). As a first approximation this view seems to be justified. Cells in the primary visual cortex (V1) are selective to local motions (Hubel and Wiesel, 1962; Schiller, Finlay, and Volman, 1976; Mikami, Newsome, and Wurtz, 1986); cells in the middle temporal cortex (MT) are selective to large motion fields (Tanaka et al., 1986); and cells in the middle superior sulcus (MST) are selective to complex optic flows like expansion and rotation (Duffy and Wurtz, 1991; Graziano, Andersen, and Snowden, 1994; Saito et al., 1986; for a review see Tanaka, chapter 10). In our framework, V1, MT, and MST would contain different models of motion in the world. However, this framework also postulates that information from models of high-level regions should feed back to low-level regions (figure 6.6). (Such a feedback would help to refine the computations performed by the lower regions.) Consistent with the framework, anatomical evidence for such feedback projections is abundant in the literature (for a review see Maunsell and Newsome, 1987).

Another way that our framework departs from the Hubel and Wiesel view and is consistent with recent literature is the prediction of long-range lateral interactions within visual cortical areas. In the framework, these interactions may mediate spatial integrations necessary to model complex motion fields (figures 6.2 and 6.3). Moreover, these interactions may mediate temporal integrations necessary to model motion occlusion and related phenomena (figures 6.8 and 6.9). Anatomical and physiological evidence for long-range lateral interactions in V1 connecting cells with similar preferred orientation
have been mounting (Gilbert and Wiesel, 1979, 1989; Martin and Whitteridge, 1984; Kitano et al., 1994). However, we still know relatively little about whether such interactions contribute to the motion domain. Of interest is the observation by Lamme and colleagues (Lamme, van Dijk, and Spekreijse, 1993, 1994; Lamme, 1995) that V1 may already participate in figure-ground segregation from motion cues. One interpretation for this observation is that long-range lateral connections in V1 may contribute to figure-ground segmentation. Another interpretation is that feedback projections from higher cortical areas, some of which are involved in figure-ground segregation, may modify V1 responses. This latter interpretation would be consistent with the MRI evidence that V2 may be crucial for this segregation (Vaina, Grzywacz, and Kikinis, 1994; also see Vaina et al., chapter 7).

As suggested by Grzywacz, Watamaniuk, and McKee (1995), a role for these long-range connections could be to facilitate responses of cells stimulated by temporally coherent motions. They suggested that a cell with a given preferred direction sends facilitatory connections to cells with roughly the same preferred direction. The pattern of these connections would be spatially asymmetric. It would prefer cells selective to positions pointed by arrows whose tails lie on our cell and whose directions are roughly in the cell’s preferred direction. There is no physiological evidence for such an asymmetric connection yet, but psychophysical data are suggestive of it (Watamaniuk, McKee, and Grzywacz, 1994; Grzywacz, Watamaniuk, McKee, 1995). Such connections lead to an intriguing neurophysiological interpretation of motion occlusion. Occlusion could be coded by cells receiving inputs from these lateral connections, but not from the lateral geniculate nucleus (LGN). The lateral inputs would indicate that a moving stimulus is probable in the positions encoded by these cells, while the lack of LGN input would indicate that no such stimulus is physically present. Therefore, this lateral input—no LGN conjunction could indicate an occlusion.

Finally, because our framework is cast in Bayesian form, it suggests an interpretation for the meaning of the code carried by cortical cells. This interpretation says that the intensity of the cells’ firing give in some sense the likelihood of a particular model. While it is not immediately obvious that the responses encode probabilities per se, they could at least provide an evaluation of them. Barlow (1989) already suggested assigning a probabilistic meaning to responses. They pointed out that such an assignment is well suited to self-organization and learning in the brain.

TECHNICAL INGREDIENTS OF THE FRAMEWORK

The first stage of our framework is a set of local motion measurements. These are performed by a set of nonlinear filters $F(x, t : \beta)$, where the Beta-parameters specify the filter’s tuning to size, spatial frequency, orientation, and velocity of the stimulus. The filter's responses can best be thought of as
a set of local velocity estimates, which are also sensitive to the intensity properties of the stimulus (such as its texture and brightness). These nonlinear filters can be implemented by a two-step process. The first step uses spatiotemporal filters, such as Gabor filters (Gabor, 1946; Daugman, 1985) tuned to the frequency components of the input and its direction of motion. Then, quadrature ensures that the filters' responses are phase-invariant as in standard motion energy models (Adelson and Bergen, 1985). Mathematical analysis (Grzywacz and Yuille, 1990) shows that if the responses of Gabor filters are plotted as functions of their preferred spatial and temporal frequencies, then under physiologically and psychophysically reasonable conditions, maximal responses will occur along a plane defined by the velocity of the stimulus. Instead of using all the motion-energy responses simultaneously to determine this velocity plane, we suggest using filters in localized regions of spatiotemporal space to specify several local estimates of velocity responses which are sensitive to the local textural properties of the stimulus (for instance, orientation and scale). The advantage of this approach is that it allows for multiple local estimates of velocity, which are needed near motion boundaries and for motion transparency (figure 6.11). Smith and Grzywacz (1993) proposed a model for motion transparency based on this approach.\footnote{Smith and Grzywacz (1993) proposed a model for motion transparency based on this approach.}

The second stage of the framework consists of grouping these local velocity estimates into regions in the measurement space. The basic idea is to group regions that have similar motion measurement properties in space and time. Motion boundaries are signaled at places where the measurement properties
differ too greatly. This grouping can be done either by a generic nonparametric approach or by a set of parametric motion models. These approaches compete to explain the measurements.

To explain these ideas in more detail, we first consider the generic nonparametric model. Suppose we want to determine whether a small region \( W \), containing measurements \( F_\omega = \{F(x, t; \beta) : (x, t; \beta) \in W \} \), should be grouped with a larger region containing measurements \( F_u = \{F(x, t; \beta) : (x, t; \beta) \in U \} \) (figure 6.12). We define a nonparametric measure of similarity \( \phi(F_u, F_\omega) \) to determine the similarity between the two measurement sets; for example, we could use the Wilcoxon sum-rank test or the Kolmogorov-Smirnoff test (for applications to computer vision see Spoerri and Ullman, 1987). These tests are used to determine whether the two measurement sets are generated by the same probability distribution or come from distributions with equal medians. If \( \phi(F_u, F_\omega) \) is small, then this is treated as evidence that \( W \) should be merged with region \( U \). Now suppose that \( W \) straddles the boundary of two regions, \( U_1 \) and \( U_2 \); then region \( W \) will feel a force (depending on \( \phi(F_u, F_\omega) \)) pulling it toward \( U_1 \) and a force (depending on \( \phi(F_u, F_\omega) \)) pulling toward \( U_2 \). So we say that \( U_1 \) and \( U_2 \) compete for ownership of \( W \).

The goal of segmentation using the nonparametric model is to divide the data into regions of homogeneity. This can be made precise by defining a global energy function criterion. The exact form of this criterion is a subject for experimentation, but we propose minimizing an energy function of form:

\[
E(\{R_i\}; N) = \sum_{i=1}^{N} \lambda \int_{\partial R_i} d\Sigma + \sum_{i=1}^{N} \int_{\partial R_i} \phi(F_{R_i}, F_\omega(\Sigma)) d\Sigma + \nu N, 
\]

(1)

where \( \{R_i\} \) is the set of regions, \( N \) is the number of regions, \( \lambda \) and \( \nu \) are positive parameters, and \( W(\Sigma) \) are small regions centered along each point.
of the surface \((\partial R_i)\) of \(R_i\). The first term in this energy function pays a penalty for the area of the boundaries of the regions.\(^2\) The second term pays a penalty for each region depending on how inhomogeneous it is according to the nonparametric similarity measure. Essentially, we compare the statistics inside the region \(R_i\) to the statistics inside \(W(\Sigma)\) and if they are similar enough then we incorporate \(W(\Sigma)\) into \(U\).\(^3\) The third term pays a penalty for the number of regions.

The region competition algorithm (Zhu and Yuille, 1996) gives an intuitive way to minimize energy functions of the form of equation 1 (figure 6.13). The algorithm starts with a large number of seed regions. The boundary of a region moves depending on how well measurement data straddling the boundary are consistent with the data inside the region. Two adjacent regions can compete to account for data on their joint boundary and the boundary will move according to the one that best explains it. The motion of a point on the boundary between two regions is given by steepest descent in equation 1:

\[
\frac{dr}{dt} = -2\lambda \kappa(r) n(r) + \{f_i(r) - f_j(r)\} n(r),
\]  

(2)

where \(\kappa\) is the mean curvature of the boundary of the region, \(n\) is the normal to the boundary, and \(f_i(r) = \phi(F_{R_i}, F_{\omega(r)})\) and \(f_j(r) = \phi(F_{R_j}, F_{\omega(r)})\), with \(R_i\) and \(R_j\) being the neighboring regions. The first term on the righthand side of equation 2 tries to keep the boundary straight. The second term attracts it into region \(R_i\) or \(R_j\), depending on which region has measurements most compatible with the boundary.

In addition to the update defined by equation 2 we will also allow large adjacent regions to be merged if this would decrease the overall energy. This enables the number of regions to be reduced.

The result of this process would be a segmentation of the image into regions of homogeneous motion. But so far, this homogeneity is defined relative to the generic nonparametric criterion only. We would also like to group regions based on parametric models, the intuition being that these parametric models correspond to certain types of motion that are particularly
important for the human visual system. We formulate these models by probability distributions (for arguments in favor of this approach see Barlow, 1989). For particular classes of motion—pure expansion, for example—we can define probability models for generating the observations. We can represent this by a set of distributions $P(F(x, t; \beta) | \alpha)$ where the parameters $\alpha$ specify the motion—for example, for pure expansions $\alpha$ specifies the rate and focus of expansion.

We now want these motion models to interact with the motion segmentation process being performed by the nonparametric model. To begin we consider a variant (Zhu and Yuille, 1996) of equation 1 to segment motion using a single parametric motion model:

$$E[\{R_i\}, \{a_i\}; N] = \sum_{i=1}^{N} \lambda \int_{\partial R_i} d\Sigma - \sum_{i=1}^{N} \int_{R_i} \log P(F(x, t; \beta) | a_i) dx \, dt \, d\beta + \nu N,$$

$$\text{(3)}$$

where $a_i$ indexes the type of motion associated with region $i$. The second term is now the cost for describing the image motion using the probability model. The logarithm of the probability is the cost of encoding the data in the region by a probability model (Shannon and Weaver, 1963) and is a consequence of using minimal length encoding (Rissanen, 1984) as a criterion for segmentation. The region competition algorithm for the energy in equation 3 includes steps for estimating the $\{a_i\}$. These steps alternate with steps for updating the positions of the boundaries by doing steepest descent on equation 3:

$$\frac{dr}{dt} = -2\lambda k(x)n(x) + \{\log P(F(x, t; \beta | a_i)) - \log P(F(x, t; \beta | a_i))\}.$$

$$\text{(4)}$$

Finally, we can generalize the model to include both the nonparametric model and all the parametric motion models (figure 6.14). This requires introducing additional binary indicator variables $V_{i, \mu} = \{0, 1\}$, where $\mu \in \Lambda$, the set of all stored motion models, and $\mu = 0$ labels the nonparametric model. Thus, $V_{i, \mu}$ indicates which model is chosen to describe region $i$. We require that each region is described by one model only, so that for all $V_{i0} + \sum_{\mu \in \Lambda} V_{i\mu} = 1$. This gives an overall energy function:

$$E[\{R_i\}, \{a_i\}, \{V_{i\mu}\}; N] = \sum_{i=1}^{N} \lambda \int_{\partial R_i} d\Sigma + \nu N + \sum_{i=1}^{N} \sum_{\mu \in \Lambda} V_{i\mu} \int_{R_i} \phi(F_{R_i}, F_{\mu}(\Sigma)) d\Sigma$$

$$- \sum_{i=1}^{N} \sum_{\mu \in \Lambda} V_{i\mu} \int_{R_i} \log P_{\mu}(F(x, t; \beta | a_{\mu, i})) dx \, dt \, d\beta$$

$$\text{(5)}$$

The region competition algorithm can be adapted to minimize equation 5. The only difference with equations 2 and 4 is that we must maintain several competing explanations for each possible region.
Figure 6.14  Competition between parametric and nonparametric models. The different models compete to explain the data. The overall result is a segmentation into subregions, each of which is associated with one of the $N$ models or has been found by the nonparametric process.

**MOTION COHERENCE AND MOTION CAPTURE**

We now give a more detailed example of one of the parametric models that can be used in the framework. This model is taken from our early work on motion capture and motion coherence.

We will not describe how this model contributes to segmentation but rather focus on how the model fits motion data (the second term on the righthand side of equation 3). The model will be presented, as it was originally, in terms of energy minimization. Yet it can be readily interpreted in terms of (Bayesian) probabilities. The standard way is to define the Gibbs distribution $P[V] = (1/Z)e^{-E[V]}$ where $E[V]$ is the energy function and $Z$ is a normalization constant. The energy function defined below will be quadratic in $V$ and hence corresponds to a multivariate Gaussian distribution.

Let the velocity measurement of the measuring stage at point $x_i$ be $M(v_i)$. The measurement operator will depend on the measurements (figure 6.15). The model (Yuille and Grzywacz, 1988, 1989) proposes that the smoothing stage constructs a velocity field, $v(x)$, such that the following function is minimized for both components of $v$:

$$E(v(x), v_i) = \sum_i (M(v(x_i)) - M(v_i))^2 + \rho \sum_{m=0}^{\infty} C_m (D^m v)^2 dx,$$

where $\rho \geq 0$ and $C_m \geq 0$ are constants. The derivative operator $(D^m)$ is a scalar operator if $m$ is even and a vector operator if $m$ is odd. For the latter case $(D^m v)^2 = (D^m v_x) \cdot (D^m v_x) + (D^m v_y) \cdot (D^m v_y)$, where $v = (v_x, v_y)$. More precisely:

$$(D^{2n} v) \cdot (D^{2n+1} v) = \nabla \cdot (D^{2n} v).$$

203  Yuille & Grzywacz: A Theoretical Framework for Visual Motion
Figure 6.15 Simulations of the psychophysics of motion coherence and motion capture. (A) Motion cooperativity (Williams and Sekuler, 1992; Williams, Philips, and Sekuler, 1986; Watanami and Sekuler, 1992). The true motion (straight lines at top) is rather noisy, because it is composed of a small velocity (vector by the side of lower grid) plus large-variance noise. In contrast, the perceived motion is smoother (straight lines at bottom). (B) How the model solves some variants of the aperture problem (Marr and Ullman, 1981; Horn and Schunck, 1981; Hildreth, 1984). (C) An explanation of this problem. In this figure, Gaussian-profile contours travel to the right with the velocity indicated by the isolated vectors in (B). In all four panels, the input motion is shown at the top, and the perceived motion at the bottom. Observe how the perception of velocity (straight lines) depends on the presence or absence of line endings (thick vertical bars represent occlusions of these endings) or of neighboring terminators (small black squares). (C) The aperture problem and our model's solution to it. In this figure, the circle is moving to the right with the velocity indicated by the isolated vector. The aperture problem (on the left) results from the impossibility of measuring locally the components of velocity parallel to the luminance gradient. The perceived motion predicted by our model (on the right) is quite different and consistent with human perception. (D) Motion capture. The true motion (on the left) consists of the components of the velocity perpendicular to the circle and the random motion of the dots inside the circle. The perceived motion (on the right) shows the circle and the dots moving rigidly together to the right. This figure was adapted from Yuille and Grzywacz (1986, 1989).

If \( m \) is even then this is a Laplacian to the power \( m/2 \). If \( m \) is odd it becomes the gradient of the Laplacian to the power \((m - 1)/2\). Smoothness operators of this type correspond to Tikhonov stabilizers (Poggio, Torre, and Koch, 1985). It was argued (Yuille and Grzywacz, 1988, 1989) that a good choice was \( c_m = \sigma^{2m}/(m!2^m) \). Then the solution of equation 6 obtained by standard calculus of variations has the form:

\[
v(x) = \sum \frac{\beta_i}{2\pi\sigma^2} \exp \left( -\frac{(x - x_i)^2}{2\sigma^2} \right)
\]  

(8)
where the $\beta_i$'s may be computed analytically by solving a linear system of equations. See Yuille and Grzywacz (1988, 1989).

This model was simulated for a number of psychophysical stimuli. Our first example showed that the model gives a possible explanation for the cooperativity of the motion system (Williams and Sekuler, 1984; Williams, Philips, and Sekuler, 1986; Watamaniuk and Sekuler, 1992). When dots move randomly, but with a slight bias, a percept of coherent motion occurs. The theory accounts for this behavior, because the small variance of the constructed velocity field may enable the mean motion to be detected (figure 6.15A).

Our theory also agrees with experiments by Nakayama and Silverman (1988a, b), which investigated variations of the aperture problem (figure 6.15B) and which were not easily explained by other current theories (for an illustration of the aperture problem, see figure 6.13C). These experiments displayed open (not straight) curves moving over time. They showed that the correct motion is not always perceived, even when the stimulus is moving rigidly but the line endings are covered; uncovering these endings makes the percepts much more veridical; and terminators positioned near, but not at, the lines and moving with the same velocity as the lines also makes the percepts veridical (figure 6.15B).

Another phenomenon accounted for by the model is motion capture (Ramachandran and Anstis, 1983b, 1986). In this phenomenon, randomly moving dots are captured and move coherently with a superimposed grating or a surrounding contour (figure 6.15D). Our theory simultaneously solves the aperture problem of the surrounding contour and captures the internal dots. The theory also predicts that capture may happen when only a few dots move. An example of this occurs when a central ambiguous motion is captured by a peripheral unambiguous motion. Psychophysical experiments with such paradigms were performed and the results were consistent with our theory (Yuille and Grzywacz, 1988, 1989). The model also made several predictions that were confirmed psychophysically by Watamaniuk, Grzywacz, and Yuille (1993). These predictions, which were somewhat not intuitive, stated that if the dot density in a random-dot cinematogram increased or decreased abruptly then the perceived speed would temporarily increase or decrease respectively.

Finally, implementations of this model, using region growing approaches, would seem to be consistent with recent experiments (Watanabe and Cole, 1995). In these experiments, the perception of coherent motion gradually spreads across the image.

**DISCUSSION AND CONCLUSION**

How does our theoretical framework relate to computer vision theories? Certain of our ingredients have been successfully exploited for a long time and are fairly well understood. Others are more novel and less explored.
Both spatial coherence and the use of filter banks to perform local measurements are ideas that appear in many existing theories. These theories can obtain good results on real images for a reasonable set of image motions (Weber and Malik, 1995). But the motions standardly dealt with in computer vision do not include cases where there are significant occlusions or outliers to be detected. To deal with these cues, our theoretical framework uses temporal coherence. This has been far less used in computer vision studies, except in the restricted cases of tracking when specific models of object motion (such as lip or hand motion, figure 6.1) are used.

The ideas of linking multiple layers and occlusions in parameter space are comparatively new and unexplored. They have appeared in the 2.1D sketch (Nitzberg, Mumford, and Shiota, 1993) and Adelson's motion layer theory (Wang and Adelson, 1994). Multiple layers simplify the explanation of transparency and dealing with occlusion.

Because our framework was designed with psychophysics in mind, it is not surprising that the framework is consistent with a large body of psychophysical literature. Perhaps more importantly, the framework raises many new experimental questions. For instance, can people detect boundaries due to Brownian motion? What classes of motion can people learn? Or, in our terminology, what new motion models can be incorporated into the framework with experience? One interesting phenomenon in this respect may be Ramachandran's and Anstis's (1986) multiple-diamond display. For diamonds presented alone, the motion is ambiguous and they can be seen rotating clockwise or anticlockwise. For a group of diamonds the percept is still ambiguous, but it is of all the diamonds rotating simultaneously in the same direction (for example, all clockwise or all anticlockwise). This seems a highly unlikely type of motion to occur in nature, and yet similar motions do occur in manmade objects, such as when the wheels of a wagon rotate in the same direction. Thus, a phenomenon of this type might describe a class of motions only found recently in evolutionary terms and learned early in ontogenetic development.

Some parts of our framework are concrete and are described by models. These include motion coherence, motion outlier detection, motion transparency detection, and temporal motion coherence. These models have been tested against experimental data. But large parts of the framework remain conceptual and still await detailed modeling. For these aspects we can only claim very general experimental agreement with our framework. Some of the phenomena involved are complex.

One exciting question is whether one can map this framework onto the anatomy of the brain. It is encouraging that many types of computations required by the models in our framework seem to map onto the known cortical organizations (Saito et al., 1986; Duffy and Wurtz, 1991; Graziano, Andersen, and Snowden, 1994). In addition, the discovery of top-down (for a review see Maunsell and Newsome, 1987) and lateral (Gilbert and Wiesel, 1979, 1989; Martin and Whitteridge, 1984; Kitano et al., 1994) cortical con-
nections seems consistent with our competitive models. The importance of these connections is emphasized by the theories of Mumford (1994) and Ullman (1994). These theories are in the same spirit as our framework, but are intended to be more general and are not specifically concerned with motion perception.

There do remain some motion perception phenomena that are out of the range of our framework. In particular, we have said nothing about how additional visual cues can be combined with motion cues. For example, we cannot explain the experiments of Kersten et al. (1996), where the perception of motion is strongly influenced by the movement of cast shadows. This would involve extending our framework, possibly as outlined by (Yuille and Bülthoff, 1996).

ACKNOWLEDGMENTS

We would like to thank Takeo Watanabe for inviting us to write this chapter. Dave Geiger's comments on this manuscript were greatly appreciated. A. Yuille would also like to acknowledge discussions with Song Chun Zhu and Robert Ogniewicz on different aspects of region competition. Moreover, we would like to thank the hospitality of Ramon Rizo and Joaquin de Juan at the University of Alicante, Spain, where some of this work was performed. A. Yuille would like to thank ONR/ARPA for funding (contract N00014-95-1-1022). NMG was funded by the Air Force grant AFOSR F49620-95-1-0265. Further support came from core grant EY-06883 from the National Eye Institute to the Smith-Kettlewell Eye Research Institute.

NOTES

1. This model also accounts for why contrast affects transparency (Adelson and Movshon, 1982) despite not being one of the variables in Beta-Space.

2. Recent experiments suggest that penalties should be paid for curvature and derivatives of curvature (Pettet, McKee, and Grzywacz, 1998).

3. This energy is a natural extension of the windows used in Zhu and Yuille (1996).

REFERENCES


Yuille & Grzywacz: A Theoretical Framework for Visual Motion